XII Triennial International Scientific Conference

Engineering TODAY

2026

CONFERENCE PROCEEDINGS

25 - 27 JUNE 2026, VRNJAČKA BANJA, SERBIA

FACULTY OF MECHANICAL AND CIVIL ENGINEERING IN KRALJEVO
UNIVERSITY OF KRAGUJEVAC

XII INTERNATIONAL TRIENNIAL SCIENTIFIC CONFERENCE

ENGINEERING TODAY
ET 2026

Conference Proceedings

25 – 27 June 2026, Vrnjačka Banja, Serbia

Supported by:

Republic of Serbia
Ministry of Science, Technological Development and Innovation

PUBLISHER: Faculty of Mechanical and Civil Engineering in Kraljevo, University of Kragujevac

EDITOR: Prof. Dr Goran Marković

TECHNICAL EDITOR AND COVER DESIGN:
Prof. Dr Mišo Bjelić

REVIEWS: All papers have been reviewed by the members of the Scientific Committee

PRINT: SaTCIP d.o.o., Vrnjačka Banja

PLACE AND YEAR OF PUBLICATION:
Kraljevo, 2026

CIRCULATION: 60 copies

ISBN 978-86-82434-15-3

CIP - Каталогизација у публикацији Народна библиотека Србије, Београд

621(082)
629.3/.4(082)
621.86/.87(082)
681.5(082)
004(048)

TRIENNIAL International Scientific Conference Engineering Today (12 ; 2026 ; Vrnjačka Banja)

Conference Proceedings / XII International Triennial Scientific Conference Engineering Today 2026, ET 2026, 25 – 27 June 2026, Vrnjačka Banja, Serbia ; [editor Goran Marković]. - Kraljevo : University of Kragujevac, Faculty of Mechanical and Civil Engineering, 2026 (Vrnjačka Banja : SaTCIP). - 1 knj. (razl. pag.) : ilustr. ; 30 cm

Na nasl. str.: Republic of Serbia, Ministry of Science, Technological Development and Innovation. - Tiraž 60. - Str. 7: Preface / Goran Marković. - Bibliografija uz svaki rad. - Registar.

ISBN 978-86-82434-15-3

a) Производно машинство -- Зборници
b) Транспортна средства -- Зборници
v) Аутоматско управљање -- Зборници
g) Информациона технологија -- Зборници

COBISS.SR-ID 201872905

INTERNATIONAL SCIENTIFIC COMMITTEE

CONFERENCE CHAIRMAN

Prof. Dr Goran Marković, FMCE Kraljevo, Serbia

Scientific Committee Chairman
Prof. Dr Slaviša Šalinić, FMCE Kraljevo, Serbia

Scientific Committee Vice-Chairman
Prof. Dr Mišo Bjelić, FMCE Kraljevo, Serbia

MEMBERS OF THE SCIENTIFIC COMMITTEE

Prof. Dr M. Alamoreanu, TU Bucharest, Romania
Prof. Dr D. Atmadzhova, VTU Sofia, Bulgaria
Prof. Dr M. Banić, FME Niš, Serbia
Prof. Dr N. I. Gergova, VTU Sofia, Bulgaria
Prof. Dr M. Berg, KTH, Sweden
Prof. Dr S. Bikić, FTS Novi Sad, Serbia
Prof. Dr K. Krastanov, VTU Sofia, Bulgaria
Prof. Dr M. Bižić, FMCE Kraljevo, Serbia
Prof. Dr M. Blagojević, FE Kragujevac, Serbia
Prof. Dr H. Bogdevicius, VILNIUS TECH, Lithuania
Prof. Dr G. Bogdanović, FE Kragujevac, Serbia
Prof. Dr N. Bogojević, FMCE Kraljevo, Serbia
Prof. Dr S. Bošnjak, FME Belgrade, Serbia
Prof. Dr I. Božić, FME Belgrade, Serbia
Prof. Dr A. Bruja, TU Bucharest, Romania
Prof. Dr R. Bulatović, FMCE Kraljevo, Serbia
Prof. Dr S. Ćirić Kostić, FMCE Kraljevo, Serbia
Prof. Dr I. Despotović, FMCE Kraljevo, Serbia
Prof. Dr M. V. Dragoi, UNITBV, Romania
Prof. Dr B. Dragović, FMS Kotor, Montenegro
Prof. Dr Lj. Dubonjić, FMCE Kraljevo, Serbia
Prof. Dr R. Durković, FME Podgorica, Montenegro
Prof. Dr R. Đokić, FTS Novi Sad, Serbia
Prof. Dr K. Ehmann, NU Chicago, USA
Prof. Dr O. Erić Cekić, FMCE Kraljevo, Serbia
Prof. Dr V. Gašić, FME Belgrade, Serbia
Prof. Dr D. Golubović, FME East Sarajevo, BiH
Prof. Dr P. Gvero, FME Banja Luka, BiH
Prof. Dr B. Jerman, FME Ljubljana, Slovenia
Prof. Dr R. Karamarković, FMCE Kraljevo, Serbia
Prof. Dr M. Karasahin, DU Istanbul, Turkey
Prof. Dr K. Kocman, TU Brno, Czech Republic
Prof. Dr S. Kolaković, FTS Novi Sad, Serbia
Prof. Dr M. Kostić, NIU DeKalb, USA
Prof. Dr M. Krajišnik, FME East Sarajevo, BiH
Prof. Dr M. Králik, FME Bratislava, Slovakia
Prof. Dr E. Kudrjavcev, MGSU Moscow, Russia
Prof. Dr Đ. Lađinović, FTS Novi Sad, Serbia
Prof. Dr D. Marinković, TU Berlin, Germany
Prof. Dr M. Marašević, FMCE Kraljevo, Serbia

Prof. Dr A. Milašinović, FME Banja Luka, BiH
Prof. Dr I. Milićević, FTS Čačak, Serbia
Prof. Dr V. Milićević, FMCE Kraljevo, Serbia
Prof. Dr D. Milković, FME Belgrade, Serbia
Prof. Dr Z. Miljković, FME Belgrade, Serbia
Prof. Dr B. Milošević, FMCE Kraljevo, Serbia
Prof. Dr V. Milovanović, FE Kragujevac, Serbia
Prof. Dr G. Minak, UNIBO, Italy
Prof. Dr D. Minić, FME Kosovska Mitrovica, Serbia
Prof. Dr A. Nikolić, FMCE Kraljevo, Serbia
Prof. Dr E. Nikolov, TU Sofia, Bulgaria
Prof. Dr V. Nikolić, FME Niš, Serbia
Prof. Dr V. Nikolov, VTU Sofia, Bulgaria
Prof. Dr M. Ognjanović, FME Belgrade, Serbia
Prof. Dr J. Peterka, FMS&T Trnava, Slovakia
Prof. Dr D. Petrović, FMCE Kraljevo, Serbia
Prof. Dr J. Polajnar, BCU Prince George, Canada
Prof. Dr M. Popović, FTS Čačak, Serbia
Prof. Dr D. Pršić, FMCE Kraljevo, Serbia
Prof. Dr Francu Catalin, UTCB, Romania
Prof. Dr N. Radić, FME East Sarajevo, BiH
Prof. Dr B. Radičević, FMCE Kraljevo, Serbia
Prof. Dr V. Radonjanin, FTS Novi Sad, Serbia
Prof. Dr M. Savković, FMCE Kraljevo, Serbia
Prof. Dr D. Sever, FCE Maribor, Slovenia
Prof. Dr B. Sredojević, FMCE Kraljevo, Serbia
Prof. Dr V. Stojanović, FMCE Kraljevo, Serbia
Prof. Dr I. S. Surovcev, VGSU Voronezh, Russia
Prof. Dr J. Tanasković, FME Belgrade, Serbia
Prof. Dr Lj. Tanović, FME Belgrade, Serbia
Prof. Dr D. Todorova, VTU Sofia, Bulgaria
Prof. Dr K. Weinert, TU Dortmund, Germany
Prof. Dr N. Zdravković, FMCE Kraljevo, Serbia
Prof. Dr D. Živanić, FTS Novi Sad, Serbia
Prof. Dr N. Zrnić, FME Belgrade, Serbia
Prof. Dr R. Vujadinovic, FME Podgorica, Montengro
Prof. Dr M. Jordović Pavlović, FMCE Kraljevo, Serbia
Prof. Dr Z. Đinović, ACMIT Wiener Neustadt, Austria
Prof. Dr Limonka Koceva Lazarova, UGD Štip, N. Maced.

ORGANIZING COMMITTEE

Chairman

Prof. Dr Vladimir Milićević, FMCE Kraljevo, Serbia

Vice-Chairman

Dr Nataša Pavlović, FMCE Kraljevo, Serbia

MEMBERS OF THE ORGANIZING COMMITTEE

Dr A. Petrović, FMCE Kraljevo, Serbia
Dr V. Grković, FMCE Kraljevo, Serbia
Dr M. Bošković, FMCE Kraljevo, Serbia
Dr M. Nikolić, FMCE Kraljevo, Serbia
Dr N. Stojić, FMCE Kraljevo, Serbia
Dr I. Franc, FMCE Kraljevo, Serbia
Dr M. Jovanović, FMCE Kraljevo, Serbia
Dr M. Nikolić, FMCE Kraljevo, Serbia
Dr V. Đorđević, FMCE Kraljevo, Serbia
Dr N. Petrović, FMCE Kraljevo, Serbia
Dr J. Bojković, FMCE Kraljevo, Serbia
Dr G. Bošković, FMCE Kraljevo, Serbia

TECHNICAL COMMITTEE

Chairman

Dr Vladimir Mandić, FMCE Kraljevo, Serbia

Vice-Chairman

Bojan Beloica, FMCE Kraljevo, Serbia

MEMBERS OF THE TECHNICAL COMMITTEE

M. Ivanović, FMCE Kraljevo, Serbia
M. Janićijević Milačak, FMCE Kraljevo, Serbia
S. Mihajlović, FMCE Kraljevo, Serbia
V. Sinđelić, FMCE Kraljevo, Serbia
M. Todorović, FMCE Kraljevo, Serbia
S. Pajović, FMCE Kraljevo, Serbia
P. Mladenović, FMCE Kraljevo, Serbia
A. Jovanović, FMCE Kraljevo, Serbia
M. Adamović, FMCE Kraljevo, Serbia
M. Rasinac, FMCE Kraljevo, Serbia
J. Perić, FMCE Kraljevo, Serbia

PREFACE

Ladies and gentlemen, dear colleagues,

Welcome to Vrnjačka Banja, to the XII International Scientific Conference *Engineering TODAY*.

Since 1993, the Faculty of Mechanical and Civil Engineering in Kraljevo has been organizing this triennial scientific conference. After eleven successful editions under the name Heavy Machinery (HM), the conference now continues under the name Engineering TODAY (ET) - a change that reflects the expansion of the faculty's study programmes and research into the areas well beyond its traditional scope, while retaining full continuity with the HM series: the same ordinal number, the same triennial cycle, and the same international character.

This year the conference takes place in Vrnjačka Banja from 25 to 27 June 2026, in the year when the Faculty celebrates 65 years of university teaching.

The goal of the Conference is to make research available and applicable within both domestic and international frames, and to give our scientists the opportunity to share results and learn from colleagues abroad. The Conference covers a wide spectrum of contemporary engineering disciplines through the following thematic areas:

- Earth-moving and transportation machinery,
- Railway engineering,
- Thermal technique,
- Environment protection,
- Automatic control and fluid technique,
- Applied mechanics,
- Computer-aided design,
- Computer-integrated processes,
- Design of machining processes,
- Entrepreneurial engineering and management,
- Civil engineering,
- Information technologies and software engineering,
- Artificial intelligence and machine learning.

The high scientific reputation of the participants and the number of papers are a guarantee that the Conference will be a success. The papers reflect the state of the art and address the topics of current interest across mechanical, civil and information engineering.

I would especially like to thank the Ministry of Science, Technological Development and Innovation of the Republic of Serbia for its support, as well as the other sponsors of the Conference. My sincere thanks also go to the members of the scientific, organizing and technical committees, the reviewers, and all participants and invited speakers.

Thank you, and see you at the next conference in three years.

Kraljevo, Vrnjačka Banja, June 2026

Conference Chairman

Prof. Dr Goran Marković

PLENARY SESSION

Ž. Ćojbašić, G. Petrović P1

TRUSTWORTHY EDGE INTELLIGENCE IN ENGINEERING TODAY:
APPLICATIONS IN INDUSTRIAL CONTROL, ROBOTICS AND TRANSPORT ENGINEERING

V. Stojanovic, H. Tao, X. Li P11

FROM DATA SCARCITY TO OPEN-SET GENERALIZATION:
A UNIFIED FRAMEWORK FOR ROBUST ROTATING MACHINERY FAULT DIAGNOSIS

SESSION A: EARTH-MOVING AND TRANSPORTATION MACHINERY

C. Sescu-Gal, M. Savaniu, C. Francu, O. Tonciu, B. Bebeselea A1

MECHATRONIC SYSTEM FOR MONITORING CRANE LIFTING OPERATIONS

M. Urošević, N. Zrnić, X. Liu A9

REMARKS ON THE INFLUENCE OF THE DYNAMIC BEHAVIOR OF BELT
CONVEYORS ON THE RESPONSE OF THE SPREADER SUPERSTRUCTURE

V. Naydenov, B. Burdin, D. Krastanov A17

EXPERIMENTAL STUDY OF ENERGY EFFICIENCY OF A HYDROGEN VEHICLE

M. Katona, A. Zelić, D. Živanić, R. Đokić, T. Jojić, N. Ilanković A25

COMPARATIVE OVERVIEW OF DETERMINING SERVICE CLASSES OF
INDUSTRIAL CRANES ACCORDING TO RELEVANT STANDARDS AND GUIDELINES

V. Jovanović, N. Petrović, J. Pavlović, B. Nikolić A33

A FUZZY AHP APPROACH FOR EVALUATING CIRCULAR EFFICIENCY
OF HYDRAULIC EXCAVATOR SWING DRIVE SYSTEMS

N. Miloradović, R. Vujanac, M. Piskulić, J. Glišović A41

COMPARATIVE ANALYSIS OF RACK LAYOUTS ACCORDING
TO THE STORAGE UNIT PLACEMENT METHOD

N. B. Zdravković, M. Savković, G. Marković, M. Todorović, P. Mladenović, G. V. Pavlović A51

THE DETERMINATION OF THE DEFLECTION OF THE STEPPED
COLUMN BY THE FINITE DIFFERENCE METHOD

D. Živanić, V. Žerajić, N. Ilanković A59

PHYSICAL ANALYSIS OF FORCES, MOMENTS, AND ENERGY EFFICIENCY OF MULTIROTOR
UNMANNED AERIAL VEHICLES FROM THE PERSPECTIVE OF FLIGHT CONTROL

V. Žerajić, D. Ružić, D. Živanić, N. Ilanković A69

SEGMENT-BASED ENERGY BALANCE AND PROPULSION FEASIBILITY
ANALYSIS OF A LIGHT PASSENGER VTOL AIRCRAFT FOR A REGIONAL MISSION

S. Trifković, N. Radić, M. Milutinović, S. Samardžić A79

ANALYSIS AND CALCULATION OF OVERHEAD CONVEYOR
FOR METAL HOUSINGS TRANSPORTATION

R. Đokić, J. Vladić, T. Jojić A85

ANALYSIS OF ZIPLINE SYSTEMS WITH LARGE SPANS AND LOW INCLINATION ANGLES

A. Stanković, G. Petrović, A. Trajković, P. Milić, M. Stanković, Ž. Ćojbašić A95

A HYBRID SWARA–MULTIMOORA APPROACH FOR PRODUCTION JOB PRIORITIZATION

SESSION B: RAILWAY ENGINEERING

V. Nikolov, F. Kirilov B1

STRENGTH-DEFORMATION ANALYSIS OF WHEEL REDUCER FOR METRO CARS

D. Milković, S. Radulović, A. K. Miličić, G. Simić, O. Obradović B11

TESTS FOR THE ASSESSMENT OF DEVICES FOR SECURING SEMI-TRAILERS DURING TRANSPORT ON A FREIGHT WAGON

D. Atmadzhova, V. Ralev B17

STRENGTH ANALYSIS OF AN ELECTRIC LOCOMOTIVE TRANSMISSION MECHANISM

E. M. Mihaylov, E. Iontchev, E. Stoyanova B29

IMPACT OF DETERIORATED TRAM TRACKS ON THE CURRENT COLLECTORS OF TRAMS

B. Gavrilović, Ž. Avramović, Z. Stjepanović, D. Stjepanović B35

MEASUREMENT OF DISTORTION POWER AND POWER FACTOR IN ELECTRIC TRACTION SUBSTATIONS “SERBIA RAILWAY INFRASTRUCTURE” WHEN POWERING ELECTRIC LOCOMOTIVES IN TRACTION AND RECUPERATIVE BRAKING

J. Glišović, D. Miloradović, J. Lukić, N. Miloradović, S. Vasiljević B41

COMPARATIVE ANALYSIS OF METHODS FOR THE INVESTIGATION OF BRAKE WEAR PARTICLES EMISSION OF RAILWAY VEHICLES

M. Bižić, D. Petrović B51

INFLUENCE OF SIGNAL-TO-NOISE RATIO ON THE ACCURACY OF EXPERIMENTAL DETERMINATION OF WHEEL–RAIL INTERACTION FORCES

V. Zochinova, V. Pavlov B59

SYSTEMS FOR PRECISION CONTROL OF MICROCLIMATE IN WORLD RAIL TRANSPORT: TECHNOLOGICAL STRUCTURE AND INTELLIGENT MANAGEMENT

V. Pavlov B65

METHODS OF INCREASING THE WEAR RESISTANCE WITH REPAIRS OF BDZ FREIGHT WAGON BOGIES

I. Petrova B69

IMPACT OF AUTOMATED SYSTEMS ON REDUCING THE HUMAN FACTOR IN RAIL TRANSPORT

Z. G. Pavlović, A. Savić, V. Radičević, R. Vasiljević B75

AN ELECTRONIC BUSINESS MODEL FOR THE REGULATION OF RAILWAY TRAFFIC AND TRANSPORT BASED ON ARTIFICIAL INTELLIGENCE

K. Trifonov B81

STATE OF DISPATCHER SYSTEMS IN BULGARIA AND THEIR ROLE
IN THE LIBERALISED PASSENGER RAIL MARKET

V. Ralev B93

STUDY ON THE WEAR OF CYLINDRICAL GEARS OF TRANSMISSION MECHANISMS

SESSION C: PRODUCTION TECHNOLOGIES, CAE AND MANAGEMENT

E. Hozdić, E. Hozdić C1

FROM INDUSTRY 4.0 TO INDUSTRY 5.0: THE TRANSITION FROM DIGITALIZATION
AND CYBERNATION TO COGNITIZATION AND HUMAN-MACHINE SYNERGY

V. Vujičić, I. Milićević, N. Mijailović, N. Dučić, V. Aleksić, M. Popović C13

ROBOT-ASSISTED LATHE MACHINE SERVICING

M. Madić, M. Trifunović, S. S. Nikolić, M. Milojković C19

ENVIRONMENTAL IMPACT ANALYSIS OF LASER CUTTING:
A CASE STUDY OF THIN STAINLESS STEEL AND ALUMINUM ALLOY SHEETS

M. Rasinac, M. Bjelić, T. Dulović, J. Perić, M. Ivanović C25

THREE-OBJECTIVE PARETO OPTIMIZATION OF GMAW PROCESS
PARAMETERS USING A MODIFIED FORCE BALANCE MODEL

S. Pajović, V. Grković, J. Perić, A. Petrović, B. Radičević C37

INVESTIGATION OF THE INFLUENCE OF CUTTING PARAMETERS
ON NOISE EMISSION DURING FACE TURNING OF BEECH WOOD

J. Perić, M. Rasinac, S. Pajović, V. Grković, A. Petrović C45

INVESTIGATION OF CUTTING PARAMETERS INFLUENCE
ON VIBRATION DURING MILLING OF HDPE

D. Milčić, M. Milčić, N. Zdravković, A. Đurić, D. Klobčar, T. Vuherer C53

ENERGY EFFICIENCY AND CARBON FOOTPRINT OF WELDED, ADHESIVE
AND HYBRID JOINING PROCESSES IN THE CONTEXT OF CIRCULAR ECONOMY

N. Petrović, V. Jovanović, J. Mihajlović, B. Nikolić, S. Marković C61

A FUZZY AHP-BASED APPROACH FOR OPTIMAL TRANSPORT MODE SELECTION

A. Jovičić, D. Veljković, A. Mitrović, J. Baralić C67

SELECTION AND ANALYSIS OF KEY OCCUPATIONAL SAFETY AND HEALTH INDICATORS:
A CASE STUDY IN THE FOUNDRY INDUSTRY

V. Vassilev, N. Ivanova Gergova C77

BALANCING BETWEEN PUBLIC SERVICE OBLIGATIONS AND MARKET EFFICIENCY:
MODERN CHALLENGES IN PASSENGER RAIL TRANSPORT

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

L. Kelić D1

HYBRID SIMULATION OF AN INVERTED PENDULUM

S. T. Radovic, D. Stavrov, G. Nadzinski, V. O. Latkoska D11

INTELLIGENT CONTROL OF A UAV USING A SENSOR GLOVE

Lj. Dubonjic, J. P. Z. Machado, D. Prsic, V. Stojanovic, X. Xie D23

EVENT-TRIGGERED ADP FOR ASYMPTOTIC OUTPUT REGULATION
OF UNKNOWN LINEAR SYSTEMS

V. Stojanovic, D. Prsic, Lj. Dubonjic, M. Defoort, X. Song, S. Song D31

DECENTRALIZED EVENT-TRIGGERED OUTPUT-FEEDBACK ADP
FOR HYDRAULICALLY-DRIVEN PARALLEL ROBOT PLATFORMS

S. Prodanovic, E. Bernardi, X. Luan, V. Stojanovic, W. Paszke D41

EVENT-TRIGGERED OUTPUT-FEEDBACK ADP FOR SINGLE-JOINT
ROBOTIC MANIPULATORS WITH UNKNOWN DYNAMICS

V. Filipović, V. Đorđević D49

MINIMUM POSITIONING-TIME FOR MECHATRONIC SYSTEMS

V. Milićević, M. Jovanović, N. Biga, M. M. Živanović, I. Franc D57

LEGAL ISSUES OF ETHICAL HACKING:
A SYSTEMATIC REVIEW OF AUTHORIZATION, COMPLIANCE, AND LIABILITY FRAMEWORKS

G. Katipoğlu, S. Utku D69

REAL-TIME OBJECT RECOGNITION AND AUDIO FEEDBACK
SYSTEM FOR VISUALLY IMPAIRED INDIVIDUALS

Z. G. Pavlović, A. Savić, V. Radičević, R. Vasiljević D75

MODEL OF APPLICATION OF ARTIFICIAL INTELLIGENCE IN HIGHER EDUCATION
BASED ON INTERACTION BETWEEN USER AND COMPUTER SYSTEM

M. K. Paneva, N. Stojkovikj, V. Kokalanov, V. Milićević D81

ML ALGORITHM FOR COVID-19 OUTCOME PREDICTION BASED ON DEMOGRAPHIC
FEATURES IN NORTH MACEDONIA

M. J. Pavlović, T. Geroski, M. Nikolić D87

DRUG CLUSTERING UPON TOXICITY EFFECTS ON LIVER
SPHEROIDS – A DEEP LEARNING APPROACH

M. Svičević, M. Pavković, N. Vučićević, A. Milenković, A. Milutinović D97

COMPARING QLORA FINE-TUNING OF A GENERAL-PURPOSE AND A
MATHEMATICALLY SPECIALIZED LLM ON SERBIAN MATHEMATICS COMPETITION TASKS

S. Jančić, M. Jovanović, E. Jovanović, I. Franc D105

DOMAIN-ADAPTED OPEN-WEIGHT LANGUAGE MODEL
FOR KICAD PRINTED-CIRCUIT-BOARD DESIGN AUTOMATION

SESSION E: THERMAL TECHNIQUE AND ENVIRONMENTAL PROTECTION

G. Bošković, Ž. Kužet, M. Bošković, M. Todorović, Z. Čepić E1

INTEGRATED RISK ASSESSMENT AND DIGITAL TWIN-BASED SAFETY MANAGEMENT IN PHOTOVOLTAIC SYSTEMS

G. Krunić, S. Vasković, S. Krile, A. Ijačić E11

ANALYSIS OF THE INFLUENCE OF THE ROTATION SPEED OF THE DFIG WIND TURBINE GENERATOR ON THE TEMPERATURE OF THE GENERATOR

M. Timotijević, O. Erić Cekić, D. Rajnović, P. Janjatović E21

INFLUENCE OF OVERHEATING EXPOSURE TIME ON THE MICROSTRUCTURAL STABILITY AND MECHANICAL PROPERTIES OF HP40-0.63NB ALLOY

V. Zochinova, M. Ivanova, T. Shalamanova E29

ASSESSMENT OF THE MICROCLIMATE AND SUBJECTIVE PERCEPTION OF THE THERMAL ENVIRONMENT IN SCHOOL

I. Kuvač, D. Stupar, M. Schwai, I. Karan, B. Mrđa E35

URBAN ACUPUNCTURE AND VISUAL ACTIVISM AS FOUNDATIONS FOR ENVIRONMENTALLY RESPONSIBLE RIVERSIDE TRANSFORMATION: THE CASE OF THE VRBAS RIVER IN BANJA LUKA

S. Knežević, M. Milojević, A. Milošević E47

APPLICATION OF ROBOTIC SYSTEMS IN INTENSIVE LIVESTOCK FARMING FROM THE ASPECT OF ENVIRONMENTAL PROTECTION

J. Belošević, S. Belošević, M. Đukić E55

CHEMICAL POLLUTANTS IN THE SOIL OF NOVI PAZAR AND ECOLOGICAL RISK ASSESSMENT

J. Belošević, S. Belošević, M. Đukić E61

MONITORING OF HEAVY METALS AND THEIR MOVEMENT THROUGH THE ENVIRONMENT OF THE CITY OF KRAGUJEVAC

SESSION F: APPLIED MECHANICS

B. Milenković, Đ. Jovanović, M. Krstić F1

VOLUME OPTIMIZATION OF THE WORM GEAR USING SSO ALGORITHM

Z. Petrović, R. Nikolić, Lj. Brzaković, M. Tufegdžić, S. Petrović Savić F7

VIRTUAL TWIN MODELING OF THE MITSUBISHI RV-2FRL INDUSTRIAL ROBOT IN THE MATLAB PROGRAMMABLE ENVIRONMENT

V. Sinđelić, A. Nikolić, S. Ćirić-Kostić, N. Bogojević, O. Cekić-Erić, N. Drvar F17

NUMERICAL INVESTIGATION OF A STAR-SHAPED AUXETIC STRUCTURE UNDER DIFFERENT IMPACT VELOCITIES

SESSION G: CIVIL ENGINEERING

B. Paleva-Kadiyska **G1**

ROAD SURFACE WEAR – ANALYSIS OF TESTS AND POSSIBILITIES
FOR APPLICATION OF NEW MATERIALS

P. Stefanova **G7**

COMPARATIVE STUDY OF BLEEDING PROCESS IN ORDINARY AND SELF-COMPACTING
CONCRETE – INFLUENCE ON THE RELIABILITY OF REINFORCED CONCRETE ELEMENTS

J. Bojković, N. Stojić, V. Bulatović, I. Despotović, S. Mihajlović **G13**

BIOCOMPOSITE BASED ON CORN COBS AS AN INNOVATIVE
THERMAL INSULATION MATERIAL

S. Bajić, S. Milošević **G19**

RECYCLED STEEL IN CONTEMPORARY ARCHITECTURE:
SUSTAINABLE POTENTIALS AND APPLICATIONS THROUGH CASE STUDIES

M. J. Milačak, S. Mihajlović, B. Milošević, J. Bojković **G27**

PUSHOVER ANALYSIS WITH VARIOUS MATERIAL NONLINEARITY MODELS

B. Milošević, N. Kojić, M. J. Milačak **G35**

EVOLUTION OF NONLINEAR ANALYSIS OF RC SECTIONS: FROM ITERATIVE
COMPUTATIONS TO RAPID MOMENT-CURVATURE PREDICTION USING ANN MODELS

S. Mihajlović, M. Bižić, I. Despotović, J. Bojković, M. J. Milačak **G45**

METHODOLOGY FOR MEASURING PRESSURE AT THE CONTACT SURFACE
BETWEEN A COLUMN FOOTING MODEL AND A SOIL LAYER LIMITED BY A
RIGID BASE USING EARTH PRESSURE CELLS

INDEX OF AUTHORS

PLENARY SESSION

PLENARY SESSION

Trustworthy edge intelligence in engineering today: Applications in industrial control, robotics and transport engineering

Žarko Ćojbašić[1*], Goran Petrović[1]

[1] University of Niš, Faculty of Mechanical Engineering, Niš, Serbia

ARTICLE INFO

[*] **Correspondence:** zcojba@ni.ac.rs

DOI: https://doi.org/10.46793/ET26.P01C **PAGES:** P1 – P10

ABSTRACT

Artificial intelligence is rapidly transforming modern engineering systems, enabling unprecedented levels of autonomy, adaptability, and real-time decision-making in industrial automation, robotics, and intelligent transportation systems. However, the increasing deployment of complex machine-learning models in safety-critical applications introduces significant challenges related to transparency, interpretability, reliability, and trustworthiness. This invited plenary paper presents a synthesis of the authors' recent research results focused on trustworthy edge intelligence for engineering applications, with particular emphasis on explainable artificial intelligence (XAI), edge computing, and resource-constrained intelligent systems. The paper presents recent developments in explainable AI approaches for industrial automation and robotics, including hybrid deep-learning and ANFIS-based architectures for interpretable industrial decision support and machine-vision applications. Recent results in TinyML- and Industrial Internet of Things (IIoT)-based intelligent edge systems for low-cost industrial quality control are also discussed. In transport engineering, special attention is devoted to edge-AI implementations for intelligent transportation systems based on YOLO and Grad-CAM approaches deployed on embedded platforms such as NVIDIA Jetson devices and Raspberry Pi systems equipped with AI accelerators. Comparative analyses of performance, explainability, energy efficiency, and deployment suitability are presented. Finally, future directions toward human-centered, explainable, and trustworthy intelligent engineering systems are discussed in the broader context of engineering today and emerging Industry 5.0 paradigms.

KEYWORDS

Trustworthy Artificial Intelligence, Explainable Artificial Intelligence, Edge Computing, Industrial Automation, Robotics, Intelligent Transportation Systems.

1. INTRODUCTION

Artificial intelligence (AI) is increasingly transforming modern engineering systems by enabling autonomous operation, adaptive decision-making and intelligent data processing across a broad range of applications. Recent advances in machine learning, computer vision and intelligent data analytics have significantly expanded the capabilities of engineering systems operating in dynamic and complex environments [1, 2]. Such developments have accelerated the adoption of AI technologies in industrial automation, robotics and transport engineering, where intelligent systems are expected to support reliable operation, improved efficiency and enhanced decision-making under real-world conditions [3, 4]. Furthermore, the growing availability of embedded computing platforms has facilitated the deployment of intelligent algorithms directly within engineering systems, bringing computational intelligence closer to sensors, actuators and end users [5, 6].

Despite their remarkable predictive capabilities, many contemporary AI models remain difficult to interpret due to their inherent complexity. This limitation is particularly important in safety-critical and mission-critical engineering applications, where understanding the rationale behind automated decisions is often as important as achieving high predictive performance [7, 8]. In industrial automation, robotics and intelligent transportation systems, engineers and operators increasingly require transparent and trustworthy decision-support mechanisms capable of providing insight into the reasoning process of intelligent systems [9, 10]. Consequently, explainable artificial intelligence (XAI) has emerged as an important research direction aimed at improving model interpretability, transparency and user trust while maintaining the advantages of modern machine-learning approaches [8, 9].

At the same time, the growing demand for real-time processing, reduced communication latency, privacy preservation and increased operational autonomy has accelerated the development of Edge AI solutions [5, 6]. By performing inference directly on embedded and resource-constrained devices, Edge AI enables intelligent operation without continuous dependence on cloud infrastructure. Recent advances in TinyML - a class of machine-learning approaches designed for execution on microcontrollers and other resource-constrained edge devices, Industrial Internet of Things (IIoT) technologies and specialised AI accelerators have further expanded the applicability of Edge AI across diverse engineering domains, supporting the deployment of intelligent systems in industrial environments, robotic platforms and transportation infrastructure [11, 12].

While XAI addresses the need for transparency and trust, and Edge AI addresses the requirements of efficient deployment and real-time operation, their simultaneous implementation remains a significant challenge. The convergence of these two paradigms gives rise to Edge Explainable Artificial Intelligence (Edge XAI), where explainability mechanisms are integrated directly into edge-deployed intelligent systems [13, 14]. Such an approach enables trustworthy edge intelligence capable of combining computational efficiency, real-time responsiveness and interpretable decision-making within practical engineering environments. As engineering systems become increasingly autonomous and interconnected, Edge XAI is emerging as a promising framework for balancing performance, transparency and trustworthiness.

This invited plenary paper presents a synthesis of the authors' recent research results in Edge AI, XAI and intelligent engineering systems [15, 16, 17, 18, 19, 20, 21, 22]. Particular attention is devoted to the emerging concept of Edge Explainable Artificial Intelligence (Edge XAI) and its application in industrial control, robotics and transport engineering. Through representative examples and experimental studies conducted on embedded platforms, the paper highlights the role of Edge XAI in enabling trustworthy edge intelligence and discusses future directions aligned with Engineering Today and the evolving paradigms of Industry 4.0 and Industry 5.0 [1]. The convergence of Edge AI and Explainable AI, referred to here as Edge XAI, provides a practical framework for achieving trustworthy edge intelligence in contemporary engineering applications and represents a promising pathway towards the development of trustworthy engineering systems operating under real-time and resource-constrained conditions.

2. EDGE EXPLAINABLE ARTIFICIAL INTELLIGENCE (EDGE XAI)

2.1. Edge AI Foundations

The rapid development of embedded computing technologies has significantly expanded the deployment possibilities of artificial intelligence in engineering applications. While conventional AI solutions often rely on cloud-based computation, many industrial and transportation systems require local processing capabilities capable of supporting real-time operation and autonomous decision-making. Consequently, increasing attention has been devoted to Edge AI approaches, where machine-learning and computer-vision algorithms are executed directly on embedded platforms located close to the data source [5, 6]. Such architectures reduce communication latency, improve operational reliability and facilitate intelligent operation in environments where cloud connectivity may be limited or undesirable [5].

Recent advances in embedded machine learning have demonstrated that intelligent algorithms can be successfully deployed across a wide spectrum of edge platforms, ranging from highly resource-constrained microcontrollers to more capable embedded computing systems. TinyML represents an important example of this trend, showing that machine-learning models can be implemented even on devices with very limited computational resources, memory and power consumption [11, 12]. Such approaches facilitate intelligent sensing, monitoring and decision-support functionalities directly at the data source while reducing communication overhead and enabling distributed intelligence. At the same time, the growing availability of low-cost embedded platforms and dedicated AI accelerators has expanded the range of Edge AI applications, allowing increasingly sophisticated machine-learning and computer-vision algorithms to be deployed in practical engineering environments [6]. As a result, modern Edge AI systems can be implemented using a variety of hardware and software solutions tailored to the performance, cost and operational requirements of specific applications.

The practical deployment of Edge AI increasingly relies on heterogeneous computing architectures combining general-purpose processors, graphics processing units and dedicated AI accelerators [6]. Such platforms enable the execution of increasingly sophisticated machine-learning and computer-vision algorithms while satisfying constraints related to computational resources, energy consumption and deployment cost. Recent developments in embedded AI hardware have therefore created new opportunities for implementing advanced intelligent systems in industrial automation, robotics and intelligent transportation applications [17, 19, 20].

2.2. Principles of Explainable Artificial Intelligence (XAI)

The growing adoption of machine-learning and deep-learning methods in engineering systems has highlighted the need for greater transparency in automated decision-making. Although state-of-the-art models frequently achieve high predictive accuracy, their internal decision processes often remain difficult to interpret [8, 9]. This limitation is particularly important in engineering environments where safety, reliability and accountability represent essential operational requirements. Consequently, understanding the factors that influence model predictions has become an important prerequisite for the successful deployment of intelligent systems in practice [7, 8].

Explainable Artificial Intelligence (XAI) encompasses a broad range of methods aimed at improving the transparency and interpretability of machine-learning models [8, 9]. By providing insights into model behaviour and decision-making processes, XAI supports validation of intelligent systems and facilitates the identification of potential errors, biases and unexpected behaviour. Furthermore, explainability contributes to the development of trustworthy AI systems by enabling engineers and operators to better understand, verify and evaluate automated decisions before they are applied in real-world environments [10].

Different explainability approaches have been proposed for engineering applications, ranging from inherently interpretable models to post-hoc explanation techniques applicable to complex deep-learning architectures [8, 9]. Figure 1 presents a taxonomy of representative XAI approaches and illustrates the diversity of methods available for improving transparency and interpretability in intelligent engineering systems [15]. In computer-vision applications, gradient-based methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) have attracted considerable attention due to their ability to provide intuitive visual explanations while maintaining compatibility with high-performance convolutional neural networks [23, 24, 25]. Such approaches enable visual interpretation of model predictions and support the practical adoption of explainable AI in engineering systems.

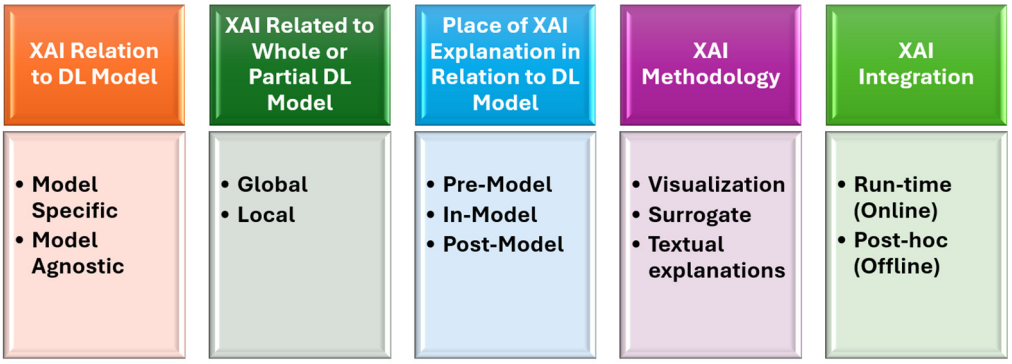


Figure 1: An approach to the taxonomy of XAI approaches [15]

2.3. Edge XAI as an Enabler of Trustworthy Intelligence

The increasing complexity of modern engineering systems has created a need for intelligent solutions that are not only computationally efficient but also transparent and interpretable. While Edge AI addresses deployment and real-time processing requirements, and XAI addresses transparency and trustworthiness requirements, many practical engineering applications require both capabilities simultaneously. As mentioned before, this convergence has led to the emergence of Edge Explainable Artificial Intelligence (Edge XAI), in which explainability mechanisms are integrated directly into edge-deployed intelligent systems [13, 14].

Unlike conventional explainability frameworks that are frequently executed offline or within cloud-based environments, Edge XAI aims to generate explanations within the operational environment itself. Such an approach enables continuous monitoring of model behaviour and supports real-time verification of automated decisions. The ability to combine local inference with local explainability is particularly relevant for example in industrial automation, robotics and intelligent transportation systems, where rapid decision-making and operational reliability are critical requirements [17, 18, 21].

By integrating explainability directly into embedded intelligent systems, Edge XAI supports the development of human-centred engineering solutions in which AI technologies function as collaborative decision-support tools rather than opaque black-box models. Through enhanced transparency, interpretability and operational trust, Edge XAI provides a practical framework for achieving trustworthy edge intelligence and represents an important step towards the next generation of intelligent engineering systems aligned with the evolving paradigms of Industry 4.0 and Industry 5.0 [1], while simultaneously emerging as an increasingly important cornerstone of Engineering Today.

3. APPLICATIONS IN INDUSTRIAL CONTROL AND ROBOTICS

This section presents representative examples from the authors' recent research illustrating how Edge XAI concepts can be applied to practical engineering problems. The selected case studies cover intelligent quality control in industrial control systems and gesture-based human–robot interaction, demonstrating different aspects of trustworthy edge intelligence in industrial and robotic environments.

3.1. Edge XAI for Industrial Quality Control

The increasing availability of low-cost embedded hardware has enabled the deployment of intelligent quality-control systems directly within industrial production environments. Among such applications, apple variety classification in the food industry represents an important quality-control task that has recently attracted increasing attention due to the growing demand for automated inspection and sorting systems. Recent advances in machine vision, Edge AI and Industrial Internet of Things (IIoT) technologies have created new opportunities for implementing automated inspection systems

capable of performing classification and monitoring tasks close to the data source [12, 25, 13]. Such approaches reduce communication requirements, improve responsiveness and support distributed intelligence, making them attractive for modern manufacturing systems [27].

To investigate the feasibility of intelligent quality control on embedded platforms, our recent research explored the integration of the TinyML framework, computer vision and IIoT technologies within an industrial inspection framework [20]. The developed system utilised embedded machine-learning models to perform real-time product classification while operating under the computational constraints typical of industrial edge devices. The proposed architecture demonstrates how intelligent inspection capabilities can be deployed directly within production environments without dependence on cloud-based processing.

The left side of Figure 2 shows the main elements of the industrial experimental setup developed in our previous research [29] and based on heavy duty industrial equipment, including the Siemens Simatic S7-1500 PLC with TM NPU artificial intelligence module, Basler camera and FlexLink conveyor. In contrast, the right side of Figure 2 shows the alternative low-cost Edge AI device for the same quality-classification task, based on an ESP32 platform. Experimental results demonstrated that compact machine-learning models can achieve satisfactory classification performance while maintaining the computational efficiency required for edge deployment on resource-constrained devices [20]. Beyond predictive accuracy, the study highlighted the importance of understanding model behaviour and assessing the suitability of deployed algorithms for practical industrial environments. From the perspective of this invited plenary paper, the presented framework represents an illustrative example of how explainability considerations can complement Edge AI and contribute to more transparent and trustworthy quality-control processes.

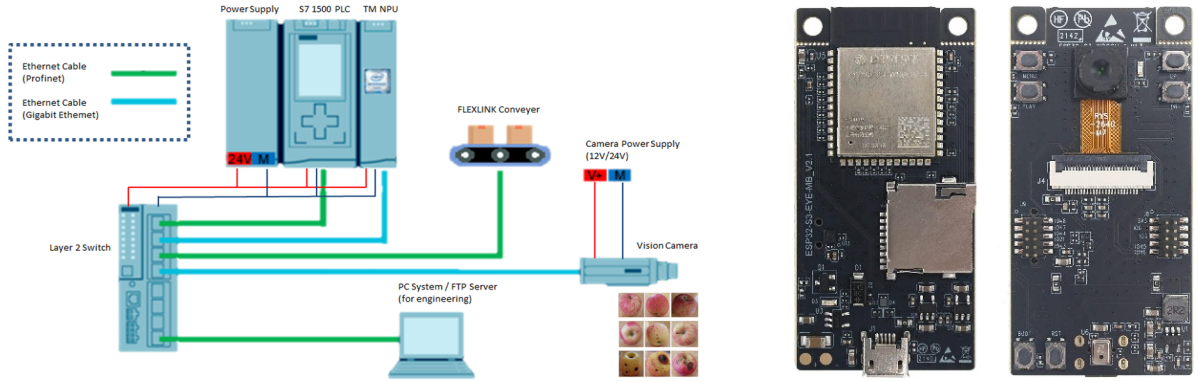


Figure 2: Elements of the experimental setup based on Siemens Simatic 1500 PLC with TM NPU artificial intelligence module, Basler 25 FPS 2 MP Camera and FlexLink conveyer (left) and ESP32-S3-EYE Main Board (right) [20]

It is further worth noticing that industrial control systems increasingly require intelligent functionalities capable of supporting monitoring, prediction and decision-making directly on embedded devices. However, the limited computational resources available on many industrial platforms impose significant constraints on the complexity of deployable machine-learning models [11, 28]. Consequently, TinyML has emerged as a promising approach for bringing machine-learning capabilities to resource-constrained industrial environments [5].

In further recent expansion of our previously mentioned results, several TinyML implementation strategies were investigated and compared within the context of an industrial control application [19]. Particular attention was devoted to analysing the trade-offs between predictive performance, computational requirements and deployment feasibility. Table 1 summarises the TinyML implementation strategies investigated and their performance characteristics. The comparison highlights the trade-offs between computational requirements, memory consumption and predictive performance that must be considered when deploying intelligent algorithms on industrial embedded platforms. The obtained results demonstrated that practical machine-learning functionality can be achieved even on highly constrained embedded platforms, confirming the potential of TinyML for intelligent industrial control.

In addition to TinyML-based approaches, our recent research also investigated a hybrid Convolutional Neural Network (CNN)–Adaptive Neuro-Fuzzy Inference System (ANFIS) framework for the same considered industrial automation problem in the food industry [26]. In the proposed architecture, CNN models were employed to extract discriminative features relevant for apple quality classification, while ANFIS provided interpretable fuzzy-rule-based reasoning capable of relating extracted features to quality classes through linguistic rules. Applied to apple classification and defect detection tasks, the hybrid approach demonstrated how predictive performance and explainability can be combined within a single intelligent system. From the perspective of trustworthy Edge XAI, such inherently interpretable models complement post-hoc explanation techniques by improving transparency, trustworthiness and understanding of AI-assisted decision-making in industrial environments.

Table 1: Model size, accuracy and inference time [19]

Model variant	Size (kB)	Accuracy (%)	S3-EYE Inference time (ms)	WROVER Inference time (ms)
MilliMobileNet	44	82	34.347	/
MicroMobileNet	16	79	29.161	/
NanoMobileNet	5.8	73	6.080	6.595
PicoMobileNet	2.7	68	2.610	2.875

The presented study illustrates that Edge AI deployment on highly resource-constrained devices involves more than achieving satisfactory predictive performance alone. Practical implementation also requires careful consideration of memory usage, computational efficiency and system responsiveness [19]. These factors directly influence the feasibility of deploying intelligent algorithms in industrial environments and therefore represent important design considerations for trustworthy edge intelligence systems.

3.2. Edge XAI for Gesture-Based Human–Robot Interaction

Human–robot interaction represents an important application domain for trustworthy intelligent systems, particularly within emerging Industry 5.0 environments where collaboration between human operators and robotic systems is becoming increasingly important [1, 29, 30]. In our recent work, a vision-based gesture-recognition framework was developed to support intuitive communication between human users and robotic systems [18]. By recognising hand gestures in real time, the proposed approach provides a natural interaction mechanism without requiring specialised control interfaces.

The developed framework combined Edge AI and explainability techniques within a gesture-recognition application deployed on embedded hardware platforms [18]. Experimental evaluation demonstrated that reliable gesture recognition can be achieved under realistic operating conditions while maintaining the computational efficiency required for edge deployment. The results further confirmed the feasibility of implementing intelligent perception algorithms directly on resource-constrained systems intended for practical robotic applications.

The left side of Figure 3 presents representative Grad-CAM and LIME explanations illustrating the image regions contributing most strongly to gesture-recognition decisions, whereas the right side of Figure 3 shows the computational time as a function of batch size for SHAP, LIME and Grad-CAM. Unlike many industrial applications where intelligent systems primarily support machine-centred processes, gesture recognition directly involves interaction with humans. Consequently, transparency and explainability become particularly important for establishing trust and user acceptance [23, 31]. The presented framework therefore represents a compelling example of human-centred Edge XAI, where intelligent perception, embedded deployment and trustworthy decision-making converge within a single robotic application [18].

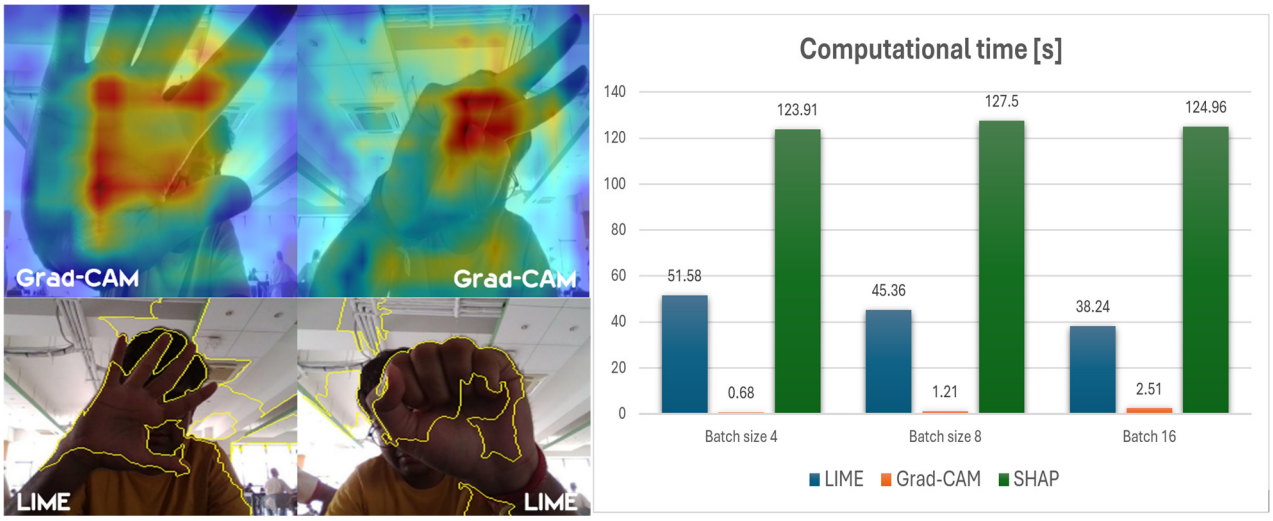


Figure 3: Grad-CAM heatmaps and LIME superpixel explanations illustrating key decision-making regions in sample images (left) and Computational time vs batch size for SHAP, LIME and Grad-CAM (right) [18]

Taken together, the presented case studies demonstrate how Edge XAI can support a broad spectrum of engineering applications ranging from industrial quality control in intelligent control systems to collaborative robotic environments. Although the investigated use cases differ substantially in their objectives and implementation details, they share common requirements for resource-efficient deployment, explainable operation and trustworthy decision-making. These characteristics collectively define the concept of trustworthy edge intelligence explored throughout this paper.

4. APPLICATIONS IN TRANSPORT ENGINEERING

This section presents representative examples from the authors' recent research demonstrating the application of Edge Explainable Artificial Intelligence (Edge XAI) concepts in intelligent transportation systems. The selected case studies address complementary aspects of trustworthy transportation intelligence, including explainable traffic monitoring, intelligent road-intersection management and the deployment of explainable deep-learning models on embedded edge platforms.

4.1. Explainable Artificial Intelligence for Trustworthy Road Intersection Management

Recent advances in computer vision and artificial intelligence have created new opportunities for intelligent traffic monitoring and road-intersection management [32, 33]. In particular, edge-deployed deep-learning systems can support automated vehicle recognition and traffic analysis while reducing dependence on cloud-based infrastructure [5, 6]. However, practical deployment of such systems requires not only accurate predictions but also transparent decision-making mechanisms capable of supporting user trust and operational reliability [8, 10].

To address these requirements, our recent research investigated the integration of explainable artificial intelligence techniques into edge-based road-intersection management systems [16]. By combining deep-learning-based vehicle recognition with visual explanation methods, the proposed framework enables both traffic analysis and interpretation of model decisions. Such an approach allows engineers and operators to verify whether classification outcomes are based on relevant visual features and therefore contributes to the development of trustworthy transportation intelligence.

The left side of Figure 4 presents representative Grad-CAM heatmaps and LIME superpixel explanations illustrating the key image regions contributing to vehicle-recognition decisions in road intersection management, whereas the right side of Figure 4 shows the computational time as a function of batch size for three comparative XAI methods - SHAP, LIME and Grad-CAM. The visual explanations identify image regions contributing most strongly to model predictions and provide insight into the internal reasoning process of the deep-learning model, while the computational performance comparison highlights the practical implications of selecting different explainability methods for deployment in resource-constrained transportation applications [34, 35]. The presented results illustrate how explainability can be incorporated into practical transportation applications and demonstrate an important example of Edge XAI in intelligent transportation systems [16].

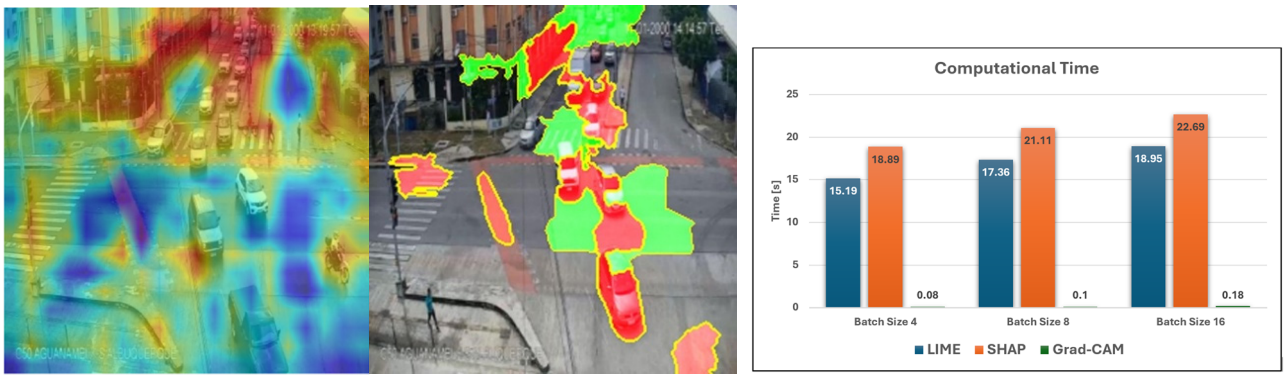


Figure 4: Grad-CAM heatmaps and LIME superpixel explanations illustrating key decision-making regions (left) and computational time vs batch size for SHAP, LIME and Grad-CAM (right) [16]

4.2. Real-Time Explainable Computer Vision for Intelligent Transportation Systems

Computer vision represents one of the most important application domains of artificial intelligence in transportation engineering [36, 37]. Recent developments in deep learning have significantly improved the performance of vehicle detection and classification algorithms, supporting applications such as traffic monitoring, emergency-vehicle recognition and intelligent infrastructure management [38]. In our recent work, YOLO-based deep-learning models were investigated as a foundation for efficient and accurate transportation intelligence systems [21].

Although deep-learning models can achieve high predictive performance, their decision-making processes often remain difficult to interpret [8, 9]. To address this challenge, Grad-CAM visualisation techniques were integrated into the vehicle-recognition framework, enabling visual interpretation of classification outcomes [24]. The resulting explanations facilitate validation of model behaviour and provide additional confidence regarding the correctness of automated decisions.

The left side of Figure 5 illustrates YOLO-based vehicle detection and classification results, whereas the right side of Figure 5 presents the corresponding Grad-CAM visual explanations highlighting the image regions contributing most strongly to the classification decisions. The results demonstrate that explainability can be incorporated into real-time transportation applications without compromising the practical usability of the system. The presented approach therefore represents a practical implementation of Edge XAI capable of combining predictive performance with transparent and interpretable decision-making [21].



Figure 5: (a) YOLO model annotated emergency and non emergency vehicles and (b) Grad-CAM generated heatmap illus-trating key decision-making regions [17]

4.3. Embedded Edge XAI Platforms for Transportation Applications

The practical deployment of intelligent transportation systems requires computational platforms capable of supporting real-time inference under constraints related to energy consumption, hardware cost and operational reliability [6, 39]. Consequently, increasing attention has been devoted to embedded AI platforms that enable local processing and low-latency operation directly at the network edge [5, 6].

To investigate the practical feasibility of Edge XAI deployment, our recent research compared two representative embedded AI platforms: NVIDIA Jetson Orin Nano and Raspberry Pi 5 equipped with the Hailo-8L AI accelerator [17]. Both platforms were evaluated with respect to deep-learning inference performance, explainability support and suitability for intelligent transportation applications. The comparison provides valuable insight into the trade-offs associated with different deployment strategies for transportation intelligence systems.

Figure 6 presents the computational performance of the investigated platforms when executing the same deep-learning model with integrated explainability mechanisms. The presented results clearly indicate that platform selection depends strongly on application requirements. While Jetson-based solutions provide superior computational performance and native support for advanced AI frameworks, Raspberry Pi–Hailo systems offer attractive cost and energy-efficiency characteristics [17].

The comparative evaluation further demonstrated that NVIDIA Jetson Orin Nano generally provides higher computational performance and a more integrated CUDA/TensorRT software environment for DL + XAI transportation tasks, particularly when explainability mechanisms are incorporated into the processing pipeline. In contrast, Raspberry Pi 5 combined with the Hailo-8L accelerator offers lower power consumption, lower cost and attractive deployment flexibility, while still maintaining real-time capability for the investigated applications. Although the two platforms differ in computational characteristics, both produced meaningful and visually consistent Grad-CAM explanations, confirming their suitability for practical Edge XAI deployment in intelligent transportation systems. The inclusion of explainability mechanisms further influences overall system behaviour, highlighting the importance of simultaneously considering performance, transparency and deployment constraints.

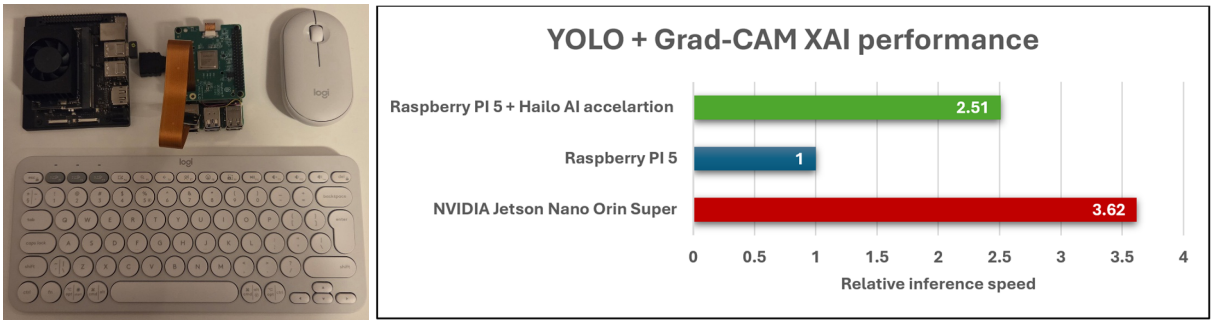


Figure 6: Experimental setup with two Edge AI devices (left) and computational performance of the same model with XAI integrated on two Edge AI devices, average from multiple runs and batch sizes (right) [17]

The presented transportation case studies demonstrate how Edge XAI can support intelligent transportation systems that are not only accurate and computationally efficient but also transparent and trustworthy. The combination of explainable deep learning, embedded deployment and practical performance evaluation illustrates a comprehensive framework for developing next-generation transportation solutions. Furthermore, the presented studies confirm that trustworthy transportation intelligence depends not only on advances in artificial intelligence algorithms but also on the ability to understand, validate and effectively deploy such algorithms in real-world operating environments. These observations reinforce the central thesis of this paper that Edge XAI represents a practical pathway towards trustworthy edge intelligence in transportation engineering.

5. CONCLUSIONS AND FUTURE DIRECTIONS

The presented industrial-control and robotics case studies demonstrate that Edge XAI can successfully support a diverse range of engineering applications, including industrial quality control, intelligent control systems and gesture-based human–robot interaction [18, 19, 20]. The investigated solutions showed that intelligent systems deployed on resource-constrained platforms can simultaneously achieve satisfactory performance and provide interpretable insights into their decision-making processes. The obtained results confirm the practical feasibility of integrating explainability into embedded intelligent systems operating in real-world industrial and robotic environments. Furthermore, the presented studies illustrate that Edge AI technologies can extend advanced machine-learning capabilities beyond cloud-based infrastructures and enable trustworthy intelligent operation directly at the point of use [5, 27].

The transport-engineering case studies further confirmed the potential of Edge XAI for intelligent transportation systems [16, 17, 21]. The combination of deep-learning-based vehicle recognition, Grad-CAM explainability and deployment on embedded AI platforms demonstrated that modern transportation solutions can benefit from both high-performance computer vision and transparent model behaviour. Comparative evaluation of different hardware architectures additionally highlighted the importance of balancing computational performance, energy efficiency and explainability requirements when designing practical transportation intelligence systems [17]. The presented results indicate that trustworthy transportation intelligence increasingly depends not only on predictive accuracy but also on the ability to understand and validate automated decisions in real-world operating conditions [16, 21].

A common observation emerging from all presented case studies is that explainability becomes increasingly important as artificial intelligence moves from laboratory environments towards practical deployment. While Edge AI provides the computational framework necessary for real-time operation [5, 6], explainability contributes transparency, interpretability and user trust [8, 10]. The convergence of these paradigms, referred to in this paper as Edge Explainable Artificial Intelligence (Edge XAI), therefore represents a practical framework for achieving trustworthy edge intelligence across diverse engineering domains. The presented examples demonstrate that the integration of explainability into edge-deployed intelligent systems is not merely a theoretical concept but a practical engineering approach capable of supporting trustworthy decision-making in industrial, robotic and transportation applications [16, 17, 18, 19, 20, 21].

Future research is expected to focus on increasingly integrated Edge XAI solutions capable of providing real-time explanations while operating under stringent resource constraints. Particular attention is likely to be devoted to human-centred intelligent systems, collaborative robotics, adaptive transportation infrastructures and emerging Industry 5.0 environments in which intelligent technologies operate in close cooperation with human users [1, 29]. Advances in embedded AI hardware, TinyML frameworks and explainability methods are expected to further expand the practical applicability of Edge XAI across engineering domains [6, 5, 27]. In this context, trustworthy artificial intelligence is likely to evolve from a desirable system characteristic into a fundamental design requirement, supporting the development of intelligent engineering systems capable of combining autonomy, transparency and human trust [1, 10]. Such developments are closely aligned with the vision of engineering today and the continuing transformation of engineering practice through intelligent technologies.

ACKNOWLEDGEMENT

This work has been supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, Grant Number 451-03-33/2026-03/200109.

REFERENCES

[1] C. Trivedi et al., “Explainable AI for Industry 5.0: Vision, Architecture, and Potential Directions”, IEEE Open Journal of Industry Applications, Vol. 5, pp. 149-159, https://doi.org/10.1109/OJIA.2024.3399057, (2024)

[2] I. Ahmed, G. Jeon, and F. Piccialli, “From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where”, IEEE Transactions on Industrial Informatics, Vol. 18(8), pp. 5031-5042, https://doi.org/10.1109/TII.2022.3146552, (2022)

[3] A. Saranya and R. Subhashini, “A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends”, Decision Analytics Journal, Vol. 7, p. 100230, https://doi.org/10.1016/j.dajour.2023.100230, (2023)

[4] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, “A review of trustworthy and explainable artificial intelligence (XAI)”, IEEE Access, Vol. 11, pp. 78994-79015, https://doi.org/10.1109/ACCESS.2023.3294569, (2023)

[5] R. Singh and S. S. Gill, “Edge AI: A survey”, Internet of Things and Cyber-Physical Systems, Vol. 3, pp. 71-92, https://doi.org/10.1016/j.iotcps.2023.02.004, (2023)

[6] I. Abdulrashid, R. Z. Farahani, S. Mammadov, M. Khalafalla, and W. C. Chiang, “Explainable artificial intelligence in transport logistics: Risk analysis for road accidents”, Transportation Research Part E: Logistics and Transportation Review, Vol. 186, p. 103563, https://doi.org/10.1016/j.tre.2024.103563, (2024)

[7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models”, ACM Computing Surveys, Vol. 51(5), pp. 1-42, https://doi.org/10.1145/3236009, (2018)

[8] L. H. Gilpin, D. Bau, B. Z. Yuan, J. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning for trustworthy systems”, Proceedings of IEEE International Conference on Data Science and Advanced Analytics, Turin (Italy), pp. 80-89, https://doi.org/10.1109/DSAA.2018.00018, (2018)

[9] L. Longo et al., “Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions”, Information Fusion, Vol. 106, p. 102301, https://doi.org/10.1016/j.inffus.2024.102301, (2024)

[10] K. Kalasampath, K. N. Spoorthi, S. Sajeev, S. S. Kuppa, K. Ajay, and M. Angulakshmi, “A Literature review on applications of explainable artificial intelligence (XAI)”, IEEE Access, Vol. 13, pp. 31412-31435, https://doi.org/10.1109/ACCESS.2025.3546681, (2025)

[11] N. Tahir and R. Parasuraman, “Edge Computing and Its Application in Robotics: A Survey”, Journal of Sensor and Actuator Networks, Vol. 14(4), p. 65, https://doi.org/10.3390/jsan14040065, (2025)

[12] M. T. Islam, K. Sepanloo, S. Woo, S. H. Woo, and Y. J. Son, “A Review of the Industry 4.0 to 5.0 Transition: Exploring the Intersection, Challenges, and Opportunities of Technology and Human–Machine Collaboration”, Machines, Vol. 13(4), p. 267, https://doi.org/10.3390/machines13040267, (2025)

[13] A. Kuznietsov et al., “Explainable AI for Safe and Trustworthy Autonomous Driving: A Systematic Review”, IEEE Transactions on Intelligent Transportation Systems, Vol. 25(12), pp. 19342-19364, https://doi.org/10.1109/TITS.2024.3411089, (2024)

[14] D. Sobrín-Hidalgo, Á. M. Guerrero-Higueras, and V. Matellán-Olivera, “Generating Explanations for Autonomous Robots: a Systematic Review”, IEEE Access, Vol. 13, pp. 2541-2563, https://doi.org/10.1109/ACCESS.2024.3521105, (2025)

[15] N. Marković, E. Ćojbašić, D. Rančić, N. Dučić, and Ž. Ćojbašić, “Explainable Artificial Intelligence in Industrial Automation and Robotics: State, Prospects and a Case Study”, Proceedings of ICPES 2025, (2025)

[16] Ž. Ćojbašić, G. Petrović, D. Rančić, E. Ćojbašić, N. Marković, and B. Lukić, “Explainable Artificial Intelligence for Trustworthy Edge-Based Road Intersection Management”, Proceedings of the Tenth International Conference Transport and Logistics (TIL 2025), Niš (Serbia), (2025)

[17] E. Ćojbašić, M. Despenić, D. Rančić, D. Stojanović, I. Ćirić, and Ž. Ćojbašić, “Performance Comparison of Explainable Artificial Intelligence Implementation on Two Edge Devices for a Transport Engineering Problem”, Proceedings of the Tenth International Conference Transport and Logistics (TIL 2025), Niš (Serbia), (2025)

[18] E. Ćojbašić, D. Stojanović, D. Rančić, N. Stojanović, and Ž. Ćojbašić, “Explainable AI and Edge AI Integration for Hand Gesture Recognition in Industrial Robotics”, Proceedings of MIEL 2025, (2025)

[19] E. Ćojbašić, D. Rančić, A. Jovanović, V. D. Sibinović, and Ž. Ćojbašić, “Comparison of Different TinyML Implementations for an Industrial Control System”, Proceedings of MIEL 2025, (2025)

[20] E. Ćojbašić, A. Jovanović, V. D. Sibinović, and Ž. Ćojbašić, “TinyML and IIoT Based Product Quality Classification for Food Industry”, Proceedings of ICEST 2025, (2025)

[21] M. Despenić, D. Rančić, D. Stojanović, I. Ćirić, E. Ćojbašić, and Ž. Ćojbašić, “Real-Time Explainable Artificial Intelligence for Intelligent Transportation Systems: YOLO and Grad-CAM on Embedded Platforms”, Proceedings of ICIST 2026, (2026)

[22] B. Lukić, G. Petrović, Ž. Ćojbašić, D. Marinković, and S. Dimić, “Hybrid Model-Based Framework for Real-Time Adaptive Traffic Signal Control”, Future Transportation, Vol. 6(3), p. 100, https://doi.org/10.3390/futuretransp6030100, (2026)

[23] T. Wang, J. Guo, B. Zhang, G. Yang, and D. Li, “Deploying AI on Edge: Advancement and Challenges in Edge Intelligence”, Mathematics, Vol. 13(11), p. 1878, https://doi.org/10.3390/math13111878, (2025)

[24] H. Zhang and K. Ogasawara, “Grad-CAM-Based Explainable Artificial Intelligence Related Medical Text Processing”, Bioengineering, Vol. 10(9), p. 1070, https://doi.org/10.3390/bioengineering10091070, (2023)

[25] M. S. Ahmmed, S. P. Isanaka, and F. Liou, “Promoting synergies to improve manufacturing efficiency in industrial material processing: A systematic review of industry 4.0 and AI”, Machines, Vol. 12(10), p. 681, https://doi.org/10.3390/machines12100681, (2024)

[26] N. Marković, M. Milošević, I. Ćirić, E. Ćojbašić, and N. Ivačko, “ANFIS Based Explainable AI Approach for Industrial Automation in the Food Industry”, Proceedings of the ICIST 2024, (2024)

[27] V. Tsoukas et al., “A Review on the emerging technology of TinyML”, ACM Computing Surveys, Vol. 56(10), pp. 1-37, https://doi.org/10.1145/3651583, (2024)

[28] S. S. Gill et al., “Edge AI: A taxonomy, systematic review and future directions”, Cluster Computing, Vol. 28(1), pp. 1-53, https://doi.org/10.1007/s10586-024-04312-w, (2025)

[29] O. Vermesan, V. Piuri, F. Scotti, A. Genovese, R. Donida Labati, and P. Coscia, “Explainability and Interpretability Concepts for Edge AI Systems”, in Advancing Edge Artificial Intelligence, River Publishers, pp. 63-92, (2024)

[30] N. Ivačko, I. Ćirić, Ž. Ćojbašić, M. Milošević, and D. Jevtić, “Machine Learning Based Hardware Accelerated Cobot Grasping in the Food Industry”, Proceedings of 2024 11th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN), (2024)

[31] N. Surantha and N. Sutisna, “Key Considerations for Real-Time Object Recognition on Edge Computing Devices”, Applied Sciences, Vol. 15(13), p. 7533, https://doi.org/10.3390/app15137533, (2025)

[32] S. Khan et al., “Towards Transparent Traffic Solutions: Reinforcement Learning and Explainable AI for Traffic Congestion”, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 16(1), pp. 503-511, https://doi.org/10.14569/IJACSA.2025.0160152, (2025)

[33] X. Liu et al., “From Black Box to Glass Box: A Practical Review of Explainable Artificial Intelligence (XAI)”, AI, Vol. 6(1), pp. 285-312, https://doi.org/10.3390/ai6010014, (2025)

[34] G. Liu et al., “Smart Traffic Monitoring System Using Computer Vision and Edge Computing”, IEEE Transactions on Intelligent Transportation Systems, Vol. 23(8), pp. 12027-12038, https://doi.org/10.1109/TITS.2021.3124233, (2022)

[35] A. Abusitta, M. Q. Li, and B. C. Fung, “Survey on Explainable AI: Techniques, challenges and open issues”, Expert Systems with Applications, Vol. 255, p. 124710, https://doi.org/10.1016/j.eswa.2024.124710, (2024)

[36] J. Wei et al., “A Review of YOLO Algorithm and Its Applications in Autonomous Driving Object Detection”, IEEE Access, Vol. 13, pp. 41102-41125, https://doi.org/10.1109/ACCESS.2025.3573376, (2025)

[37] D. K. Alqahtani, M. A. Cheema, and A. N. Toosi, “Benchmarking Deep Learning Models for Object Detection on Edge Computing Devices”, Service-Oriented Computing (ICSOC 2024), Vol. 15404, pp. 115-130, (2025)

[38] R. Santos, M. C. Silva, and R. A. Rabelo, “Real-time object detection performance analysis using YOLOv7 on edge devices”, IEEE Latin America Transactions, Vol. 22(10), pp. 799-805, https://doi.org/10.1109/TLA.2024.1071239, (2024)

[39] T. Gong, L. Zhu, L. Zhu, F. R. Yu, and T. Tang, “Edge Intelligence in Intelligent Transportation Systems: A Survey”, IEEE Transactions on Intelligent Transportation Systems, Vol. 24(9), pp. 8919-8944, https://doi.org/10.1109/TITS.2023.3283101, (2023)

PLENARY SESSION

From data scarcity to open-set generalization: A unified framework for robust rotating machinery fault diagnosis

Vladimir Stojanovic[1*], Hongfeng Tao[2], Xiaodi Li[3,4]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering, Kraljevo, Serbia
[2] Jiangnan University, Key Laboratory of Advanced Process Control for Light Industry of Ministry of Education, Wuxi 214122, China
[3] Shandong Normal University, School of Mathematics and Statistics, Ji'nan 250014, PR China
[4] Shandong Normal University, Shandong Engineering Research Center of System Control and Intelligent Technology, Ji'nan 250358, PR China

ARTICLE INFO

[*] **Correspondence:** vladostojanovic@mts.rs

DOI: https://doi.org/10.46793/ET26.P02S **PAGES:** P11 – P22

ABSTRACT

Unexpected failures in rotating machinery cause substantial economic losses and safety hazards in industrial systems, underscoring the practical need for reliable condition-based monitoring. Intelligent fault diagnosis must simultaneously address three real-world constraints in Industrial IoT deployments: severe data scarcity, dynamic distribution shifts, and the presence of previously unseen fault categories. This paper reviews a decade of methodological progress (2016–2026), tracing the development from unsupervised clustering and few-shot meta-learning to open-set single-source domain generalization. We propose a unified six-tier taxonomy classifying methods into data-level synthesis, metric/meta-learning, domain adaptation, domain generalization, open-set recognition, and an emerging category covering attention-based and self-supervised approaches. We examine the integration of wavelet-guided generative models with adversarial and prototype-aware alignment objectives, contextualize convergent findings on attention mechanisms and self-supervised pre-training that corroborate the role of structural frequency priors, and show that purely distributional alignment is insufficient under open-set conditions without latent-space structural constraints. Five open problems are identified: OS-SSDG, calibrated conformal thresholding, physics-informed class completion, zero-shot disentanglement, and foundation-model pre-training for IIoT. We conclude that Conformal Prediction combined with wavelet-structured representations offers a principled path toward bounded, interpretable fault diagnosis under realistic industrial constraints.

KEYWORDS

Intelligent fault diagnosis, Domain generalization, Open-set recognition, Wavelet packet transform, Generative models, Transfer learning

1. INTRODUCTION

Rotating machinery constitutes a critical subsystem in manufacturing, energy, and transportation infrastructure. Unplanned failures lead to significant operational disruptions and safety hazards [9-10]. While deep learning has transformed fault diagnosis through automatic feature extraction [14, 15, 19], real-world Industrial Internet of Things (IIoT) deployments differ fundamentally from controlled laboratory settings: datasets are severely imbalanced, operating conditions are non-stationary, and new failure modes emerge continuously throughout equipment lifecycles.

The development of intelligent fault diagnosis over the past decade (2016–2026) shows a consistent progression from supervised approaches that require large, perfectly labeled datasets toward methods that operate under increasing practical constraints. Early convolutional architectures from 2016–2018 demonstrated end-to-end learning on raw vibration signals [14-16, 19]. By 2020–2022 the community had largely moved toward transfer learning and domain adaptation to bridge the gap between training and deployment conditions [25, 26, 39]. Current research addresses domain generalization and open-set recognition, where neither target domain data nor a complete fault catalog is available at training time.

Practical deployment of fault diagnosis models involves three interrelated difficulties. First, fault instances are rare and annotation is costly, creating a persistent data scarcity problem. Second, variable loads, speeds, and sensor configurations produce distribution shifts that degrade models trained under fixed conditions. Third, field operation continuously introduces fault modes absent from any training catalog, requiring open-world recognition capability. Recent surveys confirm that each challenge is tractable in isolation, but their simultaneous occurrence requires unified treatment [11-13].

This paper traces the methodological evolution across these three axes, connecting data augmentation strategies, domain adaptation mechanisms, and open-set recognition frameworks into a coherent account of ten years of progress. We propose a six-tier taxonomy and identify five open problems, OS-SSDG, calibrated conformal thresholding, physics-informed class completion, zero-shot disentanglement, and foundation-model pre-training, as the agenda that follows from the surveyed decade.

Unlike prior surveys that treat data scarcity, distribution shift, and open-set recognition separately [39, 40], this paper provides a unified account that maps methodological transitions, identifies the failure modes of each approach under realistic conditions, and outlines concrete research directions. The central argument is that structural constraints, physical signal priors, prototype-based objectives, and conformal coverage guarantees, are necessary complements to purely statistical alignment in robust fault diagnosis.

The paper proceeds as follows. Section 2 formalizes the problem setting and presents the three-era historical evolution. Section 3 introduces the six-tier methodological taxonomy. Section 4 provides comparative analysis and performance benchmarking. Section 5 identifies five open problems. Section 6 concludes.

2. PROBLEM FORMULATION AND HISTORICAL EVOLUTION

2.1. Formal Problem Formulation

We formally define the four core paradigms addressed in this review. Let $\mathcal{X}$ denote the input space and $\mathcal{Y}$ the label space.

Definition 1 (Domain Adaptation (DA)). *Given a labeled source domain $\mathcal{D}_S = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and an unlabeled target domain $\mathcal{D}_T = \{x_j^t\}_{j=1}^{N_t}$, where marginal distributions differ ($P_S(X) \neq P_T(X)$). Under the standard covariate shift assumption, the conditional distributions are assumed invariant ($P_S(Y|X) = P_T(Y|X)$), and label spaces are identical ($\mathcal{Y}_S = \mathcal{Y}_T$). DA seeks a hypothesis $h: \mathcal{X} \to \mathcal{Y}$ that minimizes the target risk $\varepsilon_T(h)$ by leveraging $\mathcal{D}_T$ during training. (Note: Recent extensions address concept drift where $P_S(Y|X) \neq P_T(Y|X)$ [32].)*

Definition 2 (Domain Generalization (DG)). *Given only M labeled source domains $\{\mathcal{D}_{S_m}\}_{m=1}^{M}$, DG seeks a hypothesis h such that the risk $\varepsilon_{T_k}(h) \leq \epsilon$ for all unseen target domains $\mathcal{D}_{T_k}$ (bounded by the source domain risks and the $\mathcal{H}\Delta\mathcal{H}$-distance between domains [28]), without any access to target data during training. When $M = 1$ it is termed Single-Source DG (SSDG).*

Definition 3 (Open-Set Recognition (OSR)). *Given K known classes, OSR requires correct classification of known instances while rejecting instances from K^* unknown classes ($\mathcal{Y}_{known} \cap \mathcal{Y}_{unknown} = \emptyset$). The decision rule incorporates a rejection function based on a scoring function $s(x)$: $h(x) =$ "unknown" if $s(x) < \tau$, where τ is a calibrated threshold. This formulation avoids strict reliance on Maximum Softmax Probability (MSP) heuristics [37].*

Definition 4 (Open-Set Single-Source Domain Generalization (OS-SSDG)). *Given a single labeled source domain ($M = 1$) with known classes $\mathcal{Y}_{known}$, OS-SSDG seeks a hypothesis h such that: (a) the risk on all unseen target domains remains bounded by a theoretical margin, and (b) for any instance x from unknown classes $\mathcal{Y}_{unknown}$, $h(x) =$ "unknown" with a target coverage probability of $1 - \alpha$. **Note:** Achieving the marginal coverage guarantee $\Pr[h(x) =$ "unknown"$] \geq 1 - \alpha$ via Conformal Prediction requires a held-out calibration set that satisfies the exchangeability assumption [41]. In a strictly offline single-source setting, a source-domain proxy calibration set must be used, and the guarantee is conditional on the degree of domain similarity, see Remark 1 for a discussion.*

2.2. Era I: Foundations of Deep Learning for Fault Diagnosis (2016–2020)

The period 2016–2020 was defined by the transition from handcrafted feature pipelines to end-to-end deep learning on raw or lightly preprocessed vibration signals. One-dimensional convolutional neural networks demonstrated that raw waveforms could be processed directly without prior signal conditioning, outperforming traditional spectrum analysis baselines [14, 15]. Hierarchical feature learning via deep belief networks and stacked autoencoders established that deep architectures could capture fault-relevant representations from unlabeled data [16, 17]. The introduction of multi-scale convolutional kernels and data-adaptive preprocessing pipelines [19] further improved robustness to variable shaft speeds and sensor mounting positions, a practical concern that purely academic benchmarks had obscured.

Despite these advances, all supervised architectures in this period required large, balanced, labeled datasets that are rarely available in practice. Bearing fault events occur infrequently, and acquiring labeled data across all failure modes involves deliberate fault seeding, an expensive and operationally disruptive process. This observation motivated two parallel research threads in the second half of Era I: generative models to synthesize fault data from limited examples, and few-shot learning to generalize from small support sets.

The integration of Short-Time Fourier Transform (STFT) with Categorical GANs enabled end-to-end unsupervised clustering of fault patterns without label requirements [11]. The STFT provides localized frequency representation:

$$\mathrm{STFT}\{x(t)\}(t,\omega) = \int_{-\infty}^{\infty} x(s)\, g(s-t)\, e^{-j\omega s}\, ds \tag{1}$$

where $g(t)$ is the analysis window. The CatGAN objective minimizes conditional entropy for real samples while maximizing it for generated samples:

$$\mathcal{L}_C = \max_C \Big(H_{p_{data}(x)}[p(y \mid C)] - \mathbb{E}_{x\sim p_{data}}[H[p(y \mid x, C)]] + \mathbb{E}_{z\sim p(z)}[H[p(y \mid G(z), C)]] \Big) \tag{2}$$

Concurrently, meta-learning frameworks emerged to address cross-equipment diagnosis with minimal support sets. Building on Model-Agnostic Meta-Learning (MAML) [18] and Prototypical Networks [20], Model-Agnostic Matching Networks (MAMN) were introduced as a hybrid that applies MAML's bi-level parameter optimization at the outer loop while using cosine-similarity-based class assignment at the inner loop [2]:

$$\theta_i' = \omega - \alpha \nabla_\omega \ell(\omega), \quad \omega \leftarrow \omega - \beta \sum_{\tau_i} \nabla_\omega \ell(\tau_i; \theta_i', \omega) \tag{3}$$

with classification based on cosine similarity:

$$P(y = k|\boldsymbol{x}) = \frac{\exp(\lambda \cos(\boldsymbol{f}_\theta(\boldsymbol{x}), \boldsymbol{c}_k))}{\sum_{j=1}^{K} \exp(\lambda \cos(\boldsymbol{f}_\theta(\boldsymbol{x}), \boldsymbol{c}_j))} \tag{4}$$

While these approaches substantially reduced reliance on large labeled datasets, meta-learned knowledge degraded rapidly when equipment dynamics changed substantially, because the metric space was implicitly fitted to source-domain statistics. Alongside meta-learning, early explorations of self-supervised representation learning [9] showed that meaningful fault features could be acquired through signal reconstruction or augmentation-invariance objectives rather than supervised class boundaries, establishing a data-efficient alternative that gained momentum in Era II. The common lesson of both threads was that structural inductive biases, wavelet decompositions, metric spaces, and pretext-task constraints, are more sample-efficient than purely supervised feature learning, a principle that unified the field's subsequent progression. This limitation identified the need for explicit mechanisms that do not require any target data, motivating the domain adaptation paradigm of Era II.

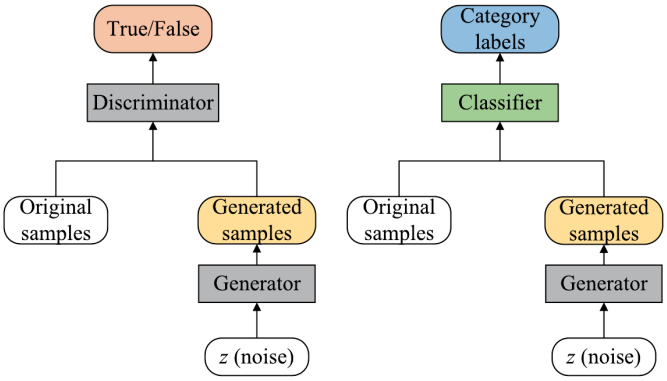


Figure 1: Architecture of STFT-CatGAN (Era I). (Reproduced from [1], Copyright 2020 Elsevier, with permission.)

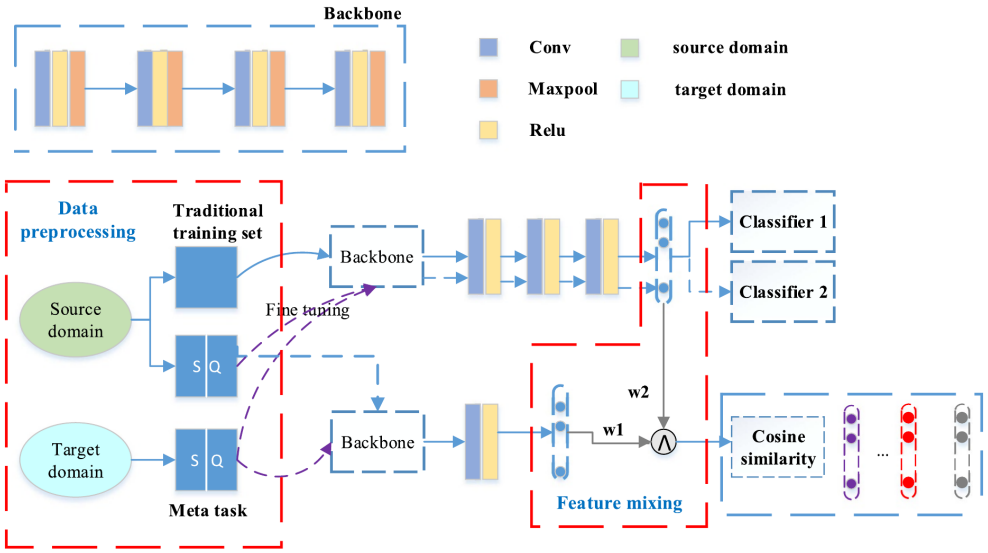


Figure 2: Architecture of Model-Agnostic Matching Network (MAMN), combining MAML-style bi-level parameter optimization with cosine-similarity metric learning [2, 18]. (Reproduced from [2], Copyright 2022, IOP Publishing, with permission.)

2.3. Era II: Generative Augmentation and Adversarial Domain Adaptation (2020–2024)

To overcome the fragility of Era I approaches, the field moved toward Domain Adaptation (DA) coupled with generative augmentation. A key development was the introduction of Cascaded Stochastic Quantization VAEs (CSQ-VAE) guided by Wavelet Packet Transform (WPT) [3]. WPT recursively decomposes a signal via a two-channel filter bank at each node (j, n):

$$W_{j+1,2n}(t) = \sqrt{2}\sum_{k} h(k) W_{j,n}(2t-k)$$
$$W_{j+1,2n+1}(t) = \sqrt{2}\sum_{k} g(k) W_{j,n}(2t-k) \tag{5}$$

where $W_{j,n}$ is the wavelet packet function at decomposition level j and node n, and $h(k)$, $g(k)$ are the low-pass and high-pass filter coefficients, respectively.

The CSQ-VAE loss combines reconstruction, codebook, and commitment terms:

$$\mathcal{L}_{\text{VQ}} = \log p(\mathbf{x} \mid \mathbf{z}_q(\mathbf{x})) + \|\text{sg}[\mathbf{z}_e(\mathbf{x})] - \mathbf{e}\|_2^2 + \beta \|\mathbf{z}_e(\mathbf{x}) - \text{sg}[\mathbf{e}]\|_2^2 \tag{6}$$

Concurrent work explored conditional GANs for minority-class oversampling [21, 22] and denoising diffusion probabilistic models for high-fidelity signal synthesis [23, 24]. The comparative advantage of WPT-guided VAEs over standard GAN approaches lies in their ability to explicitly preserve the multi-scale frequency structure of bearing vibration, where fault signatures manifest as periodic impulses at specific sub-band frequencies.

Standard adversarial DA was refined to prevent decision boundary blurring, a known failure mode of DANN [25] and DAN [26] under class imbalance. Pseudo-label Guided Dual-Classifier Domain Adversarial Networks (PG-DCDAN) addressed this through two independent classifiers and multi-criterion distribution alignment [4]:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} - \alpha \mathcal{L}_{\text{adv}} + \beta\left(\text{MK-MMD}(\mathcal{D}_S, \mathcal{D}_T) + \mathcal{L}_{\text{CORAL}}\right) + \gamma \mathcal{L}_{\text{plg}} \tag{7}$$

The target error bound formalizes the risk of negative transfer [28]:

$$\varepsilon_T(h) \leq \varepsilon_S(h) + d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda^* \tag{8}$$

where λ^* is the shared error of the ideal joint hypothesis. PG-DCDAN keeps λ^* small through dual-classifier discrepancy minimization [27].

A parallel direction, source-free DA [29], removes the need for source data at adaptation time but still assumes access to unlabeled target data, a constraint that remains problematic in safety-critical IIoT deployments. This fundamental requirement for any target data is not resolvable within the DA paradigm and motivated the shift toward domain generalization.

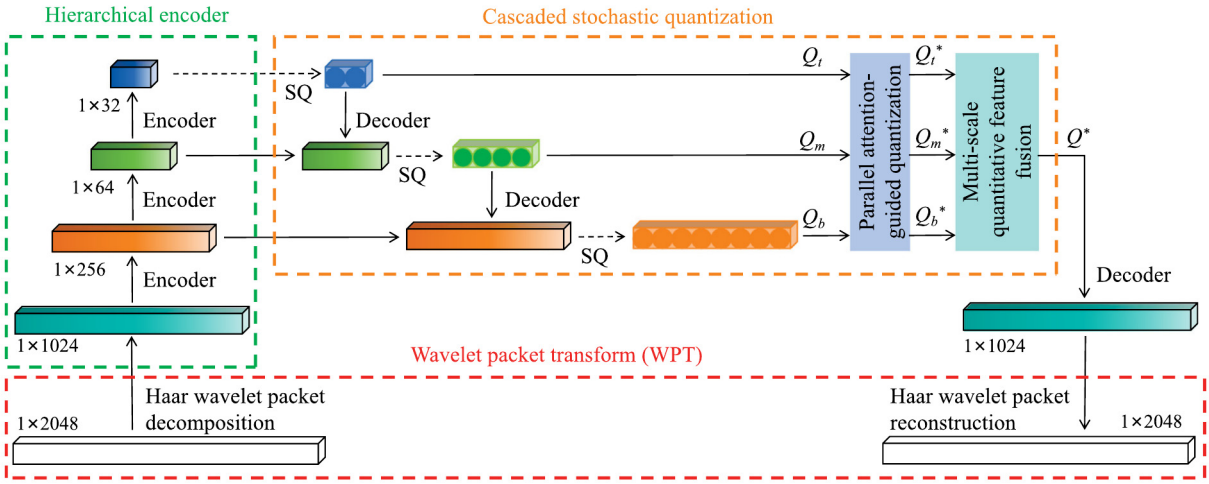
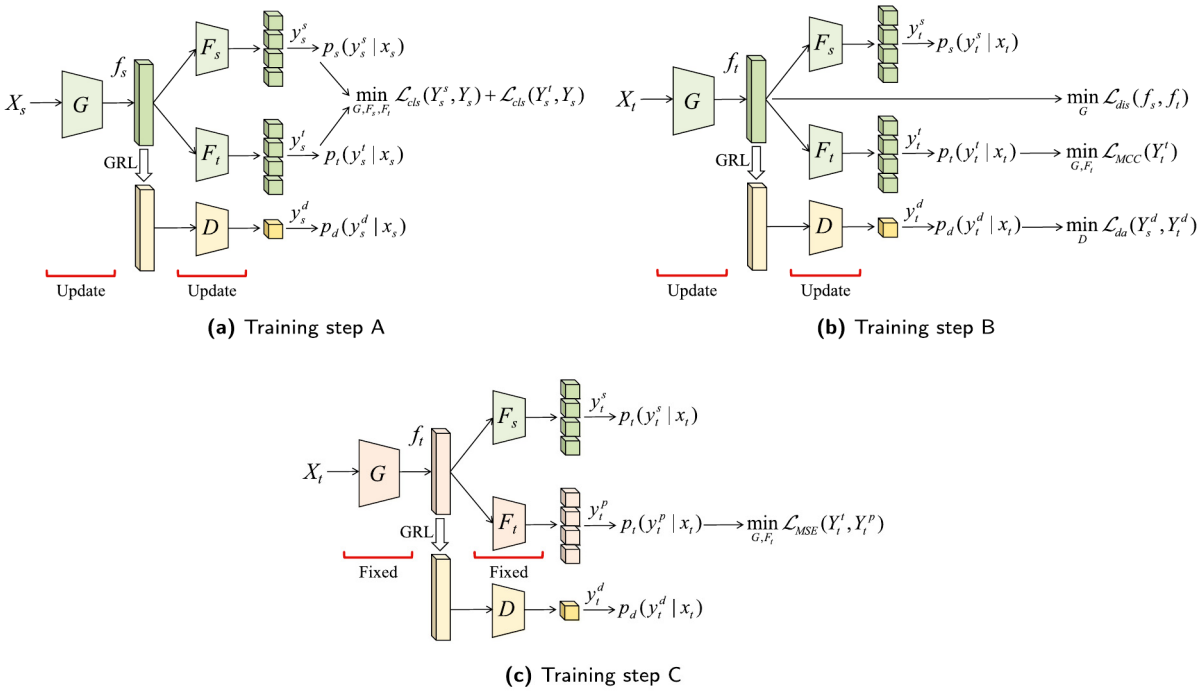


Figure 3: CSQ-VAE framework with cascaded stochastic quantization and wavelet packet guidance. (Reproduced from [3], Copyright 2024 Elsevier, with permission.)



(a) Training step A (b) Training step B (c) Training step C

Figure 4: Training pipeline of PG-DCDAN. (Reproduced from [4], Copyright 2025 Elsevier, with permission.)

2.4. Era III: Domain Generalization and the Open-Set Paradigm (2024–2026)

The requirement to operate without any target data while simultaneously detecting previously unseen fault types represents a setting for which no single existing method provides a complete solution. Recent frameworks address the target-free constraint through Wavelet Packet Augmentation (WPA) and Pseudo-Domain Generation (PDG) [5]. WPA interpolates in the sub-band coefficient space:

$$\begin{aligned} \tilde{x}_i^h &= \lambda x_i^h + (1-\lambda) x_i^{',h} \\ \tilde{x}_i^l &= \lambda x_i^l + (1-\lambda) x_i^{',l} \end{aligned} \tag{9}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$.

PDG complements WPA by constructing synthetic pseudo-target domains. Given WPA-augmented samples $\{\tilde{x}_i^{(c)}\}_{c=1}^K$ from K fault classes, a pseudo-domain $\widetilde{\mathcal{D}}_p$ is formed by drawing convex combinations across classes with independently sampled coefficients $\mu \sim \text{Dirichlet}(\mathbf{1}_K)$:

$$\tilde{x}_p = \sum_{c=1}^K \mu_c \, \tilde{x}_i^{(c)}, \quad \tilde{y}_p = \underset{c}{\text{argmax}}\, \mu_c \tag{10}$$

This populates the latent distribution space between source sub-domains, simulating the style diversity of unseen target machines without requiring any real target data [5].

For scenarios with multiple source domains, Multi-Domain Weakly Decoupled Generalization (MWDG) networks introduce intra-domain variance as a risk penalty following Risk Extrapolation [30]:

$$\mathcal{L}_{\text{MWDG}} = \sum_{k=1}^N \mathcal{L}_{\text{cls}}^k + \beta \cdot \text{Var}(\mathcal{L}_{\text{cls}}^1, \ldots, \mathcal{L}_{\text{cls}}^N) \tag{11}$$

Concurrently, the broader signal-processing community has validated the role of attention mechanisms for vibration-based diagnosis. Deep Residual Shrinkage Networks [43] introduced learnable soft-thresholding as noise-robust sub-band feature selection, demonstrating that structural channel weighting, rather than global distribution alignment, is a primary driver of accuracy gains under noisy industrial conditions. Squeeze-and-excitation channel attention [42], which forms the recalibration module of CA-MCNN, corroborated this by showing consistent intra-class variance reduction across operating conditions. These convergent results from independent research groups confirm that frequency-structured attention provides inductive biases that are complementary to, not competitive with, domain generalization objectives. Self-supervised pre-training frameworks have further extended this direction by learning fault-relevant representations through augmentation-invariance objectives, reducing the reliance on fault labels at the feature-learning stage [9].

For open-set fault diagnosis, Time-Frequency Fusion (TFF) and Latent Representation Prompting (LRP) [7] serve as the feature extraction backbone. The S2-R23 metric measures predictive confidence as the ratio of the sum of squared class probabilities to the sum of the second- and third-largest probabilities:

$$\text{S2-R23}(x) = \frac{\sum_{k=1}^K p_k^2(x)}{p_{(2)}(x) + p_{(3)}(x) + \epsilon} \tag{12}$$

where $\epsilon = 10^{-8}$ prevents division by zero.

Proposition 1 (Monotonicity of S2-R23). *Let $p = [p_1, \ldots, p_K]$ be a probability distribution with $p_1 \geq p_2 \geq \cdots \geq p_K \geq 0$ and $K \geq 3$. Then* S2-R23 *is monotonically decreasing with respect to the Shannon entropy $H(p) = -\sum_k p_k \log p_k$.*

Proof 1. *Consider the uniform residual parametrization: $p_1 = 1 - (K-1)\delta$ and $p_k = \delta$ for $k = 2, \ldots, K$, where $\delta \in [0, 1/K]$. Under this parametrization $p_{(2)} = p_{(3)} = \delta$ and the Shannon entropy $H(\delta) = -(1-(K-1)\delta)\log(1-(K-1)\delta) - (K-1)\delta\log\delta$ is strictly increasing on $[0, 1/K]$ with $H(0) = 0$, $H(1/K) = \log K$.*

Setting $N(\delta) = (1-(K-1)\delta)^2 + (K-1)\delta^2$ and $D(\delta) = 2\delta + \epsilon$, the S2-R23 metric reduces to $S(\delta) = N(\delta)/D(\delta)$. Differentiating:

$S'(\delta) = \frac{N'(\delta)D(\delta) - N(\delta)D'(\delta)}{D(\delta)^2}$ *with $N'(\delta) = 2(K-1)(K\delta - 1)$ and $D'(\delta) = 2$. For $\delta \in [0, 1/K)$ we have $K\delta - 1 < 0$, so $N'(\delta)D(\delta) < 0$, both terms of $N(\delta)$ are non-negative, so $N(\delta)D'(\delta) \geq 0$. Hence the numerator of $S'(\delta)$ is strictly negative, giving $S'(\delta) < 0$ for all $\delta \in (0, 1/K)$. Since $H'(\delta) > 0$ and $S'(\delta) < 0$, the chain rule yields $dS/dH = S'(\delta)/H'(\delta) < 0$, confirming strict monotone decrease for $K \geq 3$. 0* □

Remark 1 (Conformal Prediction Guarantees). *For any significance level $\alpha \in (0, 1)$, a valid Conformal Prediction (CP) framework guarantees marginal coverage $\Pr[y_{n+1} \in \mathcal{C}_n(x_{n+1})] \geq 1 - \alpha$ under the exchangeability assumption. In IIoT scenarios, distribution shifts between the calibration and test sets violate strict exchangeability, necessitating adaptive or weighted CP variants [41] to maintain theoretical coverage bounds.*

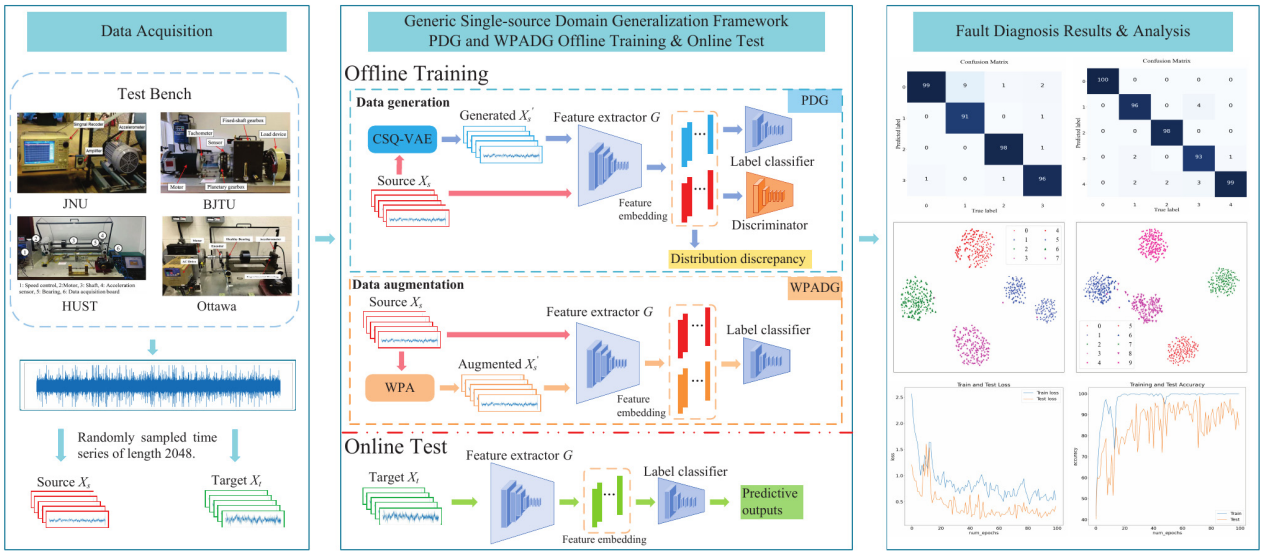


Figure 5: Generic SSDG framework (Era III), utilizing WPA and PDG without target domain data. (Reproduced from [5], Copyright 2025 IEEE, with permission.)

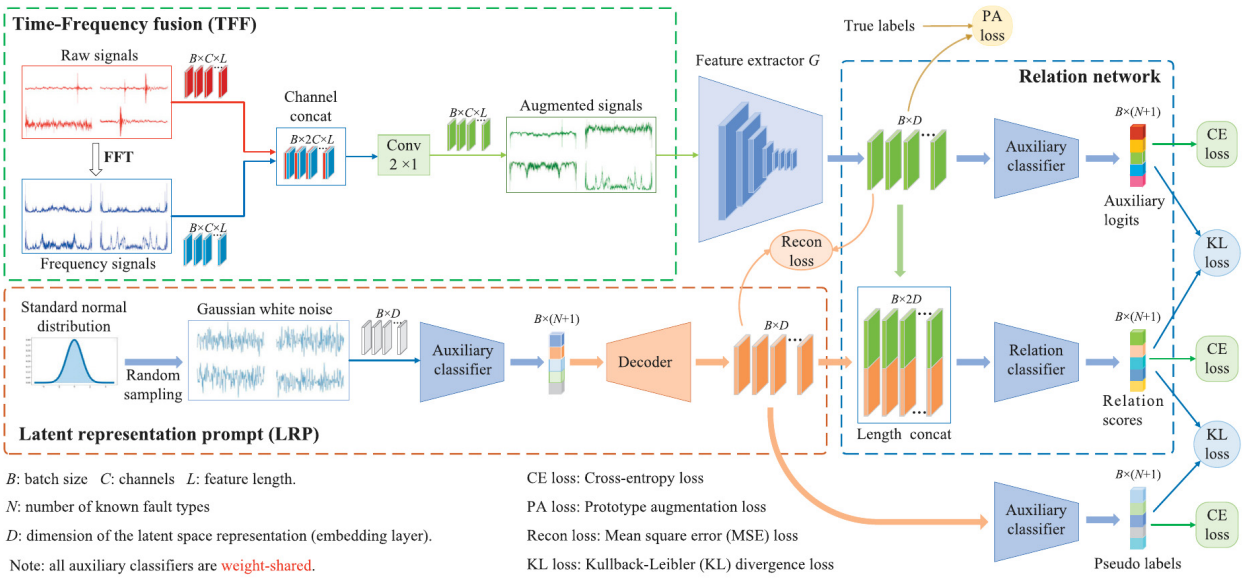


Figure 6: Architecture of LPTF-OSC featuring TFF and LRP for unknown fault rejection. (Reproduced from [7], Copyright 2025 Elsevier, with permission.)

3. METHODOLOGICAL TAXONOMY

Table 1 presents a six-tier taxonomy of methods across the full 2016–2026 decade. The taxonomy deliberately includes representative external works alongside the authors' contributions. Category VI covers attention-based and self-supervised directions that have emerged as an independent research stream since 2020.

Table 1: Comprehensive six-tier taxonomy of intelligent fault diagnosis methodologies (2016–2026)

Category	Sub-Category	Core Mechanism	Representative Works
I. Data-Level Synthesis	Unsupervised Clustering	Adversarial generation of latent representations without labels.	[1], [38], [16]
	Imbalance Augmentation	Autoregressive or wavelet-guided generation of minority-class samples. GAN-based oversampling.	[3], [21], [22]
II. Metric & Meta-Learning	Few-Shot Cross-Equipment	Bi-level parameter optimization (MAML) combined with similarity metric learning.	[2], [18], [20]
III. Domain Adaptation (DA)	Adversarial Alignment	Minimax game between feature extractor G and discriminator D via GRL.	[4], [25], [26]
	Class-Incomplete MDA	Generative class completion followed by prototype-aware conditional alignment.	[8], [33], [34]
IV. Domain Generalization (DG)	Single-Source DG	WPA and PDG without any target data.	[5], [32], [33]
	Multi-Source DG	Weakly decoupled classifiers with intra-domain variance penalty (REx-inspired).	[6], [30], [31]
V. Open-Set Recognition	Score-Based Rejection	TFF and LRP with S2-R23 open-score. OpenMax Weibull tail fitting.	[7], [36], [37]
VI. Attention & Self-Supervised	Attention Architectures	Channel and spatial attention weighting (SE, CBAM). Deep residual shrinkage for noise-robust feature extraction.	[43], [42], [19]
	Self-Supervised Pre-training	Contrastive or masked-signal objectives to reduce reliance on fault labels at feature-learning stage.	[9], [40], [12]

4. CRITICAL SYSNTHESIS AND COMPARATIVE ANALYSIS

4.1. Generative Signal Synthesis for Structural Scarcity

Generative models map noise or latent variables to the data space. For 1D vibration signals, high-frequency transient impulses are easily smoothed by standard architectures, and time-domain generation alone produces mode collapse and spectrally inconsistent samples. WPT-guided generation addresses this by imposing multi-scale frequency constraints on the generative objective, preserving the transient fault signatures that distinguish, for example, outer-race from inner-race bearing faults [3]. Recent results with Denoising Diffusion Probabilistic Models [23, 24] confirm that generation quality improves markedly when time-frequency representations are incorporated as the synthesis target rather than raw waveforms. Convergence across independent research groups toward frequency-domain priors suggests that such physical constraints are a necessary inductive bias.

4.2. Adversarial Alignment under Structural Constraints

DA relies on a minimax game where the feature extractor G minimizes classification loss while maximizing domain adversarial loss via a GRL:

$$\min_{G,C}\max_{D}[\mathcal{L}_{cls}(G,C) - \lambda\mathcal{L}_{adv}(G,D)] \tag{13}$$

Standard global alignment causes negative transfer in class-incomplete multi-source scenarios: forcing global alignment maps target samples of missing classes to the nearest available source class boundary. Generative class completion must therefore precede feature alignment, as conditional alignment requires geometric anchors in the form of valid class prototypes. Maintaining dynamic prototypes via Exponential Moving Average (EMA):

$$P_c \leftarrow \beta P_c + (1-\beta)\mu_c \tag{14}$$

strictly enforces intra-class compactness before global alignment [8].

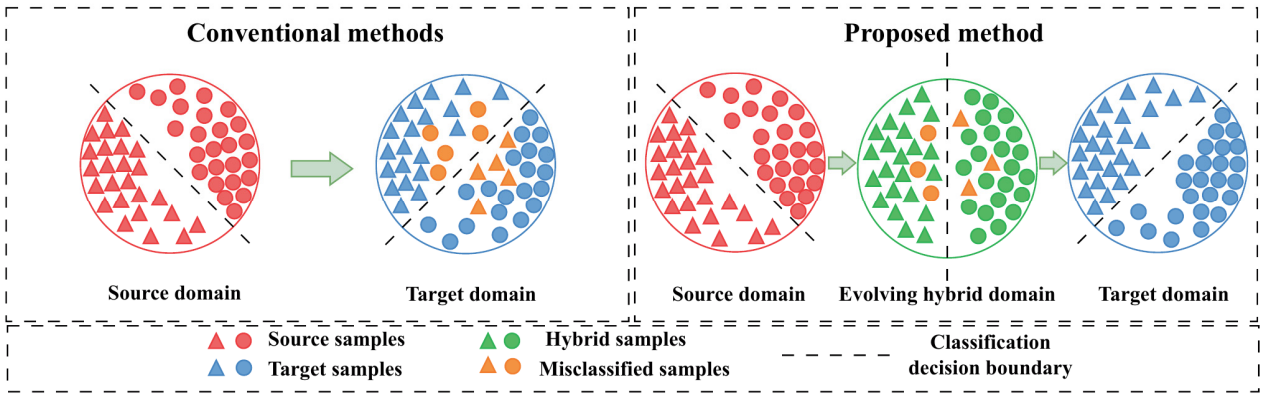


Figure 7: Comparison between conventional abrupt Domain Adaptation and the proposed Dynamic Evolution Mechanism (DEM) for Class-Incomplete MDA. (Reproduced from [8], Copyright 2026 Springer, with permission.)

4.3. Target-Free Generalization and Open-Set Rejection

DG minimizes empirical risk across source domains to enforce invariance. Multi-Source DG models add intra-domain variance as a penalty following Risk Extrapolation [30]:

$$\mathcal{L}_{DG} = \sum_{m=1}^{M} \mathcal{L}_{cls}^{m} + \beta \, \mathrm{Var}(\mathcal{L}_{cls}^{1}, \ldots, \mathcal{L}_{cls}^{M}) \tag{15}$$

For open-set rejection, the CP framework [41] replaces heuristic thresholds with a data-driven nonconformity measure. A natural nonconformity score for fault diagnosis is $A(x, y) = 1 - p_y(x)$, where $p_y(x)$ is the predicted probability of class y. Given a held-out calibration set of n source-domain samples, the empirical $(1 - \alpha)$-quantile of calibration scores $\hat{q}_{1-\alpha}$ yields the prediction set:

$$\mathcal{C}_n(x) = \{y : A(x, y) \leq \hat{q}_{1-\alpha}\} \tag{16}$$

A test sample is rejected as "unknown" when $|\mathcal{C}_n(x)| = 0$, i.e., no class achieves a conformity score above the calibrated threshold. The guarantee $\Pr[y_{n+1} \in \mathcal{C}_n(x_{n+1})] \geq 1 - \alpha$ holds marginally over the joint randomness of the calibration set and test point under the exchangeability assumption. Adaptive CP [41] extends this to mild distribution-shift settings.

4.4. Comparative Analysis and Computational Cost

Backbone architecture (CA-MCNN): All comparative results are obtained using CA-MCNN [5], a backbone comprising three parallel 1D convolutional branches with kernel sizes 16, 32, and 64, a squeeze-and-excitation channel attention module [42] for adaptive feature recalibration, and a two-layer fully-connected classifier. CA-MCNN has approximately 0.08M trainable parameters and a 2.4 MB memory footprint, supporting edge deployment on resource-constrained hardware.

Evaluation metric (H-score): For open-set evaluation the H-score is the harmonic mean of known-class accuracy $\mathrm{Acc}_{\mathrm{kn}}$ and unknown-class detection rate $\mathrm{Acc}_{\mathrm{unk}}$:

$$\text{H-score} = \frac{2 \cdot \mathrm{Acc}_{\mathrm{kn}} \cdot \mathrm{Acc}_{\mathrm{unk}}}{\mathrm{Acc}_{\mathrm{kn}} + \mathrm{Acc}_{\mathrm{unk}}} \tag{17}$$

This metric penalizes methods that achieve high known-class accuracy at the cost of poor novel-fault detection, or vice versa.

Datasets: The JNU dataset (Jiangnan University, 4 fault types at 3 rotational speeds) is used as the unified cross-paradigm benchmark in Table [2] to ensure identical evaluation protocols across all compared methods. Additional results on the CWRU (Case Western Reserve University, 10 fault classes, 4 load conditions) and BJTU (Beijing Jiaotong University, 5 gearbox fault types) benchmarks are reported in the respective source papers [5, 7, 8].

Statistical protocol: Each method is run 5 independent times. Accuracy and H-score are reported as mean ± standard deviation. Improvements over the best competitor are verified by the Wilcoxon signed-rank test ($p < 0.05$).

Table 2: Performance summary and computational cost on the JNU benchmark†

Method	Paradigm	Dataset	Arch.	Runs	Acc (%)	H-score (%)	Time (ms)	Mem (MB)
ERM (Baseline)	SSDG	JNU	CA-MCNN	5	79.26 ± 1.09	78.10 ± 1.20	1.12	1.96
V-REx [30]	SSDG	JNU	CA-MCNN	5	84.03 ± 1.11	83.50 ± 1.15	1.15	1.96
EFDMix [32]	SSDG	JNU	ResNet-1D	5	88.50 ± 0.95	87.90 ± 1.00	1.45	2.10
WPA+PDG [5]	SSDG	JNU	CA-MCNN	5	$\mathbf{92.76 \pm 0.37}$	$\mathbf{91.80 \pm 0.40}$	1.44	2.43

Method	Paradigm	Dataset	Arch.	Runs	Acc (%)	H-score (%)	Time (ms)	Mem (MB)
OpenMax [37]	OSR	JNU	CA-MCNN	5	66.83 ± 1.04	66.83 ± 1.10	1.12	1.96
ARPL [33]	OSR	JNU	CA-MCNN	5	78.57 ± 1.18	76.44 ± 1.25	1.25	2.05
LPTF-OSC [7]	OSR	JNU	CA-MCNN	5	$\mathbf{91.96 \pm 0.31}$	$\mathbf{92.08 \pm 0.35}$	1.50	2.56
DANN‡ [25]	Class-Inc DA	JNU	CA-MCNN	5	96.95 ± 0.24	96.80 ± 0.28	1.71	2.43
ICSC+PCCAN [8]	Class-Inc DA	JNU	CA-MCNN	5	$\mathbf{99.36 \pm 0.18}$	$\mathbf{99.20 \pm 0.20}$	2.18	2.80

† JNU is selected as the unified cross-paradigm benchmark to ensure identical evaluation protocols across all three paradigms. Exhaustive CWRU and BJTU results are reported in the respective source papers [5,7,8]. WPA+PDG improves over ERM by ≥8 pp on CWRU and BJTU under identical CA-MCNN backbone and 5-run protocol. Bold denotes the best result within each paradigm block.

‡ DANN achieves 96.95 % in the Class-Incomplete DA scenario on JNU because this task involves one missing class out of four (25 % incompleteness), where global adversarial alignment retains reasonable class coverage. The performance gap versus ICSC+PCCAN widens substantially (to >6 pp) under more extreme incompleteness settings (3/4 missing classes) as reported in [8].

Computational cost and IIoT scalability: Advanced frameworks introduce marginal overhead during offline training (ICSC generative component: ∼0.04M parameters, 15.36M FLOPs). At inference, all methods share the CA-MCNN backbone (∼0.08M parameters, ∼2.4 MB), supporting real-time deployment on edge devices such as NVIDIA Jetson Nano within IIoT architectures.

5. OPEN PROBLEMS AND FUTURE ROADMAP

The synthesis of three methodological eras reveals four open problems that define the near-term research agenda.

1. Open-Set Single-Source Domain Generalization (OS-SSDG). All existing OSR methods require target data for threshold calibration, and all SSDG methods assume a closed fault catalog. The natural synthesis of these two requirements — open-set rejection in a fully target-free, single-source setting, remains unsolved. The core difficulty is that generative OOD prototype synthesis, which works well with multi-source data [8], becomes ill-posed when only a single source domain is available. A candidate direction is to integrate WPA with a boundary-aware latent margin module that explicitly constructs out-of-distribution regions in feature space without requiring target samples, and to validate the resulting CP coverage guarantee using a source-domain proxy calibration set as described in Definition 4.

2. Provably calibrated rejection thresholds via Conformal Prediction. The S2-R23 metric and Weibull-based OpenMax thresholds are empirically set and degrade under distribution shift. Replacing them with CP-based thresholds gives user-specified coverage at level $1 - \alpha$ with a finite-sample guarantee. The outstanding challenge is that standard CP requires exchangeability between calibration and test points — a condition violated in non-stationary machinery data. Weighted conformal prediction [41] and localized conformal methods offer theoretical frameworks for this, but their computational feasibility on embedded edge hardware has not been studied in the fault diagnosis context.

3. Physics-informed synthesis for single-source class completion. In multi-source settings, missing fault classes can be borrowed from auxiliary source domains [8]. In strictly single-source deployments, generating physically realistic missing-class samples from data alone is ill-posed. Embedding PINN constraints or Invariant Risk Minimization priors [31] into the generative pipeline offers a pathway to synthesize missing fault signatures from rotor-bearing dynamic equations rather than data statistics. This direction connects fault diagnosis research to the broader literature on physics-constrained generative models.

4. Zero-shot cross-equipment diagnosis via disentangled representations. Few-shot meta-learning approaches [2, 18] require a small target-domain support set to adapt the metric space. Achieving zero-shot transfer requires separating machine-specific dynamic invariants (geometry, bearing kinematics, stiffness) from fault-specific transient signatures (impulse periodicity, sub-band energy) at the representation level. This decomposition is related to causal representation learning, pointing toward a convergence between the fault diagnosis and causal inference communities.

5. Foundation models and large-scale self-supervised pre-training for IIoT. The success of self-supervised pre-training in natural language processing and computer vision motivates a parallel effort for industrial machinery data: training large-parameter models on heterogeneous multi-equipment vibration corpora, then fine-tuning for specific fault types with minimal labeled examples. Two challenges are unique to this setting. First, publicly available multi-equipment vibration datasets [9] are orders of magnitude smaller than image or text corpora, making direct Transformer pre-training data-hungry. Second, foundation-model representations must capture physically meaningful sub-band fault signatures rather than statistical co-occurrence patterns. Wavelet-structured tokenization [42, 43] offers a principled interface between physical signal processing and sequence-model architectures. Addressing both challenges is likely to require federated pre-training across industry partners, raising privacy-preserving machine learning as a co-design requirement [40].

6. CONCLUSIONS

The past decade of intelligent fault diagnosis research reveals a consistent pattern: as deployment conditions become more constrained, purely data-driven approaches require progressively stronger structural regularization to remain reliable. The proposed six-tier taxonomy organizes this development by making explicit which assumptions each paradigm requires and where those assumptions break down under realistic industrial conditions. Convergent results from independent research groups on attention mechanisms [42, 43] and self-supervised pre-training [9] confirm that frequency-structured inductive biases improve fault discrimination independent of whether the training objective is supervised, adversarial, or contrastive - corroborating the CA-MCNN design and supporting Category VI as a substantive methodological direction.

The empirical record shows concrete progress: WPA and PDG improved SSDG accuracy over strong baselines on the JNU benchmark, while prototype-aware OSR frameworks elevated the H-score without requiring target-domain threshold calibration. The principal lesson from this comparative analysis is that global distributional alignment is unreliable when source label spaces are incomplete - a problem that generative class completion and prototype-anchored alignment address more directly than global domain confusion.

Looking ahead, five open problems span the research agenda: OS-SSDG under conformal coverage guarantees, calibrated rejection thresholds, physics-informed class completion, zero-shot disentanglement, and foundation-model pre-training for IIoT. Across all five, the combination of Conformal Prediction, physics-informed generative models, and wavelet-structured latent spaces represents the most principled direction for replacing empirical heuristics with bounded, interpretable uncertainty estimates in industrial fault diagnosis.

ACKNOWLEDGEMENTS

This study was supported by the Serbian Ministry of Science, Technological Development and Innovation from the grant 451-03-34/2026-03/200108.

REFERENCES

[1] H. Tao, P. Wang, Y. Chen, V. Stojanovic, and H. Yang, “An unsupervised fault diagnosis method for rolling bearing using STFT and generative neural networks”, Journal of the Franklin Institute, Vol. 357(12), pp. 7525–7545, https://doi.org/10.1016/j.jfranklin.2020.04.024, (2020)

[2] H. Tao, L. Cheng, J. Qiu, and V. Stojanovic, “Few shot cross equipment fault diagnosis method based on parameter optimization and feature metric”, Measurement Science and Technology, Vol. 33(11), p. 115005, https://doi.org/10.1088/1361-6501/ac8368, (2022)

[3] Y. Sun, H. Tao, and V. Stojanovic, “Autoregressive data generation method based on wavelet packet transform and cascaded stochastic quantization for bearing fault diagnosis under unbalanced samples”, Engineering Applications of Artificial Intelligence, Vol. 138, p. 109402, https://doi.org/10.1016/j.engappai.2024.109402, (2024)

[4] Y. Sun, H. Tao, and V. Stojanovic, “Pseudo-label guided dual classifier domain adversarial network for unsupervised cross-domain fault diagnosis with small samples”, Advanced Engineering Informatics, Vol. 64, p. 102986, https://doi.org/10.1016/j.aei.2024.102986, (2025)

[5] Y. Sun, H. Tao, Y. Ni, and V. Stojanovic, “A generic single-source domain generalization framework for fault diagnosis via wavelet packet augmentation and pseudo-domain generation”, IEEE Internet of Things Journal, Vol. 12(15), pp. 31629-31642, https://doi.org/10.1109/JIOT.2025.3573752, (2025)

[6] Y. Sun, H. Tao, and V. Stojanovic, “Multi-domain weakly decoupled domain generalization network for fault diagnosis under unknown operating conditions”, Knowledge-Based Systems, Vol. 330, p. 114452, https://doi.org/10.1016/j.knosys.2025.114452, (2025)

[7] Y. Sun, H. Tao, and V. Stojanovic, “Open-set classification method via latent representation prompt and time-frequency fusion toward unknown fault recognition”, Advanced Engineering Informatics, Vol. 68, p. 103779, https://doi.org/10.1016/j.aei.2025.103779, (2025)

[8] B. Wang, Y. Sun, H. Tao, and V. Stojanovic, “Collaborative completion and alignment for class-incomplete multi-source domain adaptation in rotating machinery fault diagnosis”, Nonlinear Dynamics, Vol. 114, p. 324, https://doi.org/10.1007/s11071-025-12203-y, (2026)

[9] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, “Applications of machine learning to machine fault diagnosis: a review and roadmap”, Mechanical Systems and Signal Processing, Vol. 138, p. 106587, https://doi.org/10.1016/j.ymssp.2019.106587, (2020)

[10] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, “Deep learning and its applications to machine health monitoring”, Mechanical Systems and Signal Processing, Vol. 115, pp. 213–237, https://doi.org/10.1016/j.ymssp.2018.05.050, (2019)

[11] M. Chen, J. Zheng, H. Pei, L. Yang, Q. Zhang, Q. Han, and H. Luo, “Noise-robust fault diagnosis of rotating machinery: a systematic review and methodological roadmap”, Measurement Science and Technology, Vol. 37, p. 192002, https://doi.org/10.1088/1361-6501/ae65b7, (2026)

[12] T. T. Feisa, H. S. Gebremedhen, F. Kibrete, D. E. Woldemichael, S. D. Ageze, and H. D. Nega, "A comprehensive review of deep transfer learning for intelligent fault diagnosis in rotating machines", Engineering Reports, Vol. 7(11), p. e70494, https://doi.org/10.1002/eng2.70494, (2025)

[13] W. Fu, S. Li, B. Wen, B. Zheng, W. Liao, and C. Tan, "A review of rolling bearing fault diagnosis: data preprocessing and model optimization", Measurement Science and Technology, Vol. 36(6), p. 062002, https://doi.org/10.1088/1361-6501/add7fb, (2025)

[14] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals", Sensors, Vol. 17(2), p. 425, https://doi.org/10.3390/s17020425, (2017)

[15] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj, "Real-time motor fault detection by 1-D convolutional neural networks", IEEE Transactions on Industrial Electronics, Vol. 63(11), pp. 7067–7075, https://doi.org/10.1109/TIE.2016.2582729, (2016)

[16] F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: a promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data", Mechanical Systems and Signal Processing, Vol. 72–73, pp. 303–315, https://doi.org/10.1016/j.ymssp.2015.10.025, (2016)

[17] Mingyue Y., Yongpeng L., Guanglei M., and Y. Wang, "An intelligent fault diagnosis for rotating machine under strong noise based on cross-attention-driven spatial-temporal feature fusion and duplexing time sequence convolution optimization", Engineering Applications of Artificial Intelligence, Vol. 153, p. 110897, https://doi.org/10.1016/j.engappai.2025.110897, (2025)

[18] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks", Proceedings of the 34th International Conference on Machine Learning, Vol. 70, PMLR, pp. 1126–1135, (2017)

[19] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method", IEEE Transactions on Industrial Electronics, Vol. 65(7), pp. 5990–5998, https://doi.org/10.1109/TIE.2017.2774777, (2018)

[20] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning", Advances in Neural Information Processing Systems, Vol. 30, pp. 4080-4090, https://doi.org/10.48550/arXiv.1703.05175, (2017)

[21] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE", Information Sciences, Vol. 465, pp. 1–20, https://doi.org/10.1016/j.ins.2018.06.056, (2018)

[22] X. Wang, C. Zhang, Y. Wang, Y. Xu, Z. Liu, and H. Jiang, "Fault diagnosis of rolling bearings with imbalanced samples using conditional temporal denoising diffusion generative adversarial network", Measurement, Vol. 283, p. 122036, https://doi.org/10.1016/j.measurement.2026.122036, (2026)

[23] X. Yang, T. Ye, X. Yuan, W. Zhu, X. Mei, and F. Zhou, "A novel data augmentation method based on denoising diffusion probabilistic model for fault diagnosis under imbalanced data", IEEE Transactions on Industrial Informatics, Vol. 20(5), pp. 7820–7831, https://doi.org/10.1109/TII.2024.3366991, (2024)

[24] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", Advances in Neural Information Processing Systems, Vol. 574, pp. 6840 - 6851, (2020)

[25] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks", Journal of Machine Learning Research, Vol. 17(1), pp. 1–35, https://doi.org/10.48550/arXiv.1505.07818, (2016)

[26] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks", Proceedings of the 32nd International Conference on Machine Learning, Vol. 37, PMLR, pp. 97–105, (2015)

[27] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3723–3732, https://doi.org/10.1109/CVPR.2018.00392, (2018)

[28] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains", Machine Learning, Vol. 79(1-2), pp. 151–175, https://doi.org/10.1007/s10994-009-5152-4, (2010)

[29] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation", International Conference on Machine Learning, ICML, Vol. 560, pp. 6028–6039, (2020)

[30] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. L. Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation", Proceedings of the 38th International Conference on Machine Learning, Vol. 139, pp. 1-30, ICLR, https://doi.org/10.48550/arXiv.2003.00688, (2021)

[31] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization", arXiv preprint, pp. 1-31, https://doi.org/10.48550/arXiv.1907.02893, (2019)

[32] Y. Zhang, M. Li, R. Li, K. Jia, and L. Zhang, “Exact feature distribution matching for arbitrary style transfer and domain generalization”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8025–8035, https://doi.org/10.1109/CVPR52688.2022.00787, (2022)

[33] G. Chen, P. Peng, X. Wang, and Y. Tian, “Adversarial reciprocal points learning for open set recognition”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44(11), pp. 8065–8081, https://doi.org/10.1109/TPAMI.2021.3106743, (2022)

[34] Z. Wang, X. Yang, T. Li, L. She, X. Guo, and F. Yang, “Intelligent Fault Diagnosis for Rotating Machinery via Transfer Learning and Attention Mechanisms: A Lightweight and Adaptive Approach”, Actuators, Vol. 14(9), p. 415, https://doi.org/10.3390/act14090415, (2025)

[35] P. Ding, Y. Shan, Z. Fei, and X.-M. Sun, “A Novel Generative Domain Generalization Network for Intelligent Fault Diagnosis Under Unknown Operating Conditions”, IEEE Transactions on Instrumentation and Measurement, Vol. 74, p. 3562613, https://doi.org/10.1109/TIM.2025.3614879, (2025)

[36] X. Yu, Z. Zhao, X. Zhang, Q. Zhang, Y. Liu, and C. Sun, “Deep-learning-based open set fault diagnosis by extreme value theory”, IEEE Transactions on Industrial Informatics, Vol. 18(1), pp. 185–196, https://doi.org/10.1109/TII.2021.3070324, (2022)

[37] A. Bendale and T. E. Boult, “Towards open set deep networks”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1563–1572, https://doi.org/10.1109/CVPR.2016.173, (2016)

[38] M. Mehralian and B. Karasfi, “RDCGAN: Unsupervised Representation Learning With Regularized Deep Convolutional Generative Adversarial Networks”, Conference on Artificial Intelligence and Robotics, 8769811, https://doi.org/10.1109/AIAR.2018.8769811, (2018)

[39] C. Li, S. Zhang, Y. Qin, and E. Estupinan, “A systematic review of deep transfer learning for machinery fault diagnosis”, Neurocomputing, Vol. 407, pp. 121–135, https://doi.org/10.1016/j.neucom.2020.04.045, (2020)

[40] D. Yang, W. Zhang, and Y. Jiang, “Mechanical fault diagnosis based on deep transfer learning: a review”, Measurement Science and Technology, Vol. 34(11), p. 112001, https://doi.org/10.1088/1361-6501/ace7e6, (2023)

[41] I. Gibbs and E. Candes, “Adaptive conformal inference under distribution shift”, Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS), Vol. 128, pp. 1660 - 1672, https://doi.org/10.48550/arXiv.2106.00170, (2021)

[42] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks”, IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141, https://doi.org/10.1109/CVPR.2018.00745, (2018)

[43] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, “Deep residual shrinkage networks for fault diagnosis”, IEEE Transactions on Industrial Informatics, Vol. 16(7), pp. 4681–4690, https://doi.org/10.1109/TII.2019.2943898, (2020)

SESSION A

EARTH MOVING AND TRANSPORTATION MACHINERY

SESSION A: EARTH MOVING AND TRANSPORTATION MACHINERY

Mechatronic system for monitoring crane lifting operations

Cristina Sescu-Gal[1], Mihail Savaniu[1], Catalin Francu[1*], Oana Tonciu[1], Bogdan Bebeselea[1]

[1] Technical University of Civil Engineering in Bucharest, Faculty of Mechanical Engineering and Construction Robotics, Bucharest, Romania

ARTICLE INFO

[*] **Correspondence:** catalin.francu@utcb.ro

DOI: https://doi.org/10.46793/ET26.A01SG **PAGES:** A1 – A7

ABSTRACT

Cranes are equipment that performs lifting and handling operations of loads with various masses and geometric configurations. Their safe use is regulated by the obligation to equip them with safety devices to limit the loads. The paper presents the development of a mechatronic system based on the integration of new technologies developed in the field of sensors and artificial intelligence in the field of lifting equipment. The implementation of the mechatronic system contributes to increasing the safety level of lifting machinery, by reducing the risk of accidents, ensuring the protection of operators and workers, as well as maintaining the integrity of the handled objects and those in the vicinity of the installation. The adoption of advanced technologies and intelligent systems to support the operation of equipment with the potential to cause accidents, accelerates the transition to automation of construction works.

KEYWORDS

Lifting machinery, Safety, Crane AI, Mechatronic cranes.

1. INTRODUCTION

Lifting plants are indispensable equipment in many industrial fields and in any construction work, used for the handling of materials, equipment and prefabricated elements. It is known that cranes are plants with a high potential for accidents, and the operational safety of these equipment is an essential priority. Statistics show that a significant proportion of work accidents are generated by technical failures, human errors in the absence of complete monitoring and warning systems. These realities justify the need to supplement the existing safety systems, by integrating intelligent solutions, capable of increasing the degree of safety and efficiency in operation. The successive stages of the transition from fully manual to automatic operation, culminating in the development of autonomous construction equipment, figure 1, are also found in the field of lifting installations.

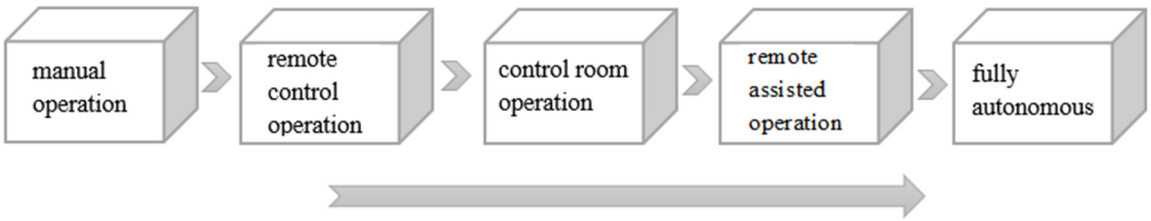


Figure 1: Construction equipment automation levels

Compared to other branches of industry, the transition of construction equipment to automation is at an early stage of development. Manufacturers are developing equipment that integrates different levels of automatic control, up to the possibility of their autonomous operation. In the field of lifting installations, research on the implementation of automation levels has mainly focused on applications that require the reproduction of predetermined sequences of movements. Among the different types of lifting installations, gantry cranes are considered among the most suitable for implementing advanced levels of automation, as they operate in a well-defined workspace, and in certain industrial environments obstacles present fixed positions.

Usually, on construction sites, the handling of loads of any kind is carried out with tower cranes, mobile cranes, crawler cranes with high lifting capacities and gantry cranes, controlled by manual control by the operator in the control cabin. The introduction of remote-control operation of lifting installations has brought numerous advantages, both in terms of operator safety and the efficiency of the work process. These include eliminating the need to carry out activities at height, the possibility of rapid evacuation of the operator in emergency situations, reducing the operator's exposure to the risks specific to the specific risks of the work environment and increasing productivity by positioning the load lifting hook faster and more precisely. The research considers the remote-assisted operation component by performing tasks autonomously.

2. TECHNOLOGICAL CONTEXT AND RECENT RESEARCH

Ensuring the safe operation of lifting installations that move on rail tracks, such as overhead cranes and new generation gantry cranes with remote control operation, is a concern determined by the situations encountered in the use of this equipment. Gantry cranes have the particularity of lifting the nominal load at any point of the workload, being frequently used in the heavy industry of mechanical takeovers, in goods warehouses, in the organizations of large infrastructure sites in order to handle materials, prefabricated (reinforced concrete beams, segments of lining rings for metro galleries, etc.), equipment, underground drilling rigs, etc. The research directions on the use of portal cranes focused on several criteria such as the management of cargo containers in a maritime terminal, monitoring the work area, the implementation of autonomous operation, etc. In the paper [1], the authors use a 3D Lidar sensor to monitor the movements of the overhead crane in an industrial workshop. The 3D Lidar sensor is placed in a fixed position inside the hall, which leads to the appearance of areas that can be obstructed by obstacles in its field of vision. The authors of the paper [2] proposed a method of three-dimensional mapping-graphing of the operating environment of overhead cranes by the combined use of a binocular camera and a laser distance measurement sensor. The solution allows for three-dimensional reconstruction of a fixed scene and improves workspace monitoring capability. Another area characterized by repetitive and intensive task manipulations is large-time terminals. The continuous increase in container flows has led to the development and implementation of self-controlled handling systems, and some operations are autonomous. As is also apparent from [3], LiDAR sensors and, in certain cases, cameras mounted on the load-bearing structure of handling equipment are frequently used in these applications for obstacle detection, monitoring the working environment and assisting autonomous operating functions.

The research results highlight both the advantages of using new monitoring and automation technologies, as well as the limitations associated with their implementation. Among the main aspects identified are:

- the high costs related to the acquisition, installation and integration of monitoring systems based on sensors and video cameras, including both the hardware components and the software applications necessary for their operation;
- the large dimensions of the workload specific to portal cranes, which require the use of efficient methods of data storage and processing, as well as formats for saving digital maps with high compression rates;
- the possibility of determining the absolute position of overhead cranes operating in enclosed spaces by using fixed markers or laser telemetry systems, similar to the location solutions used in Automated Guided Vehicles (AGVs).

These aspects outline the need to identify technical solutions that ensure an optimal compromise between the implementation costs, the accuracy of monitoring and the safety of the operation of the lifting equipment.

3. ARCHITECTURE OF THE MECHATRONIC SYSTEM

Research in the field highlights the usefulness of using sensors and smart cameras for crane workspace monitoring. Unlike the examples presented in the literature, the proposed mechatronic system concept aims to identify the existing obstacles in the workspace of gantry cranes used on construction sites, by placing the mechatronic system directly on the load lifter. This approach is based on the fact that construction sites are working environments characterized by high dynamics, in which the configuration of the cranes' workspace is constantly changing, so that temporary obstacles may arise whose presence and position are not known a priori by the operator.

The intelligent mechatronic system developed to increase safety in crane operation, Figure 2, is based on the integration of new technologies developed in the field of sensors and artificial intelligence into the field of lifting equipment. The concept of the intelligent mechatronic system comprises a 3D LiDAR sensor, SBC (Single-Board Computer) and intelligent camera, which allow detailed scanning of the hook's operating space in real time, generating a three-dimensional model of the environment, detecting and signaling any obstacle that may be in the direction of the hook's movement.

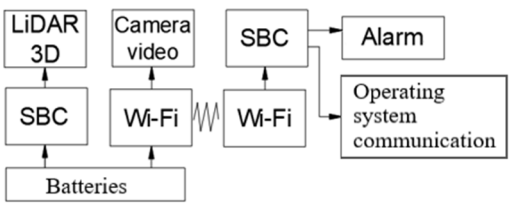


Figure 2: Mechatronic system diagram

The intelligent mechatronic system analyzes the working area in a 3D vicinity defined for the load hook, in order to achieve safe and efficient lifting/lowering operations.

A representative case is that of the portal crane used on the construction site of the M6 Bucharest metro line, figure 3, which operates from ground level and serves the workspace in the work area both on the surface and underground. In this case, the operation is done by remote control, and the operator is on the surface, in the crane running area, which does not allow him to view the underground workspace.



Figure 3: Portal crane – M6 metro line construction site, Bucharest

4. EXPERIMENTAL MODEL FOR TESTING THE FUNCTIONALITY OF THE MECHATRONIC SYSTEM

4.1. Presentation of the Laboratory Stand

In order to validate the proposed model, an experimental model of the mechatronic system was designed and made. The operation test was carried out on the laboratory stand of the Lifting Machinery Laboratory of the Faculty of Mechanical Engineering and Robotics in Constructions, U.T.C.B. The laboratory stand shown in figure 4 is equipped with all the constructive, functional elements and safety devices corresponding to the operation of a lifting installation. The overhead crane is made for educational purposes, so that its operation does not fall under the legislation of lifting installations (valid in Romania). The stand represents a functional model of a real overhead crane, having parameters from table 1:

Table 1: Parameters of the functional model of the laboratory stand

Parameter	Value
Rated load (Q)	1600 kg
Track gauge (E)	3.0 m
Track length (L)	10.0 m
Lifting Height (H)	5.0 m

The laboratory stand has the same parameters as those of the real overhead crane, the only change being the length of the track, which is 2.5 m.

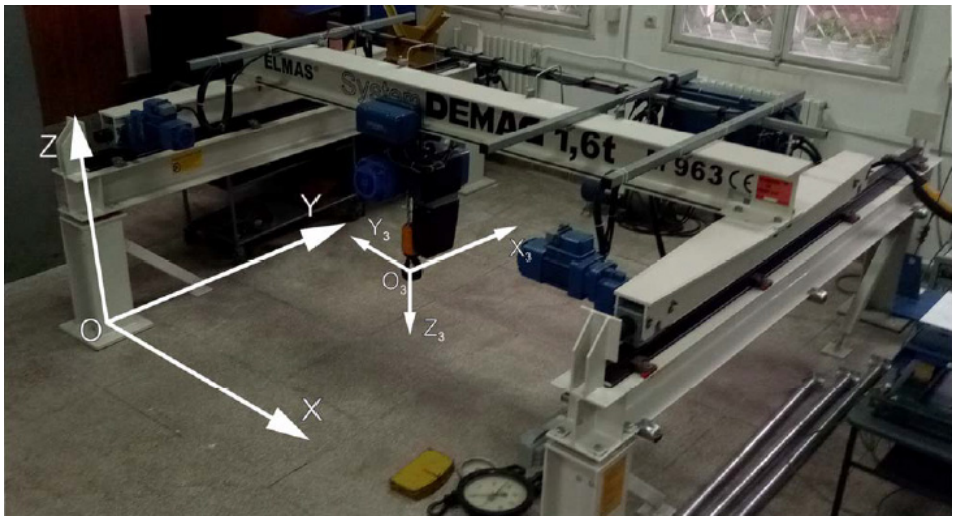


Figure 4: The laboratory stand

In order to determine the workspace of the laboratory stand, we made the kinematic diagram, figure 5.

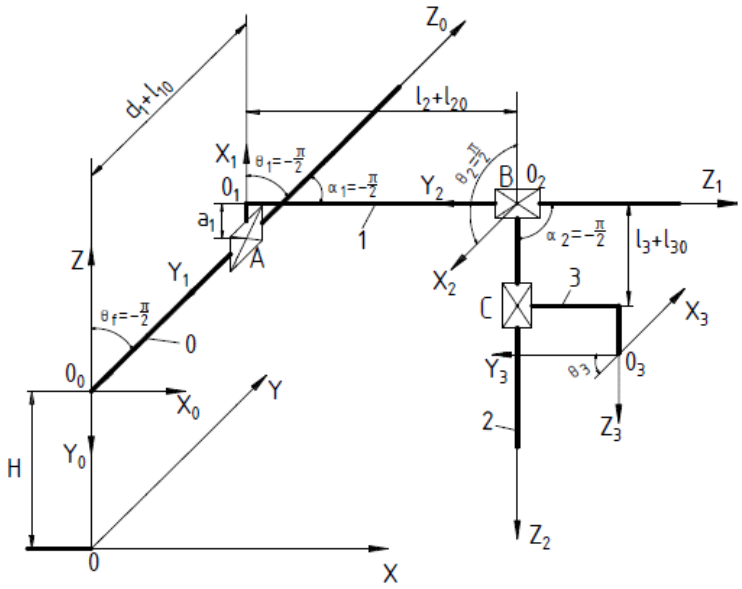


Figure 5: Kinematic scheme

The trips made by the laboratory stand are determined by the relationships (1):

$$\begin{aligned} s_x &= l_{10}, l_{10} + 60 ... (l - l_{11}) \\ s_2 &= l_{20}, l_{20} + 100 ... (l - l_{21}) \\ s_3 &= l_{30}, l_{30} + 100 ... (H) \end{aligned} \quad (1)$$

4.2. Composition of the Experimental Model

We made the experimental model using the following components:

a) The Raspberry Pi 5 platform is the central processing unit of the mechatronic system, being responsible for the acquisition and processing of data from the LiDAR sensor, as well as for the coordination and control of the peripheral components of the system, figure 6.a)

b) MS200 Lidar sensor for distance measurement that scans the environment 360° and provides information about the position of surrounding objects. It works on the principle of dToF (Direct Time of Flight), measuring the time it takes for laser pulses to reach an object and return to the receiver, figure 6.b).

c) Pan-Tilt-Zoom (PTZ) camera with 2 Degrees of Freedom (2DOF) for orientation, Figure 6.c);

d) expansion board mounted on top of the Raspberry Pi 5, with the role of connecting and controlling the robot components. It integrates an ESP32-S3 coprocessor, LiDAR interface, 6-axis IMU sensor and various communication interfaces (Wi-Fi, Bluetooth, UART);

e) The robot chassis is the mechanical support structure of the mechatronic system, having the role of integrating and fixing the electronic components, sensors and the propulsion system, while ensuring the rigidity and stability of the assembly during operation.

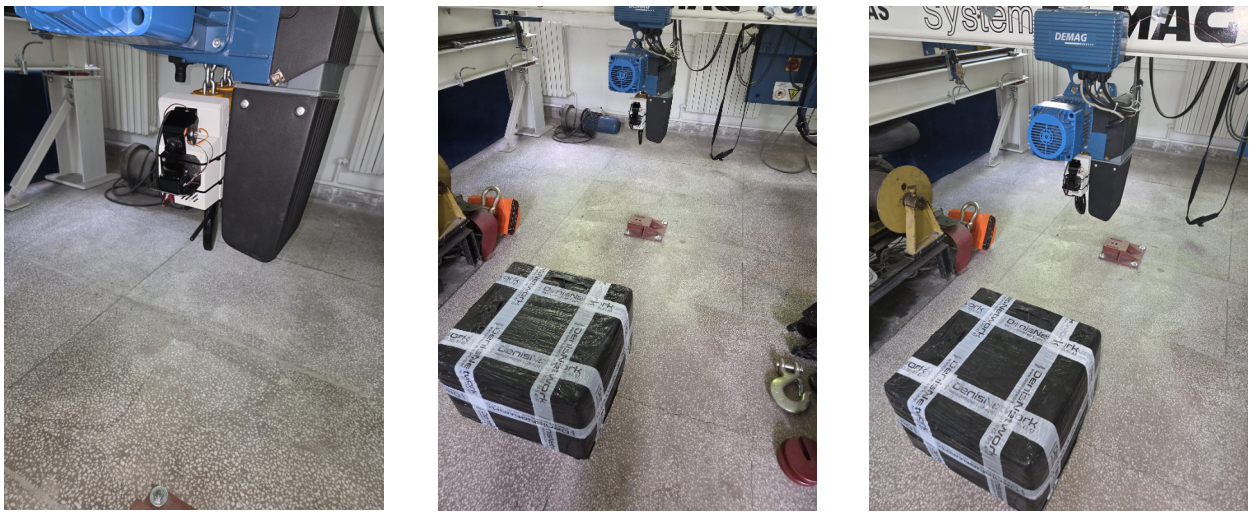


a) b) c)

Figure 6: Elements of the experimental model; a) Raspberry Pi5 – SBC; b) 2D LiDAR; c) 5MP PTZ camera

4.3. Testing the Experimental Model

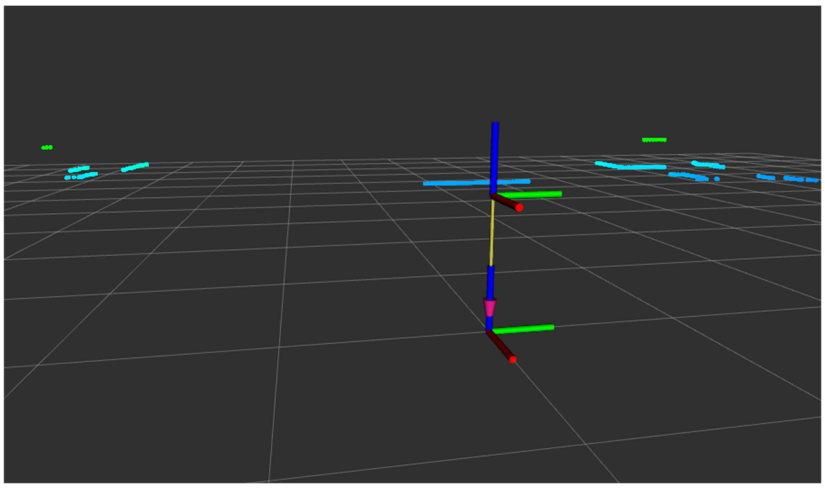
The measurements were carried out with the 2D LiDAR sensor. The experimental model was fixed on the lifting hook of the laboratory stand (Figure 7 a). Figure 7 b shows two positions of the object to be identified in the work space of the stand.



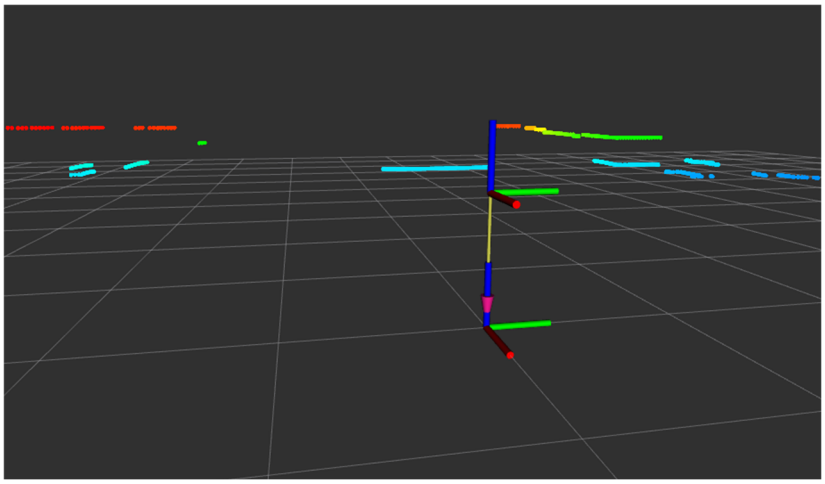
a) b)

Figure 7: Testing of the experimental model: a) the experimental model fixed on the hook of the lifting mechanism; b) test positions for the identification of the obstacle

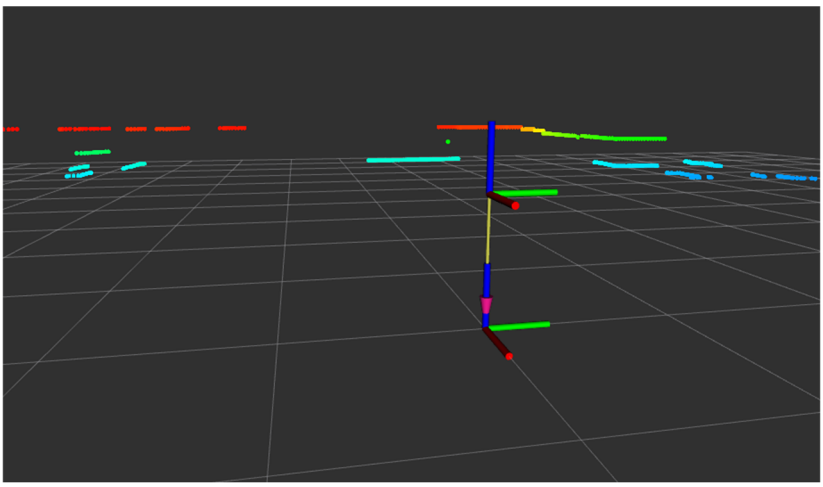
Figure 8 shows the results of the tests carried out on the previously presented object, where the change in its position with respect to the movement of the hook can be observed.



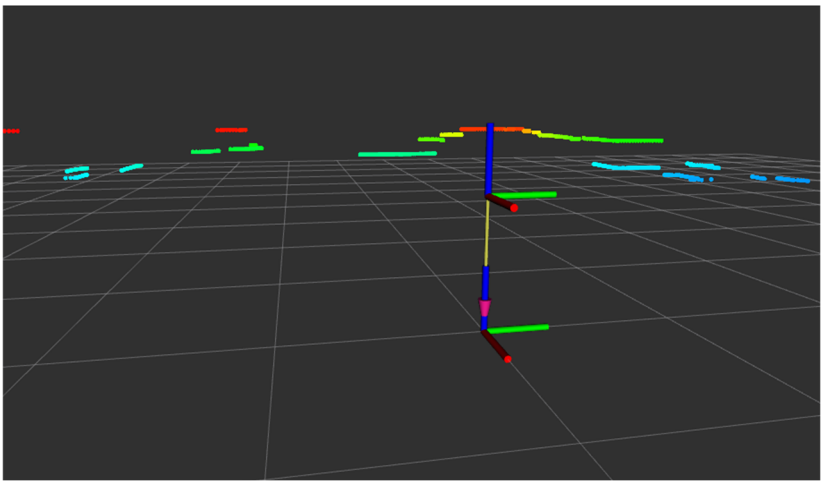
a)



b)



c)



d)

Figure 8: The point cloud generated by the 2D LiDAR sensor for identifying the object in successive positions in the booth workspace: a) the nearest position; b) intermediate position 1; c) intermediate position 2; d) distant position

Finally, the image from the video camera was superimposed on the point cloud generated by the 2D LiDAR sensor, where it is observed that the object is correctly identified.

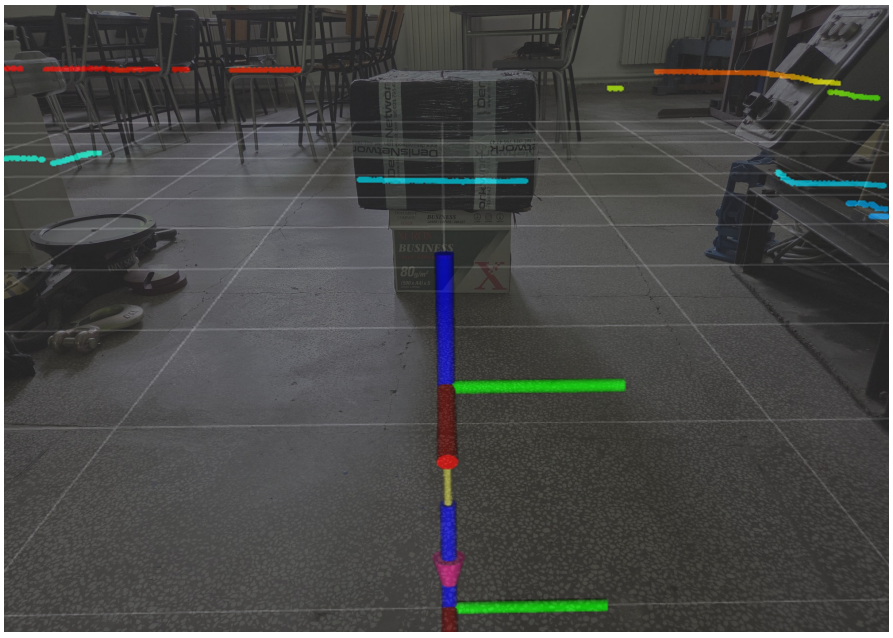


Figure 9: Overlay the images from the camcorder and the point cloud generated by the 2D LiDAR sensor and identify the object under test

4.4. Findings

As a result of the flexibility of the lifting mechanism elements, the movements of the crane cause the lifting device to oscillate. Consequently, for the implementation of the mechatronic system, it is necessary to use a smart camera stabilization device, in order to reduce the effects produced by the balance on the performance of the perception system.

5. CONCLUSIONS

The experimental model developed and tested under laboratory conditions demonstrated the feasibility of using a mechatronic system based on a 2D LiDAR sensor and a video camera to identify obstacles in the workspace of lifting installations. To validate the operation of the final solution, a 3D LiDAR sensor will be used.

The experimental results showed that the combined use of a 2D LiDAR sensor and a video camera is an effective solution for monitoring the workspace and increasing the operator's awareness of the presence of obstacles. The tests carried out confirmed the proper functioning of the system under controlled laboratory conditions and demonstrated its potential for implementation on gantry cranes used on construction sites.

The results obtained also form a basis for further developments, oriented towards the fusion of data from sensors, the estimation of distances to obstacles and the implementation of warning or automatic collision avoidance functions.

Thus, the proposed system can represent an intermediate step towards increasing the degree of automation of the lifting installations.

ACKNOWLEDGEMENTS

This work was carried out within the project funded by the competition. The authors express their gratitude for the financial and institutional support provided, which made it possible to carry out the research activities presented in this article.$GNaC_{2025}^{ARUST}$

REFERENCES

[1] J. M. García, J. L. Martinez, and A. J. Reina, “Bridge Crane Monitoring using a 3D LiDAR and Deep Learning”, IEEE Latin America transactions, Vol. 21(2), pp. 207-216, https://doi.org/10.1109/TLA.2023.10015213, (2023)

[2] K. Zhao, Q. Zhou, X. Xiong, and J. Zhao, “The construction method of the digital operation environment for bridge cranes”, Mathematical Problems in Engineering, Vol. 2021(1), https://doi.org/10.1155/2021/5528639, (2021)

[3] J. Benkert, R. Maack, and T. Meisen, “Chances and challenges: Transformation from a laser-based to a camera-based container crane automation system”, MDPI Journal of Marine Science and Engineering, Vol. 11(9), p. 1718, https://doi.org/10.3390/jmse11091718, (2023)

[4] Y. Zhang, and K. Chen, “Digital technologies for enhancing crane safety in construction: a combined quantitative and qualitative analysis”, Journal of Civil Engineering and Management, Vol. 29(7), pp. 604-620, https://doi.org/10.3846/jcem.2023.19574, (2023)

[5] T. Raj, F. H. Hashim, A.B. Huddin, M.F. Ibrahim, and A. Hussain “A survey on LiDAR scanning mechanisms”, MDPI Electronics, Vol. 9(5), p. 741, https://doi.org/10.3390/electronics9050741, (2020)

[6] M. Alămoreanu, T. Tișea, "Lifting Machines", Vol. 2, Technical Publishing House (Romania), (2000)

SESSION A: EARTH MOVING AND TRANSPORTATION MACHINERY

Remarks on the influence of the dynamic behavior of belt conveyors on the response of the spreader superstructure

Marko Urošević[1*], Nenad Zrnić[1], Xiangwei Liu[2]

[1] University of Belgrade, Faculty of Mechanical Engineering /Department for Material Handling, Constructions and Logistics, Belgrade, Serbia

[2] Shanghai Maritime University, Logistics Engineering College, Shanghai, China

ARTICLE INFO

[*] **Correspondence:** murosevic@mas.bg.ac.rs

DOI: https://doi.org/10.46793/ET26.A02U **PAGES:** A9 – A16

ABSTRACT

This review paper emphasizes the necessity of analyzing the dynamic behavior of spreaders, as high-capacity machines for continuous transport and stacking of bulk materials. Given that spreaders are self-propelled supporting structures for belt conveyors, a dominant portion of the dynamic loads originates from the conveyors themselves. A review of the existing literature shows that transverse and longitudinal oscillations of the belt represent a significant problem, typically treated from the standpoint of bearing maintenance and variable belt tensioning. Furthermore, modernization trends focused on energy efficiency increasingly employ variable frequency drives for continuous belt speed control, directly affecting the variability of dynamic loads caused by the eccentricity of rotating components. These issues are typically addressed from the standpoint of conveyor functionality, often neglecting the potential appearance of negative dynamic effects on the supporting structure, particularly in cases like spreaders, where the conveyor is not directly resting on the ground. Given the variable configuration of the spreader superstructure, with each configuration representing a unique dynamic system, a comprehensive analysis of the system dynamic behavior is fully justified. Evaluating the structural dynamic response to periodically variable excitation loads is essential for extending the service life of existing machines and optimizing the design of future units.

KEYWORDS

Belt Conveyors, Spreaders, Dynamic Behavior, Belt Oscillations.

1. INTRODUCTION

Due to their high productivity, machines for continuous transport and spreading of bulk materials represent critical points in various industries, where operational reliability plays an essential role in ensuring a stable production process. In a large number of cases, these machines operate within serially coupled systems, e.g., bucket wheel excavator - belt conveyor - spreader, in the context of open-pit coal mining. Consequently, a downtime of a single machine results in a shutdown of the entire system, leading to significant financial losses that can exceed 9,000 €/h in the case of the mining industry [1]. Within such systems, some of the largest spreaders in terms of dimensions and transport capacities are deployed. These spreaders can feature discharge booms exceeding 100 m in length and spreading capacities of over 8,000 m³/h, Figure 1.

In the most general sense, spreaders represent self-propelled machines whose structural design enables the mobility of the belt conveyor located on the discharge boom. This arrangement maximizes the duration of continuous material spreading without requiring track displacement or relocation of the entire belt conveyor system that delivers material to the spreader. To achieve continuous spreading, the operational reliability of these machines must be maintained at a high level. Current technical regulations [2], [3] provide a solid basis for stability and strength calculations, whereas the evaluation of dynamic behavior is reduced to quasi-static calculations using dynamic coefficients, from which the limit values of acceleration at reference points can be obtained. However, unplanned downtimes and structural failures still occur in practice, indicating insufficient understanding of their dynamic behavior [4], [5], [6], [7], [8], [9], [10], [11]. Due to their complexity, reflected in their geometry and the design of supporting substructures in the form of spatial or planar truss

girders, as well as the variable configuration of the superstructure where each arbitrary position represents a unique calculation case, the analysis of each machine represents a unique case study. Given the necessity of adequate maintenance and current modernization trends for this type of machinery [12] - particularly regarding the variable frequency control of their drives - with the ultimate goal of extending their service life, a comprehensive evaluation of dynamic behavior becomes imperative. Such an evaluation would minimize the occurrence of negative dynamic effects, both for existing machines and for new units during their design phase.



Figure 1: The largest manufactured spreader, produced by TAKRAF, discharge boom length of 195 m, spreading capacity of 8,800 m³/h. [13]

Due to constant demands for increased production capacities, spreaders operate 24/7 under harsh conditions, remaining continuously exposed to dynamic loads throughout their decades-long service life [14]. To the authors' knowledge, the dynamic behavior of spreaders has not been sufficiently investigated from a theoretical standpoint, but is instead generally reduced to determining dynamic characteristics through experimental approaches [15], [16], [17]. However, through indirect analysis, it can be concluded that the cause of these loads originates from the operation of the belt conveyor located on the spreader discharge boom, as well as from the uneven distribution of the transported material.

Over time, from their initial design to the modern configurations established at the end of the 19th century - upon which the current operating principles are still based [18] - belt conveyors have become some of the most efficient machines for the continuous transport of bulk materials. Through the analysis of relevant literature, it can be observed that with the advancement of mathematical and computational tools, modernization trends in belt conveyors are predominantly focused on developing analytical and numerical models to avoid negative dynamic effects during transient operating regimes, establishing optimal control strategies to increase energy efficiency, and consequently, initiating new research into behavioral analysis under non-uniform material distribution on the belt. However, it can also be noted that these studies are based on the assumption that the conveyor is an isolated system, without further consideration of the impact of the resulting operating loads on the supporting structure where the conveyor is situated, as is precisely the case with spreaders. A common feature of these studies is that operational optimization is reduced to the appropriate variation of speed, which frequently requires the implementation of variable frequency drives that enable continuous variation of the rotational speed of the drive units, and thus, the belt speed. A fact that is also neglected in modern research is the potential existence of eccentricity in the rotating components of belt conveyors. Namely, the physical nature of this problem lies in the fact that a large number of rotating elements comprising a belt conveyor may possess a certain eccentricity within specified tolerance ranges even during their manufacturing phase [19], while operational eccentricities can also occur during spreader operation due to material buildup. From the perspective of the spreader, the existence of these eccentricities results in additional periodically variable loads in the form of centrifugal forces. With the introduction of variable frequency control, these forces acquire a variable character and are transmitted to the spreader supporting structure through idler frame and drum supports. Furthermore, certain studies have demonstrated that conveyors frequently operate with a non-uniform distribution of material on the belt. Given that the dynamic behavior of a system depends on mass distribution, these studies can also be significant when analyzing the response of the spreader. Considering all the aforementioned factors, some of the relevant problems that, either independently or through their solutions, can influence the dynamic behavior of spreaders are presented below.

2. RELEVANT PROBLEMS OF BELT CONVEYORS

2.1. Belt Oscillations

Belt oscillations represent a phenomenon that can have an indirect negative impact on the service life of the belt, and consequently, on the functionality of the belt conveyor as a system. As demonstrated below, belt oscillations can lead to fluctuations in belt tension over time, non-uniform distribution of tension along the belt itself, and unfavorable dynamic effects during inadequate drive startup, among other issues. To quantify this impact, and considering that belts generally consist of upper and lower rubber covers and a wire rope core or textile carcass [20], [21], conveyor belts for bulk material transport cannot be considered rigid bodies. Instead, their mathematical formulation requires models that account for both elasticity and damping due to the viscoelasticity of the constituent material. In the literature, this problem is most frequently addressed by applying the Kelvin-Voigt mathematical model (a spring and a damper connected in parallel), Figure 2 [22], [23], [24].

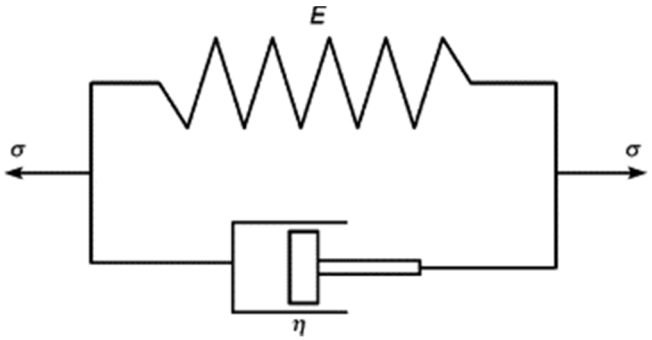


Figure 2: Kelvin-Voigt model

A procedure for establishing a constitutive model of the conveyor belt itself, along with an analysis of its dynamic parameters, is presented in [25]. This study does not consider the conveyor as a whole; instead, it focuses on the mechanical characteristics of steel-cord rubber belts and the influence of excitation parameters on the damping and elastic moduli of the rubber cover and steel cords. In this approach, the periodically variable excitation is developed into a Fourier series, while the belt is modeled using the Kelvin-Voigt model. The verification of the mathematical model was performed through experimental testing conducted on a tensile testing machine using an ST1600 belt specimen (1,100 mm in length, with 6 steel cords), where a 96% correlation between the experimental measurements and the analytical curve was achieved. Furthermore, it was demonstrated that with an increase in excitation frequency, the damping coefficient and elastic moduli of the rubber cover decrease. Conversely, with an increase in amplitude, the belt stiffness increases, while the damping initially decreases and subsequently increases. Such an approach provides a basis for selecting operating frequencies and amplitudes to prevent the degumming of the rubber cover from the steel cords, which negatively affects the service life of the belt.

In [22], an analytical procedure for solving partial differential equations was presented, where the system was modeled using the Kelvin-Voigt model, taking into account the mass of the belt, the mass of the transported material, and the masses of the idlers and drums, with the aim of obtaining a general solution for the system. During the solution process, assumptions were introduced that transverse oscillations of the belt, shear and bending stresses, as well as length changes due to vertical oscillations, could be neglected. With this approach, the conveyor was not observed as a static element, but rather as a continuous body whose dynamic behavior was described by partial differential equations. This enabled the authors to analyze the influence of various control parameters and conveyor design solutions - such as the design of the take-up device, belt speed, and drive control during transient operating regimes (startup) - on the longitudinal oscillations of the belt. It was demonstrated that with an increase in the damping coefficient, the maximum acceleration values during startup decrease, leading to more stable operation, while the total belt tension increases. Furthermore, it was determined that an increase in the take-up force effectively reduces the tension caused by dynamic belt tensioning. An increase in the material mass on the conveyor leads to an increase in tension due to dynamic belt tensioning, but this relationship is nonlinear in nature. Additionally, an increase in belt speed results in higher tension values due to dynamic belt tensioning.

From the perspective of control regimes, through the conducted simulations, it was shown that a parabolic acceleration profile during startup - with a duration 15 to 20 times longer than the fundamental vibration period - has the most favorable effect on startup uniformity, maximum acceleration and tension values, as well as on longitudinal vibration impacts throughout the entire system.

In [26], an investigation into the dynamic characteristics of long-distance belt conveyors was presented, conducted on a conveyor equipped with a gravity take-up device. As in previous studies, the mechanical model of the conveyor in this work was formulated using the Kelvin-Voigt model, while also accounting for the force due to gravity tensioning of the belt. To numerically solve the system of differential equations, the authors applied the fourth-order Runge-Kutta method. Within the scope of investigating the system's dynamic response (dynamic tension distribution), the authors' focus was placed on three points: the head, mid-span, and tail of the conveyor. By analyzing the influence of various system parameters, it was observed that increasing the speed from 2 m/s to 4 m/s, at a fixed acceleration time during startup, leads to

an increase in the maximum tension value - by 32.6% in the case of the head point, which simultaneously exhibits the highest sensitivity to this parameter. Furthermore, the conducted analysis demonstrated that extending the acceleration time during startup yields favorable results for the dynamic behavior of the system. In addition, to further improve the conveyor startup regime, the authors proposed an original drive startup method which, according to the authors, reduces the maximum tension value by approximately 9–13%.

It should be mentioned that in addition to the Kelvin-Voigt model, other methods can be found in the literature for formulating mechanical belt models that can account for the viscoelastic properties of the material. For instance, Guo and Wang in [27] proposed a method for optimizing the conveyor startup regime by modeling the belt using rigid finite elements interconnected by a spring and a damper. The verification of the method was performed through experimental field measurements, yielding satisfactory agreement between the results of the analytical method and the experiment. The authors analyzed several different startup regimes and established that extending the startup time drastically reduces the dynamic belt tension.

In addition to the analysis of longitudinal oscillations using the Kelvin-Voigt model, in studies such as [28], [29], [30], the authors addressed the problem of creating new mathematical models that also account for transverse oscillations and their impact on the functionality of belt conveyors, exploring non-linear models that encompass the mechanical, geometric, and electromechanical aspects of the conveyor. Thus, through the aforementioned studies, the influences of the belt trough geometry, the effects of belt sag between idlers on belt tension, the variability of motion resistance, and the influence of the eccentricity of rotating components were analyzed. It is interesting to mention that in [30] a drive controlled by variable frequency regulation is also considered with the aim of achieving a soft-start, as well as reaching the nominal belt load, regardless of the mass of material present on the conveyor. Such a multidisciplinary approach provides a comprehensive analysis of the conveyor as a whole, aiming to describe the dynamic response to variations in different parameters as precisely as possible.

2.2. Energy-Efficient Speed Control

When discussing conveyors, in addition to the requirement for reliable operation - which, as previously demonstrated, is significantly influenced by the startup regime - the efficiency of their drive units plays an essential role. As energy efficiency demands continue to rise, and given that electricity costs for conveyors can account for up to 40% of the total operating costs of a mine [31], a considerable number of papers on this topic can be found in the literature.

An overview of current engineering solutions aimed at improving energy savings is provided in [32]. Among other aspects, the mentioned study describes the application of variable frequency drives (VFDs) to enable belt speed control within a specific range, with the objective of achieving better real-time belt filling. According to research conducted in [33], this management approach can reduce energy consumption by up to 20%. To optimize conveyor operation from the perspective of energy consumption, Kawalec et al. [34] introduced the Specific Energy Consumption (SEC) index as a criterion for evaluating the energy efficiency analysis of belt conveyor systems at a Polish open-pit coal mine. In addition to the number of conveyors in the system, belt width, trough angle, and material flow rate, the authors varied the belt speed. Through a combination of these parameters, a reduction in electricity consumption of 27.78% was achieved in the best-case scenario.

Vasic et al. in [35] address these issues in the context of conveyors deployed in open-pit mines. Since the amount of material on the conveyor in such cases depends on the operating conditions of bucket wheel excavators, which frequently do not operate at full capacity, there is a well-founded need to control the conveyor belt speed to adapt to the current real-time material volume. This avoids operating the conveyor at maximum speed when small quantities of material are being transported, thereby providing opportunities for energy savings and a subsequent reduction in financial costs. In their research, the authors presented a mathematical model for determining the required belt speed and the times necessary for acceleration/deceleration, under the assumption that speed adjustment should only be performed if the desired speed is at least 15% higher than the current speed [36]. Through a simulation conducted on a conveyor deployed at the Drmno open-pit mine in Serbia, based on annual operating data of the bucket wheel excavator that is serially coupled with this conveyor, the authors demonstrated that within the considered time period of 360 s, the conveyor would operate below its nominal speed the entire time, inevitably leading to a more energy-efficient operation of the system.

He et al. in [37] propose a new method, Estimation-Calculation-Optimization (ECO), which determines the minimum time required to adjust the belt speed based on the amount of material present on it, aiming to reduce energy consumption and CO2 emissions while accounting for the dynamic behavior of the conveyor. Within the scope of this method, the acceleration regime (avoiding excessive belt tension), the deceleration regime (avoiding motor operation in regenerative braking mode), and the speed adjustment time (avoiding mechanical jerks) were taken into consideration. Following the analytical formulation of the method, a case study was conducted on a conveyor in the Port of Rotterdam, where the belt moves at a constant speed while the mass of material on the conveyor varies depending on the available cranes. The authors formulated a simulation model and, based on single-day operating data of the conveyor, compared specific operating parameters under constant and variable speed conditions. Through the conducted analysis, it was demonstrated that when controlling the drive using the ECO method under the same operating conditions, the belt filling would be higher by approximately 25%, energy consumption would be reduced by 12%, annual energy savings would amount to 160 MWh along with 86.9 t of CO2, while the annual financial savings would reach €13,260.

Through the research conducted in [38], He et al. extended the application of the ECO method from individual conveyors to conveyor systems by integrating an active speed control strategy using the Delft Systems Approach method [39]. Within the investigation, the operation of a two-conveyor system was simulated, where the first conveyor moves at a constant speed, while the speed of the second conveyor is constant in the first case and controlled based on the amount of material arriving from the first conveyor in the second case, ranging from 2 to 4 m/s with a step of 0.5 m/s. The results of the conducted simulation showed that the applied method yields benefits in terms of reducing electricity consumption and CO2 emissions, as well as providing financial savings, while achieving system operation under favorable dynamic conditions.

2.3. Non-Uniform Material Distribution on the Belt

According to the DIN 22101 standard [40], the material cross-section is considered a constant value along the entire conveyor path; the same assumption applies to the DIN 22261-2 standard [2]. In practice, a uniform distribution of material along the entire conveyor is rare, particularly for conveyors deployed in open-pit mines, where belt loading directly depends on the periodic discharge of the bucket wheel excavator buckets or operational conditions during excavation. For instance, Semenchenko et al. [41] conducted a 17-hour field measurement of the energy performance of a conveyor deployed at the Dovzhanska-Capital mine (Ukraine), demonstrating that non-uniformity of the material on the belt inevitably exists, which is also reflected in the uneven loading of the two conveyor drive motors. Based on the conducted analysis, the authors concluded that to save energy and extend the system's service life, it is necessary to implement belt speed regulation based on the current material flow rate.

In [42], the problem of non-uniform material distribution was analyzed from the perspective of conveyor energy efficiency. In their research, the authors modeled the conveyor belt using the Kelvin-Voigt model, where the belt was discretized into a finite number of elements whose dimension depends on the belt speed. Each of these finite elements is assigned a material mass obtained through laser scanning of the cross-section. By varying the loading levels (empty conveyor, light-load regime, and overload) and the startup duration in the conducted simulations, the authors established that a longer startup time has a favorable effect on the dynamic behavior of the system. Furthermore, the proposed method provides a basis for a drive control strategy that accounts for the actual material distribution on the belt.

3. DISCUSSION

Based on the presented literature review, it becomes evident that the majority of studies a priori consider conveyors whose supporting structures rest directly on the ground. Through this approach, the conveyor can be observed as an isolated system that does not interact with its environment. However, if the focus is shifted to spreaders, whose primary purpose is the spreading of bulk materials precisely via the conveyor located on them, the dynamic effects generated during conveyor operation are transmitted to the spreader supporting structure through their own framework.

On the other hand, the solution to the presented problems largely lies within the conveyor speed control methodology. If the issue is excessive belt tensioning, which is addressed in Section 2.1, the objective is to prevent undesirable dynamic effects during transient operating regimes of the conveyor, namely during startup and shutdown. The solutions to the problems described in Sections 2.2 and 2.3 lie in controlling the belt speed relative to the quantity and distribution of the material currently present on the conveyor. The presented control strategies aimed at energy efficiency and financial savings primarily account for transient dynamic effects within the conveyor as an isolated system, while the resulting dynamic transmission onto the wider supporting structure remains unaddressed.

When observed from the perspective of the dynamic behavior of continuous bulk material spreading machinery, a portion of the excitation load originates precisely from the operation of the conveyor [16]. The cause of this lies not only in the dynamic effects resulting from belt oscillations and inertial loads occurring during transient operating regimes, but can also be significantly induced by the eccentricities of rotating masses (idlers and drums) [43]. To the authors' knowledge, the influence of eccentricity on the dynamic behavior of the spreader supporting structure has not been sufficiently investigated. Given that spreaders are machines with a variable configuration, where each configuration represents a unique dynamic system, when combined with specific parameters such as the inclination angle of the spreader discharge boom, the amount of material on the belt, and the belt speed, even the eccentricity values defined by technical regulations [19] can lead to the occurrence of resonance, and consequently, to undesirable dynamic effects [44]. On the other hand, the conveyor belt speed has a direct impact on the periodically variable load caused by the eccentricity of rotating masses. In [45], under the assumption that no drive control strategy exists relative to the configuration of the spreader superstructure, it was demonstrated that for inertial forces occurring due to the manufacturing eccentricity of the take-up drum, which remains within prescribed tolerance limits, the response analysis of the superstructure eliminates as much as 46.6% of the belt speed spectrum ranging from 2.5 m/s to 6.5 m/s, from the standpoint of the reference point acceleration limits. If a 5% belt slip is additionally taken into account, only 29.3% of the speed spectrum would be usable. Such an analysis clearly indicates that even control strategies yielding optimal operational results for a conveyor as a closed system cannot be directly applied to conveyors on specialized machines like spreaders; instead, the dynamic response analysis must be extended to the entire system.

4. CONCLUSION

The flexibility and simplicity of belt conveyors enable their easy adaptation to various operating conditions, as well as the achievement of high material transport capacities. Through the literature review conducted in this study, some of the most common problems have been highlighted, along with modern trends aimed at improving the energy efficiency of these machines. Whether addressing the problems of belt oscillations or developing optimal control strategies regarding the amount of material on the belt, researchers predominantly focus on finding solutions through variations in belt speed, i.e., the rotational speed of the driving electric motor. However, in the case of conveyors installed on specialized machinery for continuous bulk material spreading, namely spreaders, where the mobility of the conveyor itself is enabled through a supporting structure that can weigh several thousand tons to ensure continuous material spreading at stockyards, strategic speed variations may yield favorable results for the conveyor itself, but unfavorable outcomes for the entire supporting structure of these machines. Speed parameter variation can have a particularly significant impact in the presence of eccentricity in the rotating components of the conveyor, as additional excitation loads are generated and transmitted to the supporting structure of this type of machinery. Given the large number of rotating elements (idlers and drums), the question arises as to the impact of these excitation loads if they exist only on a certain number of rotating components or across all rotating elements, and whether the eccentricity remains within manufacturing tolerance limits or occurred during operation. Furthermore, speed variation is coupled with the material volume on the belt, which can also influence the system response to such excitation. All the aforementioned aspects simultaneously represent potential directions for future research aimed at describing the impact of conveyor dynamic behavior on the dynamic behavior of the spreader supporting structure, with the ultimate goal of preventive maintenance for existing units and the elimination of potential engineering errors during the design phase of new spreaders.

ACKNOWLEDGEMENTS

The work presented here is the result of research supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, under Contract 451-03-33/2026-03/ 200105 dated 05.02.2026, and a project within the framework of bilateral scientific and technological cooperation between the Republic of Serbia and the People's Republic of China (2024-2026) entitled “Cooperative research on bulk port green & low carbon technologies” under number 002292021 2024 13440 003 000 620 026 03 005.

REFERENCES

[1] U. Bugaric, M. Tanasijevic, D. Polo Vina, D. Ignjatovic, and P. Jovancic, “Lost production costs of the overburden excavation system caused by rubber belt failure”, Eksploatacja i Niezawodność, Vol. 14(4), pp. 333–341, (2012)

[2] “DIN 22261-2:2023-06, Bagger, Absetzer und Zusatzgeräte in Braunkohlentagebauen_- Teil_2: Berechnungsgrundlagen”, DIN Media GmbH, https://doi.org/10.31030/3419208, (2023)

[3] Standards Australia, AS 4324.1:2025, “Mobile equipment for continuous handling of bulk materials - Part 1: General requirements for the design of steel structures”, Sydney, (2025)

[4] G.J. Krige, “Learning from structural failures of materials handling equipment”, Australasian Structural Engineering Conference 2012: The past, present and future of Structural Engineering, pp. 130–137, https://search.informit.org/doi/10.3316/informit.026293678460196, (2012)

[5] S. Ghosal, D. Misra, T.K. Saha, D. Chakravorty, and B. Chaudhuri, “Failure Analysis of Stacker-cum-Reclaimer in Ore Handling Plant”, Journal of Failure Analysis and Prevention, Vol. 8(6), pp. 564–571, https://doi.org/10.1007/s11668-008-9183-0, (2008)

[6] R. Morgan, “Bulk Materials Handling Machine Failure Study”, Aspec Engineering Pty., https://aspec.com.au/bulk-materials-handling-machine-failure-study/, (2023)

[7] S.M. Bošnjak, and N.Đ. Zrnić, “Dynamics, failures, redesigning and environmentally friendly technologies in surface mining systems”, Archives of Civil and Mechanical Engineering, Vol. 12(3), pp. 348–359, https://doi.org/10.1016/j.acme.2012.06.009, (2012)

[8] L. S. Araujo, L. H. De Almeida, E. M. Batista, and A. Landesmann, “Failure of a Bucket-Wheel Stacker Reclaimer: Metallographic and Structural Analyses”, Journal of Failure Analysis and Prevention, Vol. 12(4), pp. 402–407, https://doi.org/10.1007/s11668-012-9575-z, (2012)

[9] M.J. Schmidt and B.W.J. van Rensburg, “Avoiding structural failures on mobile bulk materials handling equipment”, Journal of the Southern African Institute of Mining and Metallurgy, Vol. 115(3), pp. 179–184, https://doi.org/10.17159/2411-9717/2015/v115n3a2, (2015)

[10] D. Pietrusiak, “Evaluation of large-scale load-carrying structures of machines with the application of the dynamic effects factor”, Eksploatacja i Niezawodnosc - Maintenance and Reliability, Vol. 19(4), pp. 542–551, https://doi.org/10.17531/ein.2017.4.7, (2017)

[11] S. Babiarz and D. Dudek, “Chronicle of Failures and Disasters of Basic Machinery in Polish Surface Mining”, Oficyna Wydawnicza Politechniki Wrocławskiej, https://www.oficyna.pwr.edu.pl/ksiazki/kronika-awarii-i-katastrof-maszyn-podstawowych-w-polskim-gornictwie-odkrywkowym/, (2007)

[12] E. Rusiński, J. Czmochowski, P. Moczko, and D. Pietrusiak, “Surface Mining Machines”, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-47792-3, (2017)

[13] “TAKRAF Group | For Mining with Meaning. | TAKRAF Group”, https://www.takraf.com/, (2026)

[14] D.P. Volkov and V.A. Cherkasov, “Dynamics and strength of multi-bucket excavators and stackers (in Russian)”, Mashinostroenie, Moscow, (1969)

[15] T. Smolnicki, M. Kowalczyk, and D. Pietrusiak, “Identification of dynamic characteristics of the stacker”, 6th International Conference Mechatronic Systems and Materials MSM, Opole (Poland), 5 July - 8 July 2010, (2010)

[16] Y.Z. Jiang, C.J. Liu, X.J. Li, K.F. He, and D.M. Xiao, “Low-Frequency Vibration Testing of Huge Bucket Wheel Excavator Based on Step-Decay Signals”, Shock and Vibration, Vol. 2018(1), p. 6182156, https://doi.org/10.1155/2018/6182156, (2018)

[17] D. Pietrusiak, P. Moczko, T. Smolnicki, and E. Rusiński, “Identification of low cycle dynamic loads acting on heavy machinery”, Procedia Engineering, Vol. 199, pp. 254–259, https://doi.org/10.1016/j.proeng.2017.09.017, (2017)

[18] N. Zrnić, M. Đorđević, and V. Gašić, “Historical Background and Evolution of Belt Conveyors”, Foundations of Science, Vol. 29(1), pp. 225–255, https://doi.org/10.1007/s10699-022-09894-6, (2024)

[19] CEMA, “Belt Conveyors for Bulk Materials”, Conveyor Equipment Manufacturers Association, Naples, Florida, (2007)

[20] M. Gašić, “Transportni uređaji”, Mašinski fakultet Kraljevo (Serbia), (2010)

[21] M. Đorđević, N. Zrnic, and B. Jerman, “Simplified Life Cycle Assessment of a Conveyor Belting”, Proceedings of the 21st International Conference on Material Handling Constructions and Logistics, Viena (Austria), 23 September - 25 September 2015, pp. 199-206, (2015)

[22] J. Li and X. Pang, “Belt Conveyor Dynamic Characteristics and Influential Factors”, Shock and Vibration, Vol. 2018(1), p. 8106879, https://doi.org/10.1155/2018/8106879, (2018)

[23] M. Bulíček, J. Málek, and K.R. Rajagopal, “On Kelvin-Voigt model and its generalizations”, Evolution Equations and Control Theory, Vol. 1(1), pp. 17–42, https://doi.org/10.3934/eect.2012.1.17, (2012)

[24] O. Pihnastyi and G. Kozhevnikov, “Kelvin-Voigt Model of Dynamic Stress in the Conveyors’ Belt”, Integrated Computer Technologies in Mechanical Engineering - 2020, in Lecture Notes in Networks and Systems, Vol. 188, pp. 355–365, https://doi.org/10.1007/978-3-030-66717-7_30, (2021)

[25] H. Chen and S. Liu, “Establishment of constitutive model and dynamic parameter analysis of rubber conveyor belt”, Journal of Theoretical and Applied Mechanics, Vol. 61(2), pp. 365–378, https://doi.org/10.15632/jtam-pl/163246, (2023)

[26] Z. Gao, X. Song, Z. Tian, S. Cao, Q. Jiang, and K. Ma, “Research on Dynamic Characteristics of Long-Distance Belt Conveyors”, Applied Sciences, Vol. 16(9), p. 4382, https://doi.org/10.3390/app16094382, (2026)

[27] Y. Guo and F. Wang, “Multi Body Dynamic Equations of Belt Conveyor and the Reasonable Starting Mode”, Symmetry, Vol. 12(9), p. 1489, https://doi.org/10.3390/sym12091489, (2020)

[28] P. Bortnowski, R. Król, and M. Ozdoba, “Modelling of transverse vibration of conveyor belt in aspect of the trough angle”, Scientific Reports, Vol. 13(1), p. 19897, https://doi.org/10.1038/s41598-023-46534-w, (2023)

[29] G. G. Kozhushko, D. A. Yampolsky, and M. D. Lukashuk, “Analysis of forced vibrations of conveyor belts”, IOP Conference Series: Materials Science and Engineering, Vol. 709(3), p. 033022, https://doi.org/10.1088/1757-899X/709/3/033022, (2020)

[30] B. Karolewski and P. Ligocki, “Modelling of long belt conveyors”, Eksploatacja i Niezawodnosc - Maintenance and Reliability, Vol. 16, pp. 179–187, (2014)

[31] S. Zhang and X. Xia, “Optimal control of operation efficiency of belt conveyor systems”, Applied Energy, Vol. 87(6), pp. 1929–1937, https://doi.org/10.1016/j.apenergy.2010.01.006, (2010)

[32] M. Konieczna-Fuławka, “Energy-Saving Solutions Applied in Belt Conveyors: A Literature Review”, Energies, Vol. 18(12), p. 3019, https://doi.org/10.3390/en18123019, (2025)

[33] J. Ji, C. Miao, X. Li, and Y. Liu, “Speed regulation strategy and algorithm for the variable-belt-speed energy-saving control of a belt conveyor based on the material flow rate”, PLOS ONE, Vol. 16(2), p. e0247279, https://doi.org/10.1371/journal.pone.0247279, (2021)

[34] W. Kawalec, N. Suchorab, M. Konieczna-Fuławka, and R. Król, “Specific Energy Consumption of a Belt Conveyor System in a Continuous Surface Mine”, Energies, Vol. 13(19), p. 5214, https://doi.org/10.3390/en13195214, (2020)

[35] M. Vasić, N. Miloradović, and M. Blagojević, “Speed control of high power multiple drive belt conveyors”, IMK-14 - Istrazivanje i razvoj, Vol. 27(1), pp. 9–15, https://doi.org/10.5937/IMK2101009V, (2021)

[36] D. He, Y. Pang, and G. Lodewijks, “Belt Conveyor Dynamics in Transient Operation for Speed Control”, International Journal of Civil, Environmental, Structural, Construction and Architectural Engineering, Vol. 10(7), pp. 836–841, (2016)

[37] D. He, Y. Pang, and G. Lodewijks, “Green operations of belt conveyors by means of speed control”, Applied Energy, Vol. 188, pp. 330–341, https://doi.org/10.1016/j.apenergy.2016.12.017, (2017)

[38] D. He, X. Liu, and B. Zhong, “Sustainable belt conveyor operation by active speed control”, Measurement, Vol. 154, p. 107458, https://doi.org/10.1016/j.measurement.2019.107458, (2020)

[39] H.P.M. Veeke, G. Lodewijks, and J.A. Ottjes, “The Delft Systems Approach”, Springer London, London (UK), https://doi.org/10.1007/978-1-84800-177-0, (2008)

[40] “DIN 22101:2011-12, Stetigförderer_- Gurtförderer für Schüttgüter_- Grundlagen für die Berechnung und Auslegung”, DIN Media GmbH, https://doi.org/10.31030/1821227, (2011)

[41] A. Semenchenko, M. Stadnik, P. Belitsky, D. Semenchenko, and O. Stepanenko, “The impact of an uneven loading of a belt conveyor on the loading of drive motors and energy consumption in transportation”, Eastern-European Journal of Enterprise Technologies, Vol. 4(1(82)), pp. 42-51, https://doi.org/10.15587/1729-4061.2016.75936, (2016)

[42] F. Zeng, C. Yan, Q. Wu, and T. Wang, “Dynamic Behaviour of a Conveyor Belt Considering Non-Uniform Bulk Material Distribution for Speed Control”, Applied Sciences, Vol. 10(13), p. 4436, https://doi.org/10.3390/app10134436, (2020)

[43] M. Urošević and N. Zrnić, “The Impact of the Eccentricity of the Tension Pulley on the Dynamic Response of the Stacker Superstructure”, Proceedings of the XXIV International Conference MHCL’22, pp. 61–70, (2022)

[44] N.B. Gnjatović, M.M. Urošević, A.Z. Stefanović, S.M. Bošnjak, and N.Đ. Zrnić, “Coupled and partial impacts of the inclination angle of the discharge boom and the mass of conveyed material on the dynamic behavior of the stacker superstructure”, Mechanics Based Design of Structures and Machines, Vol. 53(6), pp. 4132–4166, https://doi.org/10.1080/15397734.2024.2442555, (2025)

[45] N. Gnjatović, S. Bošnjak, A. Z. Stefanović, M. Urošević, and I. L. Milenović, “Dynamic Behavior Analysis of the Stacker Superstructure as the Basis for Determination of Favorable Belt Velocities of the Discharge Conveyor”, Proceedings of the XXV International Conference MHCL 2024, 18 September - 19 September 2024, pp. 99–112, (2024)

SESSION A: EARTH MOVING AND TRANSPORTATION MACHINERY

Experimental study of energy efficiency of a hydrogen vehicle

Vesselin Naydenov[1*], Blagoy Burdin[2], Dimitar Krastanov[1]

[1] Todor Kableshkov Higher Transport School, Department of Transport Engineering, Sofia, Bulgaria
[2] Institute of Electrochemistry and Energy Systems, Bulgarian Academy of Sciences, Sofia, Bulgaria

ARTICLE INFO

[*] **Correspondence:**v_naidenov@vtu.bg

DOI: https://doi.org/10.46793/ET26.A03N **PAGES:** A17 – A23

ABSTRACT

This study aims to determine the fuel economy of a hydrogen vehicle by analyzing its energy efficiency in real road conditions. The object of the study is a fuel cell vehicle. Hyundai NEXO, in which hydrogen is an energy carrier, from which electrical energy is obtained for the drive. The purpose of the development is to determine the hydrogen consumption and the corresponding energy intensity (kWh/km) in different driving modes, as well as to analyze the influence of speed, driving dynamics and road conditions on the efficiency of the vehicle. The experimental study is carried out in a real road environment on the territory of Sofia and the adjacent expressways. Two main road cycles are defined: urban and mixed, which reflect the typical modes of operation of cars. A GPS-based system for measuring dynamic parameters is used to collect data, and hydrogen consumption is determined by the vehicle's on-board instruments. The assessment of economy is carried out using indicators such as hydrogen consumption (kg/100 km) and energy consumption (kWh/km), which allows comparison between different driving modes. Particular attention is paid to the influence of different speeds on energy efficiency, as well as the differences between urban and extra-urban traffic. The results obtained contribute to a detailed study of the real-world operational efficiency of hydrogen vehicles and can be used to assess their role in sustainable transport. The study has practical significance in the context of the growing interest in hydrogen technologies and policies for the decarbonization of the transport sector.

KEYWORDS

Hydrogen car, Hydrogen fuel cell, hydrogen, hydrogen consumption, FCEV.

1. INTRODUCTION

The increasing demands for decarbonization of the transport sector stimulate the development of alternative propulsion systems with low or zero local emissions [1]. Among them are hydrogen fuel cell electric vehicles (FCEV), which represent a promising solution for achieving long range, short charging time and high energy efficiency under real operating conditions. These facts prove the potential of hydrogen vehicles and confirm the need for their detailed study in order to improve the control algorithms of individual powertrain nodes and the design of this type of vehicle. In-depth study of their dynamic and economic indicators will help users determine the optimal operating mode of their vehicles, which will lead to a reduction in energy costs and an extension of the service life of vehicles [2], [3].

The aim of this study is to develop and experimentally apply a methodology for assessing the energy efficiency of a hydrogen vehicle in real road conditions.

2. THE RESEARCH

Worldwide, bench tests are mainly conducted, which aim to make a comparison between different vehicles and, above all, to determine the harmful emissions. Tests in real conditions are significantly less, and they are more applicable to determining the fuel consumption of vehicles. It is important to note that the advantage of bench tests is that they can be reproduced many times for different vehicles, while road tests in real conditions cannot provide full repeatability.

The present study was developed for regional conditions in the Sofia region. In addition to the selected routes, a decisive factor for its successful implementation is the equipment used. For the accurate determination of speed regimes, overcome displacements, accelerations and distance traveled, GPS-based equipment was used, which performs detailed recordings

of various parameters in real time during the process of conducting individual studies. Due to limited access to the internal parameters of the fuel cell through standard diagnostics, the study uses a tank-to-wheel approach, based on measuring the actual hydrogen consumption and dynamic parameters of the vehicle.

3. OBJECT OF THE RESEARCH

The object of the study is a Hyundai Nexo passenger car using fuel cell (Fuel Cell Electric Vehicle – FCEV), in which electrical energy is produced on board through an electrochemical reaction between hydrogen and oxygen. The hydrogen is stored under a high pressure of 700 bar in three special vessels located in the rear of the vehicle, and with the help of a reducing valve and a pipeline, the hydrogen reaches the fuel cell in the front of the vehicle. When the fuel cell operates, electrical energy, water and heat are released without any harmful emissions. The electrical energy produced powers the vehicle's on-board network, and a high-voltage battery is used to regulate the electrical system. Its functions focus on assisting acceleration, energy recovery and stabilizing the power balance of the entire vehicle [4].

The Hyundai NEXO fuel cell is one of the most advanced series-produced PEM (Proton Exchange Membrane) systems in the automotive industry and is often used as a reference example in scientific publications on FCEV technologies. The reason for this is the relatively compact design of the stack, which is characterized by a high volumetric power density, which means that the cell generates relatively high power in a small volume. This indicator is measured in kW per unit volume and is denoted by kW/L. In the Hyundai Nexo cell, it is 1 kW/l, while in the company's previous hydrogen car, the IX 35 Hydrogen, it was 0.8 kW/L. Better efficiency of the cell is achieved through thinner membranes, better catalysts in the cell itself with high reaction activity. To increase the performance of the fuel cell, it is necessary to improve the distribution and supply of gases in the cell, as well as to improve its temperature balance. An air compressor is used to supply air to the cell. Such indicators reduce the weight of the fuel cell itself and the vehicle as a whole, freeing up additional volume for other units on board. In addition, energy efficiency and dynamic characteristics are improved. This type of cell is suitable for use in cars, as they develop a low operating temperature and start quickly, even at low temperatures [5]. The fuel cell consists of several stacks, with their total power amounting to 95 kW, while the engine has a power of 120 kW, and to reach its maximum power at peak loads, electrical energy from the battery is used. During operation of the cell and before activation, it must be well hydrated. This is a necessary condition for the effective passage of H^+protons through the membrane. Otherwise, the result is that the internal resistance increases, leading to a drop in voltage and a decrease in efficiency, but also the temperature increases locally, which can lead to cracking of the membrane and its chemical aging. To protect the cell from drying out, a system for humidifying the inlet air is used, using part of the water obtained as an end product from the operation of the cell itself. In this process, moisture is transferred without direct mixing of gases from the outlet and inlet tract. Before reaching maximum load on the fuel cell, the system monitors the temperature, humidity and activates the air system, stabilizing the cell. Cold start is also not without problems. At extremely low temperatures, freezing of the water in the cell can occur, the membrane dries out and its proton permeability decreases. Therefore, cells used in vehicles have a temperature control system, which consists of a hydrator and an electric heater. If the cell becomes overly humid and filled with water, this can block the passage of various gases through it and stop it from working [6].

These facts show that before conducting research with hydrogen cars, it is necessary to select appropriate conditions, such as bringing the fuel cell to operating temperature and ensuring its humidification.

Proton-conducting fuel cells used in cars are very sensitive to contamination and deviations in the parameters of hydrogen. Its purity must be 99.97% or higher. Carbon monoxide (CO), sulfur compounds, ammonia or hydrocarbons can lead to catalyst poisoning, reduced efficiency or accelerated degradation of the cell. In passenger cars, hydrogen is stored in special vessels at a pressure of 700 bar, while in trucks and buses it is 350 bar. Regardless of the type of car, the pressure in the cylinders must be stable and reduced when it is fed to the fuel cell, to the required levels.

The temperature during vehicle refueling can affect the density of the hydrogen fuel and safety. The rate of refueling increases the temperature in the fuel cylinders. A number of procedures have been developed for the actual refueling of vehicles with hydrogen, specifying temperature compensations, operating pressure during refueling and corresponding drop levels, as well as a number of other safety measures [8]. Other regulations specify requirements for hydrogen stations, covering their infrastructure, safety features and specialized technical requirements [7].

Table 1: Permissible levels of hydrogen contamination in fuel cells ISO 14687 [9]

Pollutant	Maximum permissible concentration	Impact on fuel cell
Carbon monoxide (CO)	0.2 ppm	Catalyst poisoning
Total sulfur compounds	0.004 ppm	Irreversible degradation
H_2S	included in sulfur limit	Catalyst deactivation
Ammonia (NH_3)	0.1 ppm	Membrane damage
Formaldehyde	0.01 ppm	Membrane degradation

Pollutant	Maximum permissible concentration	Impact on fuel cell
Formic acid	0.2 ppm	Corrosion
Water (H_2O)	5 ppm	Flooding / corrosion
Oxygen (O_2)	5 ppm	Reduced efficiency
Nitrogen (N_2) + argon	300 ppm	Fuel dilution
Hydrocarbons	2 ppm	Contamination of the electrodes
Particles	strictly limited	Blocking flow channels

4. PREPARATION OF THE VEHICLE AND EQUIPMENT

Before the test, the vehicle must be prepared according to the manufacturer's recommendations. This means checking the tire pressure, size, condition, and suitability for the season in which the test is being conducted. It is also important to check the technical condition of all bodywork components that affect the vehicle's aerodynamics and fuel efficiency. It is important to ensure that sufficient fuel is available for the test in advance. All unnecessary energy consumers should be turned off and any unnecessary load removed from the vehicle [6].

In case the vehicle has not been used immediately before the test, it is necessary to make a short trip of 10 to 15 minutes in order to reach the operating temperature of the cell and to ensure its humidification. This ensures the correct operation of the fuel cell and the recharging of the auxiliary battery. Before starting the test itself, the vehicle's on-board computer should be reset and various indicators should be recorded in the protocol, such as the temperature of the cell, the odometer reading, the setting of the air conditioning system, etc. The fuel tanks should be filled to the maximum, because the pressure in them affects the density of hydrogen, its temperature. When pre-setting it, it is also necessary to note the level of recuperation that will be used while driving. Only at the zero level is the recuperation activated when the brake pedal is pressed, while at the other three levels the recuperative mode of operation of the system is activated at the moment the driver releases the accelerator pedal.

5. THE ROUTES

The test routes are pre-determined on the territory of the city of Sofia and the expressways around it. They are of three types: urban, suburban and mixed. The leading criterion for selecting the routes is to study the fuel efficiency of the vehicle when operating on different types of roads. They are of different lengths, with the urban route being shorter and the mixed route being longer. Both routes start and end at the same place in order to eliminate the influence of elevation changes. They are also selected so that they do not pass through tunnels in which the records of the GPS-based measurement system would be disturbed.

The urban cycle is 14.9 km long, covering a section of an urban highway with a maximum permitted speed of 80 km/h. In this way, the specific study covers all classes of roads in the urban environment and the movement of the vehicle in a fixed driving mode, and not only in transient modes. This reduces the influence of energy recovery and increases the influence of speed.

The mixed cycle is 30.9 km long, covering both city streets and the Northern Speed Tangent, where the speed of movement is 120 km/h, in addition, a certain part of it passes through the ring road with a maximum speed of 90 km/h. This cycle covers a wide range of traffic modes, which makes it maximally dynamic.

6. GPS-BASED EQUIPMENT

A GPS-based VBOX type measuring system is used to record the speed, mileage and dynamic parameters of the vehicle. The device uses satellite positioning with a high sampling rate to determine the speed, distance traveled, time and coordinates of movement. The use of a GPS measuring system allows for high accuracy and independence from the vehicle's factory measuring systems.

The Vbox Racelogic used is a professional device for measurements in car racing, various road studies in real conditions and measurement of dynamic characteristics during braking and acceleration of vehicles. This type of devices is used by a number of universities, scientific organizations, automobile manufacturers, test centers and engineers in motor sports. The equipment is distinguished by particularly high accuracy in measuring speed, time and distance. According to the manufacturer, the deviation of the measuring system in speed measurements is up to ± 0.1 km/h [10]. Accuracy is higher on open country routes, due to the lack of tall buildings that would interfere with satellite signals. Suitable for the present study is the property of the device to draw the route traveled on a digital map, and it can be traced in which section of the route the car was moving at what speed and what distance it was traveling. The listed properties of the system prove that it is particularly suitable for performing measurements in real road conditions.

During the tests, in addition to the automatic recording of GPS data, manual notes were also kept on external conditions and events that could have an impact on the measurement results. Parameters such as weather conditions, traffic, temporary traffic restrictions and other factors affecting the vehicle's operating mode were recorded.

7. RESULTS

The test results are presented in tables, graphs and route profiles to visualize the operating modes and hydrogen consumption under different driving conditions. Parameters such as average speed, maximum speed, driving time and specific hydrogen consumption were analyzed.

Due to the impossibility of direct access to the data on the fuel cell parameters, the average fuel consumption was taken from the readings of the vehicle's on-board computer in the study. These data are not distinguished by high accuracy and precision, but for conducting real road conditions they provide good enough information about the fuel consumed during the test. One of the main factors that indicate the influence of the test data is traffic. Depending on its intensity, the hydrogen consumption could change significantly. Conducting multiple studies in different time ranges would provide more accurate data on hydrogen consumption.

In urban driving mode, the measured average consumption is 0.9 kg/100km, for 14,866 km traveled, this means that the actual fuel consumed is 0.134 kg of hydrogen. In terms of energy value, it is equal to 4,456 kW/h. Figure 2 shows clearly pronounced transient modes with stops and accelerations, as well as sections with established traffic.

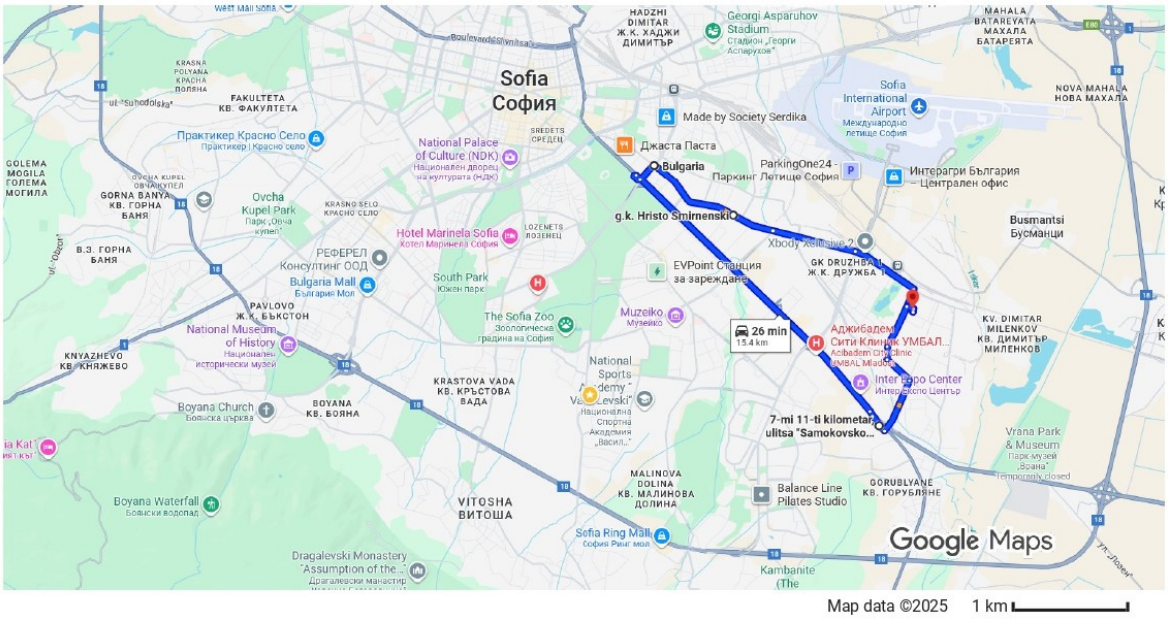


Figure 1: Map of the urban route traveled



Figure 2: Velocity change as a function of time

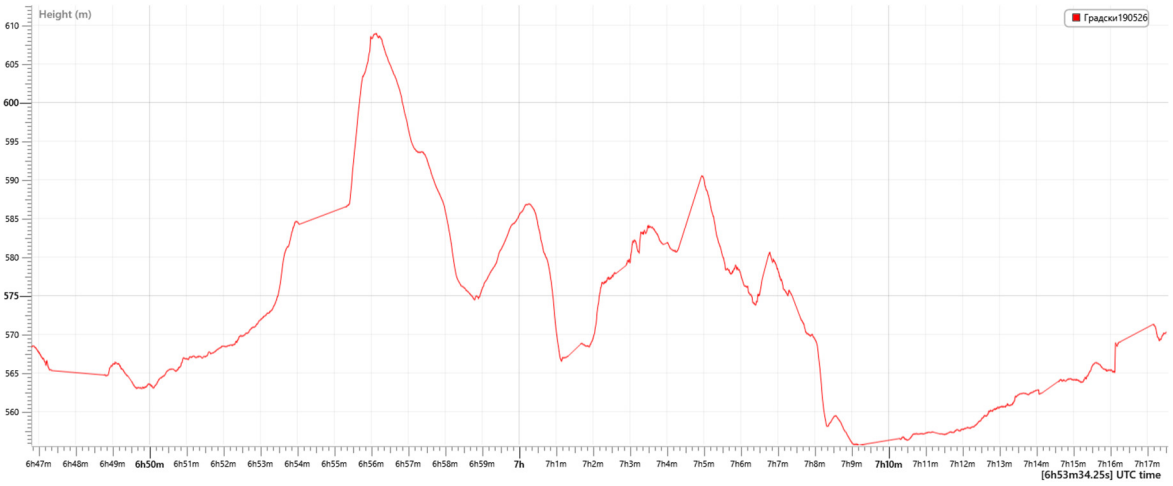


Figure 3: Overcome elevation changes as a function of time

After driving the mixed route, the measurement data shows an average consumption of 0.8 km/100km of hydrogen. For the specific trip, this is 0.246 kg of hydrogen, which is equivalent to 8.178 kW/h.

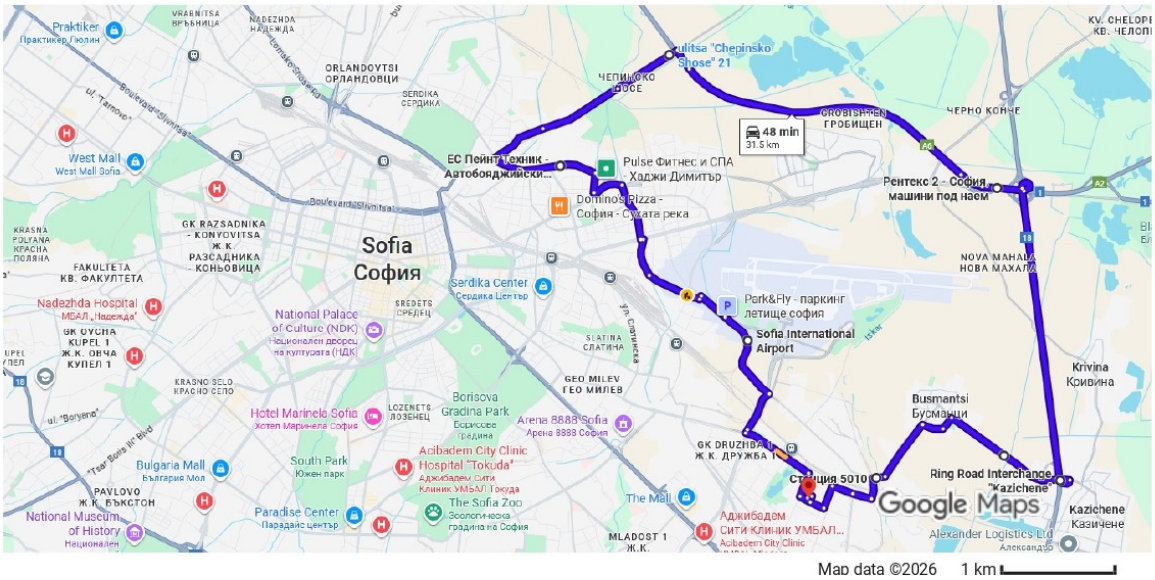


Figure 4: Map of the mixed route traveled

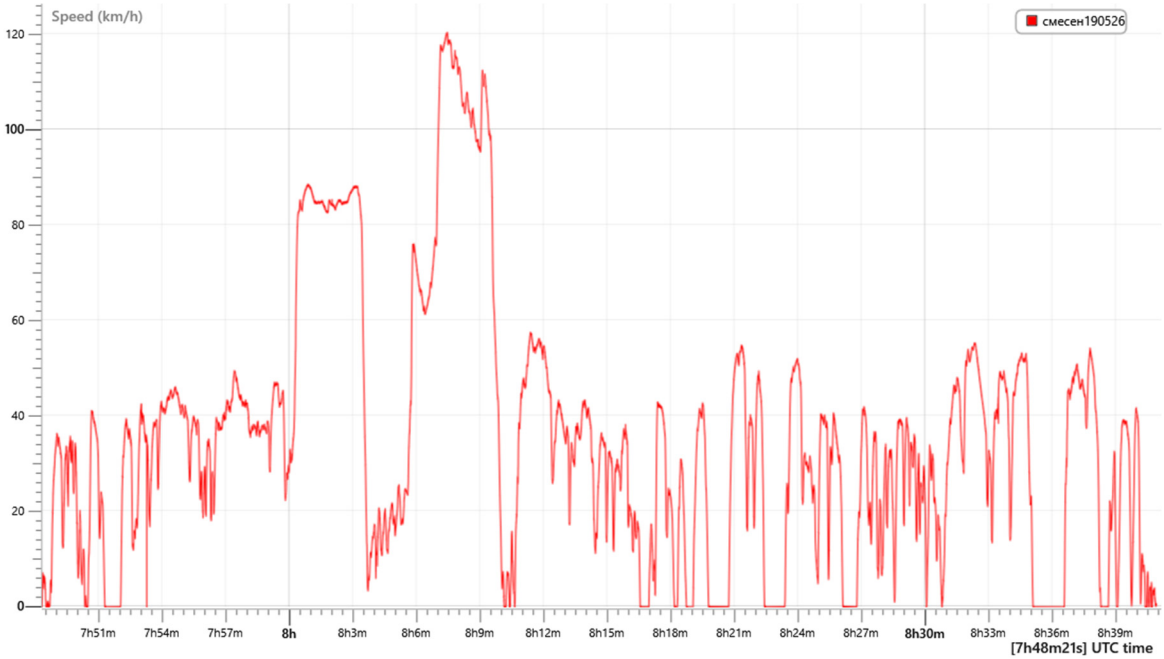


Figure 5: Speed variation on a mixed route

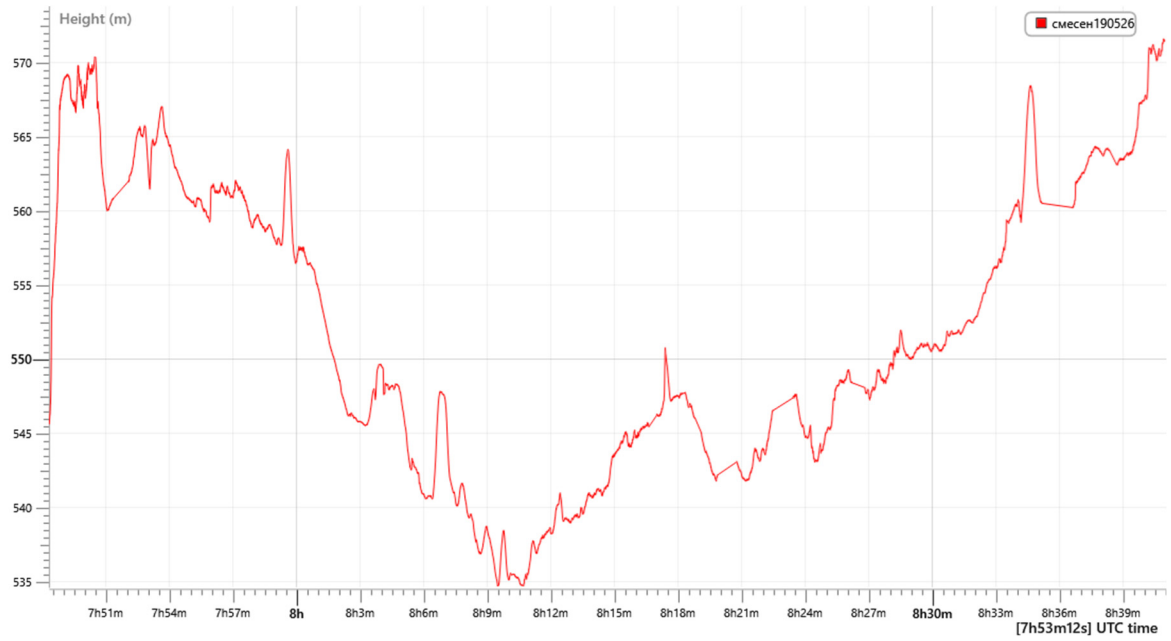


Figure 6: Variation of elevation gain over time for a mixed route

During the study, data was obtained on average fuel consumption and kilometers traveled.

By the formula:

$$C_{H2} = \frac{m_{H2}}{s} \cdot 100 \tag{1}$$

C_{H2} - average hydrogen consumption [kg/100km]

m_{H2} - hydrogen consumed [kg]

s- distance traveled [km] [11]

From which it follows that:

$$m_{H2} = \frac{C_{H2} \cdot s}{100} \tag{2}$$

The energy used for a specific trip is determined by the formula:

$$E = m_{H2} \cdot LHV_{H2} \tag{3}$$

AND- energy used [kWh]

m_{H2}- hydrogen mass [kg]

LHV_{H2}- lower calorific value of hydrogen (assumed to be 33.3 kWh/kg) [12]

$$e = \frac{E}{s} \tag{4}$$

e - energy intensity [kWh/km]

E - energy used [kWh]

s- distance traveled [km] [13]

Table 2: Comparison of hydrogen consumption in urban and mixed routes

Route	Average hydrogen consumption	Travelled distance	Hydrogen consumption (2)	Energy intensity (4)	Consumed energy (3)	Maximal distance travelled with this consumption
	kg/100km	km	kg	kWh/km	kW/h	km
City	0.90	14.866	0.133794	0.2997	4.4553402	696.67
Mixed	0.80	30.7	0.2456	0.2664	8.17848	783.75

8. EXTRACTS

The conducted research shows that the Hyundai NEXO hydrogen car demonstrates high energy efficiency when operated in real road conditions. The obtained results confirm the possibility of achieving low hydrogen consumption in urban traffic, with the measured values in the conducted tests varying between 0.8 and 0.9 kg/100 km. These differences are a result of dynamic traffic, transient traffic regimes and other factors that are reflected in real traffic conditions in a real environment.

The study shows that the nature of the speed profile has a significant impact on the energy efficiency of the vehicle. With smoother movement and a smaller number of intense accelerations, a decrease in hydrogen consumption is observed, which is associated with more stable operation of the fuel cell and lower instantaneous load on the power system. At the same time, movement at high speeds leads to an increase in energy consumption due to an increase in aerodynamic resistance.

The use of a GPS-based VBOX type measurement system allows for reliable recording of the vehicle's dynamic parameters and detailed analysis of the operating modes. The obtained speed profiles and route data allow for tracking the influence of individual sections of the route on the hydrogen consumption and the overall energy intensity of the vehicle.

Despite the limitations associated with the use of data from the on-board computer and the inability to directly access the internal parameters of the fuel cell, the applied methodology allows obtaining reliable information about the real operational efficiency of the hydrogen vehicle.

The results obtained show that hydrogen fuel cell vehicles have potential for application in sustainable transport, especially when operating in urban and mixed traffic conditions. The present study can serve as a basis for future developments related to expanding the experimental base, conducting long-term tests and comparing different operating modes and types of electric vehicles.

REFERENCES

[1] M. Evtimova, D. Todorova, K. Krastanov, B. Burdin, M. Slavova, and V. Naydenov, “European green pact and hydrogen economy”, Mechanics. Transport. Communications: scientific journal. Sofia: VTU "Todor Kableshkov", Vol. 22(3/3), p.2585, (2024)

[2] V. Naydenov, “Application of hydrogen as a source of energy in road vehicles”, Mechanics. Transport. Communications: scientific journal. Sofia: VTU "Todor Kableshkov", Vol. 22(3/3), pp. (IX-7)-(IX-12), (2024)

[3] “Assessment of the potential for the development of hydrogen technologies in the Republic of Bulgaria”, Ministry of Innovation and Growth, Sofia 2022

[4] Hyundai Motor Company, “Hyundai NEXO Specifications”, Hyundai Motor Company, Seoul, South Korea, 2023

[5] M. Slavova, I. Slavov, and V. Naydenov, “Types of fuel cells”, Mechanics. Transport. Communications: scientific journal. Sofia: VTU "Todor Kableshkov", Vol. 22(3/3), pp. (IX-1)-(IX-6), (2024)

[6] A. Szałek, I. Pielecha, and W. Cieślik, “Fuel Cell Electric Vehicle (FCEV) Energy Flow Analysis in Real Driving Conditions (RDC)”, Energies, Vol. 14(16), p. 5018, https://doi.org/10.3390/en14165018, (2021)

[7] SAE International. SAE J2601_202005: Fueling Protocols for Light Duty Gaseous Hydrogen Surface Vehicles. SAE International, Warrendale, PA, USA, 2020.7

[8] ISO 14687:2019. Hydrogen fuel quality - Product specification. International Organization for Standardization (ISO), Geneva, Switzerland, (2019)

[9] Racelogic Ltd., “VBOX Automotive: GPS Data Logging and Vehicle Testing Systems – Technical Specifications”, Buckingham, UK, (2023)

[10] T.D. Gillespie, “Fundamentals of Vehicle Dynamics”, SAE International, Warrendale, PA (USA), (1992)

[11] J. Larminie, A. Dicks, “Fuel Cell Systems Explained, 2nd ed.”, Wiley, Chichester (UK), (2003)

[12] G.J. Offer, D. Howey, M. Contestabile, R. Clague, and N.P. Brandon, “Comparative analysis of battery electric, hydrogen fuel cell and hybrid vehicles in real-world driving”, Energy Policy, Vol. 38(1), pp. 24–29, https://doi.org/10.1016/j.enpol.2009.08.040, (2010)

SESSION A: EARTH MOVING AND TRANSPORTATION MACHINERY

Comparative overview of determining service classes of industrial cranes according to relevant standards and guidelines

Mirko Katona[1], Atila Zelić[1*], Dragan Živanić[1], Radomir Đokić[1], Tanasije Jojić[1], Nikola Ilanković[1]

[1] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia

ARTICLE INFO

[*] **Correspondence:** zelic@uns.ac.rs

DOI: https://doi.org/10.46793/ET26.A04K **PAGES:** A25 – A31

ABSTRACT

The service class of a crane is one of the key parameters in crane design and can be determined according to various standards. In domestic practice, the classification according to the JUS M.D1.020 standard was prevalent for a long time, defining four crane service classes based on the number of working cycles and operating time. In the past, many renowned crane manufacturers based their classification on the DIN 15020 standard, which is predicated on FEM (Fédération Européenne de la Manutention) guidelines (which consider operating time and load spectrum as the dominant classification parameters). Older versions of ISO standards are also closely linked to FEM guidelines and provide service classes equivalent to these recommendations. With the introduction of the EN 13001 series of standards and ISO 4301-1:2016, the number of operating cycles has a more significant impact on crane classification. Unfortunately, in domestic engineering practice, the newly adopted standards for industrial cranes are still not widely applied due to various uncertainties. The aim of this paper is to resolve certain uncertainties regarding the definition of crane service classes and to present, in a concise form, a comparison and the relationship between parameters according to current standards and guidelines. A brief overview of the withdrawn JUS M.D1.020 standard is also included.

KEYWORDS

Service class, Crane design, Crane classification, Classification of parameters.

1. INTRODUCTION

The service class of a crane represents one of the most important parameters in the process of designing, sizing, and assessing the service life of industrial cranes. In addition to rated capacity, lifting height, span or outreach, the service class is determined at the very beginning of the design process. The service class defines the intensity and frequency of use, the load spectrum, and the operating conditions, all of which directly influence the selection and calculation of structural elements, drive mechanisms, as well as fatigue assessment and reliability analysis.

In domestic engineering practice, [1] was long used for crane classification. It provided a relatively simple, though limited, system based mainly on the number of working cycles and operating time. Many lifting-equipment manufacturers also applied [2], which classifies service based on operating time and load spectrum. At the same time, [3] was used for cycle-based classification, while [4] was aligned with [2]. A-classes for complete cranes, based on the total number of working cycles are introduced in [5], while retaining time-based classification for hoisting mechanisms from [2]. Standard [6] was harmonized with [5], whereas [7] classifies cranes by the total number of working cycles and defines an expanded range of A-classes. The introduction of [8] marked a major shift in service classification. These standards classify cranes according to the total number of working cycles over the design life and combine this approach with limit-state design and advanced load factors. This method provides greater accuracy and better reflects actual operating conditions, but it also requires a more thorough understanding of the relevant parameters and calculation procedures. Despite the formal adoption of EN standards, domestic engineering practice still shows inconsistent use of different methodologies and uncertainty about their relationships and equivalence.

2. THE IMPORTANCE OF CORRECTLY DETERMINING THE SERVICE CLASS OF INDUSTRIAL CRANES

Industrial cranes include several drive mechanisms, such as hoisting, travelling, and slewing, each with its own service class. Because hoisting generates the greatest load effects, the hoisting mechanism's service class also determines the service class of the supporting structure, that is, of the crane as a whole. In structural safety verification during crane design, safety and dynamic factors must be selected according to the crane type, the design method used, and the assigned service class. In addition, the dimensioning of components such as drums, sheaves, ropes, and wheels depends not only on the applicable standard but also on the crane's service class. An incorrect service class can lead to oversized components, but more importantly, it can result in undersized critical elements. As noted in [10], standards [4] and [13] do not require fatigue calculations for wire ropes; rope bending cycles are not considered in design, and sizing is based only on a static safety factor.

In addition to improving safety, consistent service classification of drive mechanisms, their components, and the supporting structure helps ensure that elements selected and sized for the assigned service class have approximately equal durability in terms of operating time or number of working cycles. Another benefit of classifying individual drive mechanisms and their components is the improved clarity and mutual understanding between users and manufacturers of lifting equipment when ordering or replacing specific elements.

In practice, used industrial cranes are often sold and refurbished for continued operation over many additional years. Such sales usually involve relocating the crane, which changes the operating conditions under which it continues service after refurbishment. Therefore, it is essential to understand the service class and the parameters it includes.

3. OVERVIEW OF STANDARDS AND GUIDELINES FOR CRANE CLASSIFICATION

Crane classification is defined by several standards, with additional guidance provided by related guidelines. The following section reviews crane classification under the most relevant standards and guidelines.

3.1. JUS M.D1.020

This standard was issued in 1964 and for decades represented the basis for crane classification in domestic engineering practice, which is why cranes classified according to it can still be found in operation today. The standard defines four service classes, determined according to two primary and four secondary parameters.

Table 1 shows two primary parameters, n – number of working cycles and T_t – theoretical operating time.

Table 1: Two primary parameters of classification according to [1]

Service classes		n	T_t [h]
name	number		
light	1	to 16	to 0.5
medium	2	over 16 to 32	over 0.5 to 1
heavy	3	over 32 to 63	over 1 to 2
very heavy	4	over 63	over 2

The theoretical operating time is calculated according to the following formula:

$$T_t = T_s \cdot \left(\frac{q_r}{100} \right)^3 \tag{1}$$

Table 2 shows the other four secondary parameters, q_r – relative load, T_s – actual working hours per day, V_t – theoretical service life and T_g – annual actual working time.

Table 2: Four secondary parameters of classification according to [1]

Service classes		q_r [%]	T_s [h]	V_t [≈ h]	T_g [h]
name	number				
light	1	to 56	2.8 to 5.6	2000	to 2000
medium	2	50 to 63	4 to 8	4000	over 2000 to 3000
heavy	3	56 to 70	5.6 to 11.2	8000	over 3000 to 4000
very heavy	4	over 63	8 to 16	16000	over 4000

Reference [9] notes that, under this standard, the service class is not determined by a combined assessment of all parameters. Instead, it is established using two essential parallel parameters, while the remaining four parameters serve only as

guidance. The standard further states that the governing parameter is the one that results in the higher, that is, more severe class. This raises a possible uncertainty as to whether all six parameters are given equal importance in classification. Reference [9] also points out the limited importance assigned to the crane's load condition, as illustrated by the relative load parameter, q_r. According to this criterion, an estimation error of only 7% may shift a crane from the 1st to the 4th service class, or vice versa. In addition, when calculating the actual annual operating time, the standard does not specify the number of days per year, whereas the calculation of theoretical service life assumes all 365 days rather than the substantially lower number of actual working days.

3.2. FEM 1.001

This is the third edition, published in 1998, and it remains valid today. The guidelines classify cranes as complete units into A-classes and serial hoisting mechanisms into M-classes, in harmonization with [6]. For A-class determination, two independent parameters are used: the number of working cycles and the load spectrum. Their combination yields a single overall class. The class based on the number of working cycles is designated U and is determined by the total number of working cycles over the crane's service life, as shown in Table 3.

Table 3: Classes U of total numbers of crane working cycles

Class of utilization	Total duration of use (number n_{max} of hoisting cycles)				
U0			n_{max}	≤	16 000
U1	16 000	<	n_{max}	≤	32 000
U2	32 000	<	n_{max}	≤	63 000
U3	63 000	<	n_{max}	≤	125 000
U4	125 000	<	n_{max}	≤	250 000
U5	250 000	<	n_{max}	≤	500 000
U6	500 000	<	n_{max}	≤	1 000 000
U7	1 000 000	<	n_{max}	≤	2 000 000
U8	2 000 000	<	n_{max}	≤	4 000 000
U9	4 000 000	<	n_{max}		

The load spectrum class is determined using the load spectrum factor, which is calculated according to:

$$k_p = \sum \left[\left(n_i / n_{\max} \right) \cdot \left(m_{li} / m_{l\max} \right)^3 \right] \tag{2}$$

where:

- m_l – loads;
- m_{max} – safe working load;
- n_i – number of hoisting cycles in respect of which the hoisted load is greater than or equal to m_l;
- n_{max} – number of hoisting cycles determining the total duration of use.

The load spectrum classes are presented in Table 4.

Table 4: Classes of load spectrum

Symbol	Cubic mean value
Q1	$k_p \leq 0.125$
Q2	$0.125 < k_p \leq 0.250$
Q3	$0.250 < k_p \leq 0.500$
Q4	$0.500 < k_p \leq 1.000$

Finally, the crane is assigned to one of the service classes listed in Table 5.

Table 5: Service classes according to [5]

Load spectrum	Class of utilization									
	U0	U1	U2	U3	U4	U5	U6	U7	U8	U9
Q1	A1	A1	A1	A2	A3	A4	A5	A6	A7	A8
Q2	A1	A1	A2	A3	A4	A5	A6	A7	A8	A8
Q3	A1	A2	A3	A4	A5	A6	A7	A8	A8	A8
Q4	A2	A3	A4	A5	A6	A7	A8	A8	A8	A8

Despite the publication of newer standards, this method remains one of the most widely used by crane manufacturers. Its main advantage is its simplicity and the clearly defined parameters for calculating the load spectrum factor. For this reason, newer ISO and EN standards, although expanded with additional classes, have retained the U and Q classifications.

3.3. ISO 4301-1:2016

This standard builds on the 1986 edition, which was harmonized with [5] recommendations, but the latest version no longer classifies hoisting mechanisms into M-classes. It also expands the A-class system: instead of the previous eight classes (A1 to A8), it defines fifteen classes, from A03 to A11. Classification is based on the number of working cycles (U-classes) and the load spectrum factor (Q_p-classes). In this edition, the number of Q_p-classes increases from four to six. The load spectrum factor is calculated in the same way as in FEM 1.001, and the Q_p-classes are shown in Table 6.

Table 6: Classes Q_p of load spectrum factors

Symbol	Load spectrum factor K_p				
Q_p0			K_p	≤	0.0313
Q_p1	0.313		K_p	≤	0.0625
Q_p2	0.0625		K_p	≤	0.125
Q_p3	0.125	<	K_p	≤	0.250
Q_p4	0.250	<	K_p	≤	0.500
Q_p5	0.500	<	K_p	≤	1.000

It should be noted that the former Q1 class in [5] has been divided into three classes: Q_p0, Q_p1, and Q_p2. The overall service class A is then determined by combining the U-class and the Q_p-class, as shown in Table 7.

Table 7: Classes A for group classification

Load spectrum class	Class of utilization									
	U0	U1	U2	U3	U4	U5	U6	U7	U8	U9
Q_p0	A03	A02	A01	A0	A1	A2	A3	A4	A5	A6
Q_p1	A02	A01	A0	A1	A2	A3	A4	A5	A6	A7
Q_p2	A01	A0	A1	A2	A3	A4	A5	A6	A7	A8
Q_p3	A0	A1	A2	A3	A4	A5	A6	A7	A8	A9
Q_p4	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Q_p5	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11

If the U-class and Q_p-class are unknown, classification may be based on the number of full-load cycles, as shown in Table 8.

Table 8: Design basis by classes

A- class	Design number of full load cycles, C_f
A03	500
A02	1 000
A01	2 000
A0	4 000
A1	8 000
A2	16 000
A3	31 500
A4	63 000
A5	125 000
A6	250 000
A7	500 000
A8	1 000 000
A9	2 000 000
A10	4 000 000
A11	8 000 000

Annex B of this standard outlines a procedure for converting the M-classes defined in the 1986 edition into the A-classes used in the current version.

In addition to A-classes, the standard also defines D-classes for average linear and angular displacements and P-classes for the average number of accelerations.

3.4. FEM 9.521

These guidelines provide a more detailed procedure for converting the time-based M-classes from [2] and [5] into the cycle-based A-classes from [7], and vice versa. They largely restate the content of [10] and include several examples of service-class conversion.

The method to transform time-based classification into cycle-based classification:

$$M \rightarrow \left(\frac{T_f}{K_m}\right) \cdot \left(\frac{1}{\alpha}\right) \cdot \left(\frac{1}{t_{av}}\right) \cdot K_p = C_f \rightarrow A \tag{3}$$

where:

- T_f – the total duration of use under full load [h];
- K_m – load spectrum factor by [5] and [6];
- α – drive utilization, this parameter describes the ratio between the duration of the cycle with load for the displacement and the average cycle time;
- t_{av} – average cycle time;
- K_p – load spectrum factor by [7].

Both K_m and K_p are load spectrum factors, but they are calculated differently.

The method to transform cycle-based classification into time-based classification:

$$A \rightarrow \left(\frac{C_f}{K_p}\right) \cdot \alpha \cdot t_{av} \cdot K_m = T_f \rightarrow M \tag{4}$$

As shown in [10], converting between M-classes and A-classes can be problematic. Changes in lifting height or hoisting speed alter the average working-cycle duration, t_{av}. For example, when converting class M5 to an A-class, reducing the lifting height from 12 m to 6 m and increasing the hoisting speed from 4 m/min to 20 m/min changes the classification from A2 to A6.

3.5. SRPS EN 13001-1:2016

In addition to crane classification, this European standard establishes the general principles for crane design using the limit-state method. It defines the U, Q (Q_p), D, and P classes in alignment with [7]. Although the standard does not provide a table combining the U and Q classes into A-classes, these classes may still be considered applicable in this context. Its main focus, however, is the definition of S-classes, which describe the stress-time history and are essential for fatigue-strength verification. S-classes are also defined in [14], which specifies requirements for verifying the structural integrity of crane steel structures. A more detailed analysis of S-classes and fatigue-strength verification is left for future research.

4. EXAMPLE OF DETERMINING THE SERVICE CLASS ACCORDING TO DIFFERENT STANDARDS

For the purposes of the example, an existing single girder workshop overhead crane shown in Fig. 1 is used.



Figure 1: Single girder bridge crane

The basic parameters of the crane are presented in Table 9.

Table 9: Basic crane parameters

rated capacity of crane	$m_Q = 3.2$ t
dead load	$m_{DSK} \approx 32$ kg
span	$L = 8.91$ m
lifting height	$h = 5$ m
load speed	$v_{max} = 11.3$ m/min
precision load speed	$v_{min} = 1.3$ m/min
operating conditions	standard

The crane is used to transport cast weights within the workshop hall. The mass of a single weight ranges from 460 to 470 kg. It operates 5 hours per day and performs 20 working cycles per hour: 5 cycles with one weight (m_1 = 468 kg), 10 cycles with four weights (m_2 = 1864 kg), 3 cycles with five weights (m_3 = 2328 kg), and 2 cycles with six weights (m_4 = 2796 kg).

4.1. Classification of the Crane in the Service Class According to JUS M.D1.020

Number of working cycles per hour:

$$n = 20 \tag{5}$$

Actual working hours per day

$$T_s = 5\,\text{h} \tag{6}$$

Number of working cycles per day:

$$p = n \cdot T_s = 20 \cdot 5 = 100 \tag{7}$$

The theoretical operating time

$$T_t = T_s \cdot \left(\frac{q_r}{100}\right)^3 = 5 \cdot \left(\frac{52.4}{100}\right)^3 = 0,72 \tag{8}$$

Relative load:

$$q_r = \frac{Q_s}{Q_n} \cdot 100\% = \frac{1678}{3200} \cdot 100\% = 52.4\% \tag{9}$$

Average load per day:

$$Q_s = \frac{Q_1 + Q_2 + Q_3 + Q_4}{p} = \frac{p_1 \cdot m_1 + p_2 \cdot m_2 + p_3 \cdot m_3 + p_4 \cdot m_4}{p} = \frac{25 \cdot 468 + 50 \cdot 1864 + 15 \cdot 2328 + 10 \cdot 2796}{100} \approx 1678\,\text{kg} \tag{10}$$

Based on the calculated parameters, the crane is classified as service class 2, which is consistent with the design documentation.

4.2. Classification of the Crane in the Service Class According to FEM 1.001

Determination of the U-class:

Number of working cycles during the service life of the crane:

$$n_{\max} = p \cdot N_{wd} \cdot N_y = 100 \cdot 260 \cdot 10 = 260000 \tag{11}$$

Number of working days per year:

$$N_{wd} = 260\,\text{days} \tag{12}$$

Design service life:

$$N_y = 10\,\text{years} \tag{13}$$

Based on the total number of working cycles over the crane's service life, the crane is classified as U5.

Determination of the Q-class:

$$\begin{aligned} k_p &= \frac{n_1}{n_{\max}} \cdot \left(\frac{m_{l1}}{m_{l\max}}\right)^3 + \frac{n_2}{n_{\max}} \cdot \left(\frac{m_{l2}}{m_{l\max}}\right)^3 + \frac{n_3}{n_{\max}} \cdot \left(\frac{m_{l3}}{m_{l\max}}\right)^3 + \frac{n_4}{n_{\max}} \cdot \left(\frac{m_{l4}}{m_{l\max}}\right)^3 = \\ &= \frac{p_1 \cdot N_{wd} \cdot N_y}{n_{\max}} \cdot \left(\frac{m_{l1}}{m_{l\max}}\right)^3 + \frac{p_2 \cdot N_{wd} \cdot N_y}{n_{\max}} \cdot \left(\frac{m_{l2}}{m_{l\max}}\right)^3 + \frac{p_3 \cdot N_{wd} \cdot N_y}{n_{\max}} \cdot \left(\frac{m_{l3}}{m_{l\max}}\right)^3 + \frac{p_4 \cdot N_{wd} \cdot N_y}{n_{\max}} \cdot \left(\frac{m_{l4}}{m_{l\max}}\right)^3 = \end{aligned} \tag{14}$$

$$= \frac{25 \cdot 260 \cdot 10}{260000} \cdot \left(\frac{468}{3200}\right)^3 + \frac{50 \cdot 260 \cdot 10}{260000} \cdot \left(\frac{1864}{3200}\right)^3 + \frac{15 \cdot 260 \cdot 10}{260000} \cdot \left(\frac{2328}{3200}\right)^3 + \frac{10 \cdot 260 \cdot 10}{260000} \cdot \left(\frac{2796}{3200}\right)^3 =$$

$$= \frac{65000}{260000} \cdot \left(\frac{468}{3200}\right)^3 + \frac{130000}{260000} \cdot \left(\frac{1864}{3200}\right)^3 + \frac{39000}{260000} \cdot \left(\frac{2328}{3200}\right)^3 + \frac{26000}{260000} \cdot \left(\frac{2796}{3200}\right)^3 =$$

$$= 0.00078 + 0.099 + 0.058 + 0.067 = 0,225$$

According to the load spectrum factor, the crane belongs to class Q2. By combining classes U5 and Q2, the crane is classified into service class A5, which is consistent with the recommended grouping based on the intended use of the crane.

4.3. Classification of the Crane in the Service Class According to ISO 4301-1:2016

Because the U-classes defined in [5] are harmonized with those in [7], the utilization class based on the total number of service-life cycles remains U5. The Q_p-class, however, differs from the Q-classes in [5], although the load spectrum factor is calculated in the same way. With $K_p = k_p = 0.225$, the crane is classified as Q_p3. Combining U5 and Q_p3 yields service class A5. This classification is consistent with the framework of [12] for bridge and gantry cranes.

5. CONCLUSION

Over the past two decades, crane service class definitions have changed, mainly due to national standardization practices. Most European manufacturers now classify cranes and their mechanisms according to ISO standards and FEM guidelines, as most national standards have been harmonized with current international standards. However, many industrial cranes designed under former Yugoslav standards remain in service across Serbia and other countries of the former Yugoslavia. Therefore, understanding technical parameters defined in these withdrawn standards, particularly service class, remains important for users and designers involved in maintenance, reconstruction, and modernization. Practical assessment of the remaining service life of such cranes primarily requires an accurate comparison between the service-class parameters defined in withdrawn national standards and those used in current international standards, which are based on a more advanced theory of material fatigue. This paper provides a concise overview of the key parameters defined in the relevant standards and guidelines and, through a classification example, demonstrates their application in determining crane service class. Because the theoretical background is much broader and more complex than can be covered here, future research will address the design approach for new cranes in accordance with the EN 13001 series of standards.

ACKNOWLEDGEMENTS

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-34/2026-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2026” (No. 01-3609/1).

REFERENCES

[1] Industrial cranes – Classification in service classes; JUS Standard M.D1.020:1964

[2] Rules for the design of series lifting equipment; Classification of mechanisms; FEM Guidelines 9.511, edition 06.1986

[3] Cranes; Steel structures; Verification and analyses; DIN Standard 15018-1:1984

[4] Lifting appliances; Principles relating to rope drives; Calculation and construction; DIN Standard 15020-1:1974

[5] Rules for design of hoisting appliances; Booklet 2; Classification and loading on structures and mechanism; FEM Guidelines 1.001-2, 3rd edition

[6] Cranes and lifting appliances – Classification – Part 1: General; ISO Standard 4301-1:1986

[7] Cranes and lifting appliances – Classification – Part 1: General; ISO Standard 4301-1:2016

[8] Cranes – General design – Part 1: General principles and requirements; EN Standard 13001-1:2015

[9] N. Babin, R. Mijajlović, and R. Šostakov, “Koncept pogonske čvrstoće u zakonskoj regulativi dizaličko – transportnih mašina u SFRJ”, Deseti naučno-stručni skup projektanata, konstruktora, proizvođača i korisnika “O transportnim procesima u industriji”, Beograd (Srbija), 28 septembar - 29 septembar 1988, pp. 171-178, (1988)

[10] M. Golder and G. Wagner, “As time goes by: Classification of hoists, cranes and lifting equipment,” Cranes today, pp. 40-41, (2016)

[11] Cranes and lifting equipment; Classification of Cranes and Lifting Equipment; FEM Guidelines 9.521, first edition

[12] Cranes and lifting appliances – Classification – Part 5: Bridge and gantry cranes; ISO Standard 4301-5:2025

[13] Cranes and lifting appliances – Selection of wire ropes – Part 1: General; ISO Standard 4308-1:2003

[14] Cranes – Proof of competence of steel structures; ISO Standard 20332:2016

SESSION A: EARTH MOVING AND TRANSPORTATION MACHINERY

A fuzzy AHP approach for evaluating circular efficiency of hydraulic excavator swing drive systems

Vesna Jovanović[1*], Nikola Petrović[1], Jovan Pavlović[1], Boban Nikolić[1]

[1] University of Niš, Faculty of Mechanical Engineering, Niš, Serbia

ARTICLE INFO

[*] **Correspondence:** vesna.jovanovic@masfak.ni.ac.rs

DOI: https://doi.org/10.46793/ET26.A05J **PAGES:** A33 – A39

ABSTRACT

Hydraulic excavators are among the most energy-intensive categories of heavy mobile machinery, characterized by intensive duty cycles, high energy demand, and significant environmental impact. Within the circular economy framework, the evaluation and design of key subsystems should not be limited to conventional performance indicators but should also include life-cycle-oriented criteria, such as durability, remanufacturing potential, resource efficiency, reliability, and user-related economic benefits. This paper proposes an integrated decision-making framework for assessing the circular efficiency of the hydraulic excavator swing drive system by combining engineering performance analysis with the fuzzy Analytic Hierarchy Process (fuzzy AHP). The application of triangular fuzzy numbers enables the modeling of uncertainty and subjectivity in expert judgments, thereby improving the robustness and reliability of criteria weighting. Based on a previously conducted energy-efficiency analysis of a 17-ton hydraulic excavator, two representative design alternatives were selected from a broader set of feasible configurations and evaluated according to hierarchically structured circular economy criteria. The results show that alternative L1 achieves higher overall circular efficiency than alternative L2, mainly due to better performance in energy efficiency, durability, and remanufacturing-related criteria. The proposed fuzzy AHP framework supports sustainable engineering design and component selection in heavy machinery systems.

KEYWORDS

Hydraulic excavator, Swing drive, Circular economy; Fuzzy AHP.

1. INTRODUCTION

In the current context of the European Union and the global industrial landscape, the heavy machinery sector is under increasing pressure to improve its energy and material efficiency in line with circular economy principles. The transition from linear to circular production models requires not only technological improvements, but also the development of decision-support frameworks capable of integrating technical, environmental, and economic aspects throughout the product life cycle. In this context, remanufacturing and improved drivetrain design represent important mechanisms for extending service life, reducing resource consumption, and lowering the total cost of ownership. According to studies by the Ellen MacArthur Foundation [1], machines with modular components, disassembly capability, and a high degree of reparability can achieve significantly lower emissions and resource consumption throughout their life cycle. Industrial companies such as Caterpillar and Engel confirm these findings through long-term remanufacturing programs, in which components are disassembled, restored, tested, and returned to service in “like-new” condition [2-4]. These practices demonstrate that engineering design decisions, especially at the subsystem and component level, have a direct influence on circular performance.

Hydraulic excavators are a representative category of heavy mobile machinery characterized by intensive duty cycles, high energy demand, and significant environmental impact. One of their key subsystems is the swing drive system, which enables the rotation of the upper platform during digging, material transfer, unloading, and return operations. The performance of this system depends on the interaction between the hydraulic motor, gearbox, axial bearing, and the operating conditions defined by load torque and platform rotational speed.

In the authors’ previous research [4], a detailed energy analysis was conducted for the swing drive system of a 17-ton hydraulic excavator, consisting of a hydraulic motor *2.3* (Figure 1a), gearbox *2.4*, and axial bearing *2.5*.

The drive hydraulic motor is supplied, through the directional control valve 2.2, by the hydraulic pump *2.1*, driven by the diesel engine 1 through the power divider *1.2* and the elastic coupling *1.1*. The rotation of the platform is achieved by the rigid connection of the gearbox housing and the non-toothed ring of the axial bearing to the rotating platform L_2 (Figure 1b), while the gear z_{24} on the output shaft of the gearbox rotates in mesh with the stationary toothed ring of the axial bearing z_{25}, connected to the moving member of the excavator L_1.

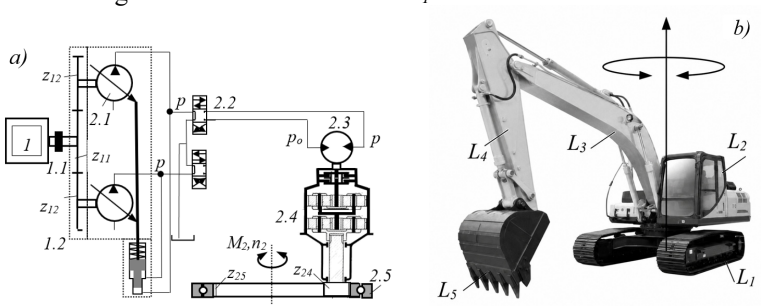


Figure 1: a) Swing drive of hydraulic excavator platforms with an open hydrostatic circuit
b) Hydraulic crawler excavator

The functional parameters of the drive mechanism of the excavator rotating platform are the output torque M_2, equation (1), and the rotational speed n_2 of the platform, equation (2), determined by the equations [5]:

$$M_2 = \frac{(p - p_o) q_{mmax} \cdot \varepsilon_m}{2\pi} \cdot \eta_{mm} \cdot i_r \cdot \eta_r \cdot i_l \cdot \eta_l \tag{1}$$

$$n_2 = \frac{q_{pmax} \cdot \varepsilon_p \cdot n_p}{q_{mmax} \cdot \varepsilon_m} \cdot \eta_{pv} \cdot \eta_{mv} \cdot \frac{1}{i_r \cdot i_l} \tag{2}$$

where: q_{pmax}, q_{mmax} – the maximum specific flow of the hydraulic pump and the hydraulic motor, p, p_o – the pressure in the extension and retraction duct of the hydraulic motor, n_p – the number of the hydraulic pump revolutions, i_r, η_r – the transmission ratio and the degree of efficiency of the reducer, η_{pv}, η_{pm}, η_{mv}, η_{mm} – the volumetric and mechanical degree of efficiency of the hydraulic pump and the hydraulic motor, $\varepsilon_p = q_p/q_{pmax}$ – the hydraulic pump regulation range, $\varepsilon_m = q_m/q_{mmax}$ – the hydraulic motor regulation range, η_l – the degree of efficiency between the reducer and the axial bearing. Equations (1) and (2) show that, for the same desired functional parameters of the platform swing drive, different drive variants can be generated through the synthesis procedure by changing the pressure and flow rate of the hydraulic pump and by changing the size of the hydraulic motor and gearbox. Based on the specified functional parameters, imposed constraints, and formed databases of available drive component models, 94 possible variant solutions of the swing platform mechanism drive of the considered excavator model were generated using the developed program. From this set, two representative variants — VE10/1 and VE12/1 — were selected for further evaluation [5]. The analysis of the obtained results showed that the energy efficiency of the drive transmission variants differs considerably and changes with the functional parameters of the excavator platform rotation. In the zones of higher load torques and lower rotational speeds of the platform, corresponding to the transfer of the excavated material from the digging plane toward the unloading plane, variant VE10/1, with a hydraulic motor of lower specific flow and a gearbox of higher transmission ratio, achieves higher efficiency values. In contrast, variant VE12/1, with a hydraulic motor of higher specific flow and a reducer of lower transmission ratio, achieves higher efficiency values at lower load torques and higher rotational speeds of the platform, corresponding to the return from the unloading plane to a new digging plane, particularly at larger swing angles when higher platform rotational speeds can be achieved (Figure 2).

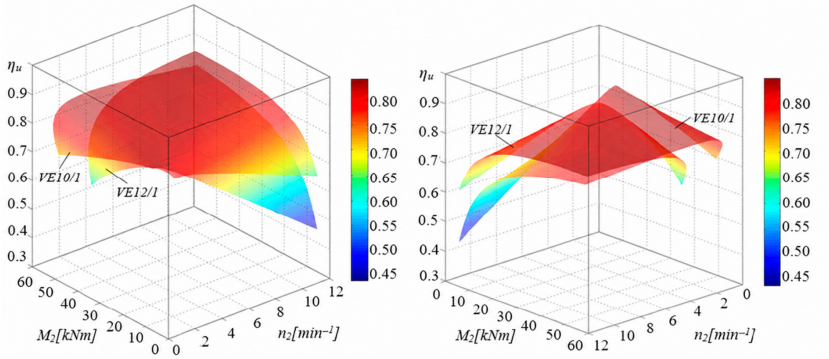


Figure 2: Efficiencies of the transmission: Both drive variants VE10/1 and VE12/1 [5]

Nevertheless, energy efficiency represents only one dimension of circular performance. A comprehensive assessment of swing drive systems must also include criteria such as durability, modularity, ease of disassembly, material quality, reparability, remanufacturing potential, resource efficiency, reliability, and user-oriented economic benefits. Many of these

criteria are inherently qualitative or uncertain, especially those related to reparability, environmental impact, and user perception, which makes their integration into conventional decision-making models difficult.

Despite the growing application of multi-criteria decision-making methods in engineering design, most existing studies rely on classical approaches, such as the Analytic Hierarchy Process (AHP), which assume precise and deterministic expert judgments. This represents a methodological limitation when dealing with complex engineering systems characterized by uncertainty and subjective evaluations, as is the case in circular economy-based assessments. To address this limitation, this paper proposes an enhanced decision-making framework based on the fuzzy Analytic Hierarchy Process (fuzzy AHP), where pairwise comparisons are expressed using triangular fuzzy numbers in order to better capture the vagueness and ambiguity of expert judgments.

Accordingly, the aim of this paper is to develop a comprehensive framework for evaluating the circular efficiency of hydraulic excavator swing drive systems by combining engineering analysis with fuzzy AHP-based multi-criteria decision-making. The proposed criteria set encompasses energy efficiency, durability and remanufacturing potential, environmental impact, resource efficiency, reliability, dynamic stability, and user-related economic benefits. Such a framework enables a more accurate and robust assessment of technical solutions and contributes to the rational design and selection of components aligned with circular economy principles in heavy machinery systems.

2. IDENTIFICATION OF CRITERIA BASED ON CIRCULAR ECONOMY PRINCIPLES

The process of defining criteria for the optimal selection of the hydraulic excavator swing drive system is based on the integration of technical, environmental, and economic aspects, in accordance with contemporary circular economy principles. The starting point in this process is the analysis of existing technical models of the swing drive system, including parameters such as energy efficiency, mechanical loads, dynamic stability, and losses in hydraulic and mechanical transmission. While these elements form the foundation of traditional engineering approaches, they must be extended by incorporating life-cycle-oriented criteria that reflect circular economy requirements. According to the European Commission's circular economy framework and the conceptual model of the Ellen MacArthur Foundation, industrial systems should be designed to enable life extension, remanufacturing, reparability, and reduced resource consumption throughout the entire life cycle. These principles are particularly relevant for heavy machinery systems, where components such as swing drives exhibit high material and energy intensity, but also significant potential for restoration through remanufacturing processes. The hierarchical structure of the decision model is defined across three levels. The first level represents the overall objective, which is the selection of a swing drive solution that maximizes circular efficiency by achieving an optimal balance between technical performance, durability, resource utilization, and environmental impact. The second level consists of seven primary criteria: energy efficiency, dynamic stability and mechanical loads, durability and remanufacturing potential, environmental impact, resource efficiency, reliability, and economic benefit to the user. The third level includes a set of subcriteria that enable a more detailed and structured evaluation of each criterion, thereby supporting consistent expert judgment within the fuzzy AHP framework. Energy efficiency represents one of the dominant evaluation domains and includes subcriteria such as hydraulic motor efficiency, transmission efficiency, losses in the gearbox, and total energy consumption per operating cycle. Given the high-frequency operation of the swing drive system, these parameters directly influence both environmental impact and operational costs over the machine's life cycle. Dynamic stability and mechanical loads constitute the second group of criteria and include parameters such as shock torques, angular accelerations, vibration levels, and thermal behavior. These factors are critical for assessing structural integrity and material fatigue, and from a circular economy perspective, they indirectly contribute to extending the service life of key components and reducing replacement frequency. Durability and remanufacturing potential form the third category and represent a central pillar of circular efficiency. Subcriteria include fatigue resistance, modularity of design, ease of disassembly, and compatibility with remanufacturing processes. Systems designed with these characteristics enable component recovery and reuse, resulting in substantial environmental and economic benefits. The environmental impact category includes parameters such as CO_2 emissions associated with manufacturing, hydraulic fluid consumption, and overall ecological footprint. These subcriteria enable the quantification of environmental performance and support the identification of solutions with reduced life-cycle impact. Resource efficiency focuses on system mass, material selection, and the use of critical raw materials. Solutions that employ recyclable materials or require less intensive processing during remanufacturing contribute to improved circular performance and reduced material footprint. Reliability and economic benefit to the user represent the final evaluation domains. Reliability is reflected through system stability, failure rates, and maintenance intervals, while economic criteria include maintenance costs, residual value, and availability of remanufacturing infrastructure. In addition, system availability is considered as a supporting indicator that directly links technical performance with operational efficiency, reflecting downtime duration and recovery capability. From a methodological perspective, it is important to emphasize that several of the defined criteria—particularly those related to remanufacturing potential, environmental impact, and user perception—are inherently uncertain and difficult to quantify precisely. For this reason, the application of the fuzzy AHP approach enables the incorporation of vagueness and subjectivity in expert evaluations by representing pairwise comparisons through triangular fuzzy numbers. This allows for a more realistic and robust modeling of decision-making conditions compared to classical deterministic approaches. Consequently, the defined criteria system provides a comprehensive and structured basis for evaluating the circular efficiency of swing drive systems, while the integration of fuzzy AHP ensures that the inherent uncertainty of engineering judgments is adequately addressed, thereby enhancing the reliability and applicability of the proposed decision-making framework.

3. FUZZY AHP EVALUATION OF CRITERIA AND ASSIGNMENT OF WEIGHT COEFFICIENTS

In order to enable a more realistic and robust comparison of criteria characterized by uncertainty and subjectivity, the classical Analytic Hierarchy Process (AHP) is extended through the application of the fuzzy AHP approach [7]. Unlike the conventional AHP method, which assumes precise numerical judgments, fuzzy AHP incorporates vagueness in expert evaluations by representing pairwise comparisons using triangular fuzzy numbers (TFNs). This allows for a more accurate modeling of decision-making conditions, particularly in the context of circular economy-based assessments, where several criteria are inherently qualitative and imprecise.

The linguistic importance of criteria is expressed using a fuzzy extension of the Saaty scale [6]. Each judgment is represented by a triangular fuzzy number:

$$\tilde{a}_{ij} = (l_{ij}, m_{ij}, u_{ij}) \tag{3}$$

where l_{ij}, m_{ij}, and u_{ij} denote the lower, modal, and upper bounds of the expert preference, respectively. The reciprocal condition is defined as:

$$\tilde{a}_{ij} = \left(\frac{1}{u_{ji}}, \frac{1}{m_{ji}}, \frac{1}{l_{ji}}\right), i \neq j \tag{4}$$

The fuzzy pairwise comparison matrix is then constructed as:

$$\tilde{A} = [\tilde{a}_{ij}]_{nxn} \tag{5}$$

where n is the number of criteria.

To derive the weights of criteria, the geometric mean method proposed by Buckley is applied. For each criterion, the fuzzy geometric mean is calculated as:

$$g_i = \left(\prod_{j=1}^{n} a_{ij}\right)^{1/n} \tag{6}$$

The fuzzy weight vector is then obtained by normalization:

$$\widetilde{w}_i = \tilde{g}_i \otimes (\sum_{i=1}^{n} \tilde{g}_i)^{-1} \tag{7}$$

where $\otimes$ denotes fuzzy multiplication.

In order to obtain crisp values suitable for further analysis, the fuzzy weights are defuzzified using the centroid method:

$$\widetilde{w}_i = \frac{(l_i + m_i + u_i)}{3} \tag{8}$$

For each criterion, a fuzzy pairwise comparison matrix of alternatives is constructed following the same procedure. Local priorities of alternatives are calculated using the fuzzy geometric mean method and subsequently defuzzified to obtain crisp values. This ensures methodological consistency across all levels of the hierarchy.

The overall priority of each alternative is determined by aggregating local priorities with the corresponding criteria weights:

$$P_k = \sum_{i=1}^{n} w_i^* \cdot p_{ik} \tag{9}$$

where p_{ik} represents the local priority of alternative k with respect to criterion i.

3.1. Results and Interpretation

The application of the fuzzy AHP procedure was carried out in several consecutive steps in order to obtain criteria weights and the final ranking of the selected swing drive alternatives. First, the linguistic judgments used for pairwise comparison were transformed into triangular fuzzy numbers according to the fuzzy Saaty scale (Table 1.). This scale enables expert preferences to be expressed not as fixed crisp values, but as intervals that include the lower, modal, and upper values of judgment. In this way, uncertainty and subjectivity in expert assessments are incorporated directly into the evaluation process.

Table 1: Fuzzy Saaty scale (linguistic judgments and tfns)

Linguistic scale	Crisp value	Triangular fuzzy number (TFN)
Equal importance	1	(1, 1, 1)
Weak importance	2	(1, 2, 3)
Moderate importance	3	(2, 3, 4)
Moderate plus	4	(3, 4, 5)

Strong importance	5	(4, 5, 6)
Strong plus	6	(5, 6, 7)
Very strong importance	7	(6, 7, 8)
Very very strong	8	(7, 8, 9)
Extreme importance	9	(8, 9, 10)

Based on the defined linguistic scale, the original crisp pairwise comparison matrix was transformed into a fuzzy comparison matrix (Table 2.). The matrix includes seven primary criteria: energy efficiency, dynamic stability, durability and remanufacturing potential, environmental impact, resource efficiency, reliability, and economic aspect. The transformation into triangular fuzzy numbers preserves the initial expert judgments, while at the same time allowing the uncertainty of these judgments to be considered in the calculation of criteria weights.

Table 2: Fuzzy pairwise comparison matrix of criteria

Criteria a_{ij}	energy efficiency a_1	dynamic stability a_2	durability/ remanufacturing a_3	environmental impact a_4	resource efficiency a_5	reliability a_6	economic aspect a_7
energy efficiency a_1	(1,1,1)	(2,3,4)	(1,1,1)	(4,5,6)	(4,5,6)	(2,3,4)	(4,5,6)
dynamic stability a_2	(1/4,1/3,1/2)	(1,1,1)	(1/4,1/3,1/2)	(2,3,4)	(2,3,4)	(1,1,1)	(2,3,4)
durability/ remanufacturing a_3	(1,1,1)	(2,3,4)	(1,1,1)	(4,5,6)	(4,5,6)	(2,3,4)	(4,5,6)
environmental impact a_4	(1/6,1/5,1/4)	(1/4,1/3,1/2)	(1/6,1/5,1/4)	(1,1,1)	(1,1,1)	(1/4,1/3,1/2)	(1,1,1)
resource efficiency a_5	(1/6,1/5,1/4)	(1/4,1/3,1/2)	(1/6,1/5,1/4)	(1,1,1)	(1,1,1)	(1/4,1/3,1/2)	(1,1,1)
reliability a_6	(1/4,1/3,1/2)	(1,1,1)	(1/4,1/3,1/2)	(2,3,4)	(2,3,4)	(1,1,1)	(2,3,4)
economic aspect a_7	(1/6,1/5,1/4)	(1/4,1/3,1/2)	(1/6,1/5,1/4)	(1,1,1)	(1,1,1)	(1/4,1/3,1/2)	(1,1,1)

After forming the fuzzy pairwise comparison matrix, the fuzzy weights of the criteria were calculated (Table 3.) and then defuzzified to obtain crisp values suitable for further ranking of alternatives. The obtained results show that the highest weights are assigned to energy efficiency and durability/remanufacturing, each with a defuzzified value of 0.2939. These two criteria together account for almost 58.8% of the total importance, which confirms their dominant role in the evaluation of circular efficiency. This is expected, since energy efficiency directly affects operating costs and environmental impact, while durability and remanufacturing potential directly influence service life extension, component reuse, and circular value preservation. Dynamic stability and reliability follow with equal defuzzified weights of 0.1296. Their moderate importance indicates that mechanical behavior, operational stability, failure rates, and maintenance intervals are also relevant for circular performance, although they have a lower influence than energy efficiency and durability. Environmental impact, resource efficiency, and economic aspect have lower, but still relevant, weights of 0.0510 each. These criteria support the broader life-cycle interpretation of the system, especially in terms of material footprint, ecological effects, and user-related economic benefits. The local priorities of the alternatives were then determined for each criterion and aggregated with the corresponding criteria weights (Table 4.). The results show that alternative L1 achieves a higher overall priority than alternative L2. The overall priority of L1 is 0.6889, while the overall priority of L2 is 0.3111. This ranking is primarily influenced by the better performance of L1 in the most important criteria, especially energy efficiency and durability/remanufacturing. In addition, L1 also shows better local priorities in dynamic stability, environmental impact, reliability, and economic aspect, while both alternatives have equal priority with respect to resource efficiency**.**

Table 3: Fuzzy weights and defuzzified values

Criteria	Fuzzy weight (l, m, u)	Defuzzified weight
Energy efficiency (a1)	(0.2001, 0.2972, 0.4306)	0.2939
Dynamic stability (a2)	(0.0821, 0.1274, 0.1998)	0.1296
Durability/reman. (a3)	(0.2001, 0.2972, 0.4306)	0.2939
Environmental impact (a4)	(0.0366, 0.0502, 0.0742)	0.0510
Resource efficiency (a5)	(0.0366, 0.0502, 0.0742)	0.0510
Reliability (a6)	(0.0821, 0.1274, 0.1998)	0.1296
Economic aspect (a7)	(0.0366, 0.0502, 0.0742)	0.0510

Table 4: Local Priorities and Overall Ranking of Alternatives

Criteria	L1	L2
Energy efficiency	0.75	0.25
Dynamic stability	0.67	0.33
Durability/remanufacturing	0.67	0.33
Environmental impact	0.67	0.33
Resource efficiency	0.50	0.50
Reliability	0.67	0.33
Economic aspect	0.75	0.25
Overall priority	0.6889	0.3111

The obtained ranking indicates that alternative L1 provides a more favorable balance between engineering performance and circular economy requirements. Its advantage is mainly associated with criteria that have the greatest influence on the final decision, namely energy efficiency and durability/remanufacturing potential. At the same time, the difference between the alternatives should not be interpreted as absolute or extreme, since the two variants are structurally and functionally similar and belong to the same group of feasible swing drive solutions.

The fuzzy AHP results also show that the obtained criteria weights are highly consistent with the classical AHP results, while the ranking of alternatives remains unchanged. This confirms the robustness and reliability of the proposed decision-making framework. The main advantage of the fuzzy AHP approach is that it captures uncertainty in expert judgments and avoids overly rigid preference assignments typical of classical AHP. As a result, the final ranking is both mathematically consistent and better aligned with real engineering decision-making conditions, where criteria are often partly qualitative, uncertain, and dependent on expert interpretation.

4. DISCUSSION AND BROADER IMPLICATIONS OF THE FRAMEWORK (FUZZY AHP)

The results obtained using the fuzzy AHP approach provide more than a simple numerical ranking of the considered drive alternatives; they offer insight into the key technical attributes that contribute to the circular value of the system under conditions of uncertainty. The high importance assigned to energy efficiency and durability with remanufacturing potential confirms that the primary drivers of circular performance in heavy machinery are directly embedded in the design of the drive system. In particular, the selection of hydraulic motor displacement, gearbox configuration, and structural modularity has a decisive influence not only on operational performance, but also on service life extension, remanufacturing feasibility, and overall life-cycle costs. The application of fuzzy AHP enables a more realistic interpretation of these results by accounting for the inherent vagueness and subjectivity in expert evaluations. Criteria such as remanufacturing potential, environmental impact and user-related benefits cannot be precisely quantified in early design stages; therefore, the use of triangular fuzzy numbers allows these aspects to be incorporated into the decision-making process in a structured and analytically consistent manner. This enhances the reliability of the obtained priorities and strengthens the link between engineering judgment and decision outcomes. From a broader perspective, the proposed framework represents a practical bridge between engineering design and circular economy principles. The selected criteria-reparability, modularity, availability, and resource efficiency-are fully aligned with European policy objectives aimed at extending product life cycles and supporting remanufacturing markets (European Commission, 2015). Through the application of fuzzy AHP, these policy concepts are translated into measurable decision variables, enabling their direct integration into engineering design processes. Industrial practices further support these findings. Remanufacturing programs implemented by companies such as Caterpillar and Engel demonstrate that design decisions made at the component level have long-term implications for resource consumption, energy efficiency, and operational performance. Systems designed with higher modularity and reparability enable more efficient refurbishment processes, reduce material usage, and improve equipment availability, thereby contributing to both economic and environmental sustainability (Caterpillar Inc., 2020; ENGEL, 2020). From the perspective of end users, the obtained results have immediate practical relevance. The alternative identified as optimal (L1) achieves superior performance in the criteria most closely related to user value, including service life, system availability, and remanufacturing capability. This indicates that the selection of an appropriate drive configuration directly affects total cost of ownership, maintenance planning, and operational reliability. Consequently, the proposed framework supports decision-making not only at the design stage, but also in the context of equipment procurement and lifecycle management. In addition, the use of fuzzy AHP provides an important methodological contribution by ensuring that decision outcomes remain stable despite the presence of uncertainty in input judgments. This is particularly relevant for complex engineering systems, where precise data are often unavailable or incomplete. By capturing uncertainty within the evaluation process, the framework increases confidence in the obtained ranking and supports more robust and defensible engineering decisions. Therefore, the proposed fuzzy AHP-based framework can be applied not only as a tool for the selection of swing drive systems, but also as a broader decision-support methodology for designers, manufacturers, and policy makers. It enables the systematic incorporation of circular economy principles into technical decision-making and provides a foundation for the development of sustainable design guidelines, eco-design standards, and procurement strategies in the heavy machinery sector.

5. CONCLUSIONS

The proposed methodology has certain limitations that should be taken into account. Although the fuzzy AHP approach enables the incorporation of uncertainty and subjectivity in expert judgments, the evaluation process still depends on the quality and consistency of expert input, which may vary depending on the industrial context and type of machinery. In addition, the assessment of specific circularity-related criteria, such as environmental impact and remanufacturing potential, is influenced by the availability and reliability of practical data, which may differ across manufacturers and operational environments. In this paper, an integrated methodological framework was developed for assessing the circular efficiency of the swing drive system of a hydraulic excavator by combining engineering analysis with fuzzy AHP-based multi-criteria decision-making. Based on a previously conducted energy analysis, a comprehensive set of criteria was defined in accordance with circular economy principles, encompassing technical, environmental, resource, and economic aspects over the entire life cycle of the system. The results of the fuzzy AHP analysis indicate that alternative L1 achieves higher overall circular efficiency compared to alternative L2, primarily due to its superior performance in the most influential criteria, namely energy efficiency and durability with remanufacturing potential. These findings confirm that key design parameters, such as hydraulic motor displacement and gearbox configuration, play a crucial role not only in improving operational efficiency but also in extending component lifespan and enhancing circular performance. At the same time, the results show that the differences between the analyzed alternatives are not extreme, reflecting their structural and functional similarity. The application of fuzzy AHP ensures that such differences are evaluated in a balanced and realistic manner, avoiding overemphasis on individual criteria and enabling a stable and consistent ranking under conditions of uncertainty. The proposed framework represents a practical and robust tool for supporting decision-making in the design and selection of heavy machinery components. By integrating circular economy principles into a structured evaluation model, it enables the translation of abstract sustainability concepts into concrete engineering criteria. Furthermore, the alignment of the most influential criteria with European circular economy priorities highlights the relevance of the framework for supporting eco-design strategies and sustainable procurement policies. From the perspective of end users, the advantage of the selected alternative (L1) indicates the potential for reduced total cost of ownership, increased system availability, and more reliable maintenance planning. These findings demonstrate that engineering decisions at the design stage have direct economic and operational implications, reinforcing the importance of integrating circularity considerations into early design phases. Although the model was developed for the swing drive system of a hydraulic excavator, the structure of the proposed criteria is sufficiently general to be adapted to other subsystems of heavy machinery, such as travel drives or working hydraulic systems. This indicates the broader applicability of the framework across different engineering domains. Future research should focus on further improving the methodology by incorporating real operational data, expanding the set of evaluated alternatives, and validating the model through expert-based and data-driven approaches. In addition, the integration of fuzzy AHP with advanced methods, such as machine learning or hybrid MCDM techniques, represents a promising direction for enhancing the accuracy, adaptability, and practical applicability of decision-support models in the context of circular economy-oriented engineering design.

ACKNOWLEDGEMENTS

This research was financially supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-33/2026-03/ 200109).

REFERENCES

[1] Ellen MacArthur Foundation. Caterpillar: Design and business model consideration for heavy machinery remanufacturing. Case Studies, Ellen MacArthur Foundation. (2021). All Rights Reserved. https://www.ellenmacarthurfoundation.org/case-studies/design-and-business-model-considerations-for-heavy-machinery-remanufacturing - retrieved on 12 Mar. 2025.

[2] Caterpillar Inc., Remanufacturing Overview, (2020)

[3] https://www.caterpillar.com/en/company/sustainability/remanufacturing.html- retrieved on 10. Jun. 2025.

[4] ENGEL Austria GmbH, Circular Economy Initiatives, (2020)

[5] V. Jovanović, D. Janošević, D. Marinković, N. Petrović, and B. Nikolić, "Energy Efficiency Analysis of Hydraulic Excavators' Swing Drive Transmission", Machines, Vol. 13(7), p. 596, https://doi.org/10.3390/machines13070596, (2025)

[6] T.L. Saaty, "Decision Making with the Analytic Hierarchy Process", International Journal of Services Sciences, Vol. 1(1), pp. 83–98, https://doi.org/10.1504/IJSSci.2008.01759, (2008)

[7] Y. Liu, C.M. Eckert, and C. Earl, "A review of fuzzy AHP methods for decision-making with subjective judgements", Expert Systems with Applications, Vol. 161, p.113738, https://doi.org/10.1016/j.eswa.2020.113738, (2020)

SESSION A: EARTH MOVING AND TRANSPORTATION MACHINERY

Comparative analysis of rack layouts according to the storage unit placement method

Nenad Miloradović[1*], Rodoljub Vujanac[1], Mirjana Piskulić[1], Jasna Glišović[1]

[1] University of Kragujevac, Faculty of Engineering, Kragujevac, Serbia

ARTICLE INFO

[*] **Correspondence:** mnenad@kg.ac.rs

DOI: https://doi.org/10.46793/ET26.A06M **PAGES:** A41 – A49

ABSTRACT

The paper presents a comparative analysis of two selective pallet rack configurations based on the storage unit placement method, using the same rack orientation within the warehouse space. Selective pallet racks are simple, efficient, and widely used systems for storing palletized goods. They provide direct and unobstructed access to every individual pallet at any time. The storage unit was defined, and the sizing of the beams and upright frames was performed. Appropriate storage and retrieval equipment was selected. The rack layout in the warehouse was determined in accordance with warehouse system design principles. Adequate structural components were chosen, and the total capacity of the rack warehouse was calculated. Within the comparative analysis of selective racking solutions in the planned warehouse space, designs with both transverse and longitudinal storage unit placement were considered. The design focus was on increasing warehouse capacity and reducing rack construction costs.

KEYWORDS

Warehouse, Pallet Rack, Forklift.

1. INTRODUCTION

Existing storage systems have the primary task of efficiently storing and distributing a large quantity of goods in the shortest possible time in a defined space.

Pallet racking storage system consists of the three basic elements: racks, pallets, and forklifts or handling equipment. Each of them is selected based on materials being stored, the storage requirements, operating and positioning systems and the applied management [1]. In general, spatial frame structure of the pallet racking system is made of the side frames and the horizontal beams, usually made of thin-walled cold-formed profiles, to make the pallet rack configuration.

Selective pallet racking systems are the most common. They can accept a large number of different unit loads and allow direct access to the pallets [2]. As an upgrade to standard racks, a double deep pallet racking is used. With this solution, each pallet shelf holds two rows of pallets instead of one. The mobile pallet racking variant maximizes storage capacity by eliminating multiple static aisles. Efficient solutions such as drive-in and drive-through racking systems are in use. Pallet flow racking systems have been developed that use gravity or wheel and roller systems to transfer and store pallets.

The primary components of all these racks include vertical elements (such as columns), horizontal elements (such as front and rear beams) and horizontal and/or diagonal braces [3]. Elements that provide the stable spatial structure are an essential part of every type of rack. They are categorized into basic and additional elements. The basic elements are: vertical frames, horizontal beams, baseplates with anchors, and spacers.

A complete analysis of the rack structure and each component is achieved by using the determined structural properties and applying European standards and regulations. [4-6].

2. DESIGN OF SELECTIVE PALLET RACKING SYSTEMS

This storage system has three basic elements: racks, pallets, and forklifts or handling equipment. Each of them is adopted based on the materials being stored, storage requirements, working and positioning systems, and the management that is being applied [7]. Since storage systems represent a large capital investment, it is necessary to properly define the layout and configuration of the storage equipment [8].

2.1. Defining a Storage Unit

When designing a warehouse, the first step is to define the storage unit, i.e., the corresponding pallet with its load. Defining the storage unit provides information regarding its length, width, height, shape, tolerances, and total weight.

2.2. Dimensioning of Storage Unit Supports

The dimensioning of the storage unit supports is carried out according to the load of the pair of the supports. The length of the storage unit supports is determined by the number of storage units in one bay, including the required clearances, defined by regulations and standards [9-10]. The aforementioned clearances are present between adjacent storage units, as well as between the outermost points of the storage units and the upright frames. In addition to the dimensions of the storage unit, the method of its placement also has an undeniable impact on determining the beam (support) length. Storage unit placement can be: transverse and longitudinal.

2.2.1. Racking Beam Length

Based on the sum of the prescribed clearances defined according to [7], both between the storage units and the upright frames, and between the storage units themselves, the beam length is adopted as a standard dimension defined by the manufacturer, which closely matches the calculated value.

2.2.2. Load Capacity of a Pair of Horizontal Beams

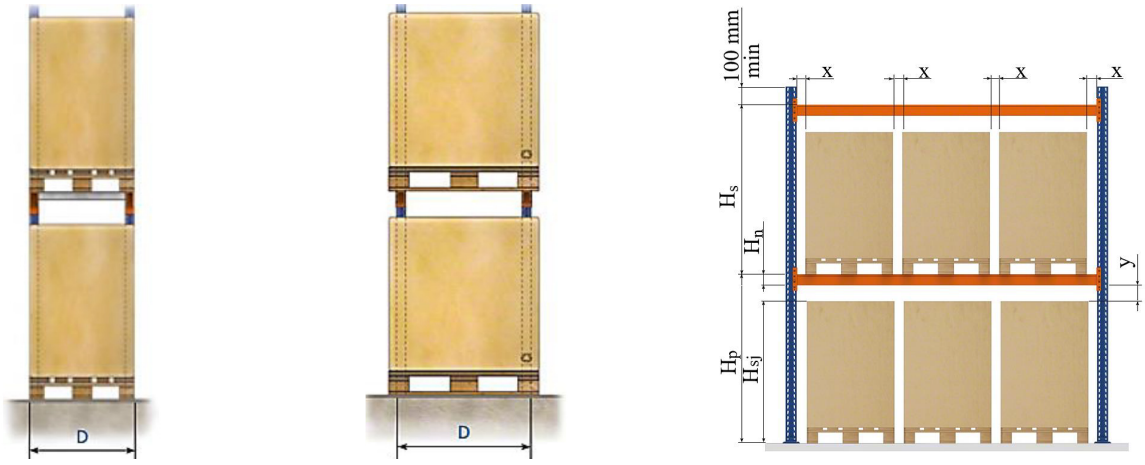
The load capacity of a pair of horizontal beams is determined by the number and weight of the storage units that need to be stored in one bay. As a product of these two parameters, the adopted load capacity of the pair of horizontal beams must be equal to or greater than the calculated value. Then, according to the length of the beams and the required load capacity, the first larger standard profile of the storage unit beam is adopted with defined characteristics equal to or greater than the calculated ones. The defined characteristics relate, in addition to the basic characteristics, to the geometry of the adopted profile.

2.3. Dimensioning of the Upright Frame

When dimensioning a vertical frame, its width, height of one floor and total height are determined, according to the recommendations in accordance with [7,8].

2.3.1. Upright Frame Width

The width of the upright frame is influenced by the method of storage unit placement and the available transport and handling equipment, so the width of the vertical frame is equal to the shortest dimension of the storage unit at the base that enters the rack in the case of transverse stacking. If the stacking is longitudinal, the width of the vertical frame is 50 mm ÷ 100 mm shorter than the length of the storage unit. In this case, its ends have an overhang of at least 50 mm over the beam.



a) b) c)

Figure 1: View of the storage unit layout: a) transverse, b) longitudinal, c) upright frame dimensions

Figure 1 illustrates a parallel view of the position of the storage unit in the case of transverse and longitudinal stacking. Dimension D in the figure indicates the width of the frame and the position of the storage unit during placement.

The dimensions indicated in Figure 1 are: H_{sj} − height of the storage unit, H_p − height from the base to the first floor, H_s − height of one floor, H_n − height of the horizontal storage unit beam, x − horizontal clearances between storage units, and between uprights and storage units, and y − vertical clearance between the top of the storage unit and the horizontal beam.

When designing, in order to achieve optimal load storage, proper placement is extremely important, especially when stacking longitudinally, because in the case of improperly distributed cargo, it can tip over.

2.3.2. Floor Height or Pitch

The height of the rack floor is determined by the height of the storage unit and the required clearances in the vertical direction defined according to [7]. It is necessary to pay attention to the position of the first floor when determining the total height of the vertical frame, whether it is located directly on the floor or is raised to a specific height above it. The resulting calculated value of the floor height must be divisible by the perforation pitch of the vertical frame column. Based on the data obtained, it is necessary to adopt a standard beam pitch value according to the relevant manufacturer's catalogue.

2.3.3. Vertical Frame Height and Column Profile

The height of the vertical frame is determined by the height of the storage unit, the necessary clearances between the storage unit and the beam, and the beam height, and it is definitely limited by the available height of the storage space. Figure 1 c) shows all the dimensions by which the vertical frame is fully dimensioned. The key elements that influence the dimensioning of the frame are: the height of the storage space, the height of the storage unit, the position of the first floor, and the capabilities of the material handling equipment.

Based on the number of floors and the number of storage units in the bay, the maximum load capacity of the vertical frame can be determined. The load capacity of the vertical frame is calculated as the sum of the load capacity of each bay, or as the product of the load capacity of one bay/floor and the number of floors, excluding the load capacity of the one located directly on the floor. After that, it is possible to adopt the appropriate standard column profile of defined dimensions and shapes.

2.4. Dimensioning of Cross Beams of a Storage Unit

In the case of transverse placement of a storage unit, in addition to dimensioning a pair of horizontal beams, it is also necessary to dimension the transverse supports of the storage units according to their load capacity. These supports are placed in a direction perpendicular to the horizontal beams and are appropriately supported by a pair of horizontal supports.

According to the above, a minimum of two cross-beams, usually of rectangular cross-section, with defined characteristics, are always adopted for one storage unit. Therefore, the height of the cross-beam must not be greater than the cross-section height of the horizontal beam of the storage unit. When it comes to longitudinal stacking, cross-beams are generally not used, except in the case of poor quality pallets. Most often, the storage unit is supported by a pair of horizontal beams via the pallet sleds, so there is no possibility of collapse through the rack structure or pallet breakage.

2.5. Definition of the Material Handling Devices

Rack warehouses are most often served by forklifts and represent the most developed form of material handling equipment, which, in addition to the ability to actively pick up and place loads, also have the ability to stack loads vertically.

2.6. Racking Layout in the Storage Space

When designing the layout of selective pallet racks, it is necessary to fully define three conditions necessary for the implementation of the design: defining the storage unit, selecting the material handling equipment, and defining the storage space. In order to achieve maximum utilization of the storage space and work efficiency, in addition to the previously covered conditions, it is necessary to take into account certain design rules. These rules are reflected in the following [11]: A single-row rack should be placed against the wall of the storage space; Placing racks in the same space in two different directions should be avoided, as this wastes a lot of space for manipulation and intersects transport routes; Racks should always be placed along the longer side of the intended storage space, which results in fewer long corridors, reduces transport aisles, and increases the economic feasibility of the racks; Corridors should not be placed along the perimeter of the walls of the space.

2.7. Total Capacity of the Rack Warehouse

After calculating and adopting adequate racking structure components, it is possible to determine the achieved warehouse capacity for different concepts of storage unit placement concepts. To determine the total warehouse capacity, it is necessary to define: the number of storage units per bay, the number of the identical bays, the number of floors and the number of continuous identical racking rows.

The number of storage units per bay differs for cases of transverse and longitudinal placement, primarily due to the dimensions of the storage unit. Also, the load of the pair of horizontal beams in one bay and the load of the vertical frame have a significant influence.

A bay is defined as the space between adjacent vertical frames and adjacent horizontal beams, in which a certain number of storage units are stored, most commonly two or three.

The number of floors is the number of bays in the height of the rack, excluding the one located directly on the basement, while the section is a vertical module located between adjacent vertical frames. A section is defined by the sum of the heights of the individual bays, including the one located on the basement.

The number of continuous identical racking rows in the case of selective pallet racks can be configured as a single-row or a double-row. In this case, a single-row rack is defined as a row of bays accessible from one working aisle. Double-row racks are a parallel row of two single-row racks whose backs are connected into one whole by spacers. Compared to the single-sided racks, the double-sided racks are accessible from two parallel operating aisles. The total capacity of the rack storage is determined according to the expression:

$$Q_u = N_s \cdot N_p \cdot N_r \cdot N_k \tag{1}$$

where:
N_s – number of storage units per bay,
N_p – number of identical bays,
N_r – number of continuous identical racking rows, and
N_k – number of floors.

3. COMPARATIVE ANALYSIS

The planned storage space has the following dimensions: length 34500 mm, width 28500 mm, height 7500 mm. The load is stored on standard EURO pallets with basic dimensions of 800 mm x 1200 mm. The height of the storage unit is 1200 mm. The maximum weight of the storage unit is 1000 daN.

The orientation of the entry-exit zones is defined so that the entry and exit are on the different sides at the same level. The racking structure is of the assembly-disassembly type. The connection between upright frames and horizontal beams is achieved via upright columns and beam end connectors.

The working aisle dimensions between each rack and the rack's height depend on the characteristics of the forklift trucks, lifting devices, dimensions of the external structure, and specific needs [8-10]

The comparative analysis will be carried out depending on the method of stacking the storage unit into the selective racks. Therefore, the design of the warehouse with transverse and longitudinal pallet stacking was considered.

3.1. Horizontal Racking Beams

In the first case, the adopted storage unit stacking method is transverse handling. To achieve a lightweight structure, a smaller number of storage units per pair of beams is preferred. A higher number of units per bay requires longer horizontal beams. However, a higher number of storage units also demands a greater load capacity for the horizontal beams, thereby requiring a stronger cross-section and a higher structural mass. Therefore, two cases will be considered when determining the horizontal beam length, with clearances within the limits of A = 50 mm ÷ 100 mm, according to [7,11].

- When a single storage unit is stacked crosswise on a pair of horizontal supports, the following applies:÷

$$L_{G1} = 2 \cdot A + B = 2 \cdot [50 \div 100] + 1200 = [1300 \div 1400] \text{ mm}. \tag{2}$$

For the first case of stacking, and according to [12], the standard beam length of L_{G1} = 1300 mm is adopted. Dimension A refers to the aforementioned clearance, while dimension B was previously represented by x in Figure 1.

- When two storage units are stacked crosswise on a pair of horizontal beams, then:÷

$$L_{G2} = 3 \cdot A + 2 \cdot B = 3 \cdot [50 \div 100] + 2 \cdot 1200 = [2550 \div 2700] \text{ mm}. \tag{3}$$

For the second stacking scenario, with a load capacity of 2000 daN, according to [12], standard beam length L_{G2} = 2700 mm, a standard beam type with defined clearances between all elements of A = 100 mm was adopted.

In the case of transverse stacking, it was assumed that the optimal solution is the storage of two storage units in the form of two standard EURO pallets. The storage unit beam is marked as $R120L$ (120 indicates the profile height, while L defines the thickness of the wall of the carrier, which is 1 mm).

In the second case, the adopted method of stacking will be longitudinal. Longitudinal stacking of one storage unit is not considered. In practice, this type of stacking is practically not applied, because it leads to a significant reduction in storage capacity, an increase in the number of racking structure components, and thus unnecessary costs. Two potential solutions will be considered when determining the length of the storage unit supports with clearances according to [7].

When two storage units are stacked longitudinally on a pair of horizontal supports, the following is valid:

$$L_{G1} = 3 \cdot A + 2 \cdot B = 3 \cdot [50 \div 100] + 2 \cdot 800 = [1750 \div 1900] \text{ mm}. \tag{4}$$

For the first case of stacking, and according to [12], the standard beam length of L_{G1} = 1800 mm is adopted.

When three storage units are stacked longitudinally on a pair of horizontal beams, then the following is valid:

$$L_{G2} = 4 \cdot A + 3 \cdot B = 4 \cdot [50 \div 100] + 3 \cdot 800 = [2600 \div 2800] \text{ mm} \quad (5)$$

For the second stacking scenario, according to [12], for the required load capacity of 3000 daN and the standard length L_{G2} = 2700 mm, a beam type designated as $R140L$ is adopted. The designation 140 denotes the height of the profile, while L defines the thickness of the support wall, which is 1 mm. In this case, the three storage units will be placed on one support, with the clearance between the observed elements being slightly smaller than in the transverse stacking, and amounting to A = 75 mm. This stacking method will be further considered for comparative analysis, because it is assumed to be a more economical solution compared to the stacking of two storage units.

3.2. Upright Frame

It is necessary to determine the width of the upright frame, the height of the floor and the total height of the frame.

3.2.1. Width of the Upright Frame

The width of the frame is equal to the shortest dimension of the pallet at the base that enters the rack, i.e. 800 mm, in the case of transverse stacking. In the case of longitudinal stacking, the width of the frame is smaller than the dimension of the loaded pallet 1200 mm, i.e. 1100 mm.

3.2.2. Floor Height or Pitch

The floor height for both stacking scenarios is:

$$h_s = h + \Delta + h_n = 1200 + [75 \div 100] + 140 = [1415 \div 1440] \text{ mm}. \quad (6)$$

The resulting calculated value of the floor height must be divisible by the perforation pitch, so that a whole number is obtained as the quotient.

$$\frac{h_s}{p} = \frac{[1415 \div 1440]}{75} = [18.86 \div 19.02] \quad (7)$$

Since the integer value of the perforated holes was not obtained, two cases are considered:

- When the number of perforated holes n_{p1} = 18:

$$h_{s1} = n_{p1} \cdot p = 18 \cdot 75 = 1350 \text{ mm} \quad (8)$$

Based on the previously obtained value, the clearance Δ is:

$$\Delta = h_{s1} - h - h_n = 1350 - 1200 - 100 = 50 \text{ mm} \quad (9)$$

- When the number of perforated holes n_{p2} = 19:

$$h_{s2} = n_{p2} \cdot p = 19 \cdot 75 = 1425 \text{ mm} \quad (10)$$

In this case, the clearance Δ is:

$$\Delta = h_{s2} - h - h_n = 1425 - 1200 - 100 = 125 \text{ mm} \quad (11)$$

For the first case, with 18 perforated holes, the clearance between the beam and the storage unit is too small, and given the most commonly used transport and handling equipment in this storage system, such precision during stacking cannot be expected. Therefore, the second case is adopted, with 19 perforated holes and a calculated clearance of at least 125 mm, where the standard value of the floor pitch hs=1500 mm is adopted, according to [12].

3.2.3. Upright Frame Height

The total number of floors is determined as the quotient of the available height of the storage space and the adopted height of one floor:

$$n_s = \frac{H}{h_s} = \frac{7500}{1500} = 5 \quad (12)$$

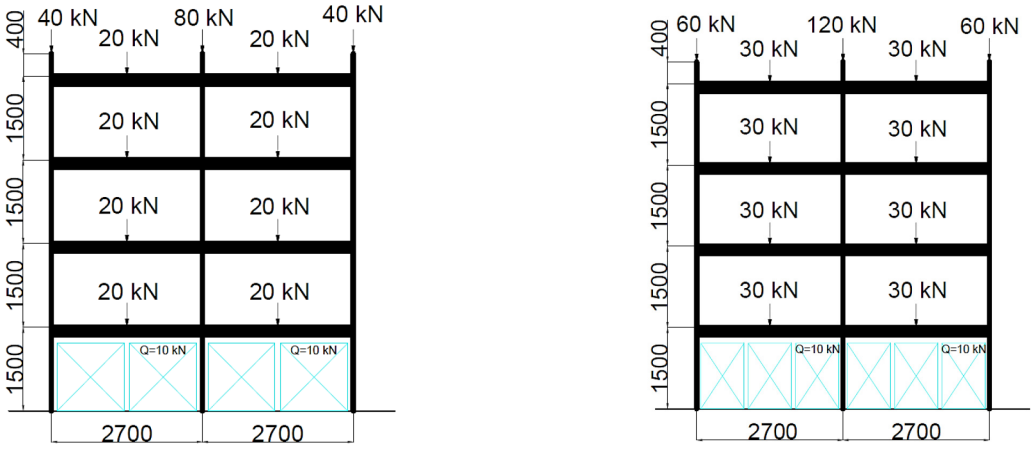
Given that the first floor is on the ground floor, and the total number of floors according to expression (12) is obtained as an integer, the final number of floors will be:

$$n_n = n_s - 1 = 5 - 1 = 4 \quad (13)$$

Taking into account the height of the storage space, the total number of floors for both placement cases will be 4+1. The last floor is located at a height of 6000 mm, which further affects the height of the load lifting when choosing the material handling equipment.

3.2.4. Dimensioning of the Upright Frame

For a defined number of floors, two storage units per bay with transverse stacking, for a floor pitch of 1500 mm, the required frame load capacity is 8000 daN. Therefore, the column profile *S*80*M* is adopted, with a total frame load capacity Q = 10300 daN, based on data from [12]. The mass of the adopted profile per unit length is 4.23 kg/m. In the designation of the adopted vertical column profile, 80 denotes the height of the profile, while M defines the column wall thickness, which is 2 mm. Variable loads, in addition to the constant load determined by the own weight of the racking structure, in the form of stacked storage units on the horizontal supports for the case of transverse stacking, are shown in Figure 2 a).



а) б)

Figure 2: Illustration of the loads on vertical columns for the case of: a) transverse stacking of the two storage units, b) longitudinal stacking of the three storage units

According to the adopted number of floors and three storage units per bay in the case of longitudinal stacking, floor pitch of 1500 mm, the minimum load capacity of the frame is 12000 daN. Therefore, the column profile *S*80*MH* is adopted, with a maximum load capacity Q=12800 daN, based on data according to [12]. The mass of the adopted profile per unit length is 5.23 kg/m. In the designation of the adopted vertical column, 80 indicates the height of the profile, while *MH* defines the column wall thickness, which is 2.5 mm. The variable loads on the rack structure in the longitudinal stacking in the longitudinal direction, caused by the weight of the stored load, are shown in Figure 2 b).

It can be seen that the central upright is loaded twice as much as the edge uprights, because it accepts the load from the bays on both the left and right sides, which is the basis for its dimensioning.

According to the defined loads acting on the rack structure, it can be concluded that the load on the rack structure is greater in the case of longitudinal stacking than in the case of transverse stacking, by the weight of one storage unit per bay. This statement is also confirmed based on the adopted profiles of upright columns and transverse beams, because in the case of longitudinal stacking, the profiles with larger dimensions and load capacity were adopted compared to transverse stacking.

3.3. Cross Beam

A cross beam with dimensions of 810 mm x 80 mm x 45 mm x 1.25 mm is adopted, with the load capacity of a pair of cross beams being 1020 kg. The mass of the beam is 1.66 kg/piece. In the case of longitudinal stacking, the presence of cross beams is not required.

3.4. Transport and Handling Device

A very narrow aisle truck STILL NXV 16 with swivel traverse fork is used as a transport and handling device, which has a flexible range of applications for different loads and warehouse conditions. Its load capacity is 1500 kg, the maximum lift height is 6350 mm, and the working aisle width is 1662 mm.

3.5. Selecting the Optimal Solution

Taking into account the included design conditions, as well as the selection of racking construction elements, the racks are arranged in the transverse direction, i.e. along the shorter side of the storage space due to the orientation of the entry-exit zones. The operational passages between the racks are defined by the characteristics of the transport and handling equipment.

The width of the space intended for movement or transport, which does not allow direct access to the unit load, is 2200 mm for both racking configurations. In the case of transverse stacking, the racking arrangement is designed as a total of 20 double-row and 4 single-row racks, according to Table 1. Two storage units are stacked in one bay, with a total of 5 floors including the "floor " located on the ground floor. Figure 3 illustrates the optimal rack layout in the designated warehouse space, which is served by a racking forklift, with the stacking being transverse.

Table 1: Warehouse capacity with transverse pallet stacking

Rack designation	Rack type	Length, mm	Width, mm	Number of identical rows, pcs	Row capacity, pcs	Total rack capacity, pcs
P1	single-deep	16790	800	2	60	120
P2	double-deep	16790	1685	8	120	960
P1	single-deep	8430	800	1	30	30
P2	double-deep	8430	1685	8	60	480
P1	single-deep	5650	800	1	20	20
P2	double-deep	5650	1685	3	40	120
P2	double-deep	2865	1685	1	20	20
Total warehouse capacity, pcs						1750

For the case of longitudinal stacking, the rack layout is designed as a total of 16 double-row and 4 single-row racks, more precisely defined in Table 2. In relation to transverse stacking, 3 storage units are stored in each rack bay in one section, with a total of 5 floors including the 'floor located on the ground floor. Figure 4 illustrates the optimal rack layout in the intended storage space, which is served by a rack forklift, where the storage unit is stacked longitudinally.

Table 2: Warehouse capacity with longitudinal pallet stacking

Rack designation	Rack type	Length, mm	Width, mm	Number of identical rows, pcs	Row capacity, pcs	Total rack capacity, pcs
P1	single-deep	16790	1200	2	90	180
P2	double-deep	16790	2500	6	180	1080
P1	single-deep	8430	1200	2	45	90
P2	double-deep	8430	2500	6	90	540
P2	double-deep	5650	2500	3	60	180
P2	double-deep	2865	2500	1	30	30
Total warehouse capacity, pcs						2100

The comparative analysis of the rack layout refers to both the achieved capacity and the weight of the rack structure. The total weight of all elements of the rack support structure with transverse stacking is 37088 kg, while with longitudinal stacking it is 28683 kg. Based on the given data, it can be concluded that, the optimal solution in terms of the weight of the observed rack settings is with longitudinal stacking.

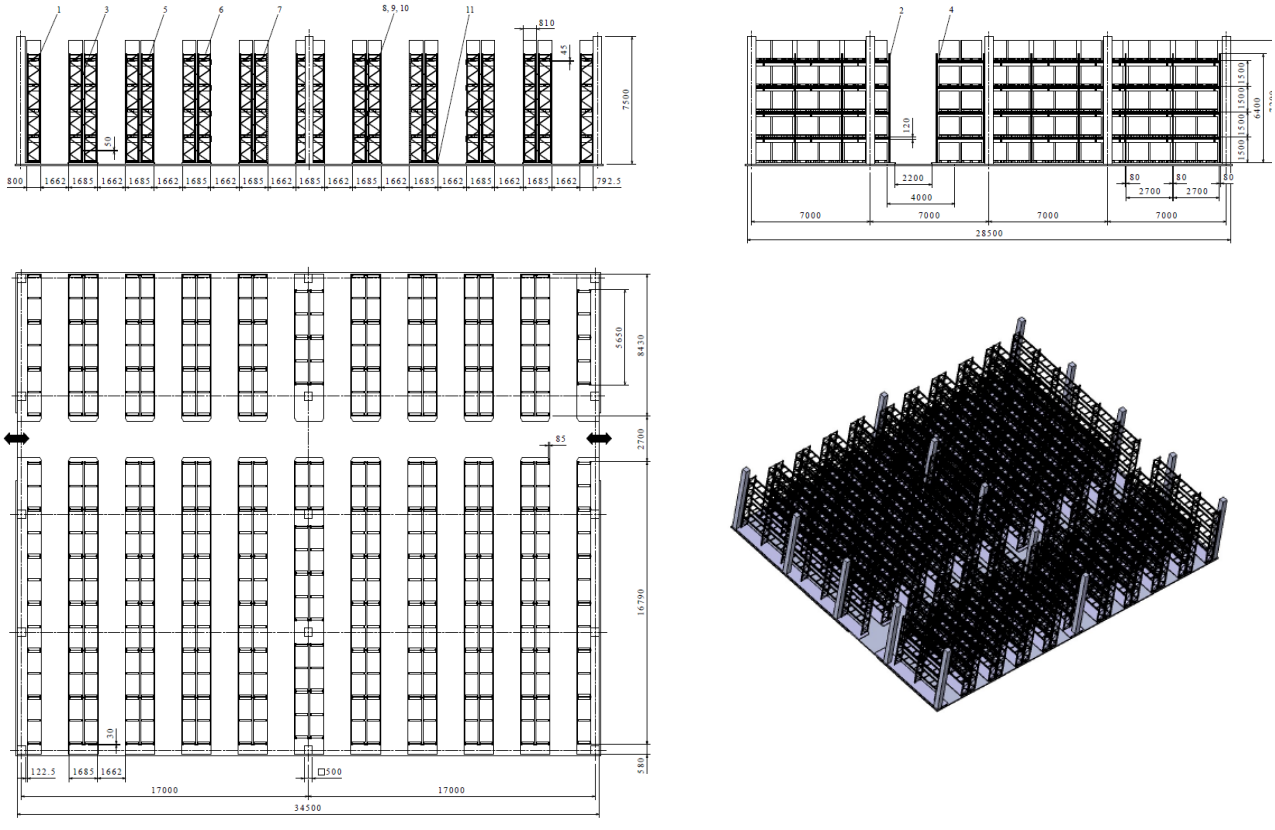


Figure 3: Warehouse layout with transverse pallet stacking

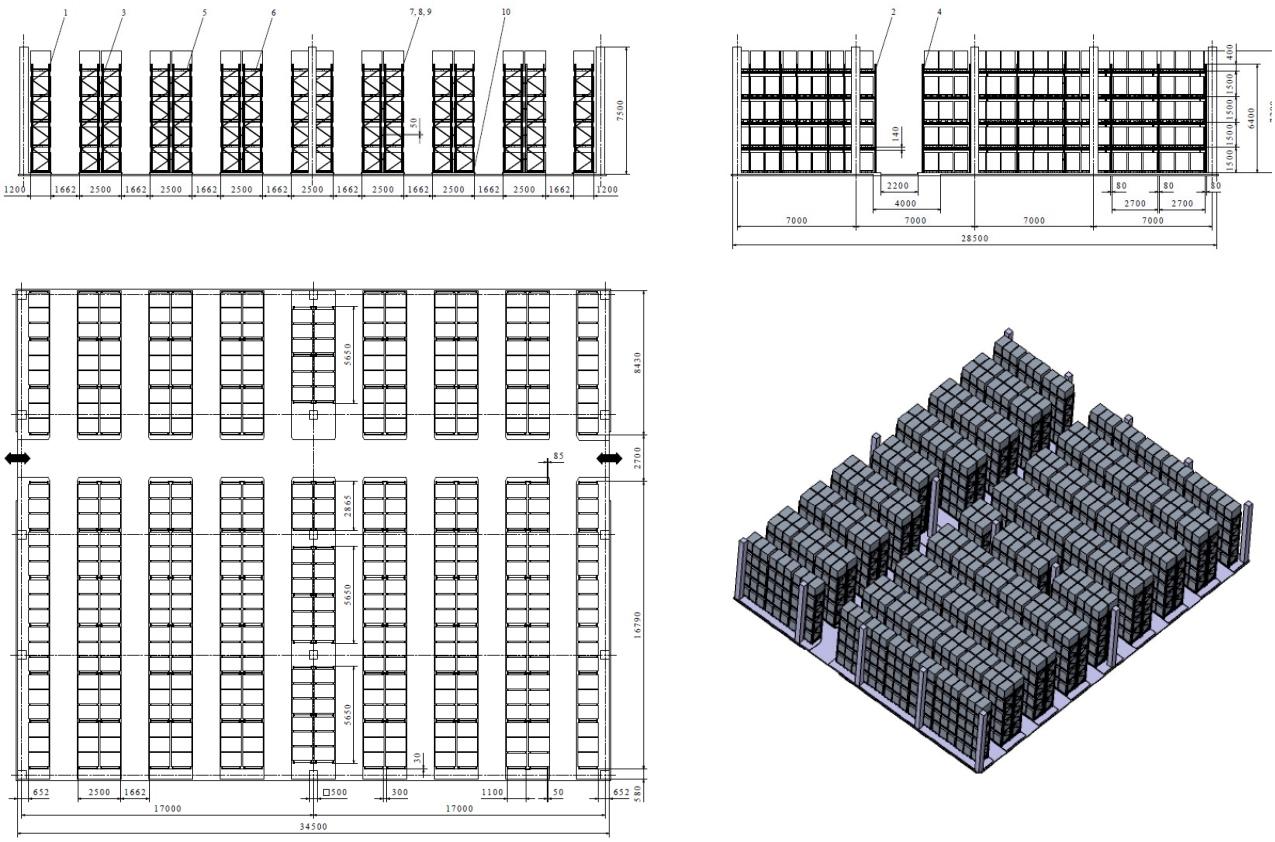


Figure 4: Warehouse layout with longitudinal pallet stacking

The key difference between longitudinal and transverse stacking is reflected in the position of the stacked storage unit, as well as in the number of units stored within a single bay. Also, compared to longitudinal stacking, transverse stacking requires the presence of cross beams supports for the storage units, which complicates and increases the cost of the rack structure to a certain extent.

The aisle width is the same for both stacking options, because the same material handling equipment is used to serve the storage space.

Considering the dimensions of the racks, the frame width is greater for longitudinal stacking. Despite the fact that the racks for transverse stacking are more compact in terms of construction, the greater storage space capacity, with the smaller number of racks, is achieved for longitudinal stacking by 20%.

Considering the number of rack elements, a smaller number of elements is certainly used in the case of longitudinal stacking, given that there are fewer rows of racks.

Given the achieved higher storage capacity and lower own weight of the rack structure elements, the rack layout with longitudinal stacking of storage units is adopted as an optimal, more economical solution.

4. CONCLUSION

Selective pallets racks, also known as adjustable or conventional racks, are the most widespread due to their economy, simplicity and efficiency. This type of rack is designed as a skeletal structure with a steel frame intended for the storage of usually palletized goods in a conventional manner.

The primary task of any rack warehouse is to maximize the available space and enable safe storage of goods. Pallet racks should be designed in such a way that they adapt to the type and dimensions of pallets with a unit load and can carry a certain weight, taking into account storage at multiple levels. Great attention should be paid to the selection of racks that best meet the requirements in terms of easy access to goods and efficient use of warehouse space. The rack, as the foundation of a warehouse, is crucial to enabling optimal and efficient flow of goods, because it is precisely this efficiency that depends on the use of an adequate racking system.

Pallet rack systems allow material handling equipment to smoothly access the storage units. Today, racks are most often designed as modular, which means they can be adapted and rearranged, which is primarily reflected in the advantage of their application. Selective pallet racks are an ideal choice for warehouses that prioritize the storage unit that is first placed in the rack, first taken out, namely FIFO. This system ensures that the earliest received storage units are retrieved first, promoting efficient stock rotation and minimizing the risk of product expiration.

Also, the correct selection of forklift truck characteristics is important for the efficient use of space within the layout of selective pallet racks.

According to the achieved capacity, the weight of the rack structure and the number of rack elements, the installation of racks with longitudinal stacking is more economical. The disadvantage of transverse stacking compared to longitudinal stacking is also reflected in the use of cross beams, which further complicate and increase the cost of the structure.

ACKNOWLEDGEMENTS

This research was supported by the Ministry of science, technological development and innovation of the Republic of Serbia, contract number 451-03-34/2026-03/200107.

REFERENCES

[1] R. Vujanac, N. Miloradović, S. Vulović, and A. Pavlović, "A Comprehensive Study into the Boltless Connections of Racking Systems", Metals, Vol. 10(2), https://doi.org/10.3390/met10020276, (2020)

[2] N. Miloradović and R. Vujanac, "Comparative analysis of applications of different storage systems", Proceedings of the 17th International Conference on Accomplishments in Mechanical and Industrial Engineering 'DEMI 2025', Banja Luka (Bosnia and Herzegovina), 29 May – 30 May 2025, pp. 389–397, (2025)

[3] M. Piskulic, R. Vujanac, S. Vulovic, N. Miloradovic, M. Blagojevic, and Z. Djordjevic, "Structural Performance of Semi-Rigid Beam-to-Column Connections in Steel Storage Racks: A Review Focused on FEM Analysis", Metals, Vol. 16(1), pp. 1–25, https://doi.org/10.3390/met16010002, (2026)

[4] "EN 1993 Eurocode 3 - Design of Steel Structures", Brussels, Belgium, (2005)

[5] "FEM 10.2.02 - Design of Static Steel Pallet Racking", Brussels, Belgium, (2000)

[6] "EN 15512:2020+A1:2022 - Steel static storage systems - Adjustable pallet racking systems - Principles for structural design", Brussels, Belgium, (2022)

[7] R. Vujanac and N. Miloradović, "Basics of storage and transport systems", The Faculty of engineering of the University of Kragujevac, Kragujevac, (2023)

[8] R. Vujanac, N. Miloradovic, P. Zivkovic, and L. Petrovic, "Basis for the design of drive-in and drive-through racking", Proceedings of X Triennial International Conference Heavy Machinery - HM 2021, Vrnjacka Banja (Serbia), pp. 57–63, (2021)

[9] "FEM 9.831-1 - Basis of calculations for storage and retrieval machines - Tolerances, deformations and clearances in the storage system - Part 1: General, single deep and double deep beam pallet racking", Brussels, Belgium, (1995)

[10] "FEM 10.2.07. - Drive-in and Drive-Through Racking Drive-in Design Code", Brussels, Belgium, (2012)

[11] M. Piskulić, R. Vujanac, N. Miloradović, and M. Blagojević, "Influence of pallet placement method on beam behavior in rack structure", Proceedings of the 41st Danubia-Adria Symposium Advances in Experimental Mechanics - DAS 2025, Kragujevac (Serbia), pp. 99–102, (2025)

[12] "Catalogue of Pallet Racking System", Dalmine Logistic Solutions, Piacenza, Italy, (2022)

SESSION A: EARTH MOVING AND TRANSPORTATION MACHINERY

The determination of the deflection of the stepped column by the finite difference method

Nebojša B. Zdravković[1*], Mile Savković[1], Goran Marković[1], Marko Todorović[1], Predrag Mladenović[1], Goran V. Pavlović[1]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Kraljevo, Serbia

ARTICLE INFO

[*] **Correspondence:** zdravkovic.n@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.A07Z **PAGES:** A51 – A57

ABSTRACT

Besides the strength criterion, steel structures must also meet the deformation criterion, which in some cases may be critical. Excessive deformation of the structure may lead to additional loads and improper operation of the installed equipment. The most common practice in steel structure design involves using the finite element method (FEM) to calculate the stress and deformation of the current model. However, this is usually limited to the case study for the particular structural model. The paper demonstrates the application of the finite difference method (FDM) to determine the deflection of a stepped column, enabling a broader parametric study and analysis of a variety of solutions that satisfy the deformation criterion. The application of FDM yields a scheme convenient for coding in MATLAB while showing the dependence between the column deflection and the design parameters. The results from the finite element analysis (FEA) are in excellent agreement with those from the FDM.

KEYWORDS

Stepped Column, Deflection, Finite Difference Method, Finite Element Method.

1. INTRODUCTION

The steel structures must meet many requirements, such as strength, deformation, elastic and dynamic stability, and connection safety. It is often the case that the whole structure and/or its parts have long lengths, so the maximum deformation criterion may be the most restrictive for its design. This criterion is defined by permissible deformation at characteristic points of the structure under maximum operational load. Within exploitation, the steel structure's lack of stiffness can lead to excessive deformation and the development of additional static and dynamic loads. Further on, this may eventually jeopardize operational safety and shorten the operational life of the subassemblies or the entire facility. So, the limitation of deformation provides good exploitation features and a long equipment lifespan.

For instance, for each crane type, there are prescribed maximum deflection values for characteristic points on its carrying structure. Permissible static deflection at the mid-span point of the bridge crane is L/600 [1], where L denotes the span. For the case of the gantry cranes, there is a prescription for the static deflection values for the mid-span point and the endpoint of the overhang: it is (1/1000÷1/600)L for the mid-span, while it is (1/400÷1/200)L1 for the overhang endpoint, where L is the distance between the gantry crane legs, and L1 denotes the overhang length [2].

If the steel supporting structure has an unchangeable geometry and consists of beams with uniform cross-sections, it is relatively easy to calculate its deflection. Nevertheless, it is a rather rare situation. The typical example of a structure with a variable-geometry configuration that changes with the working position is an articulated boom in a truck crane [3]. However, if the cross-sections of the segments are uniform, it is possible, with some simplifications, to develop an analytical model for calculating the boom tip displacements [4]. Building an analytical model to calculate displacements of characteristic points enables the use of deformation conditions to optimize the supporting structure and reduce its mass [5-9].

In more complex structures, deflection checks are typically conducted using FEM models [10-12]. It is usually done for the final design version. This occurs in the late design phase, when it is far more difficult, and sometimes even impossible,

to implement design changes to the structural model if it does not meet the displacement criteria for characteristic points. Hence, there is a need for an approach applicable in the early design phase to provide a broader perspective on possible design solutions that meet the permissible deformation criteria.

The predominant method for the structural analysis of steel structures is FEM. Nevertheless, the authors employ additional numerical methods to compute deflection. For example, the angle of rotation is represented by a power series and substituted into the derived governing nonlinear differential equation for bending of prismatic and non-prismatic inextensible beams [13]. Others used the FDM to develop the numerical scheme for solving the deformation problem of beams in bending [14-16].

The paper presents the FDM application procedure for determining the deflection of the supporting structure in a simple example of a multi-segment column subjected to a force at the top. This approach is also applicable to any load combination and number of segments.

2. THE MATHEMATICAL MODEL

The model used to present the proposed FDM approach for deflection determination is shown in Fig. 1. To make things as simple as possible, it is a three-segment column made of steel pipes with different diameters *D1, D2,* and *D3*. Nevertheless, this approach is applicable to any type of cross-section. To make the expressions simpler, but with no derogation of the generality and applicability of the method, the wall thicknesses of all pipes are taken to be equal (*t*). The column is fixed at point *A*, while points *B* and *C* are the joints between the segments.

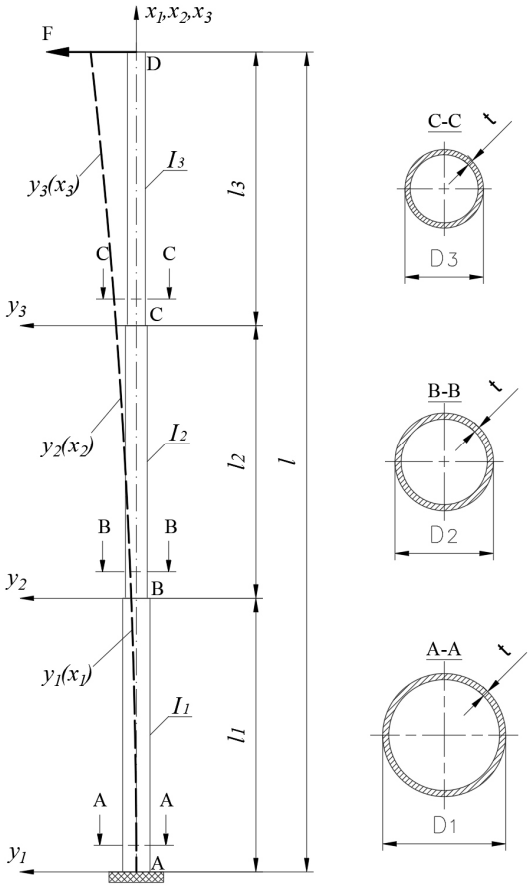


Figure 1: A column with stepped changes of cross-section

The lengths of the segments are l_1, l_2 and l_3 and the column total length is $l=l_1+l_2+l_3$. The point force F, acting transversely at the top, is bending the column and the deformed shape is shown as dashed elastic line. Each segment has its own coordinate system x_jy_j and its own deflection function $y_j=y_j(x_j)$, $j=1,2,3$. If the following ratios of diameters are introduced

$$p=\frac{D_2}{D_1},\ q=\frac{D_3}{D_2} \tag{1}$$

the moments of inertia for the z_i axes of the segments are respectively as follows:

$$I_1=\frac{\pi\left[D_1^4-\left(D_1-2t\right)^4\right]}{64} \tag{2}$$

$$I_2 = \frac{\pi\left[D_2^4 - \left(D_2 - 2t\right)^4\right]}{64} = \frac{\pi D_1^4}{64}\left[p^4 - \left(p - 2\frac{t}{D_1}\right)^4\right] \tag{3}$$

$$I_3 = \frac{\pi\left[D_3^4 - \left(D_3 - 2t\right)^4\right]}{64} = \frac{\pi D_2^4}{64}\left[q^4 - \left(q - 2\frac{t}{D_2}\right)^4\right] \tag{4}$$

The bending moment at the distance x from the column support is

$$M = F\left(l - x\right) \tag{5}$$

The governing bending equations for each segment are as follows:

$$\frac{d^2 y_j\left(x_j\right)}{dx_j^2} = \frac{M}{B_j},\ j = 1,2,3 \tag{6}$$

The boundary conditions define the values of displacements and the slopes at points A, B and C. For point A they are:

$$y_1\left(x_1 = 0\right) = 0 \tag{7}$$

$$\frac{dy_1}{dx_1}\left(x_1 = 0\right) = 0 \tag{8}$$

The boundary conditions for point B are:

$$y_1\left(x_1 = l_1\right) = y_2\left(x_2 = 0\right) \tag{9}$$

$$\frac{dy_1}{dx_1}\left(x_1 = l_1\right) = \frac{dy_2}{dx_2}\left(x_2 = 0\right) \tag{10}$$

The boundary conditions for point C are:

$$y_2\left(x_2 = l_1 + l_2\right) = y_3\left(x_3 = 0\right) \tag{11}$$

$$\frac{dy_2}{dx_2}\left(x_2 = l_1 + l_2\right) = \frac{dy_3}{dx_3}\left(x_3 = 0\right) \tag{12}$$

3. THE FDM DISCRETIZATION

Fig. 2 shows FDM grid scheme for the numerical solution of the deflection line. The length of the beam l is equally divided by $N=l/s$ grid points (nodes) into N segments with length $s = L / N$. Later transformation of the boundary conditions includes the support point A denoted as node *0* and one fictitious node on the left side of the support, denoted as node *-1*. As the displacements of fictitious nodes are designated by superscripted *f*, the displacement of this fictitious node is denoted as Y^f_{-1}.

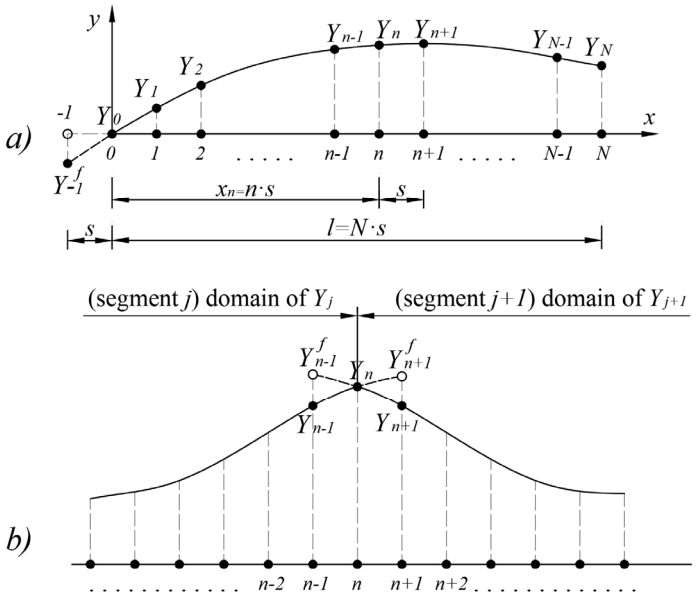


Figure 2: a) FDM grid scheme; b) fictitious nodes at adjacent segments' junction

According to the central FDM, approximations for derivatives of the deflection function *y(x)* in node *n* are as follows:

$$\left(\frac{dy}{dx}\right)_n \approx \frac{-Y_{n-1}+Y_{n+1}}{2s} \tag{13}$$

$$\left(\frac{d^2y}{dx^2}\right)_n \approx \frac{Y_{n-1}-2Y_n+Y_{n+1}}{s^2} \tag{14}$$

After the introduction of bending stiffnesses for the segments as $B_j=EI_j$, $j=1,2,3$ and discretization of the global longitudinal coordinate to $x=n\cdot s$, governing bending equations (6) can be transformed to the following algebraic formulation:

$$Y_{n-1}-2Y_n+Y_{n+1}=\frac{s^2F}{B_j}(l-ns),\ \text{j=1,2,3} \tag{15}$$

An FDM transformation of the boundary conditions in support A yields:

$$Y_0=0 \tag{16}$$

$$\frac{-Y_{-1}^f+Y_1}{2s}=0 \Rightarrow Y_{-1}^f=Y_1 \tag{17}$$

Geometrical interpretation of the discretized boundary conditions (16) and (17) is shown in Fig. 3.

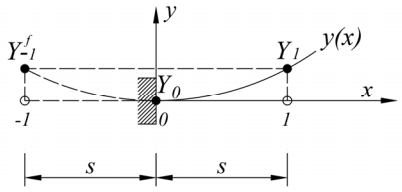


Figure 3: Deflection function around the fixed support

If the node in point B is designated as $b=l_1/s$, the algebraic boundary conditions are as follows:

$$Y_b=Y_b^f \tag{18}$$

$$-Y_{b-1}+Y_{b+1}^f=-Y_{b-1}^f+Y_{b+1} \tag{19}$$

Since displacements of all fictitious nodes must be eliminated from the discretized governing equations (15), an additional condition related to equality of bending moments is introduced:

$$B_1\frac{d^2y_1}{dx_1^2}(x_1=l_1)=B_2\frac{d^2y_2}{dx_2^2}(x_2=0)\Rightarrow B_1\frac{Y_{b-1}-2Y_b+Y_{b+1}^f}{s^2}=B_2\frac{Y_{b-1}^f-2Y_b^f+Y_{b+1}}{s^2} \tag{20}$$

Combining (18), (19) and (20) yields:

$$Y_{b+1}^f=(1-2P_1)Y_{b-1}+(2P_1-2P_2)Y_b+(1-P_1+P_2)Y_{b+1} \tag{21}$$

where $P_1=\frac{B_1}{B_1+B_2}, P_2=\frac{B_2}{B_1+B_2}$.

Similarly to the previous, the algebraic boundary conditions for the node at point C are derived. If the node in point C is designated as $c=l_1+l_2/s$, the algebraic boundary conditions are:

$$Y_c=Y_c^f \tag{22}$$

$$-Y_{c-1}+Y_{c+1}^f=-Y_{c-1}^f+Y_{c+1} \tag{23}$$

Since displacements of all fictitious nodes must be eliminated from the discretized governing equations (15), an additional condition related to equality of bending moments is introduced:

$$B_2\frac{d^2y_2}{dx_2^2}(x_2=l_2)=B_3\frac{d^2y_3}{dx_3^2}(x_3=0)\Rightarrow B_2\frac{Y_{c-1}-2Y_c+Y_{c+1}^f}{s^2}=B_3\frac{Y_{c-1}^f-2Y_c^f+Y_{c+1}}{s^2} \tag{24}$$

Combining (22), (23) and (24) yields:

$$Y_{c+1}^f=(1-2P_3)Y_{c-1}+(2P_3-2P_4)Y_c+(1-P_3+P_4)Y_{c+1} \tag{25}$$

where $P_3 = \frac{B_2}{B_2 + B_3}, P_4 = \frac{B_3}{B_2 + B_3}$.

Now, a set of N algebraic equations can be written for the grid nodes $n=0,1,\ldots, N-1$, with the node displacements Y_i, $i=1,\ldots,N$ as unknowns. Thereby, the equations of the deflection function written for the nodes *0* and *1* include displacements Y_0 and Y_{-1}, which are eliminated by using transformed boundary conditions (16) and (17). Also, the fictitious displacements in algebraic bending equations written for nodes b and c are eliminated using (18), (21), (22) and (25). Finally, a system of linear algebraic equations, very convenient for programming (e.g. in MATLAB), has the form:

$$
\begin{aligned}
&n = 0 \rightarrow 2Y_1 = \frac{s^2 F l}{B_1} \\
&n = 1 \rightarrow -2Y_1 + Y_2 = \frac{s^2 F}{B_1}(l - s) \\
&n = 2 \div (b-1) \rightarrow Y_{n-1} - 2Y_n + Y_{n+1} = \frac{s^2 F}{B_1}(l - ns) \\
&n = b \rightarrow (2 - 2P_1) Y_{b-1} + (2P_1 - 2P_2 - 2) Y_b + (1 - P_1 + P_2) Y_{b+1} = \frac{s^2 F}{B_1}(l - bs) \\
&n = b+1 \rightarrow Y_b - 2Y_{b+1} + Y_{b+2} = \frac{s^2 F}{B_2}\left[l - (b+1)s\right] \\
&n = (b+2) \div (c-1) \rightarrow Y_{n-1} - 2Y_n + Y_{n+1} = \frac{s^2 F}{B_2}(l - ns) \\
&n = c \rightarrow (2 - 2P_3) Y_{c-1} + (2P_3 - 2P_4 - 2) Y_c + (1 - P_3 + P_4) Y_{c+1} = \frac{s^2 F}{B_2}(l - cs) \\
&n = c+1 \rightarrow Y_c - 2Y_{c+1} + Y_{c+2} = \frac{s^2 F}{B_3}\left[l - (c+1)s\right] \\
&n = (c+2) \div (N-1) \rightarrow Y_{n-1} - 2Y_n + Y_{n+1} = \frac{s^2 F}{B_3}(l - ns)
\end{aligned}
\tag{26}
$$

4. NUMERICAL EXAMPLE AND COMPARISON WITH FEM MODEL RESULTS

The derived system of N algebraic equations yields the column node displacements Y_i, $i=1,\ldots,N$. The numerical example has the following set of input parameters: $t=3$ mm, $l_1=l_2=l_3=1000$ mm, $D_1=100$ mm, $F=100$ N. The spacing for creating the node grid is set to $s=10$ mm. The values of the non-dimensional ratios between the pipes' diameters p and q are varied between 0.6 and 0.9. The modulus of elasticity is taken as $E=2.1\cdot10^5$ N/mm^2. The solution of the system (26) is obtained using MATLAB, and the FEM analyses are performed in SAP. Table 1 presents the comparison of the results obtained from both (FDM and FEM) numerical approaches for the deflection at the top of the column Y_N. The relative deviations between the results (δ) are up to 0.2%.

Table 1: Comparison of the results obtained in FDM and FEM approach

p	q	Y_N [mm]		$\lvert\delta\rvert$ [%]
		FDM	FEM	
0.6	0.6	11.5996	11.61457	0.13%
	0.7	10.1390	10.15290	0.14%
	0.8	9.3530	9.36604	0.14%
	0.9	8.8932	8.90554	0.14%
0.7	0.6	8.1899	8.20309	0.16%
	0.7	7.3102	7.32244	0.17%
	0.8	6.8340	6.84545	0.17%
	0.9	6.5540	6.56498	0.17%
0.8	0.6	6.3371	6.34892	0.19%
	0.7	5.7671	5.77814	0.19%
	0.8	5.4572	5.46754	0.19%
	0.9	5.2744	5.28424	0.19%

p	q	Y_N [mm]		$\lvert\delta\rvert$ [%]
		FDM	FEM	
0.9	0.6	5.2450	5.25583	0.21%
	0.7	4.8550	4.86507	0.21%
	0.8	4.6422	4.65166	0.20%
	0.9	4.5163	4.52535	0.20%

Fig. 4 shows the 3D diagram of the FDM results, while Fig. 5 graphically shows the comparison of FDM and FEM results.

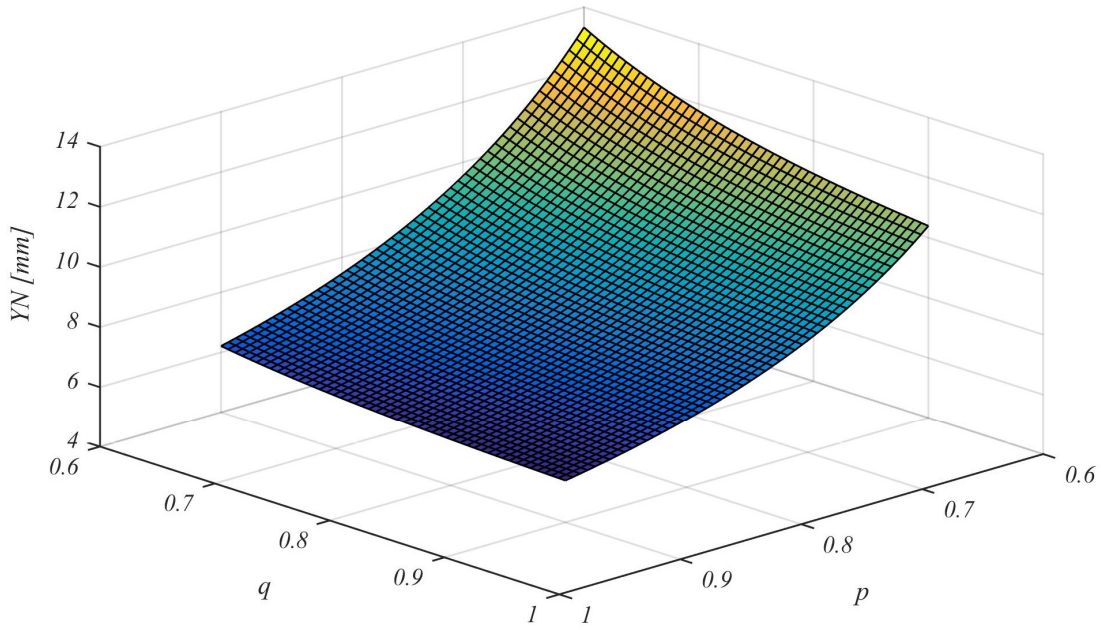


Figure 4: 3D diagram of the results

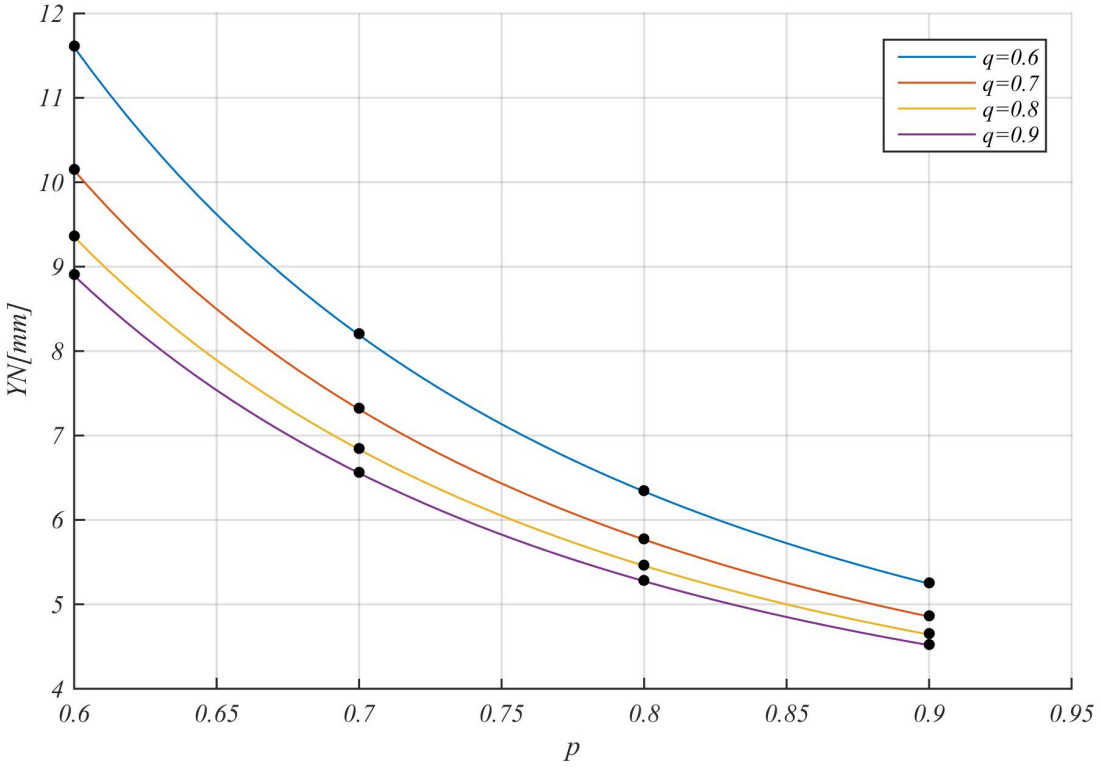


Figure 5: The comparison of the FDM (lines) and FEM (dots) results

5. CONCLUSION

As expected, the higher relative coefficients p and q the stiffer the column, and the smaller the deflection. The FDM approach for calculating the column deflection yielded results practically identical to those of the FEM approach, with relative deviations up to 0.20%. The FDM model is very convenient for coding mathematical calculations in software, e.g., MATLAB. Once established, the FDM mathematical model enables broader analysis of the structure by varying the geometrical parameters or their ratios. It is possible to track the displacement of any node in response to changes in selected design parameters, e.g., the lengths of the segments.

ACKNOWLEDGEMENTS

The paper was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia through the Contract for the scientific research activity realization and financing in 2026, ID 451-03-34/2026-03/200108. This paper refers to the UN 2030 Agenda for Sustainable Development and its Goal 9: Build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation.

REFERENCES

[1] Eurocode 3: Design of steel structures - Part 6: Crane supporting structures, EN 1993-6.

[2] Z. Petković and D. Ostrić, “Metal structures in mechanical engineering I (in Serbian)”, ISBN 86-70803-274-7, Faculty of Mechanical Engineering Belgrade, (1995)

[3] S. Bošnjak, N. Zrnić, and B. Dragović, “Dynamic Response of Mobile Elevating Work Platform under Wind Excitation”, Strojniški vestnik - Journal of Mechanical Engineering, ISSN 0039-2480, Vol. 55(2), pp. 104-113, (2009)

[4] N. Zdravković, Z. Petković, M. Gašić, and M. Savković, “Research of deflection-payload dependence of the auto crane articulated boom”, Proceedings of International Conference on material Handling, Construction and Logistics “Mhcl’12”, Belgrade (Serbia), isbn 978-86-7083-763-8, pp. 101-106, (2012)

[5] J. Farkas, “Optimum Design of Metal Structures”, Akademiai Kiado, Budapest, (1984)

[6] M. Gasic, M. Savkovic, R. Bulatovic, and R. Petrovic, “Optimization of a pentagonal cross section of the truck crane boom using Lagrange’s multipliers and differential evolution algorithm”, Meccanica, Vol. 46, pp. 845–853. doi: https://doi.org/10.1007/s11012-010-9343-7, (2011)

[7] M. Gasic, M. Savkovic, R. Bulatovic, and R. Petrovic, “Optimization of trapezoidal cross section of the truck crane boom by Lagrange’s multipliers and by differential evolution algorithm (de)”, Strojniški vestnik – Journal of Mechanical Engineering, Vol. 57(4), pp. 304-312, doi: https://doi.org/10.5545/sv-jme.2008.029, (2011)

[8] B.C. Pinca, O.G. Tirian, V.A. Socalici, and D.E. Ardeleadn, “Dimensional optimization for the strength structure of a traveling crane”, WSEAS Transactions on Applied and Theoretical Mechanics, Vol. 4(4), ISSN:1991-8747, pp 147-156, (2009)

[9] R. Selmic, R. Cvetkovic, and R. Mijailovic, “Optimization of cross-section in structures”, Monograph, ISBN 86-7395-213-1, Faculty of Transport and Traffic Engineering, Belgrade, (2006)

[10] R.G. Sreedhara, D.M.G. Comissiong, K. Jordan, and A. Sankar, “A Finite Element Method for the Beam equation Using MATLAB”, International Journal of Applied Science and Technology, ISSN 2221-0997, Vol. 2(8), pp. 80-88, (2012)

[11] N. Omar and F. De’nan, “Finite element analysis of deflection and stress in triangular web profiled steel section (triwp) cantilever beam”, DIGES PMU, ISSN 2289-6376, Vol. 3, pp. 1-6, (2016)

[12] S. Kumar, “Comparison of Deflection and slope of Cantilever Beam with Analytical and Finite Element Method for Different Loading Conditions”, International Journal of Engineering Science and Innovative Technology (IJESIT), Vol. 5(6), pp. 45-51, (2016)

[13] S.Z. Al-Sadder and M.H. Dado, “An Accurate Numerical Scheme for Large Deflection Analysis of Non-Prismatic Inextensible Slender Beams Subjected to General Loading and Boundary Conditions”, Advances in Structural Engineering, Vol. 8(6), pp. 585–594, https://doi.org/10.1260/136943305776318392, (2005)

[14] A.A. Al-Azzawi and D.M. Theeban, “Large Deflection of Deep Beams on Elastic Foundations”, Journal of the Serbian Society for Computational Mechanics, ISSN 1820-6530, Vol. 4(1), pp. 88-101, (2010)

[15] K.S. Thankane and T. Stys, “Finite Difference Method for Beam Equation with Free Ends Using Mathematica”, Southern African Journal of Pure and Applied Mathematics, ISSN 1993-8594, Vol. 4, pp. 61-78, (2009)

[16] S. AL-Sadder, R. AL-Rawi, “Finite difference scheme for large-deflection analysis of non-prismatic cantilever beams subjected to different types of continuous and discontinuous loadings”, Archive of Applied Mechanics, Vol. 75, pp. 459–473, https://doi.org/10.1007/s00419-005-0422-5, (2006)

SESSION A: EARTH MOVING AND TRANSPORTATION MACHINERY

Physical analysis of forces, moments, and energy efficiency of multi-rotor unmanned aerial vehicles from the perspective of flight control

Dragan Živanić[1*], Vladimir Žerajić[2], Nikola Ilanković[1]

[1] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia

[2] GASTRANS, Novi Sad, Serbia

ARTICLE INFO

[*] **Correspondence:** zivanic@uns.ac.rs

DOI: https://doi.org/10.46793/ET26.A08Z **PAGES:** A59 – A67

ABSTRACT

Multirotor unmanned aerial vehicles (UAVs) form a specific class of aircraft in which lift is generated directly by rotor thrust rather than by fixed wings. This principle enables vertical take-off and landing, hovering and precise low-speed positioning, but it also increases energy demand and makes the flight dynamics strongly coupled. This paper analyses the physical principles that govern multirotor UAV behaviour, with emphasis on the distribution of forces, moments and energy consumption. Hovering, vertical motion, translational motion through aircraft tilt and yaw control by counter-rotating rotors are considered from the perspective of flight control. The energy model is interpreted through disk loading, induced velocity and ideal hover power. A reference numerical example for a 4 kg UAV is used to compare propeller diameters and quadrotor, hexarotor and octorotor configurations. The results show that larger total rotor-disk area reduces induced power and can extend hover endurance, while additional rotors improve redundancy at the cost of mass, structural complexity and control allocation requirements. The analysis provides a compact basis for engineering optimisation of multirotor UAV systems.

KEYWORDS

UAV, Multirotor, Flight physics, Thrust, Moments, Disk loading, Energy efficiency, Flight control.

1. INTRODUCTION

Unmanned aerial vehicles (UAVs) are one of the most dynamic fields of contemporary aeronautics, automation and robotics. Their development has enabled aircraft of different size, mass, purpose and autonomy level, ranging from small experimental platforms to complex civil, industrial, scientific and security systems [1]. Modern UAV platforms combine aerodynamics, mechatronics, sensor systems, control algorithms, navigation and energy storage [2].

The range of UAV applications has expanded considerably in recent decades. UAVs are now used in geodesy, agriculture, infrastructure inspection, logistics, energy systems, environmental monitoring, terrain mapping, scientific research and security-related tasks [3]. This expansion has been supported by miniaturised electronics, more reliable sensors, autonomous control, improved communication systems and increasingly capable onboard computing.

The modern development of UAV systems has been further accelerated by the integration of autonomous navigation algorithms, sensor fusion techniques, and adaptive control systems. Particularly in multirotor platforms, the application of advanced state-estimation methods enables stable operation even under conditions of partial GNSS signal loss or degradation of sensor data, which is of great importance for industrial and safety-related applications in urban environments. [4].

Multirotor UAVs occupy a particularly important position within this field. Unlike fixed-wing aircraft, which generate lift primarily through aerodynamic action of the wings during forward motion, multirotor platforms generate lift directly through rotor thrust [5]. This makes vertical take-off and landing, hovering, precise positioning and manoeuvring in confined spaces possible [6].

These advantages also explain why multirotor systems are preferred when the mission requires operation at low altitude or sustained positioning above a target point.

The physical behaviour of a multirotor system is determined by the interaction of forces and moments. During flight, the aircraft is subjected to weight, rotor thrust and aerodynamic drag, while changes in the rotational speed of individual rotors modify thrust distribution and create moments about the main body axes. The basic motion modes are vertical climb and descent, roll and pitch tilting, and yaw rotation about the vertical axis [7].

A particularly important characteristic of multirotor aircraft is that translational motion is not achieved directly by a horizontal propulsive force, but rather through vehicle tilting. When the aircraft tilts, the total thrust vector is no longer directed solely in the vertical direction; instead, it is decomposed into a vertical component, which maintains altitude, and a horizontal component, which enables motion in the desired direction. Consequently, the flight control system must simultaneously regulate the vehicle attitude, total thrust, and flight stability, making a multirotor aircraft a nonlinear and dynamically coupled system.

Energy efficiency remains one of the main limitations of multirotor UAV systems. In contrast to fixed-wing aircraft, which can rely on aerodynamic lift during cruise, multirotor aircraft must continuously generate thrust to remain airborne [8]. Consequently, energy consumption is high in hover, and the endurance is strongly influenced by mass, rotor diameter, disk loading, motor efficiency, battery characteristics and mission profile.

From an engineering perspective, the design of a multirotor aircraft represents a compromise between stability, payload capacity, flight endurance, energy efficiency, and control complexity. A larger number of rotors may improve thrust distribution, increase payload capacity, and enhance system reliability; however, it also increases vehicle mass, the number of propulsion components, and overall structural complexity. On the other hand, simpler configurations reduce mass and energy consumption but provide lower redundancy in the event of a propulsion unit failure.

In addition to technical requirements, the development and deployment of UAV systems are also governed by the applicable regulatory framework. In civil and industrial applications, increasing importance is placed on flight safety, airspace restrictions, operation categorization, operator responsibility, and compliance with relevant regulations [9]. Therefore, the modern development of UAV technology cannot be viewed solely in terms of performance improvement, but also through the safe and controlled integration of these systems into real-world operational environments.

The aim of this paper is to analyse the physical principles that determine the behaviour of multirotor UAV systems, with special emphasis on force distribution, moment generation and energy efficiency. The paper connects theoretical models with practical aspects of flight control in order to clarify the stability, controllability and operational limitations of multirotor platforms.

2. PHYSICAL MODEL OF MULTIROTOR FLIGHT

The motion of a multirotor UAV can be analysed using classical mechanics and rotor aerodynamics. In a simplified engineering model, the aircraft is treated as a rigid body acted upon by external forces and moments. This approach makes it possible to interpret the system through basic physical quantities while still retaining a clear connection with practical flight control [5,8].

2.1. Forces and Rotor Thrust

The three dominant external forces are weight, rotor thrust and aerodynamic drag. Weight acts vertically downward and represents the basic load that the propulsion system must overcome:

$$W = mg \tag{1}$$

where W is the weight of the aircraft [N], m is the mass [kg] and g is the gravitational acceleration [m/s²]. The total thrust generated by all rotors is the sum of individual rotor thrusts:

$$\sum_{i=1}^{n} T_i = T_1 + T_2 + \cdots + T_n = W \tag{2}$$

In steady hover, the total thrust must balance the weight:

$$\sum T = W \tag{3}$$

If total thrust is greater than the weight, the aircraft climbs; if it is smaller, the aircraft descends. In a simplified rotor model, the thrust of the i-th rotor and its reaction torque can be approximated as follows:

$$T_i = k_T \omega_i^2 \tag{4}$$

$$Q_i = k_Q \omega_i^2 \tag{5}$$

where ω_i is the angular speed of the i-th rotor [rad/s], while kT and kQ are coefficients that depend on propeller geometry, air density and propulsion-system characteristics [5]. These relations show that thrust and reaction torque are nonlinear functions of rotor speed. In real flight, this nonlinearity is further affected by air-density variation, turbulence and rotor-rotor interference [10].

2.2. Moments and Attitude Control

Moments about the body axes define the rotational behaviour of the aircraft. Roll is rotation about the longitudinal axis, pitch is rotation about the lateral axis, and yaw is rotation about the vertical axis. Roll moment is generated by a thrust difference between the left and right sides of the aircraft, while pitch moment is generated by a thrust difference between the front and rear rotors [5].

Yaw moment originates from the reaction torques of the rotors. In a quadrotor, two rotors rotate clockwise (CW) and two counter-clockwise (CCW). In hover, the reaction torques cancel each other. Yaw motion is produced by increasing the speed of rotors rotating in one direction and decreasing the speed of rotors rotating in the opposite direction, while keeping the total thrust approximately constant [6,7].

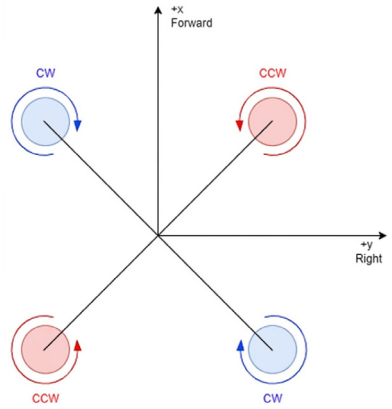


Figure 1: Quadrotor geometry and CW/CCW rotor directions

2.3. Translational Motion and Drag

A multirotor aircraft does not produce horizontal motion by using a separate horizontal propulsive force. Instead, the aircraft tilts, and the total thrust vector is decomposed into vertical and horizontal components. For a small tilt angle α, the components can be written as:

$$T_v = \sum T \cos \alpha \tag{6}$$

$$T_h = \sum_{i=1}^{n} T_i \sin \alpha \tag{7}$$

The horizontal component enables motion in the desired direction, while the vertical component must still support the weight of the aircraft. To maintain altitude during tilted flight, the required total thrust becomes:

$$\sum T = \frac{W}{\cos \alpha} \tag{8}$$

This relation explains why aggressive maneuvers and large tilt angles reduce endurance: an increase in tilt requires an increase in total thrust and therefore greater power consumption [8]. Aerodynamic drag acts opposite to the direction of motion and becomes more important in translational flight than in hover:

$$D = \frac{1}{2}\rho C_D A v^2 \tag{9}$$

where D is the drag force [N], ρ is air density [kg/m³], C_D is the drag coefficient, A is the reference area [m²] and v is the flight speed [m/s]. Drag has a direct influence on energy consumption, particularly at higher speeds and during longer missions.

Table 1: Command interpretation for a quadrotor in X configuration

Command	Increase ω	Decrease ω	Main effect
Forward	Rear pair	Front pair	Thrust vector tilts toward +X; forward motion
Backward	Front pair	Rear pair	Thrust vector tilts toward −X; backward motion
Right	Left pair	Right pair	Thrust vector tilts toward +Y; right motion
Left	Right pair	Left pair	Thrust vector tilts toward −Y; left motion
Yaw right	CW group	CCW group	Rotation about +Z with small change in ΣT
Yaw left	CCW group	CW group	Rotation about −Z with small change in ΣT

From the control perspective, the controller must transform desired translational and rotational motion into rotor-speed commands. Because every change in rotor speed affects both force and moment distribution, multirotor stabilisation requires fast sensors, robust state estimation and continuous feedback control [4,10].

The presented control allocation scheme shows that a multirotor system achieves control through differential variation of the thrust generated by individual rotors. Such a principle enables highly precise control of the aircraft attitude and position; however, it also increases the demands placed on the control algorithm, sensor response time, and the onboard energy system [4].

This model demonstrates that a multirotor aircraft does not achieve translational motion in the same manner as a conventional aircraft. In fixed-wing aircraft, horizontal motion is achieved through a combination of propulsion and aerodynamic lift generated by the wings, whereas in multirotor aircraft motion is obtained by changing the orientation of the thrust vector. Consequently, thrust in a multirotor system serves a dual purpose: it provides lift while simultaneously enabling vehicle control and maneuvering [6].

Under real flight conditions, system behavior additionally depends on aircraft mass, center-of-gravity location, rotor arrangement geometry, propeller characteristics, environmental conditions, and the quality of the sensor system. If the center of gravity is not properly positioned relative to the rotor plane, additional moments may arise, making stabilization more difficult. Therefore, the design of multirotor systems requires a symmetrical mass distribution and proper sizing of propulsion components [7].

An additional source of system complexity arises from the strong coupling between flight dynamics and control. Any change in the thrust generated by an individual rotor affects not only translational motion but also the distribution of moments about all aircraft axes. As a result, stabilization of multirotor platforms requires continuous control-command adjustments and accurate real-time processing of sensor data [10].

The basic physical model indicates that a multirotor unmanned aircraft is a complex nonlinear system in which small variations in the rotational speed of individual rotors can significantly alter the overall distribution of forces and moments. Consequently, stable flight requires continuous thrust correction, measurement of attitude and velocity, and closed-loop feedback provided by sensors such as gyroscopes, accelerometers, and state-estimation systems [5]. This close interaction between the physical model and the flight control system forms the basis for the further analysis of multirotor UAV systems presented in this paper [6].

3. CONFIGURATION AND CONTROL MODEL

The configuration of a multirotor UAV is defined by the number of rotors, their geometric arrangement and the corresponding control-allocation logic. A quadrotor is the simplest and most common configuration, while hexarotors and octorotors are used when higher payload capacity, greater stability or partial fault tolerance is required [6].

3.1. Vertical Motion and Control Allocation

Vertical motion is governed by the balance between total thrust and weight. Hover corresponds *to* $\Sigma T_i = W$ and $h = const$; climb occurs when $\Sigma T_i > W$; and descent occurs when $\Sigma T_i < W$ [5]. The flight controller coordinates rotor speeds so that total thrust and body moments satisfy the desired motion. For a quadrotor in X configuration, a simplified linear control allocation can be expressed as:

$$T_1 = \frac{1}{4}u_T + \frac{1}{2}u_\phi + \frac{1}{2}u_\theta - \frac{1}{2}u_\psi \tag{10}$$

$$T_2 = \frac{1}{4}u_T - \frac{1}{2}u_\phi + \frac{1}{2}u_\theta + \frac{1}{2}u_\psi \tag{11}$$

$$T_3 = \frac{1}{4}u_T - \frac{1}{2}u_\phi - \frac{1}{2}u_\theta - \frac{1}{2}u_\psi \tag{12}$$

$$T_4 = \frac{1}{4}u_T + \frac{1}{2}u_\phi - \frac{1}{2}u_\theta + \frac{1}{2}u_\psi \tag{13}$$

where uT is the total-thrust command and $u\varphi$, $u\theta$ and $u\psi$ are the roll, pitch and yaw control commands. This formulation is an approximation, but it clearly shows that attitude and thrust commands are distributed to individual rotors rather than applied independently.

3.2. Rotor Configurations and Redundancy

Increasing the number of rotors does not only increase payload potential; it also changes the redundancy and complexity of the system. A quadrotor has limited tolerance to a motor failure because the loss of one propulsion unit directly disturbs both force and moment balance. Hexarotors and octorotors introduce additional actuators and therefore allow partial thrust redistribution in some failure scenarios [7].

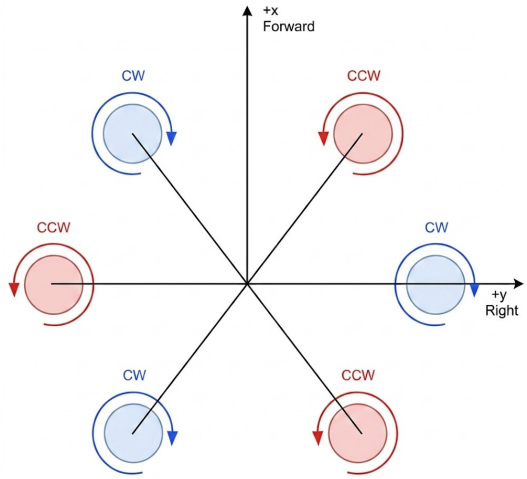


Figure 2: Hexarotor geometry and alternating rotor directions

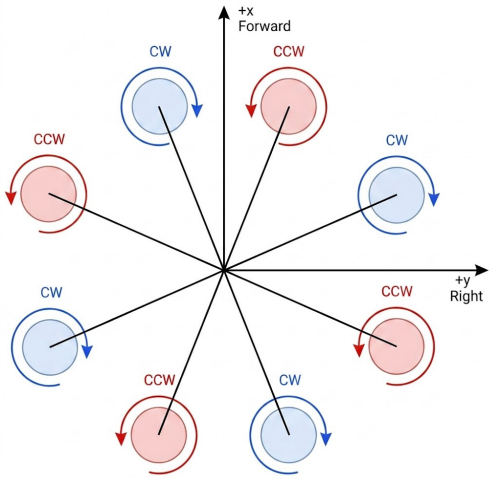


Figure 3: Octorotor geometry and alternating rotor directions

The advantage of additional rotors is not automatic. The remaining motors must have enough thrust reserve, the structure must be able to withstand redistributed loads, and the controller must solve the allocation problem in real time while respecting motor limits and available battery power.

3.3. Failure-Case Thrust Model

The preservation of flight after a propulsion-unit failure can be interpreted as a control-allocation problem. The controller must maintain the total thrust required for weight compensation and, at the same time, generate stabilising moments:

$$\sum_{i=1}^{n} T_i = W \tag{14}$$

$$M = BT \tag{15}$$

where B is the allocation matrix that maps individual rotor thrusts into the resulting moments. Its structure depends on rotor geometry and the selected multirotor configuration [7]. For a system with n rotors and one failed rotor, a minimum thrust condition is:

$$(n-1)T_{max} \geq W \tag{16}$$

For practical controllability, however, this condition is not sufficient. A thrust reserve is required, and a common engineering criterion is that the maximum thrust per rotor should be at least twice the hover thrust:

$$T_{max} \geq 2T_{hover} \tag{17}$$

This criterion provides margin for disturbance rejection, attitude correction and emergency maneuvers, but it also increases the requirements placed on motors, propellers and the battery system.

4. ENERGY MODEL AND DISK LOADING

Energy efficiency is a critical parameter that defines mission endurance, operational radius and useful payload. Multirotor platforms consume significant energy even in stationary flight because the rotors must continuously accelerate air downward in order to balance the aircraft weight. Therefore, energy modelling is central to multirotor design and mission planning [10,11].

4.1. Disk Loading

Disk loading is a basic parameter of rotor aerodynamic efficiency. It is defined as the ratio between aircraft weight and total effective rotor-disk area:

$$DL = \frac{W}{\Sigma A_i} \tag{18}$$

Lower disk loading generally improves hover efficiency because the required thrust is produced by accelerating a larger mass of air to a lower induced velocity. This is why larger propellers can reduce the power required for hover. At the same time, larger propellers increase aircraft dimensions, may require stiffer arms and can limit use in urban or confined spaces [3,12].

4.2. Induced Velocity, Ideal Power and Endurance

From momentum theory, the induced velocity through the i-th rotor disk can be approximated as:

$$v_i = \sqrt{\frac{T_i}{2\rho A_i}} \tag{19}$$

For the complete aircraft in hover, ideal induced power can be written in compact form as:

$$P_{ideal} = \sqrt{\frac{W^3}{2\rho \Sigma A_i}} \tag{20}$$

The relation shows that an increase in aircraft weight has a strong influence on hover power. Even a relatively small mass increase due to sensors, payload or batteries can noticeably reduce flight endurance [13]. A practical estimate of endurance is:

$$t = \frac{E_{bat}\eta}{P_{prosek}} \tag{21}$$

where E_{bat} is the usable battery energy [Wh], η is the total system-efficiency coefficient and $P_{el,avg}$ is the average electrical power [W]. In real operation, energy reserves must also be planned for wind, manoeuvres, landing, sensor operation and unexpected mission changes [14,15].

5. NUMERICAL ANALYSIS OF A REFERENCE UAV SYSTEM

A reference numerical analysis was carried out to quantify the influence of rotor diameter and propulsion configuration on energy efficiency. The analysed system is a medium-size multirotor platform intended for civil and industrial applications. The adopted mass is m = 4 kg, which gives $W \approx 39.24$ N.

5.1. Baseline Thrust Calculation

In steady hover, the total thrust equals the aircraft weight. For a quadrotor configuration, the required thrust per rotor is:

$$T_i = \frac{W}{4} = \frac{39{,}24}{4} \approx 9.81 \text{ N} \tag{22}$$

This value is only the theoretical minimum required to maintain altitude. It does not include thrust reserve for stabilisation, acceleration, disturbance compensation or emergency manoeuvres. For this reason, a total thrust-to-weight ratio $T/W \geq 2$ is often used as a practical design target for reliable control.

5.2. Influence of Propeller Diameter

The influence of propeller diameter was analysed for a constant aircraft mass of 4 kg and four rotors. The comparison includes propellers of 10 in, 12 in and 15 in. The numerical results are summarised in Table 2.

Table 2: Influence of propeller diameter on ideal hover power (m = 4 kg, four rotors)

Case	Propeller	A_i [m²]	ΣA_i [m²]	v_i [m/s]	$P_{ideal,tot}$ [W]
A	12 in (0.305 m)	0.07298	0.2919	7.42	291
B	10 in (0.254 m)	0.05070	0.2028	8.87	348
C	15 in (0.381 m)	0.11400	0.4560	5.93	233

Increasing the diameter from 12 in to 15 in reduces ideal hover power by about 20%, while reducing the diameter to 10 in increases it by about 19.5%. The reason is the change in total disk area and induced velocity. However, the largest propeller is not necessarily the best practical solution, because it increases the aircraft envelope and structural requirements.

5.3. Rotor Configuration and Hover Endurance

The performance of quadrotor, hexarotor and octorotor configurations was compared using 12 in propellers and a 6S 10 Ah battery system with usable energy of approximately 178 Wh. The results are given in Table 3.

Table 3: Comparison of rotor configurations (12 in propellers, m = 4 kg)

Configuration	Rotors	DL [N/m²]	$P_{el,hover}$ [W]	Hover time [min]
Quadrotor	4	134.4	≈ 420	≈ 25.4
Hexarotor	6	89.6	≈ 343	≈ 31.1
Octorotor	8	67.2	≈ 297	≈ 35.9

At identical mass and propeller diameter, the octorotor configuration gives approximately 41% longer hover time than the quadrotor. This result confirms the importance of total rotor-disk area. Nevertheless, in a real design, additional rotors increase mass, cost, wiring complexity, maintenance requirements and the complexity of the flight-control allocation problem.

5.4. Practical Interpretation

The most energy-efficient operating condition is hover or mild translational motion with small tilt angles. A tilt of 10° already creates a useful horizontal thrust component, but it also requires an increase in total thrust to maintain altitude. Therefore, aggressive maneuvers, abrupt direction changes and large tilt angles directly increase power consumption and reduce endurance. The results indicate that mass optimisation, propeller selection, rotor configuration and mission-profile planning must be treated as a coupled design problem.

For fault-tolerant operation, rotor count alone is not sufficient. After a motor failure, the remaining propulsion units must generate enough thrust and stabilising moments. For example, in a hexarotor with one failed motor and $W = 39.24$ N, the remaining five rotors must generate at least 7.85 N each only to compensate for weight. Additional margin is still necessary for attitude stabilisation and yaw-moment compensation.

6. PERFORMANCE AND APPLICATIONS OF MULTIROTOR UAV SYSTEMS

The previous analysis shows that the performance of multirotor UAV systems cannot be assessed through a single design parameter. Endurance, payload capacity, stability, controllability and operational safety are the result of a coupled interaction between aircraft mass, rotor configuration, propeller diameter, disk loading, thrust reserve, propulsion efficiency and mission profile. Therefore, the selection of a multirotor platform should not be reduced to the number of rotors or battery capacity alone, but should be treated as an integrated engineering optimisation problem.

From the performance perspective, quadrotor configurations offer the simplest and lightest solution. They have fewer propulsion units, lower structural complexity, reduced maintenance requirements and simpler control allocation. These advantages make quadrotors suitable for compact platforms, short-duration missions, educational applications, visual inspection, mapping of smaller areas and general monitoring tasks. However, their simplicity is accompanied by limited redundancy. Since each rotor has a critical role in both lift generation and attitude control, the failure of a single motor or propeller usually causes a major loss of stability and controllability.

Hexarotor and octorotor configurations provide a different balance between efficiency, redundancy and complexity. By increasing the number of rotors, the total rotor-disk area is increased, disk loading is reduced and the required induced velocity in hover becomes lower. This can improve hover efficiency and provide greater possibilities for thrust redistribution. Such configurations are therefore more suitable for missions that require higher payload capacity, improved stability, safer operation or partial tolerance to propulsion-unit failure. Typical examples include industrial inspection, professional photogrammetry, environmental monitoring, emergency response and missions above sensitive or inaccessible areas.

The numerical results obtained in this paper confirm the importance of total rotor-disk area. For the same aircraft mass and propeller diameter, configurations with more rotors can achieve lower disk loading and lower hover power. However, this result must be interpreted with caution. In a real UAV system, additional rotors also introduce additional motor mass, arms, wiring, electronic speed controllers, structural reinforcements and control complexity. Consequently, the theoretical improvement in hover efficiency can be partially reduced or even cancelled if the added components significantly increase the total aircraft mass. This is why the optimal configuration depends strongly on the intended mission rather than on aerodynamic efficiency alone.

Propeller diameter has a similarly important but constrained influence. Larger propellers reduce induced power because they increase the effective disk area and accelerate a larger mass of air at a lower velocity. This improves hover efficiency and can extend endurance. At the same time, larger propellers increase the overall aircraft dimensions, require longer and stiffer arms, reduce compactness and may limit operation in narrow, urban or indoor environments. For this reason, the most efficient propeller from the energy point of view is not always the most practical solution from the operational point of view.

Another important conclusion is that increasing battery capacity does not automatically increase endurance in a proportional way. Although a larger battery stores more energy, it also increases aircraft mass. This additional mass increases the thrust required for hover and therefore raises power consumption. A poorly balanced design with an oversized battery may have only a small endurance improvement, or in some cases no meaningful improvement at all. Effective endurance optimisation therefore requires simultaneous consideration of battery energy, aircraft mass, propulsion efficiency and mission power demand.

In practical applications, multirotor UAVs are especially valuable when the mission requires vertical take-off and landing, stable hovering, precise positioning and low-speed manoeuvring. These characteristics make them useful for infrastructure inspection, geodetic surveying, agricultural monitoring, terrain mapping, environmental observation, search-and-rescue operations and industrial diagnostics. In such missions, the main advantage of a multirotor UAV is not maximum

forward-flight efficiency, but the ability to remain accurately positioned above or near a target while carrying sensors, cameras or measuring equipment.

For infrastructure and industrial inspection, multirotor UAVs can approach bridges, roofs, towers, chimneys, power lines and other structures without the need for scaffolding, cranes or direct human access. In agriculture and environmental monitoring, they can provide spatially distributed data on vegetation, soil condition, irrigation, erosion, pollution or flood risk. In emergency-response scenarios, they can rapidly collect visual or thermal information from areas that may be unsafe for people, such as fire zones, collapsed structures, flooded terrain or chemically contaminated locations.

The use of multirotor UAVs in public-safety and emergency-response systems also highlights the importance of reliability and fail-safe behaviour. In these applications, flight failure may not only result in equipment loss, but may also create additional risk for people and property. Therefore, platforms intended for such missions require sufficient thrust reserve, robust communication, reliable navigation, emergency landing logic and, where necessary, partial propulsion redundancy. The choice between quadrotor, hexarotor and octorotor configurations should therefore be made according to the acceptable level of operational risk.

From the flight-control perspective, the practical performance of a multirotor UAV depends on how effectively the controller transforms desired motion into rotor-speed commands. Since the same rotors are responsible for lift, attitude control and yaw control, every manoeuvre affects both force and moment balance. This becomes particularly important during aggressive motion, wind disturbances, payload variation or partial actuator degradation. Advanced control algorithms, sensor fusion and adaptive control methods can improve stability and robustness, but they also require accurate modelling and reliable onboard computation.

Overall, the application potential of multirotor UAV systems is highest in missions where manoeuvrability, hovering ability, spatial precision and operational flexibility are more important than long-range aerodynamic efficiency. Their main limitations remain endurance, sensitivity to mass increase, wind influence, battery constraints and system complexity. Future development should therefore focus on lighter structures, more efficient propulsion systems, improved batteries, accurate energy prediction, fault-tolerant control and autonomous mission adaptation in realistic operating environments.

7. CONCLUSION

This paper presented a physical and engineering analysis of multirotor unmanned aerial vehicles, with emphasis on the relationship between force generation, moment control and energy efficiency. The analysis confirmed that multirotor UAVs are highly coupled dynamic systems in which lift, attitude control, yaw motion and power consumption cannot be considered independently. Every change in rotor speed affects not only the total thrust, but also the distribution of moments and the energetic state of the aircraft.

The force analysis showed that hover is governed by the balance between total rotor thrust and aircraft weight, while climb and descent are achieved by changing this balance. Translational motion is obtained by tilting the aircraft and redirecting the thrust vector, rather than by generating a separate horizontal propulsive force. This means that forward, backward or lateral motion always introduces a compromise between horizontal acceleration and vertical lift. As the tilt angle increases, additional thrust is required to maintain altitude, which directly increases power demand and reduces endurance.

The moment analysis demonstrated that roll, pitch and yaw control are generated through controlled differences in rotor thrust and reaction torque. Roll and pitch motion depend on asymmetric thrust distribution between opposite sides of the aircraft, while yaw motion depends on the torque imbalance between clockwise and counter-clockwise rotating rotors. These principles explain why the flight controller must continuously solve a control-allocation problem: it must distribute commands to individual rotors while simultaneously satisfying requirements for lift, attitude stabilisation and yaw control.

The energy analysis confirmed that disk loading is one of the key parameters of multirotor efficiency. A larger total rotor-disk area reduces induced velocity and ideal hover power, while higher aircraft mass strongly increases the required power. The numerical example for a 4 kg UAV showed that increasing propeller diameter can significantly reduce ideal hover power, and that increasing the number of rotors can reduce disk loading when the total mass is kept constant. However, these results must not be interpreted mechanically. In real UAV design, larger propellers and additional rotors also increase size, structural demands, mass, cost and control complexity.

The comparison of quadrotor, hexarotor and octorotor configurations showed that no configuration is universally optimal. A quadrotor is usually the most rational choice when simplicity, compactness, low mass and low cost are dominant requirements. Hexarotor and octorotor configurations become more appropriate when the mission requires higher payload capacity, improved stability, greater thrust reserve or partial fault tolerance. The selection of the configuration must therefore be based on the mission profile, payload, safety requirements, operating environment and available energy, rather than on rotor number alone.

One of the most important engineering conclusions is that endurance cannot be improved by isolated component changes. Increasing battery capacity, propeller diameter or rotor number may improve one aspect of performance, but each of these changes also affects mass, geometry, structural stiffness, power consumption and control allocation. For this reason, multirotor UAV design must be understood as a coupled optimisation problem involving propulsion, structure, control system, battery selection and mission planning.

The practical implication of the analysis is that multirotor UAVs are most effective in missions where hovering, vertical take-off and landing, precise positioning and low-speed manoeuvrability are more important than long-range aerodynamic efficiency. Their advantages are especially relevant in inspection, mapping, monitoring, emergency response and operations in hazardous or inaccessible environments. At the same time, their main limitations remain short endurance, sensitivity to payload increase, wind influence, battery constraints and dependence on reliable control systems.

Future research should move from simplified analytical models toward experimentally validated performance models. This includes detailed measurement of propulsion efficiency, battery behaviour under variable load, rotor-rotor aerodynamic interference, influence of wind and manoeuvres, and energy consumption during realistic mission profiles. Special attention should also be given to fault-tolerant control, adaptive control allocation and autonomous decision-making under degraded operating conditions.

Overall, the study confirms that efficient and reliable multirotor UAV design requires an integrated approach. The best solution is not the aircraft with the largest battery, the greatest number of rotors or the largest propellers, but the platform in which mass, thrust, disk area, control authority, energy reserve and mission requirements are properly balanced. This balance represents the central engineering challenge in the development of modern multirotor UAV systems.

ACKNOWLEDGEMENTS

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-34/2026-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2026” (No. 01-3609/1).

REFERENCES

[1] R. Austin, “Unmanned Aircraft Systems: UAVS Design, Development and Deployment”, John Wiley & Sons, Chichester, (2010)

[2] K. P. Valavanis and G. J. Vachtsevanos, “Eds., Handbook of Unmanned Aerial Vehicles”, Springer, Dordrecht, (2015)

[3] M. Hassanalian and A. Abdelkefi, “Classifications, applications, and design challenges of drones: A review”, Progress in Aerospace Sciences, Vol. 91, pp. 99–131, https://doi.org/10.1016/j.paerosci.2017.04.003, (2017)

[4] Y. Wang and J. Guo, “Literature Review on Autonomous Navigation of Drones”, International Core Journal of Engineering, Vol. 11(5), pp. 354–364, https://doi.org/10.6919/ICJE.202505_11(5).0042, (2025)

[5] R. Mahony, V. Kumar, and P. Corke, “Multirotor aerial vehicles: Modeling, estimation, and control of quadrotor”, IEEE Robotics & Automation Magazine, Vol. 19(3), pp. 20–32, https://doi.org/10.1109/MRA.2012.2206474, (2012)

[6] S. Bouabdallah, P. Murrieri, and R. Siegwart, “Design and control of an indoor micro quadrotor”, Proceedings of the IEEE International Conference on Robotics and Automation ICRA ‘04, New Orleans (LA, USA), 26 April – 01 May 2004, pp. 4393–4398, https://doi.org/10.1109/ROBOT.2004.1302409, (2004)

[7] P. Pounds, R. Mahony, and P. Corke, “Modelling and control of a large quadrotor robot”, Control Engineering Practice, Vol. 18(7), pp. 691–699, https://doi.org/10.1016/j.conengprac.2010.02.008, (2010)

[8] J. G. Leishman, “Principles of Helicopter Aerodynamics, 2nd ed.”, Cambridge University Press, Cambridge, (2006)

[9] European Union Aviation Safety Agency, “Operating a drone,” EASA, Cologne, available at: easa.europa.eu (accessed 26 April 2026)

[10] H. Li, R. Jia, Z. Zheng, and M. Li, “Energy-Efficient UAV Trajectory Design and Velocity Control for Visual Coverage of Terrestrial Regions”, Drones, Vol. 9(5), p. 339, https://doi.org/10.3390/drones9050339, (2025)

[11] F. Ahmed, J.C. Mohanta, A. Keshari, and P.S. Yadav, “Recent Advances in Unmanned Aerial Vehicles: A Review”, Arabian Journal for Science and Engineering, Vol. 47, pp. 7963–7984, (2022)

[12] NASA, “Advanced Air Mobility (AAM) Mission Overview,” NASA, Washington, available at: nasa.gov/aam (accessed 3 May 2026)

[13] Z. Liu, R. Sengupta and A. Kurzhanskiy, “A power consumption model for multi-rotor small unmanned aircraft systems”, 2017 International Conference on Unmanned Aircraft Systems (ICUAS), Miami (FL,USA), 13 June – 16 June 2017, pp. 310–315, https://doi.org/10.1109/ICUAS.2017.7991310, (2017).

[14] M. Jacewicz, M. Żugaj, R. Głębocki, and P. Bibik, “Quadrotor Model for Energy Consumption Analysis,” Energies, Vol. 15(19), p. 7136, https://doi.org/10.3390/en15197136, (2022).

[15] ICAO, “Unmanned Aircraft Systems (UAS) Toolkit,” ICAO, Montreal, available at: icao.int/uas (accessed 3 May 2026).

SESSION A: EARTH MOVING AND TRANSPORTATION MACHINERY

Segment-based energy balance and propulsion feasibility analysis of a light passenger VTOL aircraft for a regional mission

Vladimir Žerajić[1*], Dragan Ružić[2], Dragan Živanić[2], Nikola Ilanković[2]

[1] GASTRANS, Novi Sad, Serbia

[2] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia

ARTICLE INFO

* **Correspondence**: vladimir.zerajic@gmail.com

DOI: https://doi.org/10.46793/ET26.A09Z **PAGES:** A69 – A77

ABSTRACT

This paper presents an engineering calculation of the energy balance of a light passenger vertical take-off and landing (VTOL) aircraft for a reference regional route between Novi Sad and Novi Beograd. The mission profile is divided into vertical take-off, climb, transition, cruise, descent, vertical landing and operational reserve. For each segment, the required shaft power and energy are estimated using the adopted aircraft mass, rotor geometry, disk loading and propulsion efficiency. The analysis compares battery-electric, liquid-fuel piston and hybrid-electric propulsion concepts from the aspects of available energy, mass balance, peak power demand and practical feasibility. The numerical case study indicates that the reference mission requires approximately 189.93 kWh of shaft energy for a flight time of 53.25 min, while the vertical phases impose peak power requirements of about 265 kW. The results show that battery-electric propulsion is limited mainly by energy-storage mass, liquid-fuel propulsion offers a more favorable specific-energy basis, and hybrid-electric propulsion can represent a practical compromise if its additional system complexity is justified by mission range and power-distribution benefits.

KEYWORDS

VTOL aircraft, Energy balance, Propulsion feasibility, Disk loading, Hybrid-electric propulsion, Energy efficiency, Shaft power, Regional air mobility.

1. INTRODUCTION

The development of light passenger vertical take-off and landing VTOL (Vertical Take-Off and Landing) aircraft is one of the most active directions in advanced air mobility. Such aircraft can operate without conventional runway infrastructure and may therefore support fast regional connections, flexible transport services and operations in urban or suburban environments where conventional airports are unavailable or economically unjustified [1].

In contrast to conventional fixed-wing aircraft, a VTOL configuration spends a significant part of its mission in vertical take-off, hovering, transitioning and landing. In these regimes lift is generated directly by rotor thrust rather than mainly by wing aerodynamics. Consequently, the energy balance becomes a primary criterion for assessing technical feasibility. The propulsion system must provide high instantaneous power in vertical flight and sufficient total energy to complete the cruise segment and the required operational reserve [2].

For light passenger VTOL platforms, the relationship between aircraft mass, disk loading, rotor diameter, flight regime and available energy is critical. Any increase in mass directly raises the required hover power, while a larger rotor diameter can reduce induced power but increases overall dimensions, structural complexity and integration requirements [3]. Therefore, the propulsion architecture cannot be selected only by nominal power; it must be evaluated together with the complete mass, geometry, energy and mission profile.

The subject of this paper is the energy-balance assessment of a light passenger VTOL aircraft on the reference route Novi Sad - Novi Beograd. The route is used as a representative medium-range regional mission that makes it possible to analyze the influence of aircraft mass, flight profile, energy reserve and propulsion-system characteristics on overall feasibility.

The paper considers several propulsion architectures: battery-electric propulsion, liquid-fuel piston propulsion using Jet-A, gasoline piston propulsion and hybrid-electric propulsion. Battery-electric propulsion has advantages in simplicity, local emissions and precise power control, but is constrained by battery specific energy. Liquid-fuel propulsion provides a much more favorable stored-energy-to-mass ratio, but introduces additional mechanical integration, vibration, cooling and maintenance requirements. Hybrid-electric propulsion attempts to combine the high specific energy of fuel with the flexibility of electric power distribution [4].

Particular importance in the analysis is given to the selection of reference technical parameters. In VTOL systems, disk loading and rotor diameter have a direct influence on the power required during vertical flight, while the mass of the energy source determines the available range and payload capacity. Therefore, this study examines the interrelationship between rotor geometry, total aircraft mass, propulsion architecture, and mission energy requirements. Such an approach makes it possible to assess, already in the early stages of the design process, whether the proposed concept is technically feasible or whether modifications to the configuration are required [5].

The objective of the paper is to use a simplified but engineering-relevant model to determine the energy demand of the reference mission and to compare propulsion variants in terms of power, energy, mass and practical feasibility. The results provide a basis for further configuration optimization and for defining technical constraints in the subsequent development of a light passenger VTOL aircraft.

2. MISSION PROFILE AND INITIAL ASSUMPTIONS

The analyzed mission is a regional transport flight of a light passenger VTOL aircraft on the Novi Sad - Novi Beograd route. The selected route represents a direct connection between two urban areas without relying on conventional airport infrastructure. This mission profile includes vertical take-off, climb, cruise, descent, landing and the necessary energy reserve.

For calculation, the mission is divided into segments. The total energy consumption is therefore not estimated only by an average power value, but as the sum of the energy required in individual flight phases. This is especially important for VTOL aircraft because vertical take-off and landing impose much higher instantaneous power demands than horizontal cruise [6].

The basic phases of the mission are vertical take-off, short hovering and stabilization, climb to operating altitude, cruise along the reference route, descent, final hovering and vertical landing. In addition to the energy required for these phases, an operational reserve is included to ensure safe mission completion under deviations from the planned flight profile.

In the early design phase, the purpose of the calculation is not to define every structural detail of the aircraft. The essential task is to determine whether the adopted relationship between mass, power, rotor geometry and available energy can support the mission. For that reason, a simplified segment-based energy balance is used.

The initial assumptions include aircraft mass, number of rotors, rotor diameter, total rotor-disk area, air density, estimated propulsion efficiency and duration of individual mission segments. Aircraft mass has a direct influence on vertical-flight power, while rotor diameter and total disk area determine induced power and disk loading [5].

For a light passenger VTOL aircraft, the relationship between useful payload, propulsion-system mass and energy-source mass is particularly important. An oversized battery or fuel system reduces the payload and structural margin, while insufficient energy leads to inadequate range or reserve. Thus, the mission profile must be analyzed together with the mass balance.

$$E_{total} = E_{takeoff} + E_{climb} + E_{cruise} + E_{descent} + E_{landing} + E_{reserve} \quad (1)$$

where E_{tot} is the total energy required for the mission and the individual terms represent the energy consumed in the corresponding flight phases. This segmentation allows a clearer comparison of propulsion variants because the most energy-demanding mission parts can be identified separately.

A particularly important segment of the mission profile is the cruise phase, as it has the greatest influence on the total distance traveled and the operational range of the aircraft. However, in VTOL configurations, the vertical flight phases must not be neglected, since they require a high thrust-to-weight ratio as well as sufficient power reserve for stabilization, attitude corrections, and safe landing operations [7].

The mission profile defined in this manner serves as the basis for further energy-balance analysis and comparison of different propulsion concepts. The following section presents the fundamental theoretical relationships linking aircraft mass, disk loading, required power, and total mission energy

For the purposes of this study, the mission is not considered as a detailed operational flight plan, but rather as an engineering reference profile intended for the comparison of propulsion solutions. Such an approach enables different configurations to be evaluated under identical conditions, using total energy demand, maximum required power, energy-source mass, and the impact of the propulsion system on payload capacity as the primary assessment criteria. In this way, the results are not tied to a single aircraft design, providing a broader insight into the technical feasibility of a light passenger VTOL platform concept.

3. THEORETICAL BASIS OF THE ENERGY BALANCE

The energy balance of a VTOL aircraft is based on determining the required power in individual mission phases and the corresponding energy consumed over their duration. Unlike conventional fixed-wing aircraft, VTOL systems spend an important portion of their energy in vertical take-off, hovering and landing because lift is then produced directly by rotor thrust [2].

The primary quantity for estimating propulsion-system loading is aircraft weight:

$$W = mg \tag{2}$$

where W is the aircraft weight [N], m is the aircraft mass [kg], and g is the gravitational acceleration [m/s²]. In steady hover, the total rotor thrust must be equal to the aircraft weight:

$$\sum T_i = W \tag{3}$$

This condition is the basis for the calculation of vertical mission phases. If thrust exceeds the aircraft weight, the aircraft can climb or provide additional margin for stabilization and control. Therefore, VTOL design must consider not only minimum hover thrust but also the available thrust-to-weight ratio.

A key parameter for evaluating rotor efficiency is disk loading, defined as the ratio between aircraft weight and total effective rotor-disk area:

$$DL = \frac{W}{A_{tot}} \tag{4}$$

where DL is disk loading [N/m²] and A_{tot} is the total effective rotor-disk area [m2]. Lower disk loading generally leads to lower induced power in hover because thrust is produced by accelerating a larger mass of air to a lower velocity [3].

$$A_{tot} = \frac{n\pi D^2}{4} \tag{5}$$

where n is the number of rotors and D is the rotor diameter [m]. This relation shows that increasing the rotor diameter strongly increases disk area, but it also increases aircraft size and structural requirements.

The ideal hover power can be estimated from momentum theory:

$$P_{ideal} = \sqrt{\frac{W^3}{2\rho A_{tot}}} \tag{6}$$

where ρ is air density [kg/m3]. In real conditions, the ideal power must be corrected for losses in rotors, motors, transmissions, electric power electronics or mechanical subsystems. The required real power can therefore be approximated as:

$$P_{real} = \frac{P_{ideal}}{\eta_{tot}} \tag{7}$$

where η_{tot} is the overall propulsion-system efficiency. For cruise, the required power can be approximated as the work needed to overcome aerodynamic drag at flight speed:

$$P_{cr} = \frac{DV}{\eta_{prop}} \tag{8}$$

where D is the total aerodynamic drag [N], V is flight speed [m/s], and η_{prop} is propeller efficiency. The energy consumed in each mission segment is:

$$E_i = P_i t_i \tag{9}$$

where E_i is the energy of the i-th segment, P_i is the required power and t_i is the segment duration. The total mission energy is obtained by summing all flight segments and adding the reserve. This model is suitable for propulsion comparison because it shows which flight phase imposes the dominant energy or peak-power requirement.

4. ANALYZED PROPULSION VARIANTS

The propulsion system is one of the decisive elements in assessing the feasibility of a light passenger VTOL aircraft. Such a system must provide high instantaneous power during vertical take-off and landing, sufficient energy for cruise, and an acceptable mass balance. Four propulsion architectures are considered: battery-electric propulsion, Jet-A piston propulsion, gasoline aviation piston propulsion and hybrid-electric propulsion.

4.1. Battery-Electric Propulsion

Battery-electric propulsion is the simplest architecture in terms of power transmission. Electrical energy from the battery pack is supplied through controllers to electric motors driving the propellers or rotors. The advantages are high motor efficiency, fast command response, simple power distribution and no local exhaust emissions [6].

The main limitation is battery mass. Although batteries allow direct use of electrical energy, their specific energy is significantly lower than that of liquid fuels. In VTOL aircraft this limitation is severe because increasing battery mass also increases the required vertical-flight power. In other words, additional stored energy simultaneously increases the aircraft weight and the energy needed to fly.

4.2. Jet-A Piston Propulsion

Piston engines using Jet-A fuel offer a favorable ratio between stored energy and fuel mass. Compared with a battery system, liquid fuel can provide a substantially larger range for the same energy-source mass. This advantage is important in VTOL aircraft because every additional kilogram directly affects the vertical-flight power requirement.

However, integration of Jet-A piston engines requires additional elements such as reduction gearboxes, cooling systems, tanks, fuel lines and power-transmission components. Therefore, the total assessment must include the complete propulsion system, not only the fuel mass.

4.3. Gasoline Aviation Piston Engines

Gasoline aviation piston engines are a mature and widely used technology in light aviation. They can provide high power at acceptable propulsion-unit mass and benefit from a well-developed technical base [10]. In a VTOL aircraft, however, distribution of power to multiple rotors is a major design challenge.

If several independent engines are used, redundancy increases but so do mass, cost and maintenance complexity. If one central engine with mechanical power distribution is used, the number of prime movers is reduced, but the transmission system becomes more complex and single-point failures become critical.

4.4. Hybrid-Electric Propulsion

Hybrid-electric propulsion is a compromise between battery-electric and piston propulsion. Energy can be obtained from liquid fuel, while power is distributed electrically to the rotors. This retains the high specific energy of fuel and enables more flexible power distribution and rotor control [4].

A possible architecture uses an internal-combustion engine as a generator, while electric motors at the rotors provide thrust. Such architecture reduces the need for complex mechanical power transmission but introduces additional losses in the generator, power electronics and motors. It also adds mass and complexity. The benefit of hybrid propulsion must therefore be evaluated at the complete system level rather than only at the energy-source level.

4.5. Comparative Interpretation

No analyzed propulsion solution is universally best. Battery-electric propulsion is simple and suitable for shorter missions but is limited by battery mass. Liquid-fuel propulsion has a strong advantage in specific energy and potential range, but it requires more complex mechanical integration. Hybrid-electric propulsion attempts to combine the advantages of both groups, yet it introduces the highest system complexity. For the Novi Sad - Novi Beograd mission, the key comparison criteria are total required energy, energy-source mass, available vertical-flight power and influence on useful payload.

5. NUMERICAL ANALYSIS OF THE NOVI SAD - NOVI BEOGRAD ROUTE

The numerical analysis applies the theoretical model to the reference mission. The aim is not to design a complete aircraft but to estimate concept feasibility at an early design stage. The reference aircraft is treated as a light passenger VTOL platform for regional transport, and the mission includes vertical take-off, climb, transition, cruise, descent, vertical landing and reserve. For the calculation, engineering-realistic assumptions are adopted regarding aircraft mass, rotor diameter, number of rotors, propulsion efficiency and duration of individual flight phases.

Table 1: Adopted reference parameters for the energy-balance calculation

Parameter	Symbol	Adopted value	Comment
Maximum take-off mass	MTOM	1000 kg	Reference case for the numerical calculation
Aircraft weight	W	9810 N	Calculated as $W = mg$
Number of lifting rotors	n	4	Conceptual VTOL layout
Rotor diameter	D	2.8 m	Adopted reference rotor diameter
Total rotor-disk area	A_{tot}	24.63 m²	Calculated from $n\pi D^2/4$
Disk loading	DL	398.3 N/m²	Calculated as W/A_{tot}
Air density	ρ	1.225 kg/m³	Standard sea-level condition
Overall propulsion efficiency	η_{tot}	0.70	Includes rotor, motor/transmission and conversion losses
Total mission time	t_{tot}	53.25 min	Including operational reserve
Required shaft energy	E_{tot}	189.93 kWh	Sum of mission segments
Maximum shaft power	P_{max}	265 kW	Required in vertical flight phases

5.1. Energy Requirements of Individual Mission Phases

The highest instantaneous power demands occur during the vertical take-off, hover, and landing phases. In these operating conditions, the rotors must generate thrust equal to or greater than the aircraft weight; therefore, any increase in aircraft mass immediately results in a higher power requirement. For this reason, the vertical flight phases are particularly critical for battery-electric and hybrid propulsion configurations.

The cruise phase lasts significantly longer than the vertical flight phases and therefore has a major influence on the total mission energy consumption. Although the instantaneous power required during cruise is generally lower than that required during vertical flight, the longer duration of this phase can result in substantial overall energy consumption. Consequently, the feasibility of a given route cannot be evaluated solely on the basis of take-off power requirements, but must also take into account the duration of the cruise segment [8].

Although the wider conceptual framework considers MTOM values between 1000 kg and 1200 kg, the numerical results presented in this paper refer to the 1000 kg reference case. This value is used to calculate the reference aircraft weight, disk loading, shaft-power demand and mission energy. The 1200 kg case and larger rotor diameters are treated as part of the broader design space for future configuration refinement.

Table 2: Segmented mission profile for the Novi Sad - Novi Beograd route

Mission phase	Phase description	Influence on energy
Vertical take-off	Short high-power regime	High instantaneous power demand
Climb	Transition to operating altitude	Moderate to high energy demand
Cruise	Main route segment	Largest contribution to total energy
Descent	Reduced power requirement	Lower energy demand
Vertical landing	Stabilization and final VTOL regime	High control and reserve demand
Reserve	Safety energy	Essential for operational safety

The largest instantaneous power requirements occur during vertical take-off, hovering and landing. In these phases, the rotors must directly generate thrust equal to or greater than the aircraft weight. The cruise phase lasts longer and therefore has a major influence on total mission energy, even if its instantaneous power is lower than in vertical flight.

For the reference profile, the total flight time is approximately 52-53 min including operational reserve. For MTOM = 1000 kg, the required shaft energy is in the range of approximately 186-190 kWh. If this energy is supplied exclusively by a battery system, battery mass becomes the dominant limitation because it reduces useful payload and increases aircraft mass. Therefore, the battery-electric option is technically simple but mass- and energy-constrained for the analyzed regional route.

5.2. Comparison of Propulsion Concept

In the battery-electric configuration, the primary limitation is the mass of the battery pack [8]. Although electric propulsion offers high efficiency and simplified power management, the energy required for the entire mission may necessitate a battery system of considerable mass. This reduces payload capacity and may result in a situation where additional battery energy simultaneously increases the total energy required for flight.

Within the reference mission profile, a flight duration of approximately 52–53 minutes was considered, including an operational reserve for the safe completion of the mission. For the case of MTOM = 1000 kg, the required shaft energy was estimated to be in the range of 186–190 kWh. If this energy is provided exclusively by a battery system, battery mass becomes the dominant design constraint, as it directly reduces payload capacity and increases the overall aircraft mass. Consequently, the BEV configuration represents the technically simplest, but also the least favorable solution in terms of energy-to-mass ratio for the considered regional route.

For propulsion systems based on Jet-A fuel, the main advantage is the high specific energy of liquid fuel. For the same usable energy, the required fuel mass is significantly lower than the mass of an equivalent battery pack. However, this configuration requires the additional mass of the engine, cooling system, fuel tanks, transmission components, and auxiliary subsystems. Therefore, the overall assessment must consider not only the fuel mass, but also the complete propulsion system [9].

Gasoline piston aircraft engines represent a well-established solution in light aviation. Their advantages include widespread availability, relatively low mass, and a mature technical support infrastructure. However, in VTOL configurations, an additional challenge arises from the distribution of power among multiple rotors and operation under high loads during the vertical phases of flight [10].

The hybrid-electric configuration represents a compromise solution. A piston engine may serve as the primary energy source, while the electric system enables flexible power distribution to the rotors. The main advantage of this architecture is the ability to combine the high energy density of liquid fuel with the controllability benefits of electric propulsion. Its main disadvantage is the increased complexity and additional mass associated with the generator, batteries, power electronics, and electric motors [4].

Liquid-fuel propulsion offers a more favorable specific-energy basis. For the same useful energy, the required fuel mass is much lower than the battery mass. Nevertheless, the total assessment must include engine mass, cooling, tanks, power transmission and auxiliary systems. The hybrid-electric option is a compromise: the combustion engine can provide the main energy source, while the electrical system enables flexible power distribution to the rotors.

For the high-power phases, a conservative reference shaft-power level of 265 kW was adopted to include vertical lift, stabilization and transition margins at the preliminary design stage. The same reference value is used for vertical take-off, climb and transition, while lower values are adopted for cruise and descent according to the reduced shaft-power demand in these phases.

Table 3: Estimated power and energy by mission segment

Mission segment	Power [kW]	Time [min]	Energy [kWh]
Vertical take-off	265.0	1.00	4.42
Climb	265.0	2.50	11.04
Transition	265.0	0.50	2.21
Cruise	212.0	26.25	92.75
Descent	132.5	2.00	4.42
Vertical landing	265.0	1.00	4.42
Reserve	212.0	20.00	70.67
Total		53.25	189.93

The results show that the largest contribution to total energy comes from cruise and operational reserve, while vertical phases impose the maximum instantaneous power of approximately 265 kW. This distribution demonstrates why VTOL propulsion selection must consider both total mission energy and peak power by flight segment.

Table 4: Comparative feasibility assessment of propulsion concepts

Propulsion concept	Main advantage	Main limitation	Energy-source mass implication	Feasibility for the reference mission
Battery-electric	Simple architecture, high controllability, fast power response, no local emissions	High battery mass due to limited specific energy	Very high battery mass; may dominate the aircraft mass balance	Feasible only if battery mass remains compatible with payload and MTOM limits; critical for the analyzed regional mission
Jet-A piston	High fuel specific energy and good range potential	Engine, cooling, tank and transmission integration complexity	Low fuel mass, but additional engine, cooling and transmission mass must be included	More favorable from the stored-energy-mass perspective, but mechanically more complex
Gasoline piston	Mature light-aviation technology and acceptable engine power density	Difficult mechanical power distribution to multiple rotors	Low to moderate fuel mass, but mechanical distribution may increase total system mass	Possible, but configuration-dependent and sensitive to transmission complexity
Hybrid-electric	Combines fuel energy density with flexible electric power distribution	Additional mass, conversion losses and system complexity	Fuel mass is favorable, but generator, electric motors, power electronics and battery buffer add mass	Potentially attractive if the range and power-distribution benefits compensate for added system mass

The segment calculation gives a total shaft energy of approximately 189.93 kWh for the reference mission and a total flight time of 53.25 min. Cruise accounts for 92.75 kWh, while the reserve accounts for 70.67 kWh. Together, cruise and reserve require 163.42 kWh, which corresponds to approximately 86% of the total shaft-energy demand. This indicates that the feasibility of the reference VTOL mission is governed not only by the peak power required for vertical flight, but also by route length and reserve policy. The results therefore confirm that range, reserve and vertical-flight power must be optimized together.

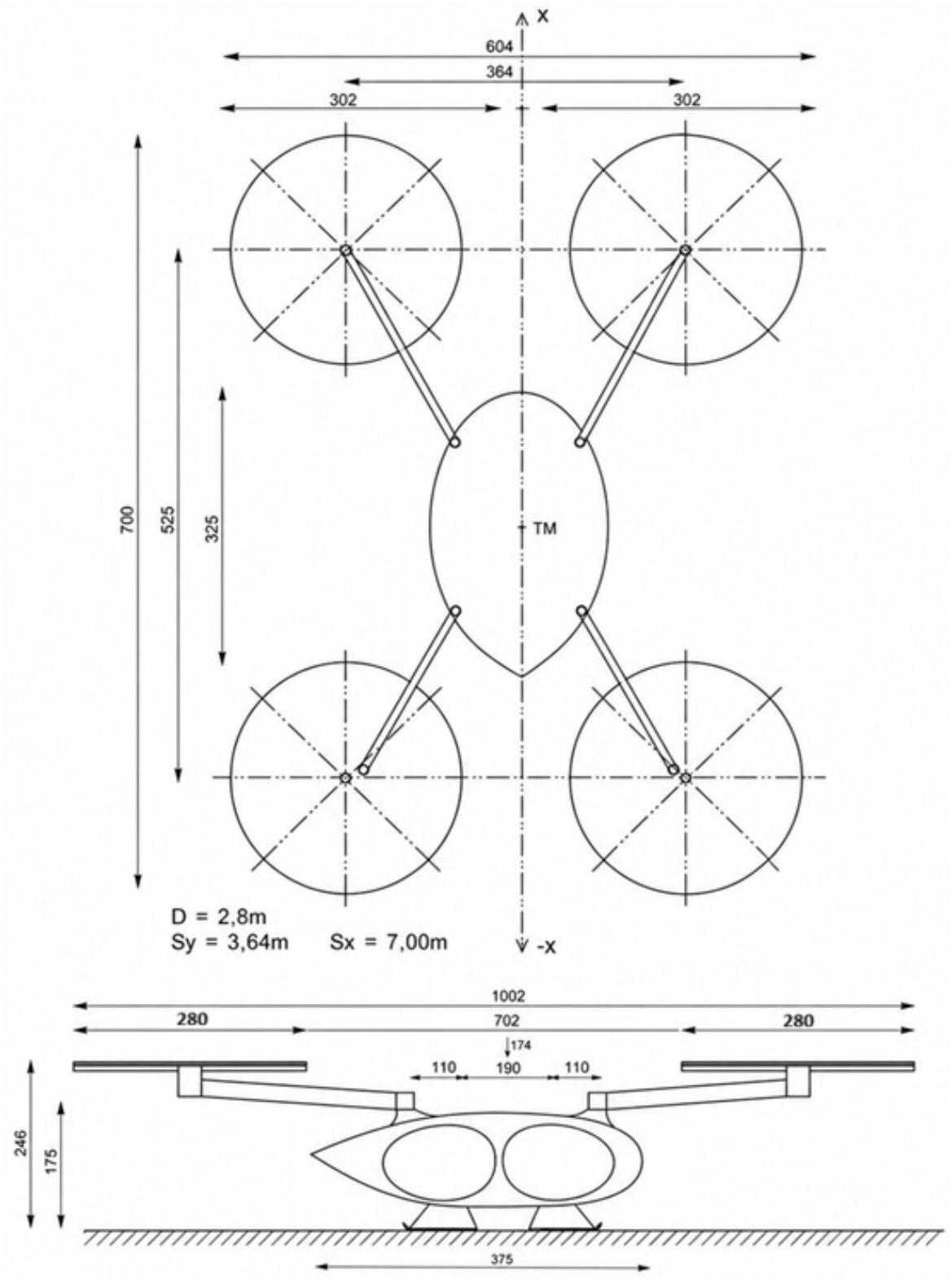


Figure 1: Conceptual layout of the analyzed light passenger VTOL configuration for the Novi Sad - Novi Beograd route

6. DISCUSSION OF RESULTS AND CONCEPT LIMITATIONS

The analysis shows that the feasibility of a light passenger VTOL aircraft cannot be evaluated using only one technical parameter. Aircraft mass, rotor geometry, propulsion-chain efficiency, mission profile, available energy and reserve requirements must be considered simultaneously. These parameters are strongly coupled because each increase in mass directly increases the required power in vertical take-off, hovering and landing. The reference mission requires 189.93 kWh of shaft energy and peak power of 265 kW in vertical flight phases. The reserve alone contributes 70.67 kWh, showing that safety requirements have a major influence on the total energy balance. This is a critical point: a concept that appears feasible without reserve may become impractical once operational margins are included.

An important result is that the analyzed VTOL concept is not limited only by the maximum shaft power required during vertical flight. A propulsion system may be capable of delivering the required 265 kW for take-off and landing, but it may still be unsuitable if it cannot provide approximately 190 kWh of usable shaft energy within an acceptable mass. This distinction is essential because VTOL feasibility is often incorrectly assessed only through peak take-off power, while the complete mission energy and reserve requirement determine whether useful payload can be carried.

The main limitation of the battery-electric option is the mass of the battery system. Although electric propulsion provides high efficiency, fast response and simple power distribution, the available energy per unit mass remains a major constraint for regional missions [11]. Battery-electric aircraft may be suitable for shorter routes and lighter platforms, but the Novi Sad - Novi Beograd mission makes the specific-energy limitation clear.

In addition to the specific energy of the battery system itself, a significant challenge is the management of available energy throughout the different mission phases. In a VTOL aircraft, the energy-management system must provide sufficient power for vertical take-off and landing, maintain the required operational reserve, and ensure stable operation of the propulsion system during cruise flight. Consequently, modern research places considerable emphasis on energy storage, distribution, and management strategies for VTOL and Advanced Air Mobility (AAM) concepts [12].

Liquid-fuel propulsion offers better range potential because fuel has much higher specific energy. However, this advantage must be balanced against the mass of engines, cooling systems, tanks, transmissions, auxiliary equipment and structural reinforcements. In a VTOL aircraft, this additional mass is not negligible because it directly affects hover and take-off power.

Hybrid-electric propulsion is technically attractive because it combines high fuel energy density with electric power distribution. It can provide flexible rotor control and may handle short peak-power demands through a battery buffer. However, it introduces additional mass, losses and system complexity. Hybrid propulsion is justified only if the gains in range, redundancy or power distribution compensate for the added complexity [4].

In regional aviation, hybrid-electric propulsion is increasingly being considered as a transitional solution between conventional propulsion systems and fully electric configurations. Its principal advantage lies in the ability to combine the high specific energy of liquid fuels with the flexibility of electric power distribution, while the main challenges include increased system mass, greater complexity, and the need to optimize the operation of all subsystems [13].

Rotor diameter and total disk area also strongly influence the concept. Larger rotors reduce disk loading and induced power, but increase aircraft dimensions, structural demands and operational-space requirements. Smaller rotors produce a compact configuration but increase power demand and propulsion loading. Therefore, rotor geometry must be selected together with mission requirements and available operating infrastructure [14].

For the analyzed route, the best solution is not necessarily the one with the lowest energy consumption in one flight phase. The preferred solution is the one that achieves the best overall balance between range, payload, reserve, propulsion mass, safety and system complexity. This conclusion points toward a multi-parameter optimization rather than a single-point propulsion selection.

7. CONCLUSION

This paper presented an energy-balance calculation and propulsion comparison for a light passenger VTOL aircraft on the Novi Sad - Novi Beograd reference route. The analysis was based on a segmented mission model including vertical take-off, climb, transition, cruise, descent, vertical landing and operational reserve.

The numerical results show that the reference mission requires approximately 189.93 kWh of shaft energy for a total flight time of 53.25 min. The largest contributions to total energy consumption are cruise and operational reserve, while the vertical phases require the highest instantaneous power of approximately 265 kW. This confirms that VTOL aircraft must be assessed simultaneously by total energy and peak-power requirements.

The key engineering finding is that the reference mission is simultaneously energy-critical and power-critical. The vertical phases define the required peak shaft power, whereas cruise and reserve dominate the total mission energy. Therefore, a feasible propulsion architecture must satisfy both requirements at the same time: short-duration high-power capability and sufficient stored energy for the complete route with reserve.

The comparison of propulsion variants shows that battery-electric propulsion is favorable in simplicity and controllability but constrained by battery mass. Liquid-fuel propulsion provides a better available-energy-to-mass ratio but requires more complex mechanical integration. Hybrid-electric propulsion is an attractive compromise because it combines fuel energy density with flexible electric power distribution, although it adds mass and complexity.

The feasibility of a light passenger VTOL aircraft for a regional route depends on the coordinated balance of aircraft mass, rotor geometry, disk loading, propulsion architecture, useful payload and reserve policy. For the analyzed Novi Sad - Novi Beograd route, battery-electric propulsion is mass-critical, while fuel-based and hybrid-electric concepts offer more favorable energy potential but require more complex system integration. Before a detailed aircraft design can be proposed, the next step must include subsystem mass breakdown, propulsion efficiency validation and sensitivity analysis with respect to MTOM, rotor diameter and reserve duration.

ACKNOWLEDGEMENTS

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-34/2026-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project "Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2026" (No. 01-3609/1).

REFERENCES

[1] D.P. Raymer, “Aircraft Design: A Conceptual Approach, 6th ed.”, American Institute of Aeronautics and Astronautics, Reston, https://doi.org/10.2514/4.104909, (2018)

[2] J.G. Leishman, “Principles of Helicopter Aerodynamics, 2nd ed.”, Cambridge University Press, Cambridge, (2006)

[3] W. Johnson, “Rotorcraft Aeromechanics”, Cambridge University Press, Cambridge, (2013)

[4] F. Yu, J. Chen, P. Gao, Y. Kong, X. Sun, J. Wang, and X. Chen, "A Review of Hybrid-Electric Propulsion in Aviation: Modeling Methods, Energy Management Strategies, and Future Prospects", Aerospace, Vol. 12(10), p. 895, https://doi.org/10.3390/aerospace12100895, (2025)

[5] W. Johnson and C. Silva, "NASA concept vehicles and the engineering of advanced air mobility aircraft", The Aeronautical Journal, Vol. 126(1295), pp. 59-91, https://doi.org/10.1017/aer.2021.92, (2022)

[6] C.A. Snyder, “More / All Electric Vertical Take-Off and Landing (VTOL) Vehicle Sensitivities to Propulsion and Power Performance”, NASA Technical Reports Server (NTRS), Document ID 20205007968, Vertical Flight Society’s 76th Annual Forum & Technology Display, (2020)

[7] European Union Aviation Safety Agency, Special Condition for small-category VTOL aircraft, EASA, Cologne, (2019)

[8] W. Li, R. Cheng, H. Huang, A. Garg, and L. Gao, “Energy Consumption Modeling and Optimization of an eVTOL Aircraft: Integrating Weight, Motor, and Battery Dynamics”, Energy, Vol. 325, p. 136229, https://doi.org/10.1016/j.energy.2025.136229, (2025)

[9] Austro Engine, E4 Series - Jet Fuel Piston Engines, Diamond Aircraft, https://www.diamondaircraft.com/en/austro-engine/e4-series/overview/, accessed 14 May 2026

[10] BRP-Rotax, Rotax 912 iS Sport / 912 iSc Sport Technical Data, BRP-Rotax, https://www.flyrotax.com/products/912-is-sport-isc-sport, accessed 15 May 2026.

[11] D. Buvarp and J. Leijon, "Electric aircraft: a review of challenges and emerging technologies", Discover Applied Sciences, Vol. 8, p. 444, https://doi.org/10.1007/s42452-026-08550-z, (2026)

[12] A. Elmeligy, N. El-Din Safwat, M. S. El-Moursi, and R. Sabatini, “Energy Production, Storage and Management for Sustainable VTOL Aircraft and Advanced Air Mobility”, Energy Reports, Vol. 15, p. 109137, https://doi.org/10.1016/j.egyr.2026.109137, (2026)

[13] G.S.K. Aidam, R. Opoku, E.A. Adjei, F. Davis, D.K. Oppong, and A. Narh, “Hybrid-Electric Propulsion Retrofits in Regional Aviation – A Review”, Next Sustainability, Vol. 7, p. 100250, https://doi.org/10.1016/j.nxsust.2026.100250, (2026)

[14] S. Gudmundsson, General Aviation Aircraft Design: Applied Methods and Procedures, 2nd ed., Butterworth-Heinemann, Oxford, (2021).

SESSION A: EARTH MOVING AND TRANSPORTATION MACHINERY

Analysis and calculation of overhead conveyor for metal housings transportation

Spasoje Trifković[1*], Nebojša Radić[1], Miroslav Milutinović[1], Srđan Samardžić[1]

[1] University of East Sarajevo, Faculty of Mechanical Engineering, Sarajevo, Bosnia and Herzegovina

ARTICLE INFO

[*] **Correspondence:** spasoje.trifkovic@ues.rs.ba

DOI: https://doi.org/10.46793/ET26.A10T **PAGES:** A79 – A84

ABSTRACT

Overhead conveyors are used to increase productivity in manufacturing processes and warehouse centers. These are continuous transport devices that operate above or over the head of both workers and other equipment. Their advantage is that they can transfer the load in any direction in space. Steel ropes and chains are used as pulling elements. The pulling element is connected to the trolley where the carriers are attached. The loads from the cargo are transferred via the wheels to the carrier conveyor track. The carrier track is most often connected to the construction of the hall. In this paper, the analysis and calculation of the overhead chain conveyor for the transportation of metal housings was performed.

KEYWORDS

Transport devices, Continuous transport systems, Overhead conveyors.

1. INTRODUCTION

Conveyors are used for the continuous transport of piece and bulk materials, along a route with horizontal, inclined and vertical sections. The routes are closed, usually flat (for example, vertical or horizontal). When transporting piece goods, overhead conveyor routes can also be spatial, where they are an integral part of a production process [1]. There are also systems based on rails that are installed according to the desired transport paths and in which the vehicles can be maneuvered individually [2]. Most of the work in this area studies overhead conveyors [3]. Various constructions of overhead conveyors were considered in the papers [4,5,6]. Papers [7,8,9] show the advantages and disadvantages of Power & Free Conveyor and their systems. More recent work [10,...,13] focuses on special vehicle designs and other systems or performance enhancements. In this work, the calculation of the basic elements of overhead conveyors used in serial production, especially in the automotive industry was performed [14].

2. OVERHEAD CONVEYORS

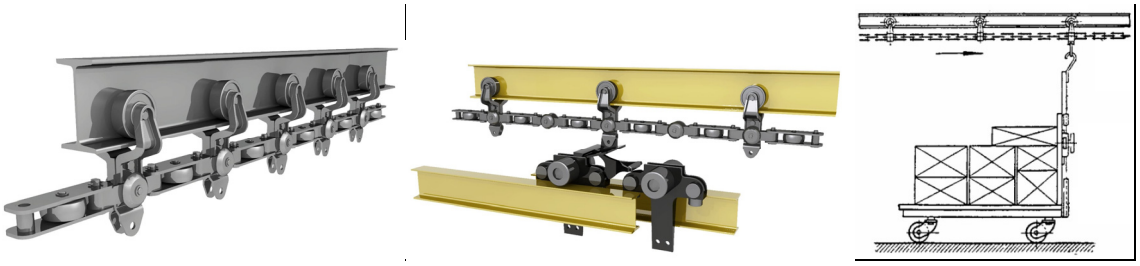
The Overhead conveyors, are typically used in manufacturing facilities to move parts from one location to another. They take advantage of the generally unused vertical space above the work area. According to the method of material transportation, they can be [1]:

- conveyors that transport load by carrying (Figure 1.a)
- conveyors that transport load by pushing (Figure 2.b)
- conveyors that transport load by pulling (Figure 3.c)

Common components of an overhead conveyor system include a chain or cable drive, either enclosed or open track, a drive unit, a take-up unit, trolleys, carriers, and supports. These components are specifically engineered to cater to the unique requirements of different applications and operations [15].

Overhead conveyor chains come in various designs, each tailored to handle unit loads or bulk materials. Choosing the appropriate chain for an overhead conveyor depends on the specific application requirements, as well as the chain's durability and reliability. Overhead conveyor chains are available in various shapes, sizes, and designs, each tailored to address the specific requirements of a transport application. The drive system in an overhead conveyor connects to the chain,

propelling it along the conveyor's track. There are two primary drive types: caterpillar and sprocket. The type of drive affects both the chain and trolley speed and the rate at which materials are delivered to their destination [15].



a) b) c)

Figure 1: Types of Overhead Conveyors

Overhead conveyor tracks are categorized into I-beam and enclosed types (Figure 2), selected based on the application's requirements and conveyor use. I-beam tracks are popular for their strength and ability to support heavy loads and trolley wheels. Enclosed tracks vary in design, with some featuring a tubular structure and others having a rectangular or square cross-section. Overhead conveyor tracks are constructed from robust steel to endure continuous operation.

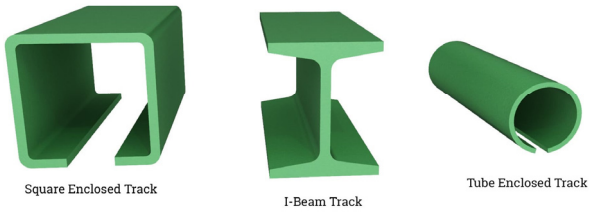


Figure 2: Types of Overhead Conveyor Tracks [15].

Trolleys, connected to a handy chain drive, glide along the track to help move heavy loads with ease. Each trolley comes with a spot to secure different kinds of load carriers. When the chain drive kicks in, it smoothly moves the trolleys along the track, carrying the load from the overhead conveyor. The trolleys' movement is managed by the chain and drive system, making it easy to start and stop whenever necessary. Plus, trolley designs are customized to fit the specific type of track they're linked to, including options for tube enclosed tracks, I-beam tracks, and other enclosed styles

Take-up units in overhead conveyors are very important for keeping the chain drive nice and taut, which helps everything run smoothly. These handy units come with a roller turn or traction wheel design and tackle any chain stretching that might happen due to temperature changes. Usually found near the drive, often at a low point in the system, take-ups are great at managing how the chain expands and contracts. You can find take-up units in different spots like roller bank turns, traction wheel assemblies, or even within spread track loops.

Pendants in an overhead conveyor system are components that connect to the chain and support the loads being transported. Constructed from durable materials, these pendants are designed to handle the weight and operational stresses of the conveyor system. They attach to load pins on the chain, enabling materials to hang below the chain itself. Pendants are compatible with various tracks, brackets, chains, and other conveyor parts. Spacing between pendants on an overhead conveyor chain differs based on the conveyor model and the type of materials being conveyed. To boost the load capacity, some items are supported by several pendants. Pendants come in various forms, including V-shaped, loading hooks, swivel fixtures, and double loading hooks, each suited for specific load needs [15].

3. CALCULATIONS

It is necessary to calculate an overhead conveyor for the transport of Metal Housings (Figure 3) weighing 40 *kg*. Other data are:

$b = 600\ \ mm\ ;\ h = 1000\ \ mm$

Drive element is welded chain $t_K = 80\ \ mm$ or detachable chain $t_K = 100\ \ mm$

The conveyor speed is $v = 0.15\ \ m/s$, one housing per pendants

The weight of the trolley 5 *kg*, weight of the pendant 15 *kg*

Overall mechanical efficiency $\eta_{mech} = 0,65$, total drag chain weight $q_K = 18\ \ kg/m$

1) What must be the minimum distance between the pendants, that there would be at least 200 *mm* of space between adjacent housings?
2) Which distance between the pendants and which pitch of the chain is obtained?

3) What are the parts and mass capacity?
4) What tractive force is needed (approximate and exact calculation)?
5) What is the maximum drive load?
6) What must be the tension force, so that the minimum force in the chain is $F_{K\min} = 0,80\ \ kN$

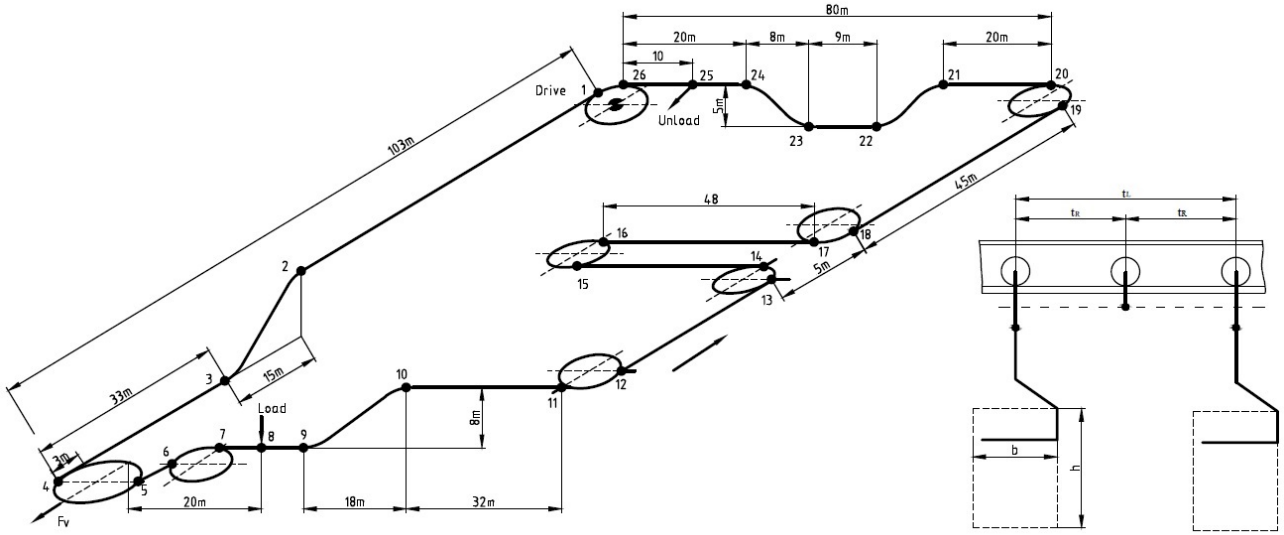


Figure 3: Overhead Conveyor for Metal Housings

Solution:

1) In order for the required condition to be met, the highest rise of the conveyor must be observed. It is between points 23 and 24 (Figure 4):

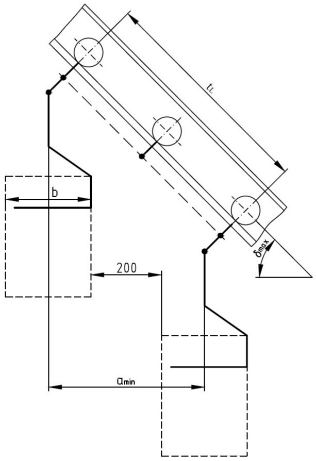


Figure 4: Determination of the minimum distance

$$\tan\delta_{\max} = \frac{5}{8} \Rightarrow \delta_{\max} = 32° \tag{1}$$

$$t_{L\min} = \frac{a_{\min}}{\cos\delta_{\max}} = \frac{b+200}{\cos 32°} \tag{2}$$

$$t_{L\min} = \frac{600+200}{0,848} = 943\ \ mm \tag{3}$$

2) Considering that: $t_L = 2 \cdot t_R$ and $t_R = (2,4,6,...) \cdot t_K$ follows:

for $t_K = 100\ \ mm \Rightarrow 6 \cdot t_K = t_R$ (4)

$t_R = 600\ \ mm \Rightarrow t_L = 1200\ \ mm$ (5)

for $t_K = 80\ \ mm \Rightarrow 6 \cdot t_K = t_R$ (6)

$t_R = 480\ \ mm \Rightarrow t_L = 960\ \ mm$ (7)

$$t_L = 960 \ ; \ t_K = 80 \ mm \qquad \text{(assumed)} \tag{8}$$

3)
$$Q_{St} = 3600 \cdot \frac{v}{t_L} \cdot z_1 = 3600 \frac{0,15}{0,96} \cdot 1 = 562 \ kom/h \tag{9}$$

$$Q_G = \frac{Q_{St} \cdot G_{St}}{1000} = \frac{562 \cdot 40}{1000} = 22,5 \ t/h \tag{10}$$

$$T = \frac{t_L}{z_1 \cdot v} = \frac{0,96}{1 \cdot 0,15} = 6,4 \ s \tag{11}$$

4)
$$f_{uk} \approx f \cdot \left(1 + f_F \cdot f_S^x \cdot f_{WK}^y \cdot f_{wR}^z\right) \tag{12}$$

Since most of the climbs are in the area $25° < \delta < 35°$, chosen $f_S = 1,02$.

$$f_{uk} \approx 0,025 \cdot \left(1 + 0.4 \cdot 1,02^8 \cdot 1.025^7 \cdot 1.03^3\right) \approx 0.4 \tag{13}$$

$$q = q_K + \frac{G_{Lw}}{t_R} + \frac{G_T}{t_L}$$

$$q = 18 + \frac{5}{0,48} + \frac{15}{0,96} = 44 \ kg/m \tag{14}$$

$$G_G = \frac{G_{St} \cdot z_1}{t_L} = \frac{40 \cdot 1}{0,96} = 41,7 \ kg/m \tag{15}$$

$$L = 2 \cdot 80 + 2 \cdot 100 + 2 \cdot 3 + 2 \cdot 48 = 462 \ m \tag{16}$$

$$L_T = L - (10 + 103 + 6 + 20) = 323 \ m \tag{18}$$

$$F_u = f_{uk} \left(L \cdot q + L_T \cdot G_G\right) \cdot g + G_G \cdot g \left(H_2 - H_1\right) \tag{19}$$

$$F_u = 0,04 \cdot \left(462 \cdot 44 + 323 \cdot 41,7\right) \cdot g + 41,7 \cdot g \cdot 8 = 16,90 \ N \tag{20}$$

$$P_E = \frac{F_u \cdot v}{\eta_u} = \frac{16,90 \cdot 0,15}{0,65} = 3,82 \ kW \tag{21}$$

The minimum force is at point 1 or 3:

$$f \cdot L_{1/3} = 0,025 \cdot 55 = 1,37 \ m, H_{1/3} = 8 \ m \tag{22}$$

$$f \cdot L_{1/3} < H_{1/3} \Rightarrow F_{K3} = F_{K\min} \tag{23}$$

$$F_{K3} = F_{K\min} = 0,80 \ kN \tag{24}$$

$$F_{K4} = F_{K3} + q \cdot g \cdot f \cdot L_{3/4}$$

$$F_{K4} = 0,80 + 44 \cdot g \cdot 0,025 \cdot 33 = 1,164 \ kN \tag{25}$$

$$F_{K5} = F_{K4} + f_{WK} = 1,164 \cdot 1,03 = 1,20 \ kN \tag{26}$$

$$F_{K6} = F_{K5} + q \cdot g \cdot f \cdot L_{5/6}$$

$$F_{K6} = 1,20 + 44 \cdot g \cdot 0,025 \cdot 3 = 1,233 \ kN \tag{27}$$

$$F_{K7} = F_{K6} + f_{WK} = 1,233 \cdot 1,025 = 1,265 \ kN \tag{28}$$

$$F_{K8} = F_{K7} + q \cdot g \cdot f \cdot L_{7/8}$$

$$F_{K8} = 1,265 + 44 \cdot g \cdot 0,025 \cdot 20 = 1,485 \ kN \tag{29}$$

$$F_{K9} = F_{K8} + \left(q + G_G\right) \cdot g \cdot f \cdot L_{8/9}$$

$$F_{K9} = 1,485 + 85,7 \cdot g \cdot 0,025 \cdot 10 = 1,70 \ kN \tag{30}$$

$$F_{K10} = \left[F_{K9} \cdot f_S + \left(q + G_G\right) \cdot g \cdot f \cdot L_{9/10} + \left(q + G_G\right) \cdot g \cdot H_{9/10}\right] \cdot f_S$$

$$\delta_{9/10} = 24° \Rightarrow f_S = 1,015 \tag{31}$$

$$F_{K10} = \left[1,70 \cdot 1,015 + 85,7 \cdot g \cdot 0,025 \cdot 18 + 85,7 \cdot g \cdot 8\right] \cdot 1,015 = 9,11 \ kN \tag{32}$$

$$F_{K11} = F_{K10} + \left(q + G_G\right) \cdot g \cdot f \cdot L_{10/11}$$

$$F_{K11} = 9{,}11 + 85{,}7 \cdot g \cdot 0{,}025 \cdot 32 = 9{,}80 \ \ kN \tag{33}$$

$$F_{K12} = F_{K11} \cdot f_{WK} = 9{,}80 \cdot 1{,}025 = 10{,}05 \ \ kN \tag{34}$$

$$F_{K13} = F_{K12} + \left(q + G_G\right) \cdot g \cdot f \cdot L_{12/13}$$

$$F_{K13} = 10{,}05 + 85{,}7 \cdot g \cdot 0{,}025 \cdot 50 = 11{,}125 \ \ kN \tag{35}$$

$$F_{K14} = F_{K13} \cdot f_{WK} = 11{,}125 \cdot 1{,}025 = 11{,}40 \ \ kN \tag{36}$$

$$F_{K15} = F_{K14} + \left(q + G_G\right) \cdot g \cdot f \cdot L_{14/15}$$

$$F_{K15} = 11{,}40 + 85{,}7 \cdot g \cdot 0{,}025 \cdot 48 = 12{,}43 \ \ kN \tag{37}$$

$$F_{K16} = F_{K15} \cdot f_{WK}^3 = 12{,}43 \cdot 1{,}03^3 = 13{,}55 \ \ kN \tag{38}$$

$$F_{K17} = F_{K16} + \left(q + G_G\right) \cdot g \cdot f \cdot L_{16/17}$$

$$F_{K17} = 13{,}55 + 85{,}7 \cdot g \cdot 0{,}025 \cdot 48 = 14{,}58 \ \ kN \tag{39}$$

$$F_{K18} = F_{K17} \cdot f_{WK} = 14{,}48 \cdot 1{,}025 = 14{,}95 \ \ kN \tag{40}$$

$$F_{K19} = F_{K18} + \left(q + G_G\right) \cdot g \cdot f \cdot L_{18/19}$$

$$F_{K19} = 14{,}95 + 85{,}7 \cdot g \cdot 0{,}025 \cdot 45 = 15{,}92 \ \ kN \tag{41}$$

$$F_{K20} = F_{K19} \cdot f_{WK} = 15{,}92 \cdot 1{,}025 = 16{,}32 \ \ kN \tag{42}$$

$$F_{K21} = F_{K20} + \left(q + G_G\right) \cdot g \cdot f \cdot L_{20/21}$$

$$F_{K21} = 16.32 + 85.7 \cdot g \cdot 0.025 \cdot 20 = 16.75 \ \ kN \tag{43}$$

$$F_{K22} = \left[F_{K21} \cdot f_S + \left(q + G_G\right) \cdot g \cdot f \cdot L_{21/22} - \left(q + G_G\right) \cdot g \cdot H_{21/22}\right] \cdot f_S$$

$$\delta_{21/22} = 29° \Rightarrow f_S = 1.020 \tag{44}$$

$$F_{K22} = \left[16.75 \cdot 1.02 + 85.7 \cdot g \cdot 0.025 \cdot 9 - 85.7 \cdot g \cdot 5\right] \cdot 1.02 = 13.24 \ \ kN \tag{45}$$

$$F_{K23} = F_{K22} + \left(q + G_G\right) \cdot g \cdot f \cdot L_{22/23}$$

$$F_{K23} = 13.24 + 85.7 \cdot g \cdot 0.025 \cdot 23 = 13.73 \ \ kN \tag{46}$$

$$F_{K24} = \left[F_{K23} \cdot f_S + \left(q + G_G\right) \cdot g \cdot f \cdot L_{23/24} - \left(q + G_G\right) \cdot g \cdot H_{23/24}\right] \cdot f_S$$

$$\delta_{23/24} = 32° \Rightarrow f_S = 1.020 \tag{47}$$

$$F_{K24} = \left[13.73 \cdot 1.02 + 85.7 \cdot g \cdot 0.025 \cdot 8 + 85.7 \cdot g \cdot 5\right] \cdot 1.02 = 18.84 \ \ kN \tag{48}$$

$$F_{K25} = F_{K24} + \left(q + G_G\right) \cdot g \cdot f \cdot L_{24/25}$$

$$F_{K25} = 18.84 + 85.7 \cdot g \cdot 0.025 \cdot 10 = 19.06 \ \ kN \tag{48}$$

$$F_{K26} = F_{K25} + q \cdot g \cdot f \cdot L_{25/26}$$

$$F_{K26} = 19.06 + 44 \cdot g \cdot 0.025 \cdot 10 = 19.17 \ \ kN \tag{50}$$

$$F_{K2} = \left(F_{K3} \cdot \frac{1}{f_S} - q \cdot g \cdot f \cdot L_{2/3} + q \cdot g \cdot H_{2/3}\right) \cdot \frac{1}{f_S}$$

$$\delta_{2/3} = 28° \Rightarrow f_S = 1.020 \tag{51}$$

$$F_{K2} = \left(\frac{0.80}{1.02} - 44 \cdot g \cdot 0.025 \cdot 15 + 44 \cdot g \cdot 8\right) \cdot \frac{1}{1.02} = 4.07 \ \ kN \tag{52}$$

$$F_{K1} = F_{K2} - q \cdot g \cdot f \cdot L_{1/2}$$

$$F_{K1} = 4.07 - 44 \cdot g \cdot 0.025 \cdot 55 = 3.46 \ \ kN \tag{53}$$

$$F_{WA} = 0{,}04\left(F_{K1} + F_{K26}\right) = 0.04\left(3.46 + 19.17\right) = 0.90 \ \ kN \tag{54}$$

$$F_O = F_{K26} - F_{K1} + F_{WA} = 19.17 - 3.46 + 0.90 = 16.61 \ \ kN \tag{55}$$

$$P_E = \frac{F_O \cdot v}{\eta_u} = \frac{16.61 \cdot 0.15}{0.65} = 3.76 \ \ kW \tag{56}$$

5) The maximum driive load occurs when the upper take-up unit passes under load. In this case the upper take-up unit is from 23 to 24. Since almost maximum chain pull occurs here, the load will be set to 100% of the allowed value. Thus, according to equation [1]., a required radius of curvature of $R = 5 \ \ m$.

$$F_{L\max} = G_E \cdot g + G_L \cdot g + F_{K24} \cdot \frac{t_R}{R}$$

$$F_{L\max} = (5+15) \cdot g + 40 \cdot g + 18.84 \cdot \frac{0.48}{5} \approx 2.50 \ \ kN \tag{57}$$

6) $$F_P = F_{K4} + F_{K5} = 1.164 + 1.20 \approx 2.37 \ \ kN \tag{58}$$

4. CONCLUSION

This paper aims to illustrate the necessary calculations for overhead conveyor systems used for piece goods. In this case, the system is designed for metal housings. For the given conveyor parameters, the minimum distance between the carriers and the capacity ware determined. The calculation of the traction force is performed by an approximate and exact method (the contour method). The difference between these forces is less than 2%, which means that the approximate method gave quite good results. Based on the results obtained, we can continue the further calculation of the conveyor and choose its elements.

REFERENCES

[1] M. Gašić, “Transportni uređaji, neprekidni transport”, Faculty of Mechanical Engineering Kraljevo, Kraljevo (Serbia), (2010)

[2] L. Schlerholz, “Elektrohängebahn. System 624 / 634”, TU Dortmund University Logistics Department, Louis Schierholz GmbH, (2014)

[3] VDI 3643, “Elektrohängebahn; Anforderungsprofil an ein kompatibles System”, VDI-Verlag, Düsseldorf, Germany, (1998)

[4] M. Broggi, “Sonderheft Lagertechnik: Fahrerlos vom Lager bis in die Produktion”, Fördermittel Journal, pp. 80-85, (1987)

[5] D. Pfeifer, “Flexible Automatisierung im Automobilbau, Beispiel: Omega-Fertigung”, Fachtagung Praxis der Fertigungsautomatisierung, Bad Soden, Germany, (1987)

[6] D. Kappeler, “Sonderausgabe Fördertechnik: Die Endmontage ist nicht das Schlusslicht,” Fördermittel Journal, pp. 48-53, (1990)

[7] H. Fraissl, “Elektrohängebahn: Schneller als Power+Free”, Materialfluss, pp. 42-44, (1985)

[8] F. Eggenstein, “Elektrohängebahnen, Technik, und Entwicklung,” Fördern und Heben, (1982)

[9] H. Giessler, “Ein System wird zum Standard,” Logistik im Unternehmen, (1987)

[10] M. Kunzmann, “Ein konsequent wirtschaftliches Containertransportsystem”, Transmatik 85. Kongress, Karlsruhe, Germany, (1985)

[11] J. Schieleit, “ Mobile Roboter in Elektrohängebahnsystemen“, Verlag TÜV Rheinland, Köln, Germany, (1990)

[12] J. Krug, “Automatiklager für fertigungsgerechte Flexibilität, -verketteter Materialfluß für Druckzylinder”, Maschine und Werkzeug, pp. 28-34, (1989)

[13] G. Uttscheid, “Elektrohängebahn”, European Patent Office EP 0 567 122 A1, April 22, (1993)

[14] G. Böttner, “Technische Berichte: Förderanlagen für die Automobilindustrie”, Thyssen MAN, Neuhausen, Germany, (1986)

[15] https://www.iqsdirectory.com/articles/conveyors/overhead-conveyors.html

SESSION A: EARTH MOVING AND TRANSPORTATION MACHINERY

Analysis of zipline systems with large spans and low inclination angles

Radomir Đokić[1*], Jovan Vladić[1], Tanasije Jojić[1]

[1] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia

ARTICLE INFO

[*] **Correspondence:** djokic@uns.ac.rs

DOI: https://doi.org/10.46793/ET26.A11DJ **PAGES:** A85 – A93

ABSTRACT

This paper presents an overview of the engineering challenges associated with long-span zipline systems characterized by low inclination angles. In such installations, particularly when combined with shallow terrain profiles, wind and ice loads can cause significant rope sagging and lateral displacements. Consequently, it is necessary to verify the required safety clearances between the rope and the surrounding terrain, as well as any obstacles located along the route. Since there are currently no dedicated standards specifically governing zipline systems, the assessment was performed in accordance with standards applicable to cableways. Accordingly, relevant excerpts from these standards are presented and discussed. A numerical verification was then carried out using the example of a zipline constructed on Fruška Gora. In addition, the paper provides an overview of the wire ropes and trolleys commonly used in ziplines of this type.

KEYWORDS

Zipline, Large Span, Safety Clearance, Wind Load, Ice Load

1. INTRODUCTION

Although the zipline, which was sporadically used in the past, is a simple type of transport that can be used to bridge canyons and basins, the expansion of its usage and construction is associated with the end of the twentieth century, primarily due to the adrenaline experience of the user. At that time, a larger number of ziplines, which primarily served the purpose of leisure, began to be built, and the places where they were built, were not strictly defined. They appear in hilly areas, as well as in and within large economic facilities such as dams and all the way to the city centres. Lately, there is a noticeable "race" among builders who will build longer ziplines, and that's how *Jebel Jais Flight* (2832 m), *Volo dell'angelo* (2213 m), *The Eye of the Jaguar* (2130 m) and *Crimean Hawk* (2022 m) were made, [1]. However, there are still no appropriate regulations and standards for designing a device of that type. In addition, there is a relatively small number of scientific papers dealing with load determination in ziplines.

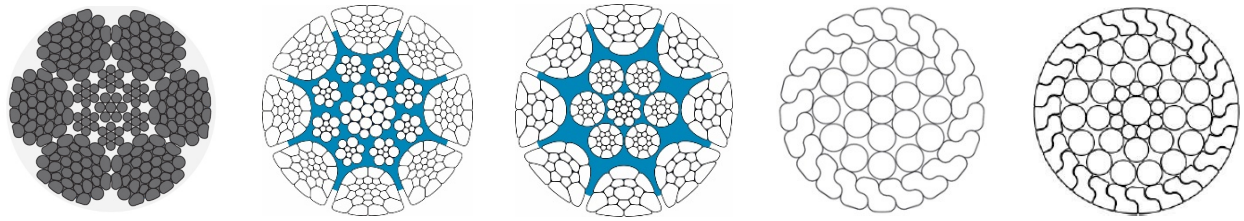
Authors of the paper "Traveling along a zipline" [2], analyzed a heavy rope. Using geometric relationships, they determined the rope configuration for different positions of a concentrated load (the rider). Furthermore, they presented the trajectory curve for different mass ratios between the rider and the rope. However, all considerations are limited to the static case.

On the other hand, authors of the paper "Structural safety evaluation of zip line due to dynamic impact load" [3], use the finite element method to analyze the stress in the rope of an existing zipline with a span of 280 m. The steel rope has a diameter of 12 mm, and a load of 150 kg moves along it at a constant speed of 54 km/h. As a result, they obtain stress values in the rope that remain within the allowable limits for the selected rope type.

Paper "Dynamic computation and design of a world record death-ride" [4], describes the construction of a temporary zipline with a span of 1163 m in Belgium. The rope is discretized using 16 finite elements. The analysis is carried out in two stages. In the first stage, the rope is considered under self-weight and tension force, and its equilibrium configuration is determined. In the second stage, external loads are applied to the previously obtained rope configuration. In this study, the external loads include the person's weight and aerodynamic drag forces. The authors place particular emphasis on the selection of an appropriate tension force, since it directly influences the natural frequency of cable oscillations and, consequently, the comfort and perceived quality of the ride.

2. ANALYSIS AND OVERVIEW OF STEEL ROPES USED IN ZIPLINE SYSTEMS

Standards for ziplines do not prescribe a mandatory steel rope construction for application in zipline systems. In short-span ziplines, general-purpose wire ropes (e.g. 6x7, 6x19, etc.) are commonly encountered; however, in long-span ziplines, there is a tendency to use locked-coil (closed) rope constructions. Locked-coil rope constructions are characterized by extremely high bending stiffness and relatively high self-weight, which necessitates the application of significant pretensioning forces. This, in turn, affects the required structural capacity of both the starting and terminal stations. As an intermediate solution, the use of compacted wire ropes has emerged. These ropes are characterized by a relatively smooth outer surface with reduced grooving, while at the same time being significantly more flexible and considerably lighter than locked-coil rope constructions, [5]. Examples of rope constructions used in ziplines are shown in Fig. 1.



a b c d e

Figure 1: Rope constructions used for ziplines

Figures 1a, 1b and 1c show compacted wire ropes, whereas the Figs. 1d and 1e show locked spiral ropes.

3. ANALYSIS AND OVERVIEW OF TROLLEYS USED IN ZIPLINE SYSTEMS

Movement of persons along the rope is enabled by wheeled trolleys to which safety harnesses are attached. In practice, different types of wheels are used depending on the field inclination angle and applied loads. Figure 2a-c shows common types of wheels available in the product range of various manufacturers. The use of wheels made of plastic materials, or steel wheels with plastic or rubber inserts, is advantageous from the perspective of reducing the contact pressure on the steel rope, thereby positively influencing its service life. The use of such trolleys is common in shorter ziplines with steeper gradients.



a b c d

Figure 2: Types of trolleys

For this specific zipline system, specially designed trolley was developed and manufactured (Fig. 2d), featuring significantly larger wheel diameters in order to reduce the influence of bearing resistance.

In ziplines with a low field inclination angle, it often occurs that persons, under certain conditions, do not reach the end of the route. Among other factors, the riding position (Fig. 3) has a significant influence. In ziplines with a low inclination angle, even a prone position is not always sufficient to ensure reaching the end of the track. For example, simply extending the arms - an action commonly performed by users experiencing increased adrenaline - significantly increases the exposed surface area to airflow and wind. This leads to a substantial increase in aerodynamic drag and, consequently, to stopping before reaching the end of the route.



Figure 3: Various riding positions

If conventional trolleys were used in such ziplines, frequent manual retrieval would be required, which is time-consuming. Therefore, in such cases motorized trolleys with electric drive systems powered by batteries are used. Examples of such trolley systems from various international manufacturers are shown in Fig 4.

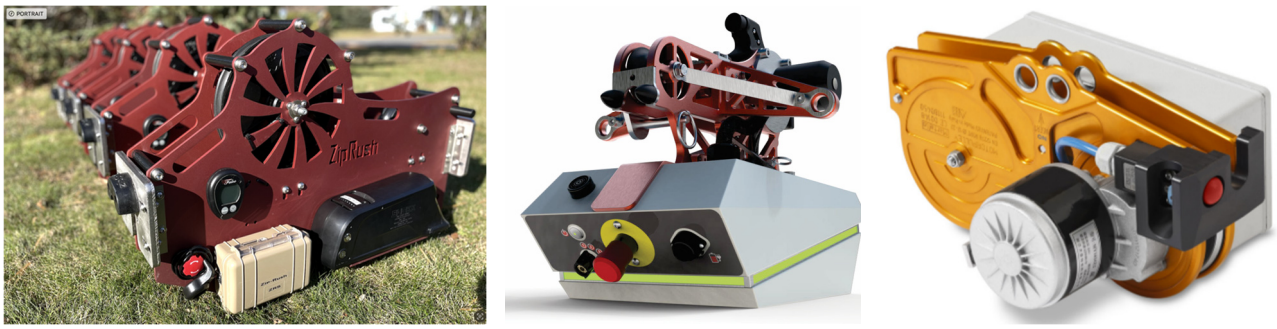


Figure 4: Motorized trolleys

4. DETERMINATION OF STEEL ROPE SAFETY CLEARANCES DURING OPERATION AND OFF-SEASON CONDITIONS

In order to determine the required safety clearances between the steel rope and the surrounding terrain, it is necessary to know both the terrain configuration and the loads acting on the rope. Defining the loads for zipline systems is particularly challenging due to the absence of regulations that directly address this type of installation. The standards EN 15567-1:2015+A1 and EN 15567-2:2015 [6-7] are not applicable, although they mention ziplines; however, this refers only to short installations within adventure parks. Consequently, the assessment of long-span zipline systems must be based on a combination of other relevant regulations. The most comparable installations to which zipline designers may refer are cableway systems. However, the actual operating conditions differ significantly in certain aspects, and strict application of cableway standards often leads to inappropriate design parameters for ziplines. The most significant differences are outlined below:

- Cableways are typically designed for operation in winter conditions in high-altitude regions with significantly low temperatures and strong winds, whereas ziplines are used under favorable meteorological conditions and are taken out of service during the winter season.
- Cableways are designed for the simultaneous transport of a large number of persons (high capacity), whereas ziplines are used for individual transport of persons; therefore, the safety requirements and operating conditions are significantly different.
- Ziplines are often located in relatively sheltered areas (urban environments or between tree canopies), whereas cableways are generally more exposed to wind action, as at least part of their route is typically located in open terrain (“clear areas”).

In the following sections, an overview and application of the regulations defined in the cableway standards EN 12930:2015 and EN 12929-1:2015+A1 will be presented [8-9], while section 4.3 will provide a more detailed determination of the safety clearances themselves. The determination of safety clearances will be demonstrated on the example of a specific zipline constructed on Fruška Gora, considering the case of installing a steel rope “Fatzer Integra” with a diameter of 16 mm, whose cross-section is shown in Fig. 1e. The terrain configuration of the specific zipline is shown in Fig. 5.

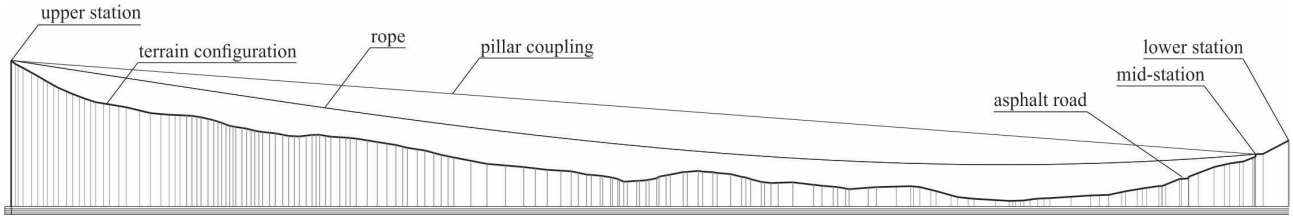


Figure 5: Terrain configuration

4.1. Loads According to Standard EN 12930:2015 (E)

4.1.1. Actions from Wind

The resultant wind force shall generally be calculated in accordance with EN 1991-1-4 on the basis of formula (1):

$$F_w = q \cdot c_f \cdot A_{ref} \tag{1}$$

where:

q - dynamic pressure;

c_f - force coefficient;
A_{ref}- reference area for c_f

The values specified in the EN 1991-1-4 and in the national construction standards can differ in accordance with cableway-specific circumstances. Provided no other regulations have been specified in EN 1709, EN 1908, EN 1909, EN 12397, EN 12408, EN 12927 (all parts), EN 12929 (all parts), EN13107, EN 13223, EN 13243 and EN 13796 (all parts), the following dynamic pressure values shall apply:

a) in operation, a minimum value for dynamic pressure q = 0.25 kN/m^2 shall be used. For specific cableways a higher value may be necessary.
b) out of operation, the minimum value for dynamic pressure is q = 1.20 kN/m^2, provided that particular local installation circumstances of the cableway or national cableway-specific provisions or directives do not stipulate other values.
c) the dynamic pressure acts as a uniformly distributed load over the total slope length /*. Out of operation, the dynamic pressure may be reduced in accordance with equation (2).

$$q_{red} = \beta \cdot q \tag{2}$$

where:

q_{red} - reduced dynamic pressure acting on the slope length l^* of the span;
β - reduction coefficient;
q - dynamic pressure.

The reduction coefficient β is:

1) 1.00 at l^* = 0 m;
2) 0.65 at l^* = 600 m;
3) 0.50 at $l^* \geq$ 2 000 m.

With a slope field length l^* between 0 m and 600 m or between 600 m and 2 000 m, β shall be established by linear interpolation.

For slope field length l^* = 1430 m, the reduction coefficient amounts to:

$$\beta = 0.5 + \left(0.65 - 0.5\right) \cdot \frac{2000 - 1430}{2000 - 600} = 0.56$$

so the reduced dynamic pressure amounts to:

$$q_{red} = 0.56 \cdot 1200 = 672 \text{ N/m}^2$$

4.1.2. Actions from Ice

An ice load on the rope shall be taken into account if the ice curtain cannot be removed in good time by moving or driving the rope in another manner. Where an expert opinion establishes no other ice load on the rope, an ice sheet with a density of 600 kg/m^3 and an ice thickness of:

- 20 mm for ropes with a nominal rope diameter ≤ 10 mm;
- 25 mm for ropes with a nominal rope diameter ≥ 100 mm;

shall be assumed.

With a nominal rope diameter between 10 mm (corresponding to an ice load of 11.1 N/m) and 100 mm (corresponding to an ice load of 57.8 N/m), the ice load shall be established by means of linear interpolation. If required by local climatic conditions, a greater ice thickness or greater density shall be assumed. In the case of favourable climatic conditions, lower values may also be taken.

For a Ø16 mm rope, the ice load amounts to q_l = 14.2 N/m. This load corresponds to an ice accretion thickness of δ_l = 20.3 mm, yielding an overall iced rope diameter of $d_l = d + 2\times\delta_l$ = 56.6 mm. Based on the previously determined values, the total distributed load of the iced Ø16 mm rope is:

$$q_z^u = q + q_l = 14.8 + 14.2 = 29 \text{ N/m}$$

where:

q = 14.8 N/m - weight of the bare rope,

q_l = 14.2 N/m – weight of ice.

Based on the previously defined values, the wind load can be determined:

In operation:

$$q_v = q \cdot c_f \cdot A_u = 250 \cdot 1.2 \cdot 0.016 = 4.8 \text{ N/m}$$

For reduced dynamic pressure (out of service):

$q_v^u = q_{red} \cdot c_f \cdot A_u = 670 \cdot 1.2 \cdot 0.016 = 12.9$ N/m - wind load acting on the bare rope,

$q_v^l = q_{red} \cdot c_f \cdot A_l = 670 \cdot 1.2 \cdot 0.057 = 45.8$ N/m - wind load acting on the iced rope.

4.1.3. Combined Actions from Wind and Ice

The following cases with simultaneous action from wind and ice shall be assumed:

a) in operation:
   1) wind force in accordance with 2.1, ice load at 40 % of the value in accordance with 2.2, and;
   2) wind force at 80 % of the value in accordance with 2.1, ice load in accordance with 2.2;
b) out of operation:
   1) wind force at 65 % of the value in accordance with 2.1, ice load at 40 % of the value in accordance with 2.2, and;
   2) wind force at 40 % of the value in accordance with 2.1, ice load in accordance with 2.2.

Zipline systems are not designed to operate under icing conditions; therefore, the analysis/check is carried out only for case b) as follows:

1) wind force at 65 % and ice load at 40 %

$q_k^h = 0,65 \cdot 45.8 = 29.8$ N/m - horizontal wind load,

$q_k^v = q_u + 0.4 \cdot q_l = 14.8 + 0.4 \cdot 14.2 = 20.5$ N/m - vertical load.

2) wind force at 40% and full ice load

$q_k^h = 0.4 \cdot q_v^l = 0.4 \cdot 45.8 = 18.3$ N/m - horizontal wind load,

$q_k^v = q_u + q_l = 14.8 + 14.2 = 29$ N/m - vertical load.

Thus, the resultant loads are:

1) wind force at 65 % and ice load at 40 %

$$q_r = \sqrt{(q_k^v)^2 + (q_k^h)^2} = \sqrt{20.5^2 + 29.8^2} = 36.2 \text{ N/m}$$

$$\text{tg}\alpha_r = \frac{q_k^h}{q_k^v} = \frac{29.8}{20.5} = 1.45 \rightarrow \alpha_r = \text{arctg}(1.45) = 55.5°$$

2) wind force at 40% and full ice load

$$q_r = \sqrt{(q_k^v)^2 + (q_k^h)^2} = \sqrt{18.3^2 + 29^2} = 34.3 \text{ N/m}$$

$$\text{tg}\alpha_r = \frac{q_k^h}{q_k^v} = \frac{18.3}{29} = 0.63 \rightarrow \alpha_r = 32.2°$$

4.2. Loads According to Standard EN 12929-1:2015+A1:2022 (E)

4.2.1. Actions from Wind

The lateral deviation of the ropes as a result of the wind is to be calculated with a reduced dynamic pressure *q'*:

$$q' = q \cdot \left(\frac{l^{*\prime}}{l^*} \right)^2 \tag{3}$$

where

q - minimum dynamic pressure

- in operation: $q = 0.20$ kN/m^2;
- out of operation: $q = 1.20$ kN/m^2;

l^* - slope length of span;
$l^{*\prime}$ - national field length.

The dynamic pressure q' acts as a uniformly distributed load over the total slope length of span l^*. The notional field length $l^{*\prime}$ is calculated from the slope field length l^* by means of the following equation:

$$l^{*\prime} = l^* \cdot \delta \tag{4}$$

where δ is the reduction coefficient and amounts to:

- 1.00 at $l^* \leq 200$ m;
- 0.65 at $l^* = 900$ m;
- 0.50 at $l^* \geq 2\,000$ m.

With a slope field length l^* between 200 m and 900 m or between 900 m and 2 000 m, δ shall be established by linear interpolation. When establishing the lateral deviation, the increase in the lateral windage areas of the ropes as a result of an ice curtain shall be taken into consideration. Where an expert opinion establishes no other values, an ice sheet with an ice thickness of:

- 20 mm for ropes with a nominal rope diameter ≤ 10 mm;
- 25 mm for ropes with a nominal rope diameter ≥ 100 mm;

shall be assumed.

For the given parameter values, the reduction factor is obtained by linear interpolation:

$$\delta = 0.5 + (0.65 - 0.5) \cdot \frac{2000 - 1430}{2000 - 900} = 0.58$$

the lateral deviation (deflection) of the rope due to wind action should be calculated using the reduced dynamic pressure:

$$q' = 1200 \cdot \left(\frac{1430 \cdot 0.58}{1430} \right)^2 = 403.7 \text{ N/m}^2$$

4.2.2. Actions from Ice

With a nominal rope diameter between 10 mm and 100 mm, the ice thickness shall be established by linear interpolation. If required by local climatic conditions, a greater ice thickness shall be assumed. In the case of favourable climatic conditions, lower values may also be applied.

4.2.3. Combined Actions from Wind and Ice

To establish the lateral deviation of the ropes, the following cases shall be assumed:

a) in operation
   1) 100 % dynamic pressure without ice sheet;
   2) 100 % dynamic pressure at 40 % of the ice thickness;
   3) 80 % of the dynamic pressure at 100 % of the ice sheet;
b) out of operation
   1) 100 % dynamic pressure without ice sheet;
   2) 65 % dynamic pressure at 40 % of the ice thickness;
   3) 40 % of the dynamic pressure at 100 % of the ice sheet.

According to calculations for the simultaneous action of wind and ice, the most unfavorable resultant load is determined as follows:

$q_v^u = q_{red} \cdot c_f \cdot A_l = 403.7 \cdot 1.2 \cdot 0.016 = 7.8$ N/m - - wind load acting on the bare rope,

$q_v^l = q_{red} \cdot c_f \cdot A_l = 403.7 \cdot 1.2 \cdot 0.057 = 27.6$ N/m - wind load acting on the iced rope.

For the simultaneous action of ice and wind, the following cases shall be considered:

a) in operation: usage of zipline is not foreseen under icing conditions.
b) out of operation

1) 100 % dynamic pressure without ice sheet

$q_k^h = 7.8$ N/m - horizontal wind load,

$q_k^v = q_u = 14.8$ N/m - vertical load,

thus, the resultant loads are:

$$q_r = \sqrt{(q_k^v)^2 + (q_k^h)^2} = \sqrt{7.8^2 + 14.8^2} = 16.7 \text{ N/m}$$

$$\text{tg}\alpha_r = \frac{q_k^h}{q_k^v} = \frac{7.8}{14.8} = 0.53 \rightarrow \alpha_r = \text{arctg}(0.53) = 27.8°$$

2) 65 % dynamic pressure at 40 % of the ice thickness

$q_k^h = 0.65 \cdot 27.6 = 17.9$ N/m - horizontal wind load,

$q_k^v = q_u + 0.4 \cdot q_l = 14.8 + 0.4 \cdot 14.2 = 20.5$ N/m - vertical load,

thus, the resultant loads are:

$$q_r = \sqrt{(q_k^v)^2 + (q_k^h)^2} = \sqrt{20.5^2 + 17.9^2} = 27.2 \text{ N/m}$$

$$\text{tg}\alpha_r = \frac{q_k^h}{q_k^v} = \frac{17.9}{20.5} = 0.87 \rightarrow \alpha_r = \text{arctg}(0.87) = 41.1°$$

3) 40 % of the dynamic pressure at 100 % of the ice sheet

$q_k^h = 0.40 \cdot 27.6 = 11$ N/m - horizontal wind load,

$q_k^v = q_u + q_l = 14.8 + 14.2 = 29$ N/m - vertical load,

thus, the resultant loads are:

$$q_r = \sqrt{(q_k^v)^2 + (q_k^h)^2} = \sqrt{29^2 + 11^2} = 31.1 \text{ N/m}$$

$$\text{tg}\alpha_r = \frac{q_k^h}{q_k^v} = \frac{11}{29} = 0.38 \rightarrow \alpha_r = \text{arctg}(0.38) = 20.8°$$

4.3. Determination of the Safety Clearances

In the present case, the safety clearances will be verified at five characteristic points along the zipline route, shown in Fig. 6. For points T1-T4, the safety clearances during operation shall be greater than 1.5 m, as follows:

- For the operational condition of the zipline, considering a person length of 2 m, a person mass of 140 kg, and a snow cover thickness of 0.5 m, the calculated sag values shall be increased by 10%.
- During periods when the zipline is out of service (off-season conditions), no safety clearance is required.
- The clearance above the road shall be greater than 6 m at all times.

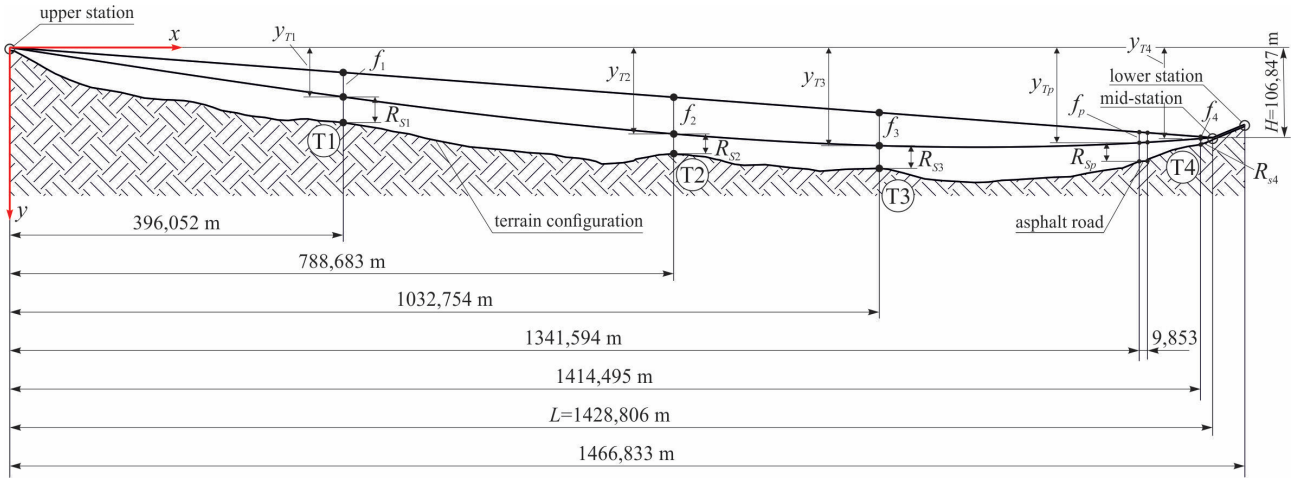


Figure 6: Terrain profile with safety clearances

The terrain parameters for the characteristic points were obtained from the topographic survey shown in Fig. 5.

The value of the safety clearance, (R_S), is determined as the difference between the vertical distance from the horizontal axis (x-axis) to the terrain elevation and the value of (y), where the value of (y) is calculated as follows, [10]:

$$y = x \cdot tg\beta + \frac{x \cdot (l - x)}{2 \cdot H} \cdot \left(\frac{q_r}{\cos\beta} + 2\frac{Q}{l} \right) \quad (5)$$

where:

y - vertical distance from the horizontal x-axis,
x - horizontal distance from the upper support,
β - field inclination angle,
l - span length,
H - horizontal component of the rope force,
q_r - self-weight of the rope including ice and wind loads,
Q - person load.

The horizontal component of the rope force (H), for the case of a rope tensioned by a weight, is determined from the following system of equations:

$$H - \left[G - \left(f_{\max} - \frac{h}{2} \right) \cdot q_r \right] \cdot \cos\beta = 0 \tag{6}$$

$$f_{\max} = \frac{l^2}{8H} \cdot \left(\frac{q_r}{\cos\beta} + \frac{2Q}{l} \right) \tag{7}$$

where G is weight of the tensioning weight and f_{max} is maximum rope sag.

In the case of a rope anchored at both ends, the value of the horizontal component (H) is determined from the following system of equations:

$$H - \left[S_A - \left(\frac{h}{2} + f_{\max} \right) \cdot q_r \right] \cdot \cos\beta = 0 \tag{8}$$

$$f_{\max} = \frac{l^2}{8H} \cdot \left(\frac{q_r}{\cos\beta} + \frac{2Q}{l} \right) \tag{9}$$

where S_A represents the maximum force at the upper support, which is typically determined as follows:

$$S_A = \frac{S_{kid}}{\nu} \tag{10}$$

where v is safety factor and S_{kid} is breaking load.

By substituting the values for the example zipline on Fruška Gora, the values of the safety clearances are obtained and are given in Table 1, while Table 2 presents the values for the descent of a person with the maximum design mass (140 kg) under operational conditions.

Table 1: Safety clearances (R_S [m]) for iced rope

v	T_1	T_2	T_3	T_p	T_4
3.15	-13.535	-23.136	-6.820	13.831	5.751
2.25	7.053	2.276	13.768	19.720	6.770

Based on the obtained results, it can be concluded that, for a rope safety factor of v = 3,15, in icing conditions according to the cableway standard EN 12930:2015, the rope would penetrate into the ground at Point 1 by 13.5 m, at Point 2 by 23.1 m, and at Point 3 by 6.8 m, which is unacceptable. Since this standard allows, for iced rope conditions, a reduction of the safety factor to v = 2,25, the minimum safety clearance between the rope and the ground is obtained at Point 2, where it amounts to 2.3 m. This ensures sufficient clearance even in the presence of snow accumulation not exceeding values typical for the Fruška Gora area (0.2÷0.3 m).

Regarding the safety clearance during the movement of a person with a trolley (Q=140 kg), the obtained values are significantly higher than the prescribed requirements. For example, above the road, considering a rider height including the safety harness of 2 m, the clearance amounts to 19.3 m.

Table 2: Safety clearances (R_S [m]) for operational conditions (Q=140 kg)

T_1	T_2	T_3	T_p	T_4
12.462	8.952	19.177	21.268	7.038

For the assessment of extreme conditions involving the simultaneous action of wind and ice, the lateral deflections of the rope will be determined in accordance with EN 12929-1:2015+A1. The maximum lateral deflection of the bare rope under wind load amounts to:

$$b_u = f_{\max} \cdot \sin\alpha_r = 52.7 \cdot \sin 27.8° = 26.7 \text{ m}$$

while the maximum lateral deflection of the iced rope under wind load amounts to:

$$b_u = f_{\max} \cdot \sin \alpha_r = 85.9 \cdot \sin 41.1° = 56.5 \text{ m}$$

5. CONCLUSIONS

Although at first aspect it seems that any „attractive“ location can be used for the installation of a zipline, it often turns out to be extremely unfavorable from a technical point of view, mainly due to:

- ensuring sufficient inclination angles, which is often the case with ziplines of long lengths (leading to frequent or constant stay of people on the zipline),
- very „shallow“ terrain of the route (problem of ensuring safety distances),
- a significant problem arises due to climatic conditions and meeting the „harsh conditions“ for ice and wind loads when the zipline is not in use (out of season), because most often the method of tensioning with a weight with an extremely large stroke must be applied, which is why it is necessary to create a tensioning station with a large depth of the „well“ for the movement of the weight.

Even more unfavorable conditions are the limit values of lateral deflections of the rope due to the effect of wind, according to the EN standard 12929-1:2015+A1, which can have large values for long ziplines. This paper shows that at the mid-span of the considered zipline, the lateral deflections are 26.7 m for a steel rope without ice and as much as 56.5 m for an iced steel rope. As already mentioned, in ziplines with large spans and small height differences between the starting and ending stations, the problem of persons remaining on the line arises, even up to several hundred meters. In such situations, in the design phase, it is always a good idea to propose the use of motorized trolleys as an option (two pieces are enough for one zipline line), which enable:

- safe arrival of persons to the exit station,
- return of the trolley with safety belts to the starting station without the need to transport the trolley and the equipment by special means of transport.

In cases where designers have favorable conditions from the point of view of the zipline route (large inclination angles), then problems arise with high incoming velocities of persons. Then the use of brakes is necessary, which requires a separate analysis that is not an integral part of this work.

ACKNOWLEDGEMENTS

This research has been supported by the Ministry of Science, Technological Development and Innovation (Contract No. 451-03-34/2026-03/200156) and the Faculty of Technical Sciences, University of Novi Sad through project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2026” (No. 01-3609/1).

REFERENCES

[1] T. Jojić, J. Vladić, and R. Đokić, “Zipline mathematical model forming, experimental verification and analysis of influential parameters”, Archive of Applied Mechanics, Vol. 93(11), pp. 4157-4180, https://doi.org/10.1007/s00419-023-02487-0, (2023)

[2] C. E.Mungan, and T. C Lipscombe, “Traveling along a zipline”, Latin-American Journal of Physics Education, Vol. 5(1), pp. 5-8, (2011)

[3] B. S. Ju, J. S. Kim, and W. Y. Jung, “Structural Safety Evaluation of Zip-Line Due to Dynamic Impact Load”, International Journal of Civil, Environmental, Structural, Construction and Architectural Engineering, Vol. 7(10), pp. 783-785, (2013)

[4] V. Denoël, T. Canor, and J. Demanet, “Dynamic Computation and Design of a World Record Death-Ride”, Proceedings of the 4th ECCOMAS Thematic Conference on Computational Methods in Structural Dynamics and Earthquake Engineering, Kos Island (Greece), 12-14 June 2013, pp. 3289-3299.

[5] J. Vladić, R. Đokić, and T. Jojić, “Elaborate: Analysis and Equipment Selection for Zipline System in Vrdnik”, Faculty of Technical Sciences, Novi Sad, (2026)

[6] EN 15567-1:2015+A1 - Sports and recreational facilities - Ropes courses - Part 1: Construction and safety requirements, European Committee for Standardization, Brussels

[7] EN 15567-2 - Sports- and recreational facilities - Ropes courses - Part 2: Operation requirements, European Committee for Standardization, Brussels

[8] EN 12930:2015 - Safety requirements for cableway installations designed to carry persons - Calculations, European Committee for Standardization, Brussels

[9] EN 12929-1:2015+A1 - Safety requirements for cableway installations designed to carry persons - General requirements - Part 1: Requirements for all installations, European Committee for Standardization, Brussels

[10] J. Vladić, “Razvoj računskih metoda za analizu statičkog i dinamičkog ponašanja žičara”, (in Serbian), Fakultet tehničkih nauka, Novi Sad, (1989)

SESSION A: EARTH MOVING AND TRANSPORTATION MACHINERY

A hybrid SWARA–MULTIMOORA approach for production job prioritization

Aleksandar Stanković[1*], Goran Petrović[1], Aleksandar Trajković[1], Predrag Milić[1], Milica Stanković[2], Žarko Ćojbašić[1]

[1] University of Niš, Faculty of Mechanical Engineering, Serbia

[2] University of Niš, Faculty of Economics, Serbia

ARTICLE INFO

[*] **Correspondence:** aleksandar.stankovic@masfak.ni.ac.rs

DOI: https://doi.org/10.46793/ET26.A12S **PAGES:** A95 – A102

ABSTRACT

Efficient prioritization of production jobs is a critical task in manufacturing systems, especially in environments with limited resources, multiple customer orders, and conflicting operational objectives. Production managers must determine which jobs should be processed first while considering several criteria such as due date urgency, profit margin, processing time, customer importance, and setup time. This decision- making problem can be effectively addressed using multi-criteria decision-making (MCDM) methods. In this paper, a hybrid SWARA–MULTIMOORA model is proposed for prioritizing production jobs. The SWARA (Step-wise Weight Assessment Ratio Analysis) method is applied to determine the relative importance of evaluation criteria based on expert judgment, while the MULTIMOORA (Multi-Objective Optimization by Ratio Analysis plus the Full Multiplicative Form) method is used to rank production jobs. The proposed approach is illustrated through a case study involving five production orders in a manufacturing environment. The results demonstrate that the model provides a systematic, transparent, and practical decision-support tool for production managers, enabling more effective scheduling decisions and improved utilization of manufacturing resources. The proposed methodology can be easily integrated into existing production planning and scheduling systems.

KEYWORDS

Job prioritization, Production planning, MCDM, Decision support systems.

1. INTRODUCTION

In today's highly competitive manufacturing environment, efficient production planning and scheduling have become essential for achieving operational excellence and maintaining customer satisfaction. Manufacturing companies are constantly challenged by increasing product variety, fluctuating customer demands, limited production capacities, and strict delivery deadlines. Under such conditions, production managers must continuously make complex decisions regarding the prioritization and sequencing of production jobs in order to optimize the utilization of available resources and minimize production delays. Production job prioritization represents one of the most important activities in manufacturing systems because it directly affects delivery performance, production efficiency, operational costs, and customer relationships [1]. In practice, production managers are often required to determine which production order should be processed first while simultaneously considering several conflicting criteria. These criteria may include due date urgency, profit margin, customer importance, processing time, setup time, machine availability, and production costs. Since these criteria are usually conflicting in nature, the decision-making process becomes significantly more complicated. Traditional production scheduling approaches frequently rely on simple dispatching rules such as First Come First Served (FCFS), Shortest Processing Time (SPT), or Earliest Due Date (EDD). Although such methods are easy to implement, they often fail to provide satisfactory solutions in complex manufacturing environments where multiple decision criteria must be considered simultaneously [2]. Consequently, manufacturing organizations increasingly apply multi-criteria decision-making (MCDM) methods to improve the quality and consistency of scheduling decisions.

Multi-criteria decision-making methods provide systematic and structured approaches for evaluating alternatives according to multiple criteria. These methods help decision-makers analyze complex problems, reduce subjectivity, and improve

the transparency of decision-making processes. Various MCDM techniques have been widely applied in manufacturing systems, including AHP, TOPSIS, VIKOR, ELECTRE, MOORA, and MULTIMOORA methods [3],[4], [5]. Among these methods, MULTIMOORA has attracted considerable attention due to its robustness, simplicity, and ability to provide reliable ranking results through the integration of multiple evaluation mechanisms. In addition to alternative ranking methods, determining the relative importance of evaluation criteria is also a crucial step in MCDM problems. The SWARA (Step-wise Weight Assessment Ratio Analysis) method has become increasingly popular for criteria weighting because it enables experts to express their judgments in a simple and flexible manner. Unlike some traditional weighting methods, SWARA requires fewer pairwise comparisons and allows decision-makers to determine criteria importance through a step-wise evaluation process. This paper proposes a hybrid SWARA–MULTIMOORA model for prioritizing production jobs in manufacturing systems. The SWARA method is employed to determine the weights of evaluation criteria based on expert opinions, while the MULTIMOORA method is used to rank production jobs according to their overall performance. The proposed model aims to support production managers in making more rational, transparent, and efficient scheduling decisions. To demonstrate the applicability of the proposed methodology, a case study involving five production jobs is presented. The jobs are evaluated according to several important production criteria, including due date urgency, profit margin, customer importance, processing time, and setup time. The obtained results illustrate the effectiveness of the hybrid model in identifying production priorities and improving decision-making quality in manufacturing environments.

The main contribution of this study is the development of an integrated decision-support framework that combines the advantages of SWARA and MULTIMOORA methods for production job prioritization. The proposed approach provides practical benefits for manufacturing organizations by enabling better resource allocation, improved production scheduling, and enhanced operational performance. The remainder of this paper is organized as follows. Section 2 explains the proposed SWARA–MULTIMOORA methodology. Section 3 presents the case study and computational procedure. Section 4 discusses the obtained results, while Section 5 provides concluding remarks and recommendations for future research.

2. METHODOLOGY

This section presents the proposed hybrid SWARA–MULTIMOORA methodology for prioritizing production jobs in manufacturing systems. The methodology consists of two main phases. In the first phase, the SWARA method is used to determine the relative importance of evaluation criteria based on expert judgments. In the second phase, the MULTIMOORA method is applied to rank production jobs according to the obtained criteria weights [6].

The proposed methodology enables decision-makers to evaluate production jobs systematically while considering multiple conflicting criteria simultaneously. The overall framework of the proposed model is illustrated through the following steps:

- Definition of evaluation criteria and production jobs,
- Determination of criteria weights using SWARA,
- Construction of the decision matrix,
- Application of the MULTIMOORA method,
- Ranking of production jobs according to final results.

2.1. Applications of SWARA and MULTIMOORA Methods

Numerous studies have applied SWARA and MULTIMOORA methods to solve complex industrial and manufacturing problems. Researchers have demonstrated that these methods provide reliable and practical solutions in environments characterized by uncertainty, conflicting objectives, and multiple evaluation criteria. The SWARA method has been applied in supplier selection, risk assessment, logistics management, personnel evaluation, and manufacturing strategy development [7]. Due to its flexibility and simplicity, SWARA is particularly useful in situations where expert judgment is required to determine the relative importance of criteria. Several studies have shown that SWARA provides consistent and realistic criteria weights while reducing the computational burden associated with traditional weighting approaches. Similarly, MULTIMOORA has been successfully utilized in various engineering and manufacturing applications, including machine selection, facility location, project evaluation, material selection, and production optimization [8]. Researchers have emphasized that the integration of multiple ranking mechanisms within MULTIMOORA improves ranking stability and increases the reliability of final decisions. In manufacturing environments, hybrid MCDM models combining weighting and ranking methods have become increasingly common [9], [10], [11].

The integration of SWARA with ranking methods such as TOPSIS, VIKOR, and MOORA has demonstrated promising results in solving complex production and operational problems. Hybrid approaches allow decision-makers to benefit from the strengths of multiple methods simultaneously. Despite the growing number of studies involving MCDM methods in manufacturing systems, limited research has focused specifically on the application of hybrid SWARA–MULTIMOORA models for production job prioritization. Most existing studies primarily concentrate on supplier selection, machine evaluation, or strategic decision-making problems, while operational scheduling decisions remain relatively underexplored. Therefore, this study aims to fill this research gap by proposing a hybrid SWARA–MULTIMOORA model for prioritizing production jobs in manufacturing systems [12]. The proposed approach contributes to the existing literature

by providing a practical and systematic framework for improving production scheduling decisions under multiple conflicting criteria.

The proposed hybrid model combines the strengths of both SWARA and MULTIMOORA methods. SWARA enables accurate determination of criteria importance based on expert knowledge, while MULTIMOORA provides a robust ranking mechanism through multiple evaluation perspectives. The integrated methodology offers several advantages:

- systematic evaluation of production jobs,
- consideration of multiple conflicting criteria,
- reduction of subjective bias,
- improved transparency of decision-making,
- practical applicability in manufacturing environments.

The proposed model can be effectively integrated into existing production planning and scheduling systems to support production managers in prioritizing jobs and improving manufacturing performance.

2.1.1 SWARA Method

The SWARA (Step-wise Weight Assessment Ratio Analysis) method is an MCDM weighting technique developed for determining the relative importance of criteria through expert evaluation. The method is based on the opinions and experience of decision-makers and enables a simple and flexible weighting process. One of the main advantages of SWARA is that it requires fewer comparisons than traditional methods such as AHP. Furthermore, the method allows experts to express their judgments gradually according to the significance of criteria [13].

The SWARA procedure consists of the following steps.

Step 1: *Ranking of Criteria*: First, experts rank the evaluation criteria according to their importance in descending order. The most significant criterion is placed in the first position. Assume that the criteria are represented as $C_1, C_2, C_3, \dots, C_n$. Where: $C_j$represents the j-th criterion, n is the total number of criteria.

Step 2: *Determination of Comparative Importance*: Starting from the second criterion, experts determine the comparative importance of criterion $j$relative to the previous criterion $(j-1)$. The comparative importance coefficient is denoted as s_j. If the criterion has equal importance compared to the previous one, then:

$$s_j = 0 \tag{1}$$

Higher values of $s_j$indicate lower importance of the current criterion relative to the previous criterion.

Step 3: Calculation of the Coefficient k_j:The coefficient $k_j$is calculated using the following equation:

$$k_j = s_j + 1 \tag{2}$$

For the first criterion:

$$k_1 = 1 \tag{3}$$

Step 4: Calculation of Recalculated Weights: The recalculated weight $q_j$is obtained recursively. For the first criterion:

$$q_1 = 1 \tag{4}$$

For the remaining criteria:

$$q_j = \frac{q_{j-1}}{k_j} \tag{5}$$

Step 5: Determination of Final Criteria Weights The final normalized weights are calculated using:

$$w_j = \frac{q_j}{\sum_{j=1}^{n} q_j} \tag{6}$$

where: $w_j$represents the final weight of criterion j, $\sum q_j$is the total sum of recalculated weights. The obtained weights are later used in the MULTIMOORA ranking procedure.

2.1.2 MULTIMOORA Method

MULTIMOORA (Multi-Objective Optimization by Ratio Analysis plus Full Multiplicative Form) is an advanced MCDM method developed as an extension of the MOORA technique [14], [15], [16]. The method integrates three different evaluation approaches:

- Ratio System
- Reference Point Approach
- Full Multiplicative Form

The integration of these components improves the robustness and reliability of ranking results. Assume that there are m production jobs and $n$evaluation criteria. The decision matrix is represented as:

$$X = [x_{ij}] \tag{7}$$

where: x_{ij}represents the performance value of alternative $i$under criterion j, $i = 1,2,\ldots,m,\ j = 1,2,\ldots,n$.

Step 1: Construction of the Decision Matrix: The decision matrix includes all production jobs and their corresponding values according to each evaluation criterion. The matrix can be represented as:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{8}$$

Step 2: Normalization of the Decision Matrix: The decision matrix is normalized to eliminate differences in measurement units. The normalized value is calculated as follows:

$$x_{ij}^{*} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^{2}}} \tag{9}$$

where: $x_{ij}^{*}$is the normalized value, $x_{ij}$is the original value.

Step 3: Ratio System Approach In the Ratio System, beneficial criteria are maximized while non-beneficial criteria are minimized. The overall assessment value for each alternative is calculated as:

$$Y_i = \sum_{j=1}^{g} w_j\, x_{ij}^{*} - \sum_{j=g+1}^{n} w_j\, x_{ij}^{*} \tag{10}$$

where: $g$is the number of beneficial criteria, $w_j$is the weight of criterion j, Y_i is the performance score of alternatives i.

Higher values of $Y_i$indicate better alternatives.

Step 4: Reference Point Approach: In the Reference Point Approach, the best performance values for each criterion are identified as reference points. The deviation from the reference point is calculated using:

$$D_i = \max_{j} \mid r_j - w_j x_{ij}^{*} \mid \tag{11}$$

where:r_j represents the reference value, D_i is the maximum deviation for alternative i. Smaller deviation values indicate better alternatives.

Step 5: Full Multiplicative Form: The Full Multiplicative Form evaluates alternatives by multiplying beneficial criteria and dividing them by non-beneficial criteria. The overall utility value is calculated as:

$$U_i = \frac{\prod_{j=1}^{g} x_{ij}^{w_j}}{\prod_{j=g+1}^{n} x_{ij}^{w_j}} \tag{12}$$

Higher values of $U_i$represent better alternatives.

Step 6: Final Ranking of Alternatives: The final ranking of production jobs is obtained by combining the rankings derived from the three MULTIMOORA components:

- Ratio System,
- Reference Point Approach,
- Full Multiplicative Form.

The production job with the best overall ranking is considered the highest priority for processing.

3. CASE STUDY

This section presents the application of the proposed hybrid SWARA–MULTIMOORA model in a manufacturing environment. The objective of the case study is to prioritize production jobs by considering multiple operational and managerial criteria simultaneously. The case study involves five production jobs processed in a manufacturing company operating under limited production capacity and strict delivery requirements. Production managers must determine the order in which jobs should be processed in order to improve resource utilization, minimize delays, and satisfy customer demands. A manufacturing company receives multiple customer orders that must be processed on shared production resources. Due

to limited machine availability and time constraints, production managers need to establish priorities among production jobs. In this study, five production jobs are considered:

$$J_1, J_2, J_3, J_4, J_5 \tag{13}$$

The prioritization process is based on five evaluation criteria frequently used in manufacturing decision-making:

Table 1: The prioritization process is based on five evaluations

Criteria	Description	Type
C_1	Due date urgency	Benefit
C_2	Profit margin	Benefit
C_3	Customer importance	Benefit
C_4	Processing time	Cost
C_5	Setup time	Cost

The first three criteria are beneficial criteria, meaning that higher values are preferred. The last two criteria are non-beneficial (cost) criteria, where lower values are desirable. A team of production experts participated in the evaluation process and provided judgments regarding the relative importance of criteria.

Step 2: Determination of Criteria Weights Using SWARA: The SWARA method was applied to determine the weights of evaluation criteria according to expert opinions. The experts ranked the criteria in the following order of importance:

$$C_1 > C_2 > C_3 > C_4 > C_5 \tag{14}$$

where: C_1: Due date urgency, C_2: Profit margin, C_3: Customer importance, C_4: Processing time, C_5: Setup time.

The comparative importance coefficients (s_j), recalculated coefficients (k_j), and normalized weights (w_j) were calculated according to the SWARA procedure. The results are presented in Table 2.

Table 2: Criteria Weights Obtained Using SWARA

Criteria	s_j	k_j	q_j	w_j
C_1	–	1.00	1.000	0.300
C_2	0.20	1.20	0.833	0.250
C_3	0.25	1.25	0.666	0.200
C_4	0.33	1.33	0.500	0.150
C_5	0.50	1.50	0.333	0.100

The obtained results indicate that due date urgency (C_1) is the most significant criterion with a weight of 0.300, followed by profit margin (C_2) and customer importance (C_3). Processing time and setup time received relatively lower importance weights. These findings are consistent with real manufacturing environments where meeting delivery deadlines and maintaining customer satisfaction are considered primary operational objectives.

*Step 3: Construction of the Decision Matr*ix: After determining the criteria weights, the decision matrix was constructed based on the evaluations of production jobs. The decision matrix is presented in Table 3.

Table 3: Initial Decision Matrix

Jobs	C_1	C_2	C_3	C_4	C_5
J_1	9	8	7	5	4
J_2	7	9	8	6	5
J_3	10	6	9	4	3
J_4	6	7	5	8	6
J_5	8	8	6	5	4

The values in the matrix represent the performance of each production job according to the corresponding criterion.

4.1 Application of the MULTIMOORA Method

The MULTIMOORA method was applied to rank the production jobs according to the weighted criteria. The decision matrix was normalized using vector normalization in order to eliminate differences in measurement scales. The normalized matrix is shown in Table 4.

Table 4: Normalized Decision Matrix

Jobs	C_1	C_2	C_3	C_4	C_5
J1	0.495	0.467	0.438	0.388	0.396
J2	0.385	0.525	0.501	0.466	0.495

Jobs	C_1	C_2	C_3	C_4	C_5
J3	0.550	0.350	0.564	0.310	0.297
J4	0.330	0.408	0.313	0.621	0.594
J5	0.440	0.467	0.376	0.388	0.396

The Ratio System approach was applied by maximizing beneficial criteria and minimizing cost criteria. The obtained performance scores are presented in Table 5.

Table 5: Ratio System Results

Jobs	Y_i	Rank
J_1	0.255	2
J_2	0.228	3
J_3	0.289	1
J_4	0.111	5
J_5	0.226	4

According to the Ratio System results, production job J_3 achieved the highest score and was ranked as the best alternative. The Reference Point Approach was applied by determining the maximum deviation of each alternative from the ideal reference point. The results are shown in Table 6.

Table 6: Reference Point Results

Jobs	D_i	Rank
J_1	0.025	1
J_2	0.050	4
J_3	0.044	3
J_4	0.066	5
J_5	0.038	2

Smaller deviation values indicate better alternatives. According to the Reference Point Approach, job J_1 achieved the best performance. The Full Multiplicative Form was then applied to evaluate the utility of each production job. The obtained results are presented in Table 7.

Table 7: Full Multiplicative Form Results

Jobs	U_i	Rank
J_1	3.281	2
J_2	3.063	4
J_3	3.527	1
J_4	2.351	5
J_5	3.071	3

The results again confirm that production job J_3 represents the best alternative among the evaluated jobs. The final ranking was determined by combining the rankings obtained from the three MULTIMOORA components. The obtained ranking indicates that job J_3 should receive the highest production priority due to its urgent delivery requirement, high customer importance, and relatively low processing and setup times. The proposed SWARA–MULTIMOORA model successfully integrates multiple criteria into a unified decision-making framework and provides production managers with a practical tool for improving scheduling decisions in manufacturing systems. The final results are presented in Figure 1.

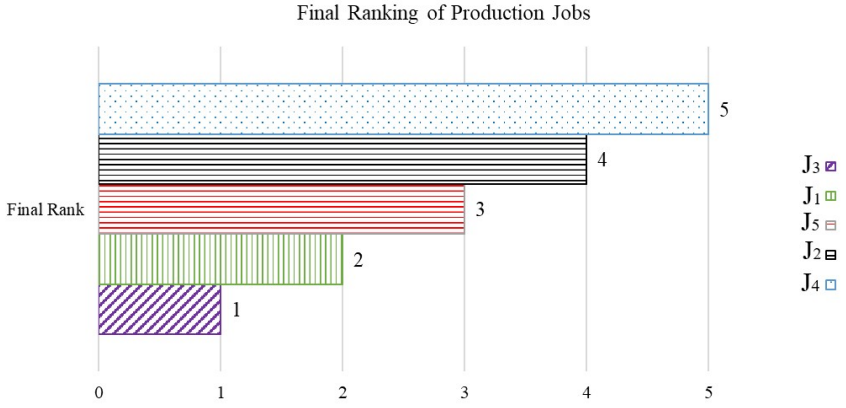


Figure 1: Final Ranking of Production Jobs

4. RESULTS AND DISCUSSION

The results obtained from the proposed SWARA–MULTIMOORA model demonstrate its effectiveness in prioritizing production jobs in a manufacturing environment. The final ranking showed that job J_3achieved the highest priority among all evaluated alternatives. This result was consistent across all three MULTIMOORA components, confirming the reliability and stability of the proposed methodology. Job J_3 obtained the best performance mainly due to its high due date urgency and customer importance, which were identified as the most significant criteria through the SWARA weighting process. In addition, the job had relatively low processing and setup times, positively influencing its final ranking. In contrast, job J_4received the lowest priority because of its longer processing and setup times. This finding highlights the importance of evaluating multiple criteria simultaneously rather than relying on a single operational factor. The SWARA results indicated that due date urgency was the most important criterion, followed by profit margin and customer importance. This reflects the practical requirements of modern manufacturing systems, where on-time delivery and customer satisfaction are critical for maintaining competitiveness. The MULTIMOORA method provided a comprehensive and systematic evaluation framework by integrating different ranking approaches. The proposed model reduces subjectivity in production scheduling decisions and supports more objective and efficient resource allocation. Furthermore, the methodology is flexible and can be adapted to different manufacturing environments by including additional criteria according to operational requirements. The model may also be integrated into modern production planning and Industry 4.0 systems to support real-time scheduling decisions. Overall, the obtained results confirm that the hybrid SWARA–MULTIMOORA model represents an effective decision-support tool for production job prioritization and can significantly improve manufacturing scheduling performance.

5. CONCLUSION

This paper proposed a hybrid SWARA–MULTIMOORA model for production job prioritization in manufacturing systems. The SWARA method was used to determine criteria weights, while the MULTIMOORA method was applied to rank production jobs according to multiple criteria. The results showed that the proposed methodology provides a systematic and reliable framework for production scheduling decisions. The model successfully identified the most suitable production priority by considering both operational and managerial factors. The proposed approach can help manufacturing companies improve resource utilization, scheduling efficiency, and customer satisfaction. In addition, the methodology is flexible and can be integrated into existing production planning systems. Future research may focus on extending the model using fuzzy logic, artificial intelligence techniques, and real-time scheduling applications in Industry 4.0 environments.

ACKNOWLEDGEMENTS

This research was financially supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03- 136/2025-03/200109).

REFERENCES

[1] A. Stanković, “Razvoj modela za optimalno planiranje i raspoređivanje resursa u malim i srednjim preduzećima u uslovima neizvesnosti”, PhD Thesis, University of Niš (Serbia), (2025)

[2] A. Stanković and G. Petrović, “Priority decision rules with a fuzzy MCDM approach for solving flexible job shop problem: A real case study of optimizing manufacturing”, Acta Polytechnica Hungarica, Vol. 22(1), pp. 143–162, https://doi.org/10.12700/APH.22.1.2025.1.8, (2025)

[3] M., Stanković, M.I. Đukić, A. Stanković, and S. Stefanović, “Multicriteria decision-making in the evaluation of public services: Application of MCDM methods in a real case study”, Journal of Operational and Strategic Analysis, Vol. 3(3), pp.172–181, https://doi.org/10.56578/josa030302, (2025)

[4] A. Stanković and G. Petrović, “Comparative analysis of job priority rules in single-machine scheduling under uncertainty”, Innovative Mechanical Engineering, Vol. 4(3), pp. 65–77, (2025)

[5] M. Madić, G. Petrović, D. Petković, J. Antucheviciene, and D. Marinković, “Application of a robust decision making rule for comprehensive assessment of laser cutting conditions and performance”, Machines, Vol. 10(2), p. 153, https://doi.org/10.3390/machines10020153, (2022)

[6] M. Madanchian and H. Taherdoost, “Applications of multi-criteria decision making in information systems for strategic and operational decisions”, Computers, Vol. 14(6), p. 208, https://doi.org/10.3390/computers14060208, (2025)

[7] G. Petrović, J. Mihajlović, Ž. Ćojbašić, M. Madić, and D. Marinković, “Comparison of three fuzzy MCDM methods for solving the supplier selection problem”, Facta Universitatis, Series: Mechanical Engineering, Vol. 17(3), pp. 455–469, https://doi.org/10.22190/FUME190420039P, (2019)

[8] S. Vesković, Ž. Stević, G. Stojić, M. Vasiljević, and S. Milinković, “Evaluation of the railway management model by using a new integrated model DELPHI-SWARA-MABAC”, Decision Making: Applications in Management and Engineering, Vol. 1(2), pp. 34–50, https://doi.org/10.31181/dmame1802034v, (2018)

[9] R. Mavi, M. Goh, and N. Zarbakhshnia, “Sustainable third-party reverse logistic provider selection with fuzzy SWARA and fuzzy MOORA in plastic industry”, The International Journal of Advanced Manufacturing Technology, Vol. 91, pp. 2401–2418, https://doi.org/10.1007/s00170-016-9880-x, (2017)

[10] G. Petrović, J. Mihajlović, D. Marković, S. Hashemkhani Zolfani, and M. Madić, “Comparison of aggregation operators in the group decision-making process: A real case study of location selection problem”, Sustainability, Vol. 15(10), p. 8229, https://doi.org/10.3390/su15108229, (2023)

[11] A. I. Maghsoodi, A. I. Maghsoodi, A. Mosavi, T. Rabczuk, and E.K. Zavadskas, “Renewable energy technology selection problem using integrated H-SWARA-MULTIMOORA approach”, Sustainability, Vol. 10(12), p. 4481, https://doi.org/10.3390/su10124481, (2018)

[12] A. I. Maghsoodi, G. Abouhamzeh, M. Khalilzadeh, and E.K. Zavadskas, “Ranking and selecting the best performance appraisal method using the MULTIMOORA approach integrated Shannon’s entropy”, Frontiers of Business Research in China, Vol. 12(1), pp. 1–21, https://doi.org/10.1186/s11782-017-0022-6, (2018)

[13] S. Mishra, A.K. Sahu, S. Datta, and S.S. Mahapatra, “Application of fuzzy integrated MULTIMOORA method towards supplier/partner selection in agile supply chain”, International Journal of Operational Research, Vol. 22(4), pp. 466–514, https://doi.org/10.1504/IJOR.2015.068562, (2015)

[14] A. Farhad, K. Ullah, Z. Ali, and D. Pamucar, “Integration of MULTIMOORA algorithm combined with circular q-rung orthopair fuzzy information for optimizing player positioning”, Scientific Reports, Vol. 15(1), p. 36162, https://doi.org/10.1038/s41598-025-18795-0, (2025)

[15] H. C. Liu, X. J. Fan, P. Li, and Y. Z. Chen, “Evaluating the risk of failure modes with extended MULTIMOORA method under fuzzy environment”, Engineering Applications of Artificial Intelligence, Vol. 34, pp. 168–177, https://doi.org/10.1016/j.engappai.2014.04.011, (2014)

[16] A. Hafezalkotob, A. Hafezalkotob, and M. K. Sayadi, “Extension of MULTIMOORA method with interval numbers: An application in materials selection”, Applied Mathematical Modelling, Vol. 40, pp. 1372–1386, https://doi.org/10.1016/j.apm.2015.07.019, (2016)

SESSION B

RAILWAY ENGINEERING

SESSION B: RAILWAY ENGINEERING

Strength-deformation analysis of wheel reducer for metro cars

Vasko Nikolov*[1], Filip Kirilov[1]

[1] Todor Kableshkov University of Transport, Department Transport Equipment, Sofia, Bulgaria

ARTICLE INFO

* **Correspondence:** va_r_nikolov@abv.bg

DOI: https://doi.org/10.46793/ET26.B01N **PAGES**: B1 – B10

ABSTRACT

Railway transport requires high reliability and safety in operation, which places high demands on the elements of the gear boxes. A particularly important role is played by the wheel reducers, which ensure the transmission of torque from the traction electric motor to the wheelsets of the rolling stock, as well as the braking torque during electrodynamic braking applied to the same wheelsets. In this work, a strength-deformation analysis of a wheel reducer for metro cars used in the Sofia Metropolitan is carried out. The main objective of the study is to find weak points in the structure through qualitative and quantitative analysis of the acting loads to increase the reliability and operational stability of the transmission mechanism of the bogies in the rolling stock of the Sofia Metropolitan. A review of the transmission mechanisms used in the rolling stock of the metro trains is carried out, considering their design features and operational characteristics. The operational condition of the considered node was analyzed, in which characteristic failures in the elastic membrane coupling EZK-MEM 250-1, related to stress concentrations and installation inaccuracies, were identified and analyzed. The conclusion provides recommendations for improving the design, installation, maintenance, repair and operation of this important for traffic safety node.

KEYWORDS

Wheel reducer, Metro cars, Strength-strain analysis, Finite element method, Operational reliability, EZK-MEM coupling.

1. INTRODUCTION

Transmission mechanisms are the mechanisms that transmit to the wheelsets a driving torque from the shaft of a traction electric motor or from the output shaft of a mechanical, hydraulic or other transmission system, without violating the freedom necessary for movement for relative movement of the wheels in relation to the bogie frame and of the bogie frame in relation to the wagon frame. The structural elements of the transmission mechanisms are:

- Reducers;
- Shafts;
- Joints;
- Supports;
- Couplings and others.

The technical and economic parameters of the railway vehicle significantly depend on the transmission mechanisms. This is the reason for a few requirements to be made to them. In summary, the most important of them are expressed in the following:

- The operation of the transmission mechanism should not be accompanied by the appearance of high-value dynamic forces and should not limit the operating modes;
- The reliability of the mechanism should be high and meet the requirements for economic efficiency;
- Ensure high operational economy;
- Do not reduce the space intended for installation of the traction motor as much as possible;
- The design should be simplified and allow accessibility to the mechanism for inspection and repair. The design should allow for the location of the traction motors, in which they are accessible for inspection and repair, as well as protected from snow, moisture and dust. Transmission mechanisms transmit torque from the electric motor to the wheels it drives (when the motor is in traction mode) and vice versa (when the motor is used for braking).

The power transmissions of the railway rolling stock must meet the following general requirements:

- To be safe in operation;
- To have as simple a design as possible;
- To be accessible for inspection and repair;
- To be light and compact;
- To provide optimal operating conditions for the engine.

The type and design of the power transmission is determined by the type, power and properties of the electric motor, the method of its suspension, its location in relation to the longitudinal axis of the metro car and the type, purpose and maximum speed of the metro car itself.

2. CLASSIFICATION OF TRANSMISSION MECHANISMS

2.1. According to the Number of Wheelsets, Which Are Simultaneously Driven:

- **Power Transmissions for Individual Drive of the Wheelsets;**
- **Power Transmissions for Group Drive of the Wheelsets.**

2.2. According to Their Regulation Properties:

- Non-adjustable (i.e. constant gear ratio) power transmissions – light, simple and cheap (in production, operation and repair) power transmissions. They belong to the category of mechanical power transmissions, as they transmit torque through gearing, lever systems or power chains;
- Adjustable (i.e. with variable gear ratio) power transmissions – relatively complex and expensive, but unlike the non-adjustable ones, they allow the adjustment of the engine characteristics in accordance with the changing operating conditions of the railway vehicle. They are mechanical (created based on a multi-speed gearbox with gear engagements); hydraulic and combined.

2.3. According to the Value of Unsuspended Masses:

- First class transmission mechanism – axle-mounted traction motor and axle-mounted reducer;
- Second class transmission mechanism – bogie frame-mounted traction motor and axle-mounted reducer;
- Third class transmission mechanism – frame-mounted traction motor and frame-mounted reducer. The traction motor, together with the wheel reducer, transmits its mass to the bogie frame and is part of the sprung parts.

3. STRUCTURAL LAYOUT AND FEATURES OF THE AXLE REDUCER OF ROLLING STOCK TYPE 81 – 740.2/741.2 AND 81 – 740.2B/741.2B

These wagons have two motorized two-axle bogies with individual drive of the axles - front, rear and one non-motorized two-axle bogie - intermediate. Motor bogies are an element of the running gear of the wagons and guide the movement along the rail track, ensuring minimal resistance and the necessary smoothness of the run, distributing and transmitting all loads from the body to the track, and forming and realizing the traction and braking forces for their propulsion. The suspension of the traction drive is of the third class, i.e. frame-mounted traction motor and frame-mounted reducer.

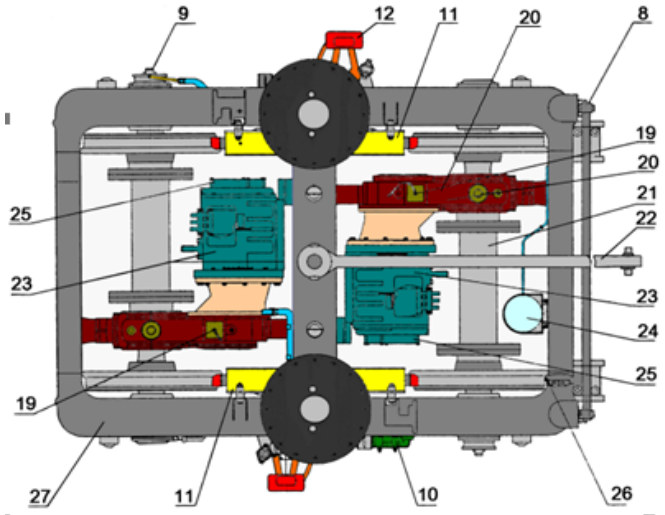


Figure 1:

It consists of the following main components, assemblies and mechanisms: frame 27; two axles 29; traction electric motor 23, reducers 20, transmission mechanism 21 and other devices for transmitting rotation (Fig. 1). The axles belong to the unspring masses and are one of the main units of the running gear and all undercarriage equipment. The axles of the motor bogie consist of an axle 8, two wheels 2 and 10, axle boxes 1 and 11, transmission mechanism 7 and reducer 4 (Fig. 2).

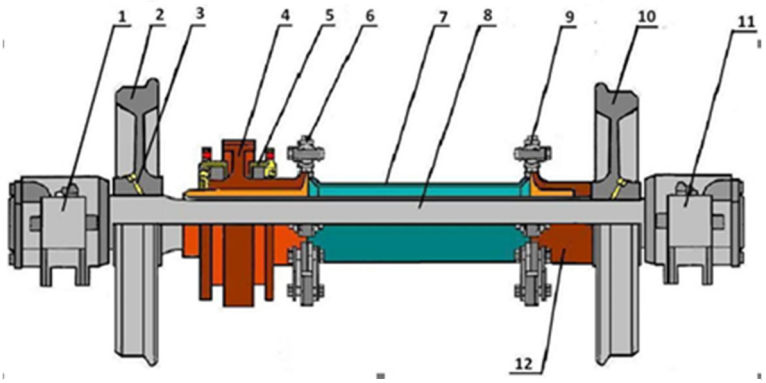


Figure 2:

The traction drive consists of a traction electric motor, a reducer and other elements that provide the transmission of torque from the electric motor to the wheelsets. The kinematic diagram of the drive from the electric motor to the wheels is shown in Fig. 3. The torque transmission is carried out according to the following scheme: electric motor rotor shaft 1, drive hub 2, gear housing 3, elastic ring 4, gear housing 5, driven hub 6, small gear shaft 7, small gear 8, parasitic gear 9, output shaft with large gear 10, flange of the output shaft of the reducer 11, flange of the hollow shaft of the transmission mechanism 12, hollow shaft 13, flange of the transmission mechanism 14, flange of the second wheel 15.

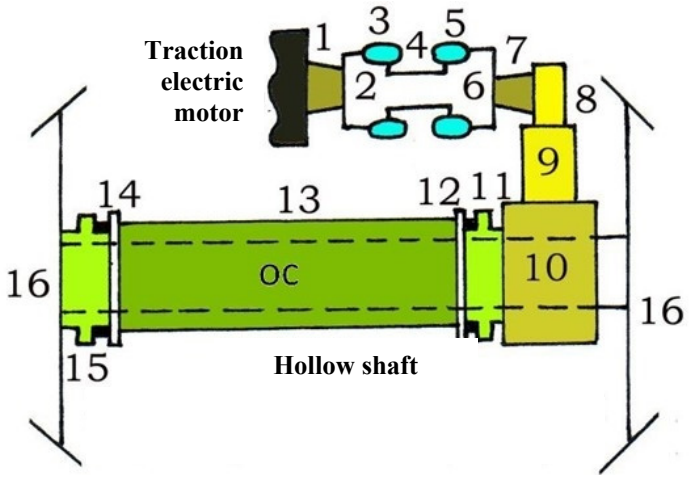


Figure 3:

The gearbox is two-stage and consists of a detachable housing of two parts - upper and lower crankcase. It is equipped with a shaft with a small gear 22, a shaft 20 with an intermediate gear 11 and an output shaft 16 with a gear ring. The shafts with the bearings are closed on both sides with covers. All gears of the gearbox have straight teeth. This prevents axial loading of the bearings and the housing (Fig. 4).

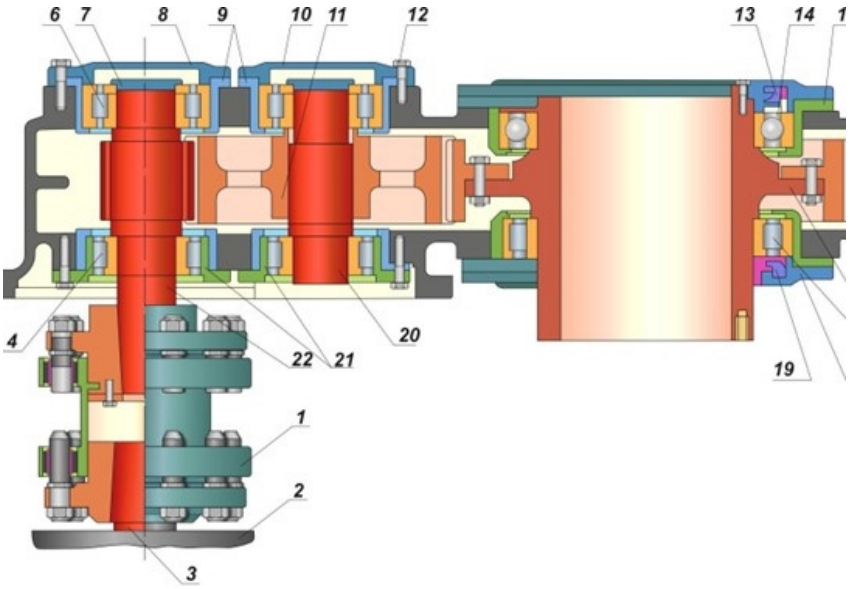


Figure 4:

4. ANALYSIS OF THE OPERATIONAL CONDITION OF THE UNIT

4.1. Unit Status Snapshot

After a detailed and thorough analysis of the structural layout and features of the transmission system of the metro cars, as well as after discussions with employees responsible for the maintenance and repair of the metro trains, it was established that the power transmissions of the trains are generally reliable in operation. The power transmissions in the metro cars have a simple design, are accessible for inspection and repair and provide optimal conditions for engine operation.

4.2. Problems Encountered in the Operation of the Unit

After in-depth research and statistical processing of the results obtained, it was found that for the period of the last four years, the most frequent registered malfunctions were in the coupling of the metro cars under consideration.

The problems are mainly concentrated in the elements that make up the elastic coupling, and more specifically in the flanges connecting the coupling halves: breakage of the housing of one coupling half, presence of cracks on the coupling half or a twisted coupling half.

4.3. Elastic Membrane Coupling EZK-MEM 250-1

The clutch is manufactured by the German company “KWD – Kupplungswerk Dresden” with No. 9981 5400. Its purpose is to transmit torque from the electric motor to the gearbox by means of opposing teeth meshing with each other through an elastic gear wheel that smooths out dynamic shocks.

4.3.1. Clutch Structure

The clutch consists of two half-clutches and an elastic gear ring. The half-clutches have the same design, each of them consists of a hub 2, a gear housing 3. The torque is transmitted from the traction motor through hub 2, the gear housing 3, the elastic ring 1 and the shaft of the small gear wheel of the reducer (Fig. 5÷10).

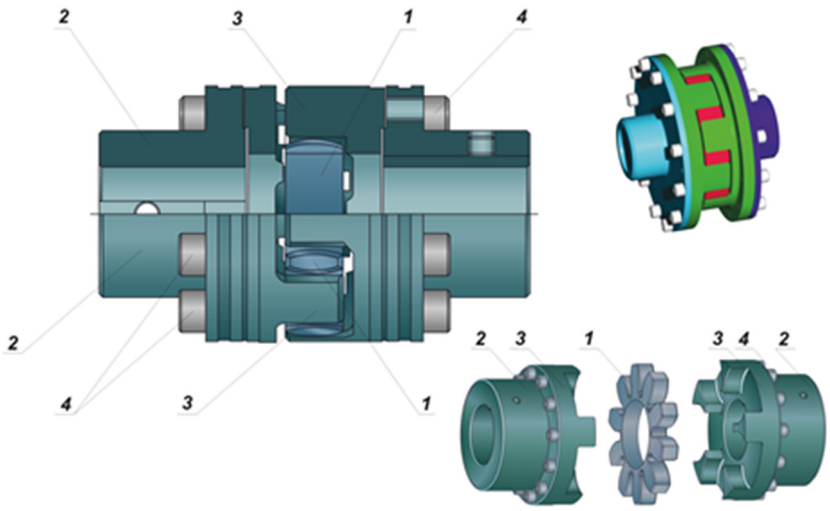


Figure 5:



Figure 6:



Figure 7:



Figure 8:



Figure 9:



Figure 10:

4.4. Common Faults in the EZK-MEM 250-1 Elastic Diaphragm Coupling

According to the conducted statistical study and observations of the studied unit, it was found that the most common damages to the elastic membrane coupling are rupture of the housing of one coupling part (Fig. 11), presence of cracks on the housing part (Fig. 12) and twisting of the coupling part (Fig. 13).



Figure 11:



Figure 12:

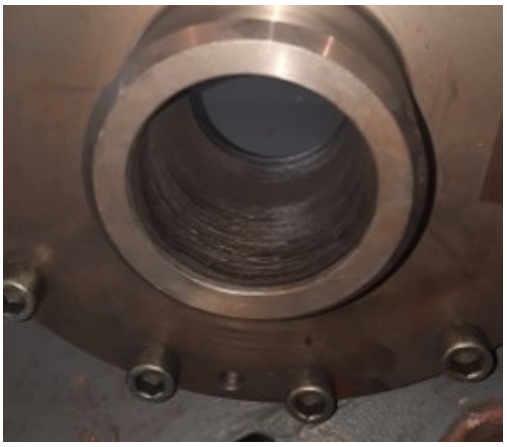


Figure 13:

4.5. Causes of Clutch Damage

4.5.1. Damage to the EZK-MEM250-1 Elastic Coupling Caused by the Suspension of the Traction Motor and Gearbox

The suspension of the traction motor and the reducer is a support arm with transmission of torque to the axles through a hollow shaft with articulated guides with sleeves. The TED is fastened to the reducer by a transition flange around the small gear (Fig. 14). The flange is fastened to the TED and the reducer with bolts.



Figure 14:

In the lower part of the transition flange there is an opening connected by a tube to the lower part of the reducer. This helps to return excess oil from the housing of the compensation sleeve to the reducer. The suspension of the reducer has a similar bracket in its rear part on the side of the small gear; it is attached with the universal bracket with a suspension to the central cross member of the bogie frame. On the side of the output shaft of the reducer to its housing there is a bracket mounted by bolts. The bracket with the help of two bolts passing through it, nuts and lock nuts is connected to the metal-rubber joint of the universal bracket. The metal-rubber joint in its upper part has a thread on which an adjustment sleeve is screwed (Fig. 15).



Figure 15:

In the upper part of the adjustment sleeve, the thread of an analogous metal-rubber hinge is wound, which, through two detachable "stones", is attached to a special thickening of the end crossbeam of the bogie frame. The adjustment of the vertical distance between the output shaft of the gearbox and the axle of the wheel axle is carried out using an adjustment sleeve (Fig. 16).

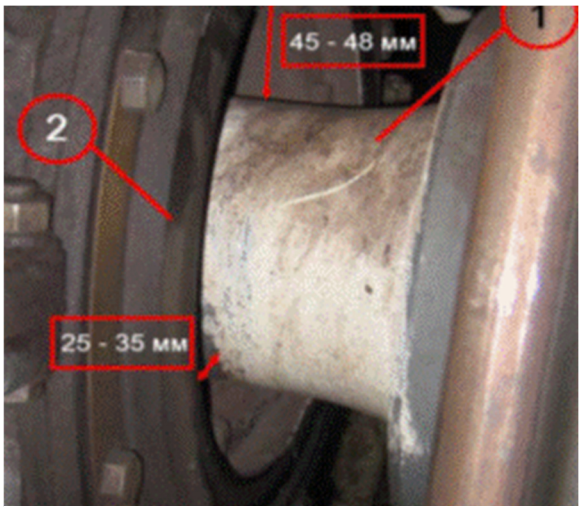


Figure 16:

4.5.2. Damage to the EZK-MEM250-1 Flexible Coupling Caused by Incorrect Assembly of the Coupling's Connecting Parts.

After conducting observations and analyzing the operational condition of the unit under consideration, it was found that the main cause of damage to the elastic membrane coupling is the coupling installation method. This is most likely due to the practice in the depot of lowering the traction motor with a crane. This method leads to inaccuracy and imprecision, as the positions of the shaft, gearbox and motor are not within the tolerances for each installation. When installing with unacceptable tolerances, misalignment occurs, with one half pressing the other, which during operation causes breakage or cracks to appear on the housing of one half. When installing the half coupling with the shaft, good contact must be obtained between the coupling hub and the traction motor shaft or the gearbox shaft. If the required contact spot is not achieved, the fitting (rubbing) is carried out with grinding paste. Clutch slippage is due to a contact spot with an area of less than 75% and poor fit (grinding) with the grinding paste on one clutch half.

5. STRENGTH-DEFORMATION ANALYSIS OF CONNECTING FLANGES USING THE FINITE ELEMENT METHOD

5.1. Method of Performing Analysis

The strength-strain analysis was performed using the finite element method. This is a numerical method for solving a wide range of problems in mechanics, in which, based on known relations between forces and displacements for individual elements of the structure, an analysis of the behavior of the entire structure is carried out. The goal of any such study is to determine the main characteristics of the field, namely the distribution of stresses or displacements, the values of stresses and displacements at individual points in the field, the extreme values of these quantities or their gradient, etc. Three-dimensional finite elements were used for the study.

5.2. Creating a Model

To perform a qualitative and realistic analysis using the finite element method, it is first necessary to create a 3D model that recreates the geometric dimensions and physical characteristics of the studied element as accurately as possible. For the study, a model was created that is as close as possible to the actual operating part, namely the coupling plates that connect the two halves of the coupling. The first plate is attached to the rotor shaft of the traction electric motor, and the second - to the shaft of the small gear from the reducer of the transmission mechanism.

5.3. Defining the Structural Load

The accuracy of the solution largely depends on the skillful development of the computational model. This requires the structure to be analyzed, the main supporting elements to be identified and a decision to be made on their appropriate representation by means of appropriate finite elements. The model is developed based on the load scheme of the real structure. This brings it as close as possible to it and the results are as close as possible to those operating in a real environment. The actual design of the coupling consists of two coupling plates, fastened with 12 bolts to the two half-couplings, one of which is loaded with the torque created by the traction electric motor, and another which is transmitted to the input shaft of the reducer. The two plates are pressed onto the rotor shaft of the traction motor and the input shaft of the reducer, respectively, by means of a conical press fit. The first plate, mounted on the rotor shaft of the traction

motor, is loaded with the torque that the motor creates and transmits it to the following elements of the reducer, and the second plate receives this torque and in turn transmits it to the input shaft of the reducer.

Thus, in the modeling, the first plate is loaded with torque in the hub and is fixed along the periphery in the bolt holes (Fig. 17), and the second is loaded with torque in the bolt holes and is fixed in the hub (Fig. 17). This also determines the differences in the results obtained in the analysis of the two elements.

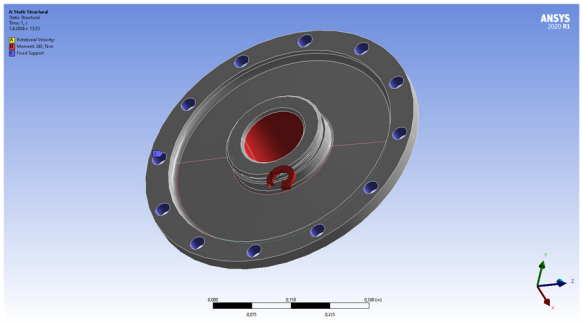


Figure 17:

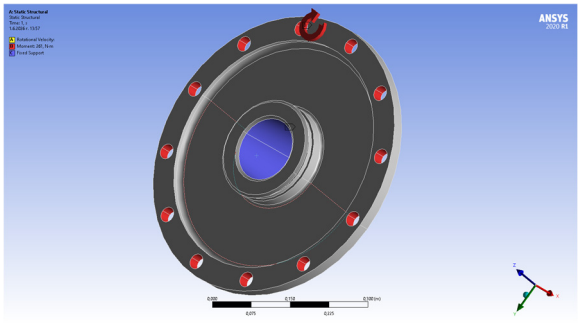


Figure 18:

The torque applied to the plates is maximum at start-up and decreases with increasing speed, with its minimum value at the design speed of the train. This dependence is key to determining operating conditions that are as close as possible to the actual ones. The maximum torque realized between the two plates is 261 Nm, realized at start-up and up to about 40 km/h, and the maximum angular velocity of the plates is 613.66 s^{-1} at the design speed. This somewhat relieves the load on the structure and must be considered.

5.4. Results Obtained

5.4.1. Load with Maximum Torque When Starting and Driving at Relatively Low Speed:

Guide plate: the maximum stress obtained in the guide plate is σ=939.18 MPa (Fig. 19), obtained in the transition from the hub to the membrane on the inside of the plate. The maximum relative elastic elongation in turn is $\varepsilon=4.71.10^{-3}$ mm/m in the same section. The safety factor is sufficiently high, uniform throughout the entire volume, which indicates that the part is sufficiently overprotected throughout its volume. The resource is evenly distributed throughout the volume and guarantees failure-free operation for 1.10^6 cycles (Fig. 20).

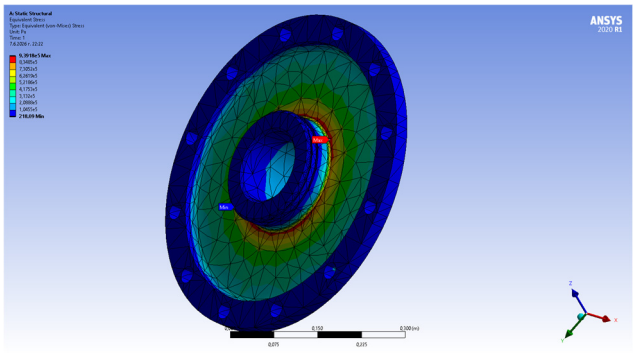


Figure 19:

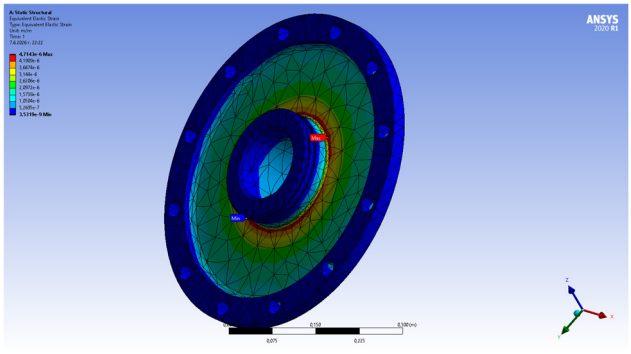


Figure 20:

Guided plate: the maximum stress is σ=953.25 MPa, again obtained in the transition from the hub to the membrane on the inside of the plate, i.e. slightly higher than that in the guide plate (Fig. 21). The maximum relative elastic strain is also located in the same place, its value being $\varepsilon=4.782.10^{-3}$ mm/m, also close in value to the guide plate. The maximum total deformation is obtained along the periphery of the plate, in the flange where the connecting bolts are located, and its value

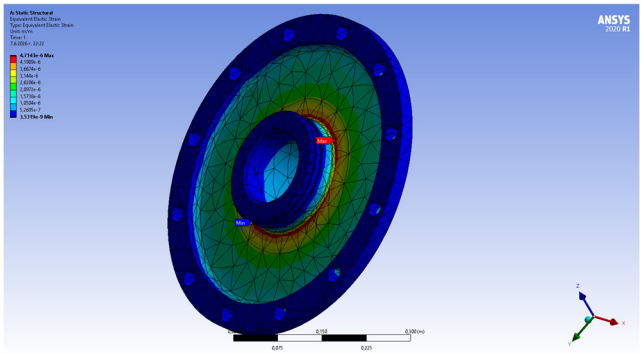


Figure 21:

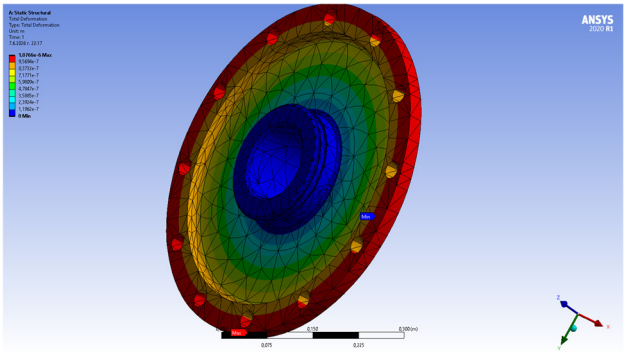


Figure 22:

is $1.0766.10^{-3}$ mm, a value significantly higher than that in the guide plate (Fig. 22). The values of the safety factors do not differ from the first plate, i.e. the guided plate is also sufficiently well overprotected.

5.4.2. Load When Driving at Maximum Speed:

Guide plate: the maximum stress is σ=5378 MPa, located at the transition between the hub and the membrane on the inside of the plate (Fig. 23). This value cannot be considered authoritative because it appears at a node and is a result of excessive accumulation of values. More important in this case is the average stress value, which is 937.14 MPa. However, it can be responsibly said that it is the maximum value that has the greatest impact on the damageability of the plate (Fig. 24) and on its resource (Fig. 25). The maximum value of the relative elastic strain is $2.74.10^{-2}$ mm/m, and the maximum total deformation is concentrated in the hub and has a value of $1.5315.10^{-2}$ m. The minimum safety factor is $4.6.10^{-2}$, obtained at the transition between the hub and the membrane on the inside of the plate, and the fatigue safety factor is $1.6.10^{-2}$, obtained at the same location (Fig. 26).

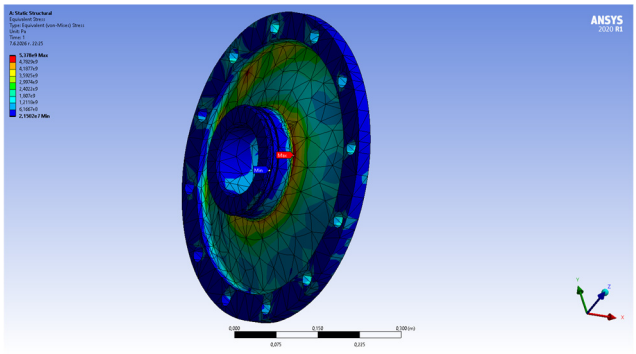


Figure 23:

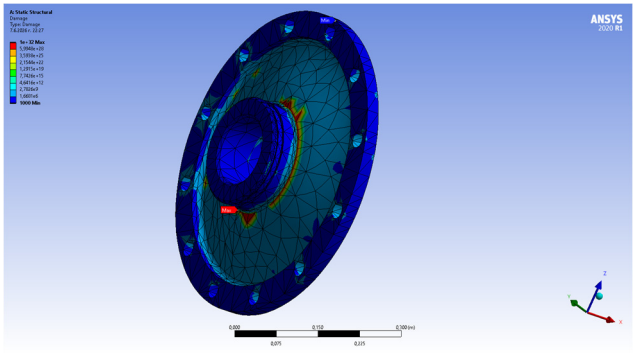


Figure 24:

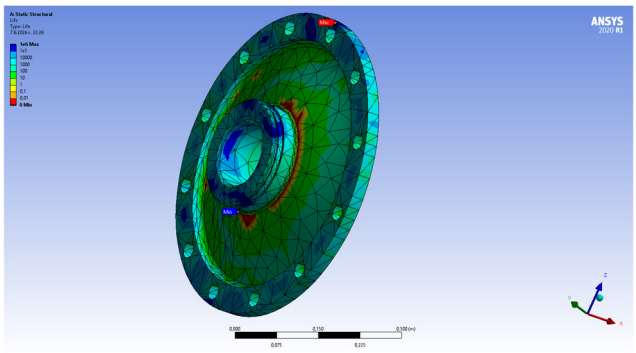


Figure 25:

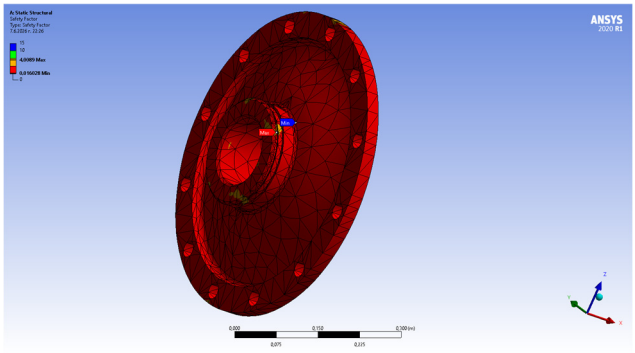


Figure 26:

Guided plate: the maximum stress is 8639 MPa, located at the same location (Fig. 27). The relative elastic elongation is $4.38.10^{-2}$ mm/m. The safety factor is $2.89.10^{-2}$, and the fatigue safety factor is $9.978.10^{-3}$ (Fig. 28). The damageability and resource reach minimum values at the same location.

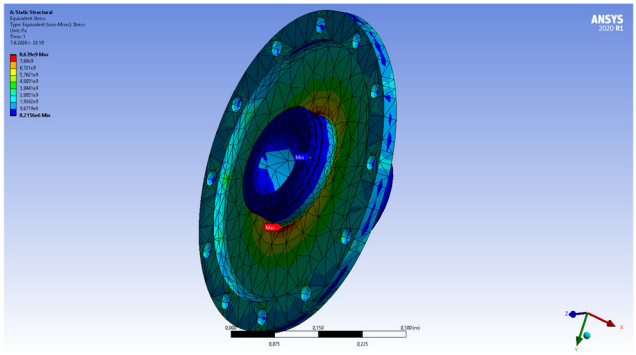


Figure 27:

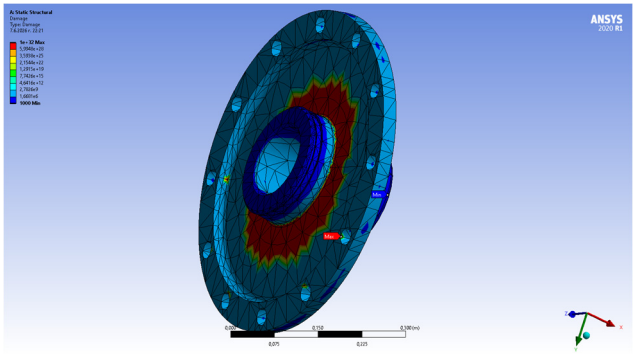


Figure 28:

5.5. Conclusions

- Obviously, the weakest section of the two coupling plates is the transition between the hub and the membrane on the inside of the plate;
- The maximum values of stress, respectively the safety factors, are not reached in starting process when the torque is the greatest, therefore the torque transmitted by the traction motor is not the most loaded element, mainly due to its relatively low value;

- The maximum values of all indicators of the state of the two elements are reached at maximum speed of movement, therefore the speed of movement, respectively the angular velocity of the rotating parts, is the most loaded indicator affecting the plates.
- When the train starts, the torque is maximum, the more loaded plate is the leading one, with the differences in the loads being minimal compared the driven one;
- When moving at maximum speed, the more loaded plate is the driven one, with the values of the loading indicators being significantly higher than those of the leading one.

5.6. Recommendations

- The elasticity of the material from which the plates are made can be increased, in order to withstand mainly the fatigue load, which is caused by their cyclic loading;
- In the comparative analyses conducted, it was found that the radius of curvature in the most endangered section does not have a significant impact on the results;
- Additional research is needed to improve the qualities of the connecting plates;
- The installation work performed when inspecting and replacing some of the elements must be carried out precisely to avoid frontal and radial runout when driving at high speeds.

5.7. Interference

The unit considered shows sufficiently high reliability in operation, but additional measures are needed to improve it, which include constructive and technological improvements in the design, as well as precision in performing assembly work on the unit.

REFERENCES

[1] Техническо описание на механичното оборудване на вагони серия 81-740, 81-741 – Метрополитен София, 2001.

[2] Стойчев Г, Метод на крайните елементи. Якостен и деформационен анализ. ТУ, С, 2000.

[3] Гуревич, В. И. и др., Руководство по эксплуатации вагонов метрополитена моделей 81-714.5 и 81-717.5, Московский метрополитен, 1993.

[4] Сементовский, Э. А., Устройство и ремонт электропоездов метрополитена, М., Транспорт, 1991.

[5] Добровольская, Э. М., Электропоезда метрополитена, М. Академия, 2003.

[6] Калиничев, В. П., Метрополитены, М. Транспорт, 1988.

SESSION B: RAILWAY ENGINEERING

Tests for the assessment of devices for securing semi-trailers during transport on a freight wagon

Dragan Milković[1*], Saša Radulović[2], Aleksandra Kostić Miličić[1], Goran Simić[2], Ognjen Obradović[2]

[1] University of Belgrade, Faculty of Mechanical Engineering, Rail Vehicle Department, Belgrade, Serbia

[2] University of Belgrade, Faculty of Mechanical Engineering, Laboratory of Rail Vehicle, Belgrade, Serbia

ARTICLE INFO

[*] **Correspondence:** dmilkovic@mas.bg.ac.rs

DOI: https://doi.org/10.46793/ET26.B02M **PAGES**: B11 – B16

ABSTRACT

The WAG TSI defines requirements for devices used to secure semi-trailers on freight wagons. The device must hold the semi-trailer in a safe position by withstanding longitudinal, lateral, and downward vertical forces in accordance with the design operating state. Following a 2019 accident, in which a semi-trailer detached during a severe storm and struck an oncoming passenger train, Denmark introduced controls and restrictions that have now been accepted and included in the WAG TSI. Additionally, the device must lock semi-trailers by withstanding upward vertical forces caused by crosswinds at the minimum crosswind stability; in the absence of this data, the upward force must be at least 85 kN. This paper presents real-world tests of a device for securing a semi-trailer to a freight wagon, along with an assessment of its conformity with EU regulations.

KEYWORDS

Semi-trailer, Kingpin, Securing device, Vertical upward force

1. INTRODUCTION

During rail transport, cargo must be adequately secured. Depending on the wagon type and freight characteristics, various measures and solutions are implemented to ensure safety in compliance with relevant European standards [1, 2]. In intermodal transport, particularly regarding semi-trailer shipments, ensuring the secure engagement of the kingpin within the locking device (hitch) is of paramount importance. Modern hitch designs must possess sufficient structural strength and feature clear, visually identifiable indicators of proper positioning and locking. Although the volume of semi-trailer and complete truck transport by rail has increased due to environmental benefits and economic efficiency, inadequate cargo securing has previously led to severe railway accidents. Specifically, due to insufficient locking forces under intense lateral loads, the connection disengaged, causing semi-trailers to encroach into the clearance gauge of the adjacent track and collide with oncoming trains. A critical example occurred in 2019 on the Great Belt Bridge in Denmark, resulting in eight fatalities, followed by a similar non-fatal incident in 2021. Consequently, the Danish authorities [3] banned semi-trailer transport during stormy weather, tightened operational controls, and strictly defined responsibilities for cargo securing. These events initiated a comprehensive re-evaluation of locking device designs; in addition to the conventional longitudinal, lateral, and downward vertical load cases, a new requirement was introduced for upward vertical resistance [4].

To evaluate the structural integrity and compliance of the locking device with the updated WAG TSI requirements, a comprehensive testing methodology was developed. The experimental framework combines laboratory-controlled, non-destructive load testing with real-world field trials under operational conditions. The primary objective is to verify that the locking mechanism can withstand the newly mandated upward vertical force of 85 kN, in addition to standard longitudinal and lateral operational loads. Conceptually, this device is engineered not only to ensure a secure connection but also to protect the payload (i.e., the semi-trailer) from excessive longitudinal accelerations during uncontrolled or improper shunting operations. This protection additionally accounts for standard braking and acceleration phases, during which lower levels of longitudinal acceleration occur.

To evaluate these conditions, buffing impact tests are conducted in accordance with standard [5]. Further investigations [6, 7] revealed that a complex interplay of extreme crosswinds, cyclic aerodynamic pressure waves generated during train passings, and vertical dynamic oscillations of the wagon traversing bridge spans induces a severe leverage effect. This phenomenon generates an upward vertical force capable of physically extracting the kingpin from its housing, provided the mechanism lacks positive mechanical locking across all degrees of freedom. In the absence of section-specific aerodynamic data, the locking device must withstand a minimum upward vertical lifting force of 85 kN to guarantee semi-trailer stability. Figure 1 illustrates the hitch devices, crash elements, and the air brake valve activator, which triggers emergency braking in the event of hitch displacement resulting from crash element deformation.

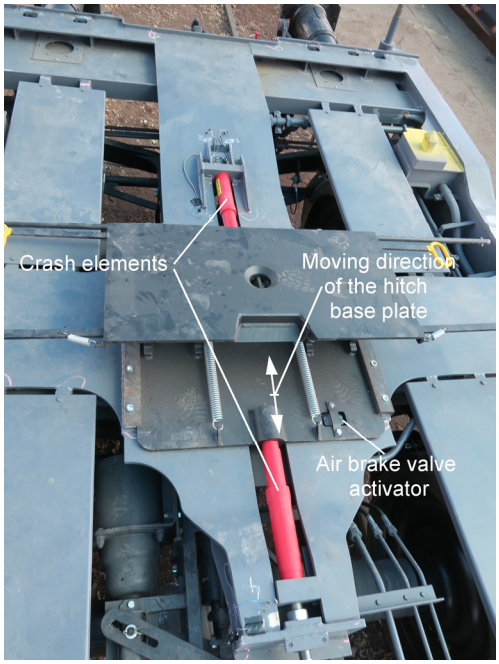


Figure 1: Hitch on the pocket wagon [8]

2. TEST OF THE POCKET WAGON LOADED WITH A SEMI-TRAILER

2.1. Buffing Impact Test

In accordance with structural strength requirements for freight wagons, buffing impact tests shall be conducted using a fully loaded configuration with a semi-trailer weighing 40 tons. Given that the vehicle belongs to the F-II category for combined transport, it is specifically designed for careful shunting operations. Consequently, the baseline certification tests were performed at impact speeds up to 7 km/h. As the partner (impact) wagon was used Eaos wagon loaded to a total mass of 80 tons (Figure 2). During these baseline trials, no structural damage or permanent deformations were observed on either the wagon frame or the trailer hitch mechanism. The peak longitudinal accelerations recorded on both the wagon and the semi-trailer chassis stabilized at approximately 1g.

Within the scope of research regarding the development of sacrificial crash elements [8], further experimental trials were carried out at higher impact velocities to evaluate cargo protection under improper shunting conditions. The integrated crash elements within the trailer hitch are designed to deform plastically and dissipate kinetic energy when shunting speeds exceed 7 km/h, thereby capping the acceleration transmitted to the cargo at a maximum threshold of 2g. To determine the ultimate structural margins of the coupling, extreme testing scenarios were conducted using compromised or rigid elements. Under these severe conditions, the longitudinal interface between the trailer hitch and the wagon was subjected to extreme accelerations reaching up to 3g. Despite these highly elevated dynamic loads and the resulting severe pitching moments, the structural integrity of the connection was maintained, and no unlocking or slippage of the kingpin from the trailer hitch occurred, demonstrating compliance with the safety margins [5].

When a pocket wagon undergoes a buffing impact, the primary load is longitudinal. However, because the semi-trailer's center of gravity is located high above the hitch level, this sudden deceleration generates a severe pitching moment (rotational leverage). This moment manifests as a structural reaction at the coupling point. During buffing, impact kinetic energy transforms into a simultaneous dynamic upward vertical force and a high bending moment on the locking jaws.

Based on the conducted analyses and operational experience, the most critical scenario occurs under strong crosswinds combined with an empty semi-trailer. This condition stands in stark contrast to the scope of the conventional buffing impact test, where the semi-trailer is evaluated under a maximum gross weight of 40 tons. Consequently, the supplementary tests applying an 85 kN upward vertical force to the hitch connection are designed to simulate the quasi-static equivalent load induced by aerodynamic effects and the vertical dynamic oscillations of a wagon carrying an unladen semi-trailer [6, 7].



Figure 2: Buffing impact test [8]

2.2. Tensile Vertical Upward Force Test

The validation test on the Sdggmrss wagon involved applying a vertical tensile load of 85 kN to a secured, ECE Regulation 55-compliant H50-type kingpin, with load monitoring and testing continuing until the wagon lifted from the rails [7]. The procedure, conducted with the securing device in the mandated position, verified that the locking capacity exceeds the required minimum [1, 7].



Figure 3: Tensile upward force [9]

To ensure measurement accuracy, the dead weight of the test adapter was excluded from the tensile force data by taring the load cell system while the adapter was suspended but not yet locked [9]. The experimental procedures were conducted at the specific operational positions designated for the securing device. Prior to and following each test cycle, a thorough visual inspection of the locking mechanism, locking pin, and securing device was performed to verify proper functionality and ensure the absence of visible damage or permanent deformation. To guarantee reproducibility, the test was repeated three times for each designated configuration. Vertical tensile load tests were systematically executed at three distinct securing device heights: 880 mm, 980 mm, and 1130 mm (Figure 4).



Figure 4: Hitch positions during testing (880, 980,1130 mm) [9]

3. TEST RESULTS

3.1. Buffing Impact Test Results

Category A long-stroke buffers with a 150 mm stroke were tested on the wagon, with measurements, for this purpose, focusing on acceleration responses on the wagon and semi-trailer structure during impacts. Tests were conducted with the trailer hitch base plate in both upper and lower positions, using sensors A1–A3 as shown in Figure 5.

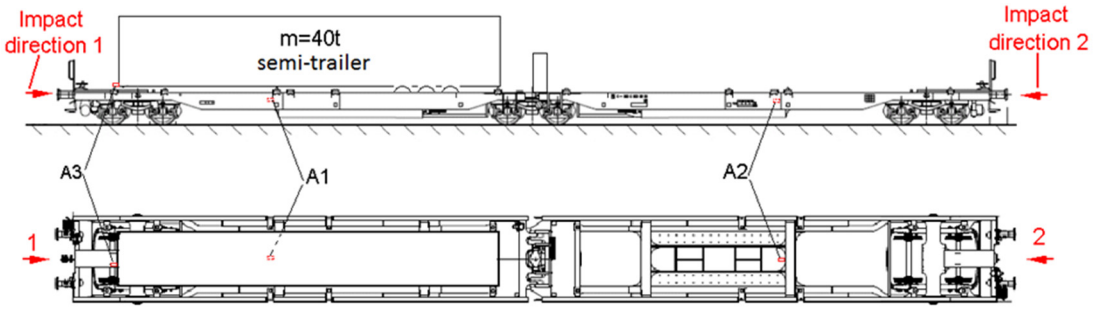


Figure 5: Acceleration transducers displacement [8]

The diagram in Figure 6 illustrates the measured accelerations as a function of the wagon's impact speed. The experimental data indicate that at standard shunting speeds of up to 7 km/h, the recorded accelerations remain at approximately 1.0 g, while elevated impact speeds of around 10 km/h result in values between 2.0 g and 3.0 g [8]. Furthermore, the results demonstrate that variation in hitch height has a negligible effect on the intensity of the measured accelerations.

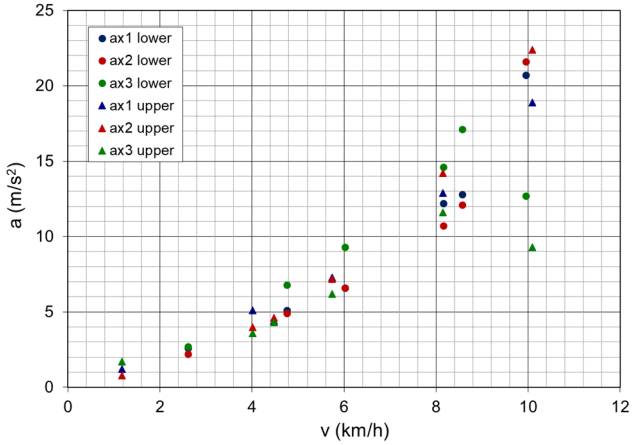


Figure 6: Accelerations on the wagon and semi-trailer vs. impact speed [8]

3.2. Tensile Vertical Upward Test Results

Vertical tensile load tests were systematically conducted across three distinct securing device heights: 880 mm, 980 mm, and 1130 mm. For each configuration, three experimental trials were executed to ensure reproducibility. During each test run, the tensile force was gradually increased until reaching a maximum recorded load of approximately 114 kN, which corresponded to the physical lifting of the wagon frame from the rails. This peak load was held in equilibrium for approximately 10 to 20 seconds. The experimental results demonstrated a highly repeatable quasi-static mechanical response of the securing device across all three tested positions, with no evidence of autonomous unlocking or anomalous structural behavior under load.

As a representative example, Figure 7 illustrates the force-time history recorded at the 1130 mm hitch configuration. No autonomous disengagement or loss of operational functionality was observed either during or after the completion of the testing cycles [9].

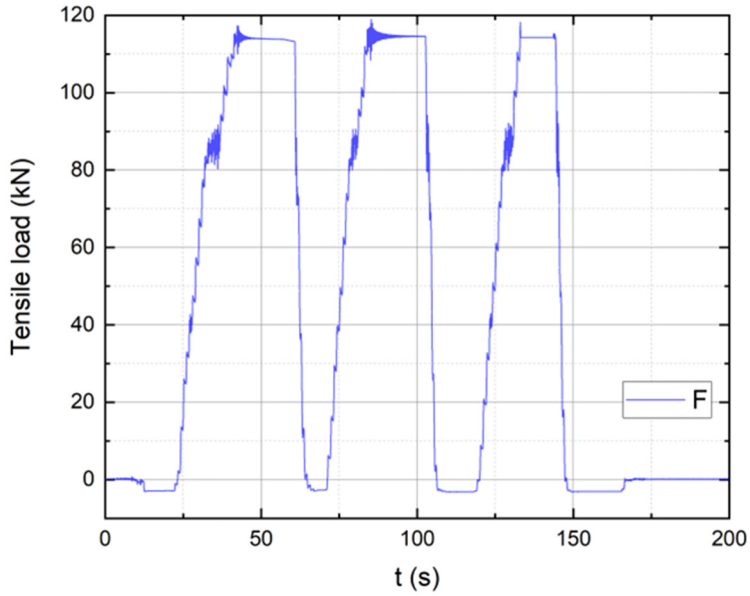


Figure 7: Vertical tensile force vs. time for test at hitch position 1130 mm [9]

Figure 8 illustrates the post-test condition of the kingpin and locking mechanism, confirming the preservation of structural integrity with no visible damage or permanent deformation. Inspection revealed only minor, superficial paint degradation, indicating that the component's functionality was not compromised.



Figure 8: Kingpin and locking mechanism after testing –no visible damage [9]

Furthermore, the locking device must incorporate a robust mechanical-visual indicator system designed to unequivocally verify full mechanism engagement. Validation of this safety subsystem requires intentional obstruction of the kingpin entry to confirm that the external visual indicators fail-safe to a distinct "Not Locked" status. The operability, contrast, and legibility of these mechanical indicators under degraded light levels and adverse environmental conditions must be optimized to ensure that trackside inspection personnel can reliably and efficiently verify locking integrity.

4. CONCLUSIONS

Based on buffing impact and vertical tensile tests [8, 9], the semi-trailer securing device on the Sdggmrss wagon withstood loads exceeding mandated thresholds [1, 4, 7] with no autonomous unlocking or damage observed. These validation results provide technical assurance that, with proper operational locking, the tested type of securing device will prevent the recurrence of similar structural failures.

ACKNOWLEDGEMENTS

The results presented in this paper are the results of research supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia under the Agreements No. 451-03-34/2026-03/200105 and No. 451-03-33/2026-03/200105, dated 05.02.2026.

REFERENCES

[1] European Union Agency for Railways (ERA), "Commission Implementing Regulation (EU) 2023/1694 amending Regulation (EU) No 321/2013 concerning the technical specification for interoperability relating to the 'rolling stock — freight wagons' subsystem (WAG TSI)," Official Journal of the European Union, Brussels, Belgium, (2023)

[2] European Committee for Standardization (CEN), "Railway applications - Cargo securing on rail wagons - Requirements and general principles," EN 16860:2020, Brussels, Belgium, (2020)

[3] Accident Investigation Board Denmark (AIB DK), "Safety Alert: Structural failure and locking limits of king-pin pocket wagon hitches following the Great Belt Bridge accident," Danish Transport Authority (Trafikstyrelsen), Copenhagen, Denmark, Tech. Rep. NSA-DK-2021-01, (2021)

[4] European Union Agency for Railways (ERA), "NSA-DK Risk Mitigation Measures: Revised technical requirements on intermodal transport with semi-trailers on pocket wagons," ERA Safety Assessment Unit, Valenciennes, France, Rep. ERA-SAF-2021-R04, (2021)

[5] European Committee for Standardization (CEN), "Railway applications - Structural requirements of railway vehicle bodies - Part 2: Freight wagons," EN 12663-2:2010+A1:2023, Brussels, Belgium, (2023)

[6] A. Kwasiborska and I. Grabarek, "Analysis of safety requirements for securing the semi-trailer on the freight wagon under crosswind conditions," Energies, Vol. 14(20), pp. 6539-, (2021)

[7] ERA Technical Document on the procedures to demonstrate compliance with the requirements in point 4.2.2.4 of the WAG TSI on securing semi-trailers: ERA/TD-2025/SECURING OF SEMI TRAILERS version 1.0 (released on 11.9.2025)

[8] D. Milković et al. Buffing impact test of Sdggmrss wagon, model S06, Test report No: LSV-I-14/23, University of Belgrade, Laboratory of Rail Vehicles, Belgrade, (2023)

[9] D. Milković et al. Vertical tensile load test of king-pin securing device type MAZ 80800 of Sdggmrss wagon, model S01, Test report No: LSV-I-2/26, University of Belgrade, Laboratory of Rail Vehicles, Belgrade, (2026)

SESSION B: RAILWAY ENGINEERING

Strength analysis of an electric locomotive transmission mechanism

Dobrinka Atmadzhova[1*], Vanyo Ralev[2]

[1] Todor Kableshkov University of Transport, Faculty of Machinery and Construction Technologies in Transport, Department of Transport Equipment, Sofia, Bulgaria

[2] Todor Kableshkov University of Transport, Faculty of Transport Management, Department of Machine Elements and Chemistry, Sofia, Bulgaria

ARTICLE INFO

[*] **Correspondence:** dba65@abv.bg

DOI: https://doi.org/10.46793/ET26.B03A **PAGES**: B17 – B27

ABSTRACT

The transmission of traction electric motor torque to the electric locomotive wheelset is carried out through a wheelset reducer with cylindrical gears with straight or inclined teeth. The wheelset reducer is one of the most important units of a modern electric locomotive. It bears significant alternating loads when the locomotive starts, accelerates and stops. It is of paramount importance to choose the right drive option, which the locomotive movement and dynamics of its running depend on. At high running speed, the system must be reliable in terms of the oscillations of locomotive unstrung masses. The thirty-year operation of electric locomotives in Bulgaria has shown a number of shortcomings of wheel reducers that cause difficulties with operation and repair activities in locomotive depots. The paper presents a strength analysis of a transmission mechanism, that has shown low operational reliability leading to locomotive derailment. The results of analysis have indicated the areas, which need monitoring in operation to prevent the occurrence of incidents and accidents due to the transmission mechanism operation.

KEYWORDS

Railway, electric locomotive, transmission mechanism, strength analysis

1. INTRODUCTION

The mechanisms that transmit drive torque from the traction electric motor shaft or the output shaft of a mechanical, hydraulic or other transmission system to wheels, without violating the freedom of movement necessary for relative displacement of wheels in regard to the bogie frame and of the bogie in regard to the frame, are marked as transmission mechanisms [1]. Their structural elements are reducers, shafts, joints, supports, couplings, etc.

The locomotive technical and economic parameters significantly depend on transmission mechanisms. This is the reason for a number of requirements put to them. Summarising, the more significant ones are expressed in the following:

- the transmission mechanism operation should not be accompanied by appearance of high-value dynamic forces and should not limit the operating modes;
- the mechanism reliability should be high and meet the requirements for economic efficiency;
- it should provide high operating economy;
- it should not possibly reduce the space intended to install the traction motor;
- the design should be simplified and and should allow the arrangement of traction motors in a way to be accessible for inspection and repair and protected from snow, moisture and dust.

The transmission mechanisms can be divided into two groups depending on whether they drive a separate wheelset or a group of wheelsets of mechanisms for individual drive and for group drive. The constructive implementation of transmission mechanism depends on whether it transmits torque from the shaft to a traction electric motor, or from the shaft at the output of a hydraulic or mechanical transmission system, therefore the transmission mechanisms are divided into two groups on this basis: transmission mechanisms for drive with a traction electric motor and transmission mechanisms with hydraulic or mechanical transmission systems.

The construction practice has shown that usually the transmission mechanisms for drive traction electric motors belong to the group of mechanisms for individual drive, and the transmission mechanisms in hydraulic or mechanical transmission systems – to mechanisms for group drive. The opposite is rare, but the group drive of wheels with an electric motor is perspective for development.

The kinematics scheme and design of transmission mechanisms depend on the method of mounting the traction motor on a bogie or, as it is accepted to call that, on the method of traction motor suspension. Due to the above mentioned, the method of suspension is taken as a classification feature and based on that, the transmission mechanisms for drive wheelsets by traction motors are divided into two groups: mechanisms with a support-axle suspended traction motor and mechanisms with a support-arm suspended traction motor.

The locomotive transmission mechanisms transmit torque from the locomotive engine to the wheelsets, driven by it (when the engine is in traction mode) and vice versa (when the engine is used to stop the locomotive).

The type and structure of power transmission is determined by the locomotive engine type, power and properties, the method of its suspension, its location in relation to the locomotive longitudinal axis and type, purpose and maximum speed of the locomotive itself [2].

2. CLASSIFICATION OF TRANSMISSION MECHANISMS

The locomotive power transmissions can be classified as follows [2 - 7]:

1. According to the number of wheelsets, which are simultaneously driven, they are power transmissions for individual drive of locomotive wheelsets and power transmissions for group drive of locomotive wheelsets.

2. According to their regulating properties, they are non-adjustable (i.e. constant gear ratio) power transmissions and adjustable (i.e. with variable gear ratio) power transmissions.

The non-adjustable gears are the lightest, simplest and cheapest locomotive power transmissions (in terms of manufacturing, operation and repair). They belong to the category of mechanical power transmissions, since they transmit torque through gear meshing , lever systems or power chains.

The aadjustable power transmissions are relatively complex and expensive, but unlike non-adjustable ones, they allow the adjustment of locomotive engine features according to the changing locomotive operating conditions. They are: mechanical (created on the basis of a multi-speed gearbox with gear meshing); hydraulic and combined.

Usually, the locomotive reversion is also implemented through the adjustable power transmissions, while the direction of its engine rotation remains unchanged.

3. According to the value of unstrung masses there are three types of transmission mechanisms that can be distinguished:

- a first class transmission mechanism with a support-arm traction motor and a support-axle reducer;
- a second class transmission mechanism with a support-arm traction motor and a support-axle reducer;
- a third class transmission mechanism with a support-arm traction motor and a support-arm reducer.

In the first class transmission mechanisms, the traction motor is supported on one side by means of an elastic element on the bogie frame, and on the other side it lies on the wheelset by means of motor-axle bearings. The pinion is mounted on the traction motor shaft, thus forming a pair of gears with the large gear wheel mounted on the wheelset.

A variety of support-axle suspension is the support-center suspension, where the traction motor lies on a hollow shaft by means of sliding bearings, which in their turn are connected to the wheel by means of elastic elements.

Such is the suspension of locomotives of series 06, 07 and 51operated in Bulgaria.

With the second class of transmission mechanisms, the traction motor transmits entire mass to the bogie frame, and the reduction gearbox transmits its mass through the axle of weelset and thus it is the only one that belongs to the unsprung parts.

All electric mainline locomotives "Skoda" in BDZ have a second class of transmission mechanism, as well as series 46.

With the third class of transmission mechanisms, the traction motor, together with the wheel reducer, transmit their mass to the bogie frame thus joining to the sprung parts. In its turn, the third-class transmission mechanism is divided into two groups:

- First group – the movement is transmitted by the electric motor to the large gear wheel, which is mounted on a hollow shaft that transmits the drive torque to the wheelset using a clutch. This is called "Third class transmission mechanism" with one-sided transmission of torque, but it is where oscillations appear.
- Second group – the movement is transmitted by the electric motor to the large gear wheel, which is mounted on a hollow shaft that transmits the drive torque using a clutch on the both sides of wheelset. This is called "Third class

transmission mechanism" with two-sided transmission of torque. In this case, washers on the hollow shaft are out of phase by 90°, which minimizes the oscillations in the entire system.

Another structure of transmission mechanism with a hollow shaft is the “Jacque main” system where the torque transmission is one-sided (as in Siemens). The large gear wheel of the wheelset reducer is fixedly mounted on a shaped sleeve, rotating with it on the bearings of the reducer. Two pins are placed on the shaped sleeve, located to each other at 180°. Through them, the torque is transmitted to the driven ring, in turn connected to the hollow shaft by means of pins located at 90° to the above-mentioned two pins. On the other side, the shaft is also connected to the drive ring by the same pins, and the ring transmits the torque to the wheel by means of pins with rubber-metal hinges mounted on them.

In mechanisms equipped with an Alstom type clutch, the torque transmission is bilateral.

The driven pins are connected to the wheel by 4 bolts and to relieve the shear force, the pin is tightly mounted in the hole of wheel.

In mechanisms equipped with an Alstom type clutch, the torque transmission is bilateral.

The driven pins are connected to wheel by 4 bolts and to relieve the shear forces, the pin is tightly mounted in the hole of wheel.

First, the spacer ring is mounted on the pin, and then the rubber bushing is pressed into the head of the rein. After the rubber bushing fits snugly, a pressing washer is placed reliably fastened with a bolt and a washer. The rubber parts of bushings are relieved of the compressive forces.

The drive pins are pressed into the hollow shaft disk and are fixed with a special pin. The connection of drive reins with the drive pins and the floating frame is carried out in the same way as with the drive pins. The torque of traction motor through the drive pins is transmitted to the driven pins, through the reins and the floating frame suspended on them. This frame can be shifted with disturbances between the wheelset and hollow shaft.

The connection of the reins with pins and the floating frame is carried out by eight hinges, filled with rubber bushings, reinforced inside and outside with steel bushings. The inner bushing is pressed onto the pin, and the outer one is pressed to the rein lug or floating frame. The disturbances in coupling appear on the account of rubber deformation. However, a hinge of this type can provide simultaneous rotation and tilting in any direction in the area of small deformations. The great advantage of this hinge is that it does not need lubrication, since it has no friction parts, thanks to which it works silently.

The rubber deformation determines the angle of rotation of the reins relative to the spokes, the radial load and sagging of the locomotive sprung parts. With simultaneous displaced centers of the hollow shaft and wheelset up to 30 mm, the angle of rotation reaches 8°.

3. ADVANTAGES AND DISADVANTAGES OF TRANSMISSION MECHANISMS

3.1. First Class Transmission Mechanism

A significant disadvantage of the support-axle suspension is the increased weight of unsprung parts and the dynamic impact of raiway track on the traction motor and transmission mechanism.

The strengthening of motor on wheelset limits the application of this type of suspension only to locomotives with average speeds of up to 100 km/h. For this reason, it is used only in locomotives intended for freight service.

3.2. Second Class Transmission Mechanism

The advantage of the support-arm suspension is that the traction motor is protected from vibrations by bogie spring suspension and the mass of unsprung parts on one wheelset is significantly reduced. However, it has a more complex design than the support-axle suspension.

3.3. Third Class Transmission Mechanism

The advantage of the third class transmission mechanism is that the traction motor, together with the wheelset reducer, transmits their mass to the bogie frame and refer to sprung parts of the system. With one-sided transmission of torque, vibrations and torsional moment appear in the mechanism, while with two-sided transmission of torque for one revolution of wheel, the floating frame center makes two revolutions. This is the main disadvantage of the Alstom type coupling.

4. DAMAGES TO THE TRANSMISSION MECHANISMS OF ELECTRIC LOCOMOTIVES IN BULGARIA

The electric locomotives of “Bulgarian State Railways” (BDZ), series 41, 42, 43, 43P are characterized by transmission of torque from the traction electric motor to wheelsets carried out by means of a single-stage wheelset reducer with straight-teeth cylindrical gears. The wheel reducer is suspended axially hinged to the bogie frame and is fixed to the wheelset. The torque from the traction motor to the reducer is transmitted by means of an original cardan joint of

"Hadji" type, manufactured by "Skoda" – Pilsen factories, where the torsion-cardan shaft acts as a damper of torsional vibrations. There is no fundamental difference in the design of wheelset reducers in the each locomotive series.

For the thirty-year period of operation of “Skoda” electric locomotives in Bulgaria, their wheelset gearboxes have shown a number of shortcomings that create difficulties in operation and repairs at locomotive depots. Some of the transmission mechanism elements characterized with unsatisfactory reliability were examined paper [8].

In 1979-1980, scientists from the Todor Kableshkov University of Transport and Technical University of Sofia developed various design options to readjust the connection of the pinion with the shaft. The developed suggestions covered practically all possible cases of cylindrical and conical press joints depending on the degree of using old parts. The main problem with the options was the welding of the splined shaft, made of chrome-nickel steel, which required a special technological mode and electrodes and mandatory heat treatment after that operation. One of the options provided extension of the spline joint durability by means of its bilateral centering in a manner similar to the design solution for the first delivery of electric locomotives of series 43. The most promising suggestion was to readjust the spline joint into a cylindrical-pressed one using shafts identical to those in locomotives of series 43 and и и 43^{P}, delivered as spare parts or manufactured in Bulgaria [8]. It was proposed to use the old pinions, including those scrapped because of damaged bearing journals or splines.

The transmission mechanism of electric locomotives of series 43/44 and 42/45 have low operational reliability. The damage occurs mainly in the cardan coupling, gears, gearbox lubrication, noise, vibrations, etc.

The damage in wheelset reducers is caused by the exceptional influence of vibrations [9, 10]. They have been studied experimentally, and it has been found that the obtained spectra are close in configuration and can be presented in one generalized case. The amplitudes located in the frequency range of about 4500 Hz belong to the reducer box. Analogous results have been obtained also for the traction gears of electric rolling stock. All main types of damage in gears lead to dynamic load on gears and bearings, and to vibrations in the reducer, respectively, as 45% of the rolling bearings are not repairable.

The structure of wheelset reducer of electric locomotive series 44 and 45 is the same for both series, but differs in the gear ratio, which is achieved by equipping them with different gears at a constant center distance. The largest amount of damage has been found in the parts of gearbox housing, square, bearings of driving wheel (pinion) and driven wheel (gear). These are the main transmission mechanism components that are subjected to both dynamic load and wear in operation.

Damage in gears and gear meshing leads to appearance of oscillations of different nature. These may be kinematic forces caused by deviations in teeth shape in the contact area, shock impacts determining the change of clearance in the gear meshing area, and friction forces between the contact surfaces.

The monitoring on locomotives owned by BDZ – Freight Transportation and BDZ – Passenger Transportation carried out in 2013-2025 showed the most common failures of transmission mechanisms, presented by percentage in Table 1.

Table 1: Damage to the elements of transmission mechanisms of the second and third classes

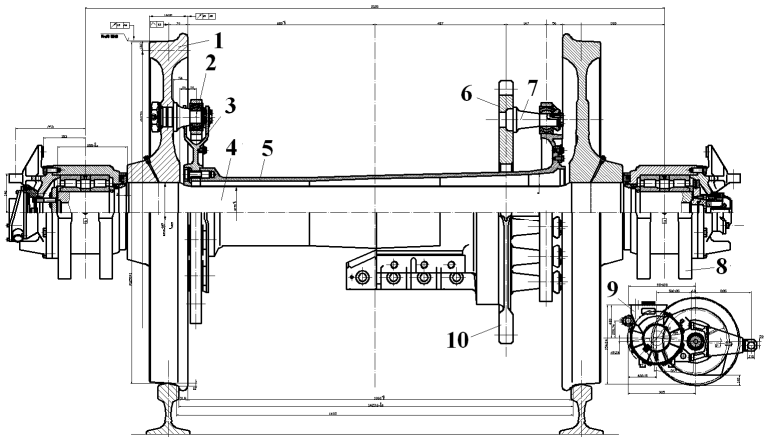
Elements	**Electric locomotives with transition mechanisms**		
	Of second class		**Of third class with Alstom type clutch**
	Series 43 and 43^{p}	**Series 44, 45, 46**	
Reducer housing	25,22%	27,37%	15%
Transmission square	29,98%	28,42%	-
Pinion bearings	16,74%	18,95%	18%
Gear bearings	1,56%	2,11%	5%
Pinion	4,02%	8,42%	8%
Gear	9,38%	8,42%	8%
Pinion flange	6,03%	0%	7%
Pinion shaft	1,34%	1,05%	2%
Traction motor	1,78%	0%	0%
Clutch	-	-	30%
Other	6,70%	4,21%	7%

From the observation done, the highest percentages of failures were found:

- for transmission mechanisms of the second class – the gearbox housing and the square;
- for transmission mechanisms of the third class – damage to the clutch.

5. SELECTION OF A TRANSMISSION MECHANISM TO BE STUDIED

The present study considers a third-class transmission mechanism with a support-arm suspended traction motor and a support-arm suspended reducer (Figure 1) and an Alstom type clutch (Figure 2), which has the most favourable dynamic indicators. One of the main features is that both the engine and the reducer belong to the sprung parts of the locomotive.



1- wheel; 2- “hollow shaft-wheel” connection; 3 – shift fork; 4 - axle; 5- hollow shaft; 6 - large gear wheel; 7- “gear wheel-hollow shaft” connection; 8 - bushing box; 9 - reducer; 10 - reducer box.

Figure 1: Third class transmission mechanism with ALSTOM type clutch.

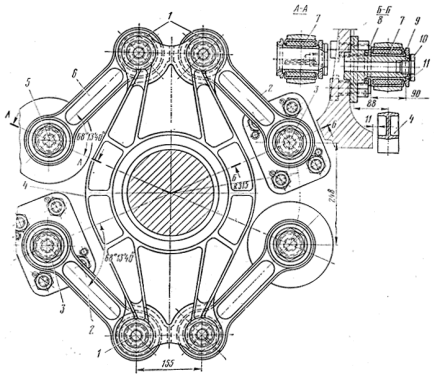


Figure 2: Alstom type clutch [3-7]

1- rubber-metal joints; 2 - guides; 3 - driven pins; 4 - floating washer; 5 –driving pins; 6 - connecting rods; 7 - rubber bushing; 8 - spacer ring; 9 - washer; 10 - bolt; 11 - securing plate.

To calculate the strength and deformation features of the transmission mechanism main elements, it is necessary to indicate the technical specifications of the locomotive type where they will be installed.

Basic technical data :Wheel formula –mm Bo - Bo; Power supply system – 25 kV/50 Hz; Track gauge – 1435 mm; Maximum speed – 175 km/h; Traction power (in the rolling circle of wheels) - 4 MW; Traction force at start-up – 260 kN; Traction force at 73 km/h – 196 kN; Braking force with electrodynamic (regenerative) brake – 118 kN; Total mass (traction mass) – 80 t; Minimum curve entry radius – 80 m; Mass of the locomotive superstructure (body) – 49 t; Mass of the bogies - 31 t; Length between buffers – 19380 mm; Distance between end axles – 12740 mm; Distance between center bolts – 9940 mm; Distance between axles in the bogie – 2800 mm; Wheel diameter – new wheels – 1250 mm; Wheel diameter – worn wheels – 1170 mm; Max. height with pantograph removed (above head rail) – 4590 mm; Maximum width – 3150 mm; Voltage/frequency for train heating – 1.5 kV /50 Hz; Gear ratio of axle gearboxes 22:106; Traction motors with air cooling 4 pcs. of 1020 kW each; Air compressor-screw /oil cooling 3560 l/min at 3000 l/min; Brake mass – brake mode – R 92 t; Brake mass – brake mode – P 63 t.

6. SELECTION OF A TRANSMISSION MECHANISM TO BE STUDIED

The strength and deformation analysis of a transmission mechanism will be done using SolidWorks Professional software. One of the modules applied to develop parts is Cosmos Works (in the new versions it is called SolidWorks Simulation).

The modeling of a third-class transmission mechanism with an Alstom type clutch is shown in Figure 3.

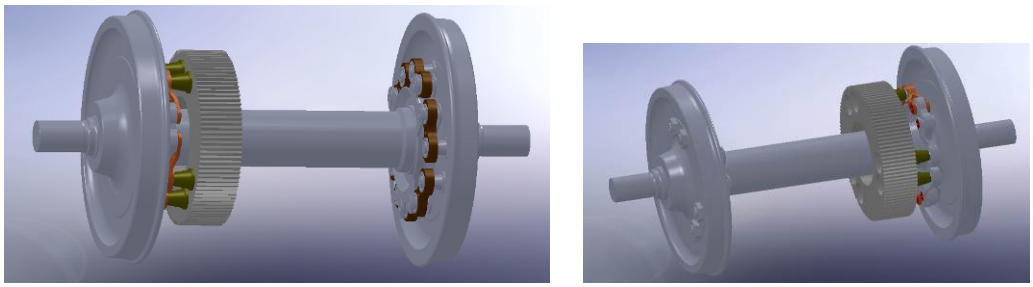


Figure 3: Model of the assembled mechanism

Its main elements, which have been studied in this paper, are the wheelset with a wheel, an axle, a hollow shaft; a hub; a large gearwheel, a shift fork and an eight-like part.

6.1. Strength-Deformation Analysis of a Wheel

The wheel model is shown in Figure 4.

The wheel is loaded by torque in the area of pressing with the axle and is stuck in the holes for attachment to the transmission mechanism. The part is covered with Solid Mesh (Figure 5).



Figure 4: Model of a transmission mechanism wheel



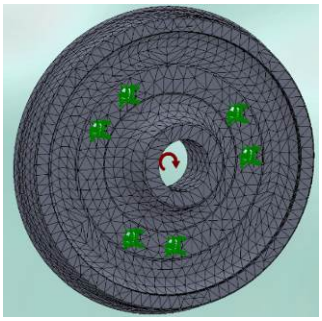


Figure 5: Solid Mesh

The results of study on the wheel are presented in Table 2.

Table 2: Results of the study on wheel

Name	Type	Max	Location
Stress1	VON: von Mises Stress	435.184 MPa Node: 8074	(-765.64 mm, -277.8 mm, -196.9 mm)
Displacement1	URES: Resultant Displacement	0.676519 mm Node: 22315	(-694.07 mm, -434.59 mm, 461.3 mm)
Strain1	ESTRN: Equivalent Strain	0.00146932 Element: 3572	(-749.68 mm, -14.07 mm, 404.49 mm)

Figure 6 shows the distribution of stresses on the wheel.The maximum value ofloading isin the area of holes connecting the wheel to the drive mechanism,the maximum load value is435.184 MPa.

Figure 7 shows the wheel deformationbyloading, andFigure8 presentsthe distribution of the relative elongation coefficient of the wheel due to loading.

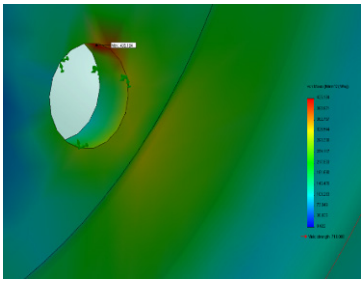


Figure 6: Stress distribution on the wheel

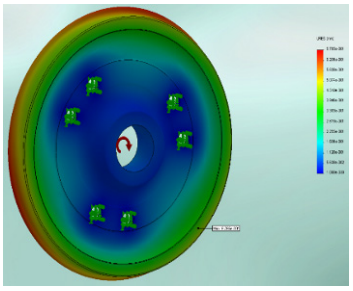


Figure 7: Deformation of the wheel by loading

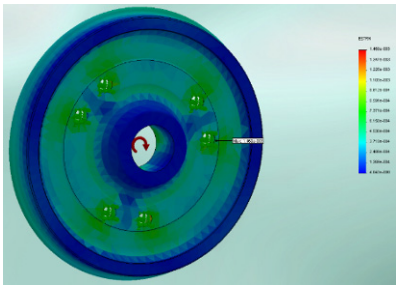


Figure 8: Distribution of the relative elongation coefficient of the wheel

6.2. Strength and Deformation Analysis of an Axle

The axle model is shown in Figure 9. The axle is loaded by torque in the central area and by vertical and horizontal forces in the axle journals, and locking occurs in the parts under hub.

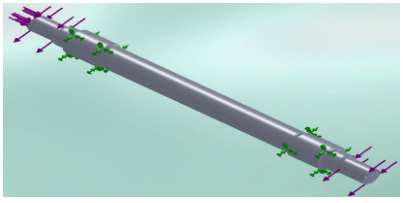


Figure 9: Axle model and load

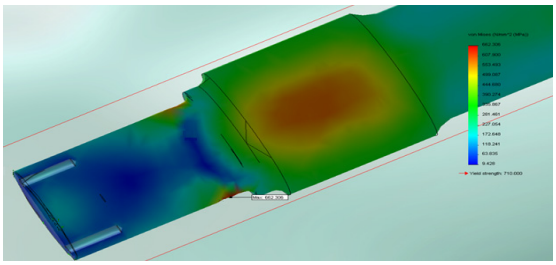


Figure 10: Areas of maximum stress

Figure 10 shows the areas of maximum stress along the axle.

It is observed that the concentration of stresses appears in the transition area between the journal of axle assembly and in the area of wheel pressing-fix, as well as in the area of pressing itself. The maximum load value is 662.306 MPa.

Figure 11 shows the axle deformationby the applied load, and Figure 12 presentsthe distribution of the relative elongation coefficient.

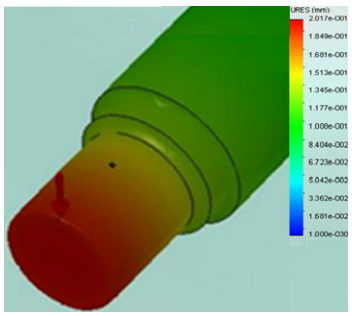


Figure 11: Deformation of the axle by load

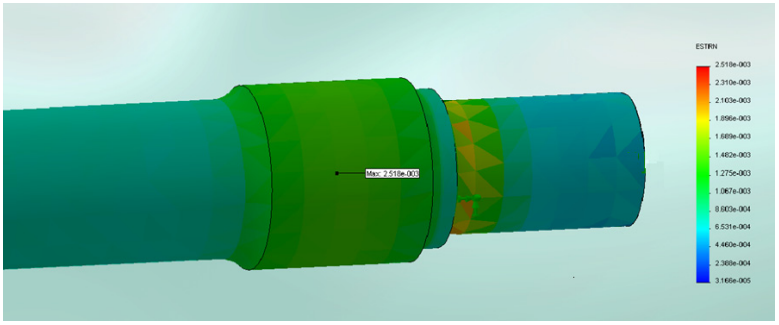


Figure 12: Distribution of the relative elongation coefficient

The results of the axle study are presented in Table 3.

Table 3: Results of the study on axle

Name	Type	Max	Location
Stress1	VON: von Mises Stress	662.306 MPa Node: 902	(-977.943 mm, -80 mm, 0 mm)
Displacement1	URES: Resultant Displacement	0.201699 mm Node: 18269	(767.464 mm, 80.84 mm, -50.07 mm)
Strain1	ESTRN: Equivalent Strain	0.00251756 Element: 9325	(827.235 mm, -0.438461 mm, 7.36 mm)

6.3. Strength and Deformation Analysis of a Hollow Shaft

The model of the hollow shaft is shown in Figure 13.

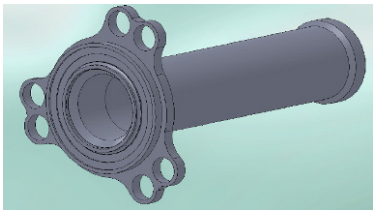


Figure 13: Model of hollow shaft

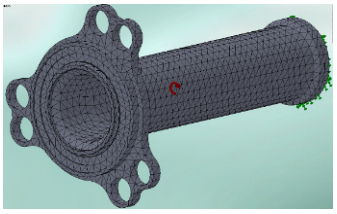


Figure 14: Loading and Solid Mesh

The loading on hollow shaft is carried out by torque concentrated in the hub area, and the locking occurs at the other shaft end. The part is covered with Solid Mesh (Figure 14).

The distribution of stresses along hollow shaft and the concentration of stresses are in the hub area between the lugs for the holes (Figure 15). The maximum load value is 597.821 MPa.

Figure 16 shows the deformation of hollow shaft due to the applied load, andFigure17 shows the distribution of the relative elongation coefficient.

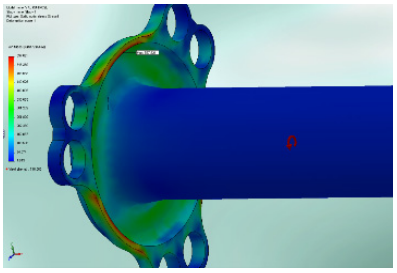


Figure 15: Areas of the maximum stress

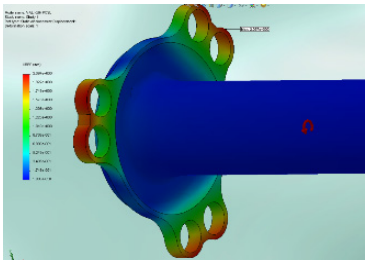


Figure 16: Deformation of the hollow shaft

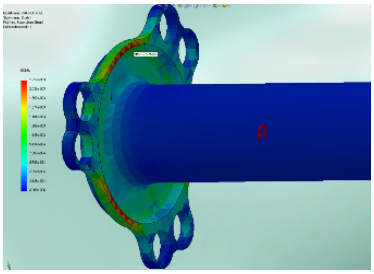


Figure 17: Distribution of the coefficient of relative elongation

The results of study on the hollow shaft are presented in Table 4.

Table 4: Results of study on the hollow shaft

Name	Type	Max	Location
Stress1	VON: von Mises Stress	597.821 MPa Node: 3245	(-1237.3 mm, 259.9 mm, 0.013 mm)
Displacement1	URES: Resultant Displacement	2.09714 mm Node: 19926	(-1239.4 mm, 153.5 mm, -369.2 mm)
Strain1	ESTRN: Equivalent Strain	0.002226 Element: 9885	(-1240 mm, 263.2 mm, 12.5 mm)

6.4. Strength and Deformation Analysis of a Hub

The hub model is shown in Figure 18. The hub is loaded by the torque concentrated in the hub central part, and the locking occurs in lug holes. The part is covered with Solid Mesh (Figure 19).

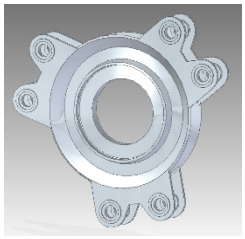


Figure 18: Hub model

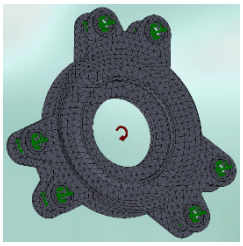


Figure 19: Hub loading and Solid Mesh

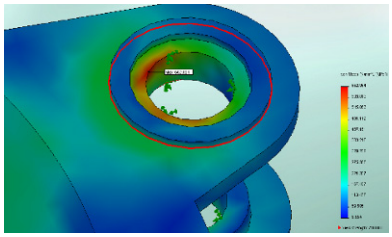


Figure 20: Areas of maximum stress in the hub

Figure 20 shows the distribution of stresses in the hub. The stress concentration is in the area of holes in the lugs. The maximum load value is 652.964 MPa.Figure 21 shows the hub deformationby loading, andFigure 22 shows the distribution of relative elongation coefficient.

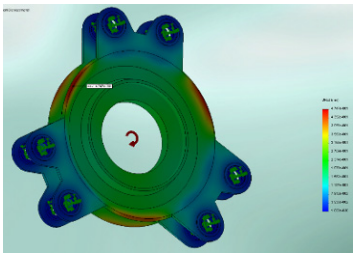


Figure 21: Deformed hub state state

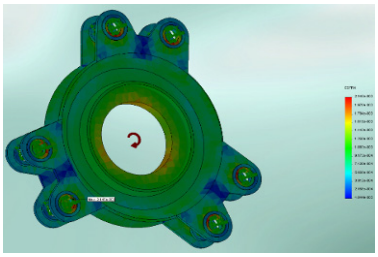


Figure 22: Distribution of the coefficient of relative elongation

The results of the hub study are presented in Table 5.

Table 5: Results of the hub study

Name	Type	Max	Location
Stress1	VON: von Mises Stress	652.964 MPa Node: 2285	(-21.672 mm, 47 mm, -330.51 mm)
Displacement1	URES: Resultant Displacement	0.474714 mm Node: 519	(-176.693 mm, -47.2506 mm, -192.6 mm)
Strain1	ESTRN: Equivalent Strain	0.00214 Element: 8301	(164.355 mm, -44.51 mm, -272.18 mm)

6.5. Strength and Deformation Analysis of a Gear Wheel



Figure 23: Gear wheel

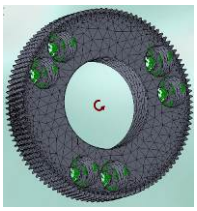


Figure 24: Load on the gear and Solid Mesh

The model of a large gear wheel is shown in Figure 23.

The gear wheel is loaded by the torque concentrated in the wheel central part, and locking occurs in the holes for gripping the mechanism. The part is covered with Solid Mesh (Figure 24).

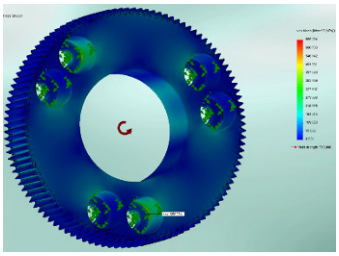


Figure 25: Areas of maximum stresses

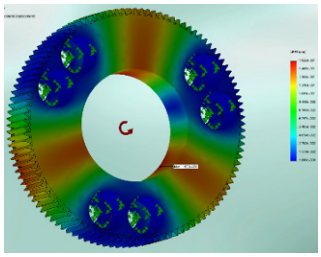


Figure 26: Areas of gear wheel deformation

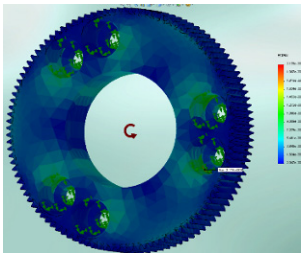


Figure 27: Distribution of the relative elongation coefficient

The distribution of stresses in gear wheel and their concentration are in the transitional part of holes for gripping the wheel (Figure 25). The maximum load value is 655.264 MPa.

Figure 26 shows the gear wheel deformation by the applied load, and Figure 27 shows the distribution of the coefficient of relative elongation.

The results of study on a large gear wheel are presented in Table 6.

Table 6: Results of the study on a large gear wheel

Name	Type	Max	Location
Stress1	VON: von Mises Stress	655.264 MPa Node: 63103	(53.0869 mm, 158.766 mm, 292.226 mm)
Displacement1	URES: Resultant Displacement	0.16216 mm Node: 65881	(-0.0049592 mm, -18.7 mm, 214.01 mm)
Strain1	ESTRN: Equivalent Strain	0.00217841 Element: 32280	(53.5429 mm, -201.8 mm, 227.4 mm)

6.6. Strength and Deformation Analysis of a Shift Fork

The shift fork model is shown in Figure 28. The load on the shift fork occurs with concentrated forces in the part lugs and the clamping occurs in the hole for gripping the gear wheel (Figure 29) and the part is covered with Solid Mesh (Figure 30).

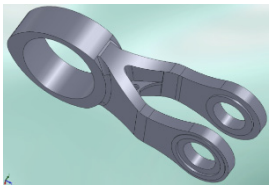


Figure 28: Model of a shift fork

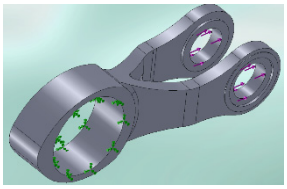


Figure 29: Load on the shift fork

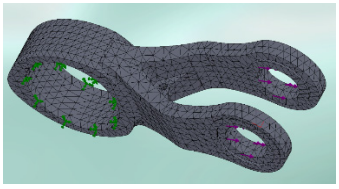


Figure 30: Solid Mesh

Figure 31 shows the distribution of stresses, which are concentrated in the reinforcement area between the two forklugs. The maximum load value is 126.809 MPa.

Figure 32 illustrates the distribution ofrelative elongation coefficient, and Figure 33is deformationby the appliedload.

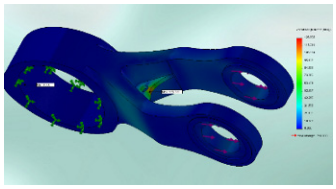


Figure 31: Areas of the maximum stresses

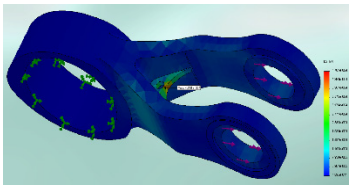


Figure 32: Distribution of the relative elongation coefficient

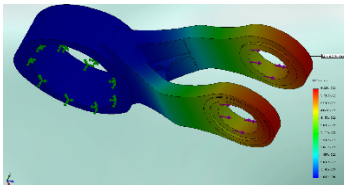


Figure 33: Deformed state of the part

The results of the study on a shift fork are presented in Table 7.

Table 7: Results of the study on a shift fork

Name	Type	Max	Location
Stress1	VON: von Mises Stress	126.8 MPa Node: 1596	(0.00029 mm, 0.728 mm, 98.13 mm)
Displacement1	URES: Resultant Displacement	0.0622768 mm Node: 105	(-0.000413 mm, 46.9 mm, 252.0 mm)
Strain1	ESTRN: Equivalent Strain	0.000476066 Element: 4889	(-2.500 mm, -0.169 mm, 98.73 mm)

6.7. Strength and Deformation Analysis of an Eight-Like Part

The model of an eight-like part is shown in Figure 34.

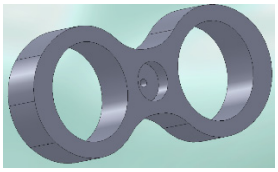


Figure 34: Model of an eight-like part

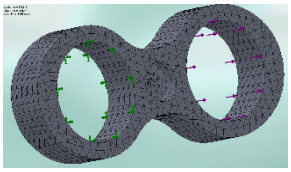


Figure 35: Loading and Solid Mesh

Loading on the eight-like part is caused by concentrated forces in one of the part holes, and the clamping occurs in the other hole (Figure 34). The part is covered with Solid Mesh (Figure 35).

Figure 36 shows the distribution of stresses that are concentrated in thecentre between the twoholes ofeight-like part. The maximum load value is 4.360 MPa.

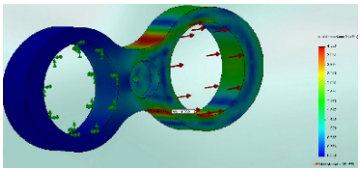


Figure 36: Distribution of the maximum stresses

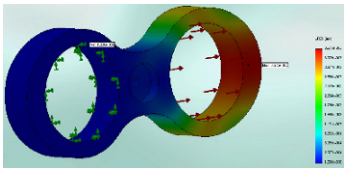


Figure 37: Deformation zone of the eight-like part

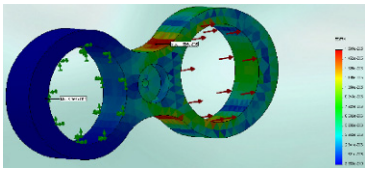


Figure 38: Distribution of the relative elongation coefficient

Figure 37 shows the eight-like partdeformationby the applied load.

The distribution of the relative elongation coefficient is shown in Figure 38.

The results of study on the eight-like part are presented in Table 8.

Table 8: Results of the study on the eight-like part

Name	Type	Max	Location
Stress1	VON: von Mises Stress	4.360 MPa Node: 6509	(4.299 mm, -38.31 mm, -111.86 mm)
Displacement1	URES: Resultant Displacement	0.003573 mm Node: 11441	(21.5 mm, -3.5e-006 mm, -226.9 mm)
Strain1	ESTRN: Equivalent Strain	1.5845e-005 Element: 6908	(-6.6527 mm, 36.3 mm, -111.26 mm)

6.8. Assessment of the Strength Condition of Main Transmission Mechanism Elements Studied

Table 9 shows a sample of the maximum stresses obtained according to the von Mises Stress criterion and indicates the areas necessary to be monitored in operation and repair of the main elements of the studied transmission mechanism.

Table 9: Maximum stresses

Element	Max Stress	Areas to be monitored in operation and repair
Wheel	435.184 MPa	In the area of e holes for connecting the wheel to the drive
Axle	662.306 MPa	In the area of transition between the neck and hub part and in the pressed-fix wheel
Hollow shaft	597.821 MPa	In the area between lugs of the hollow shaft hub holes
Hub	652.964 MPa	In the area of holes in the lugs
Gear	655.264 MPa	In the area of holes for gripping the wheel
Shift fork	126.8 MPa	In the area of reinforcement between the two fork lugs
Eight-like part	4.360 MPa	In the centre between the two holes of the eight-like part

Table 9 shows that the most loaded main elements of transmission mechanism studiedin this paper are the wheelset, hollow shaft and hub, especially the “hollow shaft-wheel” connection.The large gear wheel should be monitoredin operation and repair as well as theteeth profile and parameters and the areas of connectionholes.

The shift fork and eight-like part are slightly loaded elements.

In 2013, an electric locomotive with a third-class transmission mechanism and a cliutch of ALSTOM type suffered a derailment because of damage caused to “hollow shaft-wheel” connection and the wheelset breakage (photos 1 and 2). That fact confirmed the results that had been obtained during the strength and deformation study on the selected transmission mechanism.



Photo 1



Photo 2

By calculating strength and deformation features of the main elements, using CAE software, it has become clear that the theoretically endangered sections coincide with the actually endangered ones.

7. CONCLUSION AND RECOMMENDATIONS

The presented paper highlights advantages and disadvantages of the types of transmission mechanisms indicated in the structure of electric locomotives operated by the Bulgarian railways companies.

The observations of their operation, carried out in 2013-2025, established the personages of damage caused to elements of transmission mechanisms of the second and the third classes.

This study put special accent to the third-class transmission mechanism with a clutch of Alstom type. The modeling of selected structures performed by Solid Works software determined the strength and deformation feachers of its main elements. The established areas of maximum stresses should be set under monitoring both in operation and repair.

REFERENCES

[1] D. Petrovic, V. Aleksandrov, "Zeleznicka vozila. Osnove", MF Kraljevo, ISBN 978-86-82631-66-8 p.223, (2013)

[2] Т. Ружеков, Д. Стоянов, Г. Стайков, А. Стефанов, „ Конструкция, Теория и Проектиране на локомотиви" С. ВМЕИ "Ленин", (1987)

[3] Drive concepts for standard vehicles, https://railsystem.net/traction-systems/

[4] https://tsa.at/drive-concepts/

[5] Drive systems, https://evolution.skf.com/new-drive-systems-for-mass-transit/

[6] Railway Drive System Design: Traction Motor Drive System Basics, https://theengineeringblog.com/railway-drive-system-desing-traction-motor-suspension-tube-online-calculator/

[7] Traction drive system, https://sjsutst.polsl.pl/archives/2019/vol105/101_SJSUTST105_2019_Konowrocki.htm

[8] Ж. Димитров, „Проблеми на ремонта и надеждността на предавателния механизъм при електрическите локомотиви" списание „Железопътен транспорт", бр. 5, (1981)

[9] Andreas Steimel, "Electric Traction – Motive Power and Energies Supply", http://dnb.ddb.de, (2008)

[10] R. Konowrocki, "Modelling of dynamic aspects of operation in railway vehicle traction drive system including the electromechanical coupling". Scientific Journal of Silesian University of Technology. Series Transport. 2019, 105, 101-111. ISSN: 0209-3324. https://doi.org/10.20858, (2019)

SESSION B: RAILWAY ENGINEERING

Impact of deteriorated tram tracks on the current collectors of trams

Emil M. Mihaylov[1*], Emil Iontchev[2], Elisaveta Stoyanova[2]

[1] Department of Lifting-transport and construction machines and systems, „Todor Kableshkov“ University of Transport, Sofia, Bulgaria

[2] Department of Telecommunications and Safety Equipment and Systems, „Todor Kableshkov“ University of Transport, Sofia, Bulgaria

ARTICLE INFO

[*] **Correspondence:** emil.mihaylov@vtu.bg

DOI: https://doi.org/10.46793/ET26.B04M **PAGES**: B29 – B34

ABSTRACT

This article presents the results of a measurement campaign conducted during the summer of 2025 in Sofia. The study aims to determine the impact of deteriorated tram tracks on the current collectors of trams. Measurements were carried out on sections of the tram network where the tracks are in emergency or pre-emergency condition. The measured parameters include linear accelerations, angular velocities, and magnetic field values along the three axes. The sensors were positioned on the tram roof and on the collector strip holder. The analysis of the findings includes a comparison of the parameter values recorded at both locations, a general assessment of their impact on the current collector, and a comparison with results from a previous study conducted in 2019. Based on the obtained data, conclusions have been drawn regarding the influence of the tram's lateral sway on the current collector.

KEYWORDS

Trams, Electrical current collector, Railroad, accelerometer, Gyroscope, Magnetometer, Kalman filter

1. INTRODUCTION

The measurement conducted in 2025 is part of a long-term and comprehensive study on the impact of tram track conditions on the current collectors of tramcars in the city of Sofia [1, 2]. To determine the causes of failures and emergency breakdowns of the current collectors, a series of measurements were performed in track sections under emergency and pre-emergency conditions. It was found that current collectors are most vulnerable to the lateral rolling of the tramcars caused by track subsidence. The primary cause of failure is the destruction of the middle joint of asymmetric current collectors and the rods above the joint. The novelty lies in the fact that the measurement was carried out using two independently operating systems located at the base of the current collector and on the collector head frame.

2. MATERIALS AND METHODS

The measurement was conducted on August 22, 2025, along the route shown in Figure 1. This section provides detailed information regarding the track segment on Hristo Botev Boulevard, extending from Struga Street to Pirotska Street. In this section, the tram track underwent major overhaul in 1998, and currently exhibits severe rail subsidence with a random distribution. The operation of tramcars in this segment is characterized by significant lateral rolling.

A T6A2-BG type tram with inventory number 3027 was used for the measurement. The measurement was conducted at night under a regime corresponding to actual operating conditions. The tramcar is a single-section, four-axle vehicle with rigid axlebox guiding. The current collector is a symmetrical type KE28 - original for this tram model - featuring a modified design of the collector head frame. Following the installation of the sensor unit, the current collector was adjusted to bring the contact wire force to its standard value of 70 *N*.

The accelerations generated within the current collector during the tramcar motion are used to determine the forces acting upon it. In the method described in [2], the accelerations were measured on the ceiling of the tramcar passenger compartment and analytically mapped to its individual components.

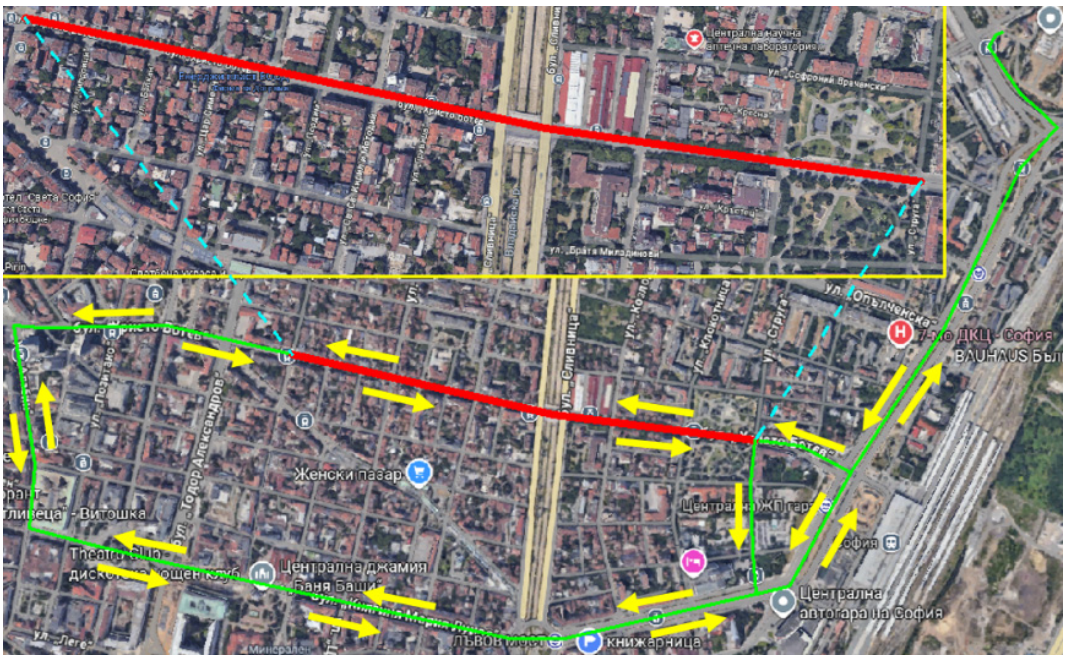


Figure 1: Test rout

To avoid this mapping procedure and its associated inaccuracies, two fully autonomous sets of measurement systems were utilized. One system was installed on the cab roof (Photo 1.), while the second was mounted on the collector head frame (Photo 2.). The specific positioning of the second module imposes additional requirements on its design. To provide protection against the magnetic field of the contact wire, it is housed in an enclosure fabricated from a steel section.

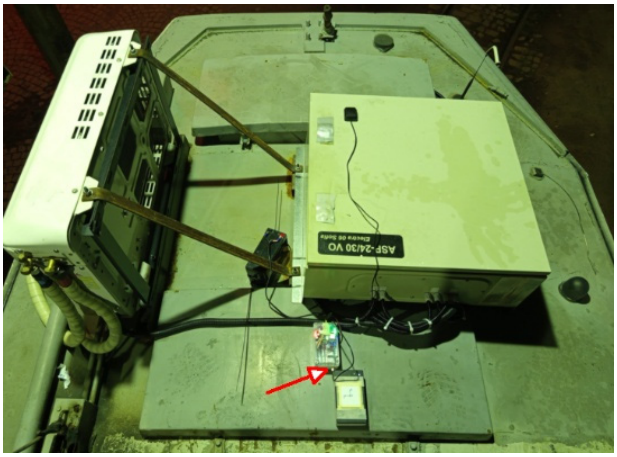


Photo 1



Photo 2

Figure 2: Sensor placement

Both systems are identical and comprise: a triaxial accelerometer, a triaxial gyroscope, a triaxial magnetometer, a navigation receiver, and a microcontroller for data processing and storage. The BNO055 inertial measurement module from Bosch [4] was selected, in which all three sensors are integrated into a single package. It is manufactured using MEMS technology. To fuse the data from the three sensors, it features an embedded 32-bit microcontroller running the proprietary BSX3.0 Fusion Lib software. This enables the sensors to be programmed in various modes to determine either relative or absolute orientation, outputted as quaternions or Euler angles. The ESP32-WROOM-32D [5] module was utilized to read and store the information from the BNO055. Out of the diverse interfaces it provides, two serial interfaces were used: I2C for communicating with the BNO055 and SPI for connecting the SD card module. The system is integrated onto a SparkFun IoT RedBoard-ESP32 [6] development board.

A navigation receiver was utilized for the geographic referencing of the measurements. The sensor data, sampled at a frequency of 40 Hz, were appended to the coordinates received every second. The information integrated in this manner was recorded onto an SD card. Since the navigation receiver located inside the steel enclosure could only acquire a time synchronization signal, this resource was utilized to synchronize its time with that of the receiver on the roof. Consequently, the corresponding geographic coordinates were determined by aligning these timestamps.

Photo 3. shows the orientation of the axes in the coordinate system of the instruments. The Y-axis is directed forward along the direction of travel of the tramcar, while the X-axis is directed to the right relative to the direction of travel.

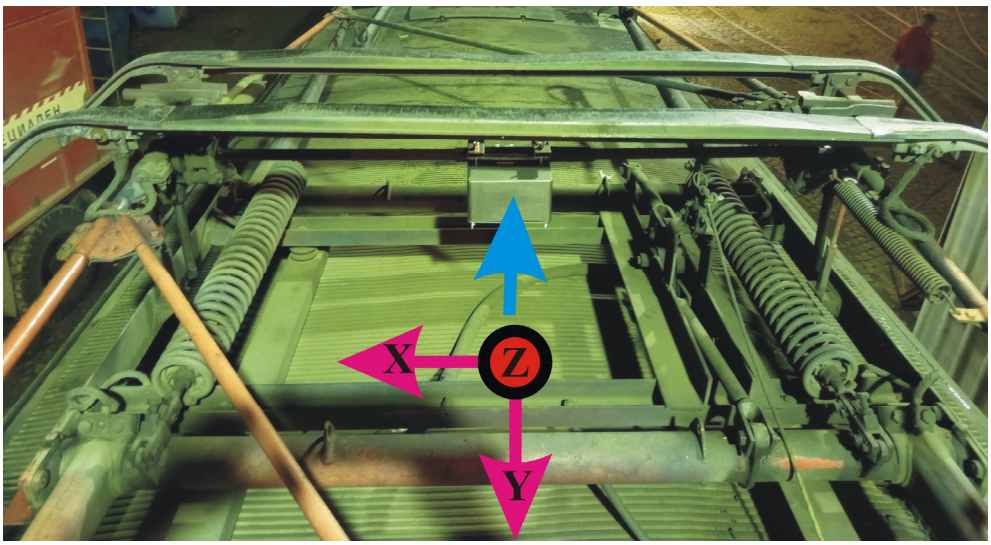


Photo 3. Orientation of the axes in the measuring instruments coordinate system

3. MEASUREMENT RESULTS

The sensors generate 40 readings per second along all three axes for all monitored parameters. To track the impact on the current collector in the lateral direction, the accelerations along the transverse axis of the vehicle were taken into account - specifically, the acceleration values and the root-mean-square (RMS) acceleration values. Table 1. presents the acceleration values obtained from the sensors on the cab roof and the current collector head, as well as the calculated root-mean-square (RMS) acceleration values.

Table 1: Measurement Results

	on the roof	on the current collector
Maximum acceleration values on the X-axis, ms^{-2}	-1,36 ÷ 2,42	-11,6 ÷ 12,67
Equivalent values of acceleration on the X-axis, ms^{-2}	2,84	24,27
Average equivalent values of acceleration along the X-axis, ms^{-2}	1,06	5,08

Figure 3 shows the diagram of the accelerations measured on the roof of the tram.

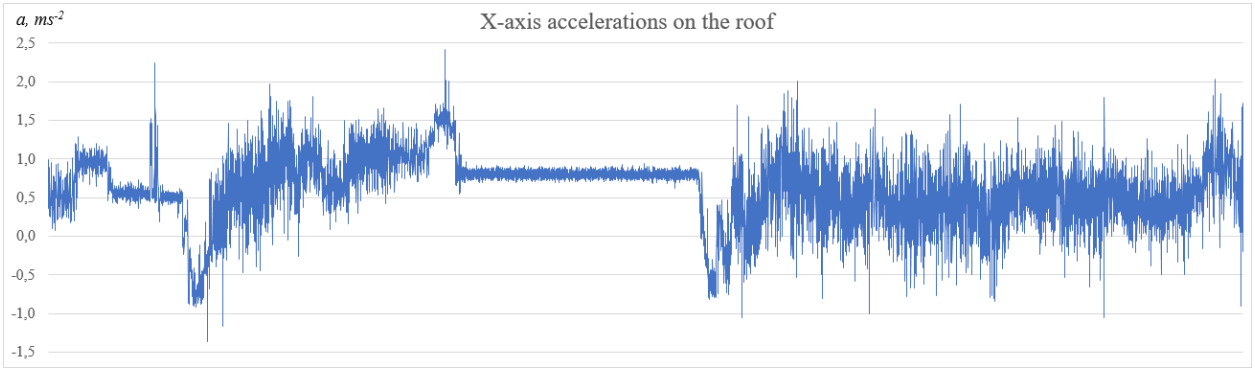


Figure 3: The diagram of the accelerations measured on the roof of the tram

Figure 4 shows the diagram of the accelerations measured on the collector head frame.

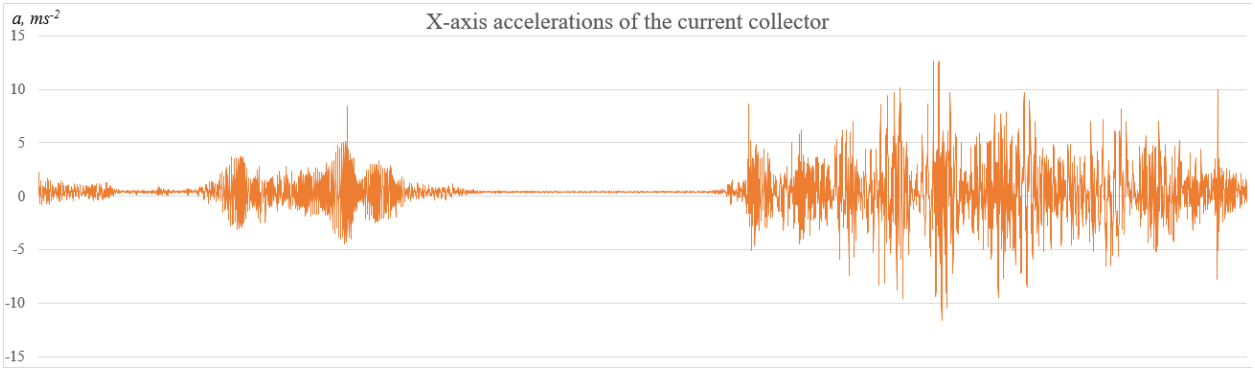


Figure 4: The diagram of the accelerations measured on the collector head frame

The diagram presented in Figure 5. shows the results of the measured accelerations - specifically, on the roof in blue and on the current collector head in orange.

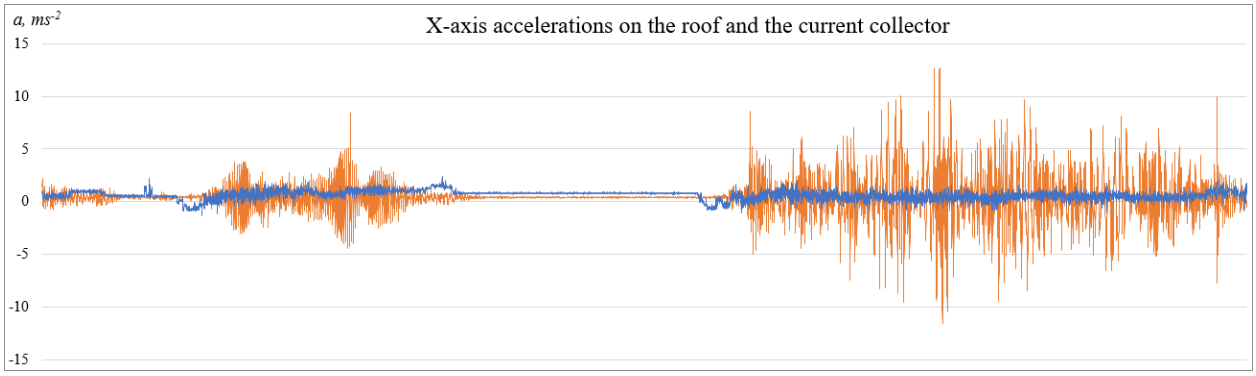


Figure 5: The diagram of the accelerations measured on the roof and on the collector head frame

The intervals with very low values recorded by the sensors correspond to the tram's dwell time at traffic lights. Significantly higher acceleration values were registered in the second part of the section, along Hristo Botev Boulevard from Slivnitsa Boulevard to Pirotska Street. Figure 6 shows the diagram for this section, presenting the results from both measurement points.

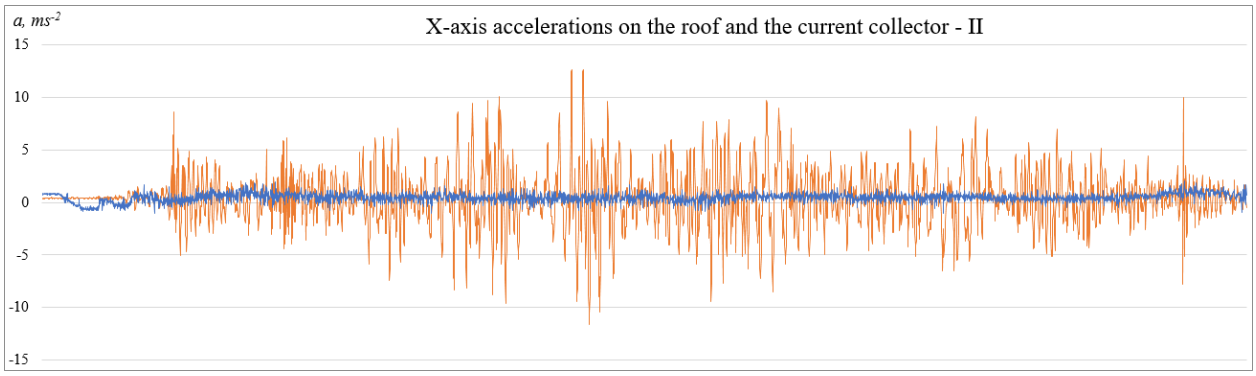


Figure 6: The diagram of the accelerations measured on the roof and on the collector head frame in the second part of the section, along Hristo Botev Boulevard from Slivnitsa Boulevard to Pirotska Street

The diagram with the calculated root-mean-square (RMS) acceleration values for this section is shown in Figure 7.

The presented diagrams clearly illustrate the discrepancy in acceleration magnitudes between the two measurement points. The acceleration magnitude at the current collector head is significantly greater than that at its base. A detailed review reveals that, in addition to exhibiting a larger amplitude, the accelerations at the current collector head display more frequent directional fluctuations. Furthermore, they are phase-shifted, indicating a certain delay in changing direction.

Considering the worst-case scenario, where consecutive maximum accelerations occur in both directions along the X-axis, summing the peak values yields 3,78 ms^{-2} at the base and 24,27 ms^{-2} at the current collector head.

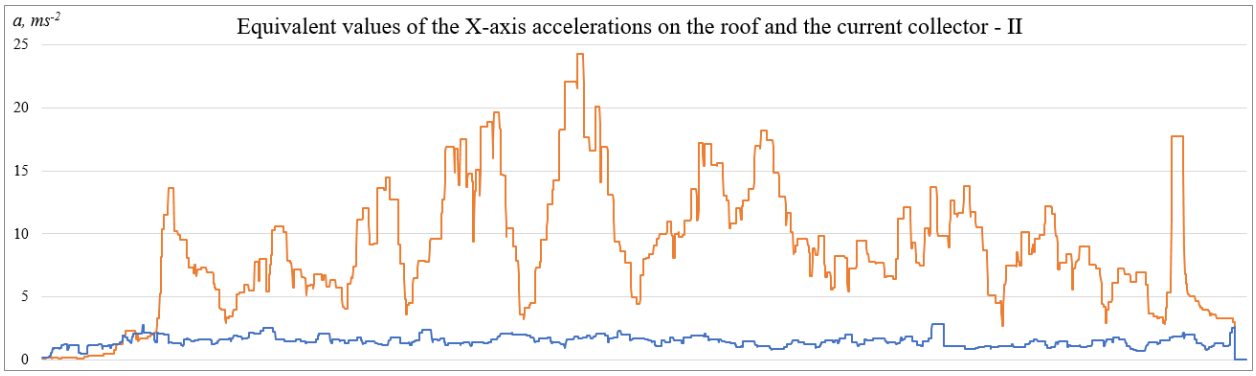


Figure 7: Diagram of the Root-Mean-Square (RMS) Acceleration Values for the Second Part of the Section along Hristo Botev Boulevard from Slivnitsa Boulevard to Pirotska Street

4. IMPACT ON THE CURRENT COLLECTORS

The prevailing acceleration values measured at the current collector head are on the order of ± 5 ms^{-2}, corresponding to a peak-to-peak amplitude of 10 ms^{-2}. Taking into account the discrepancy between the values obtained at the two measurement points, this value can be considered close to the effective acceleration acting on the current collector head during most of the travel time along track sections under emergency or pre-emergency conditions.

To determine the degree of impact on the current collector and its components, the methodology developed in [1, 2] can be applied with minimal modifications. Instead of mapping the parameter values to the current collector head, they can be calculated using the acceleration values obtained during the 2025 measurement.

Figures 8. and 9. show the geometric dimensions of the current collector involved in the calculations. In the figures: B_{bi} - represents the center of gravity locations for each component of the current collector; S_i - denotes the centers of the joints; C_i –indicates the sensor placements; G_i – represents the common centers of gravity for the individual stages of the current collector; d_i – represents the distances from the centers of gravity to the current collector base; ρ_i – denotes the distance of the characteristic points to the rotation axis of the tram; and x_ρ – represents a point on the X-axis of the tram rotation.

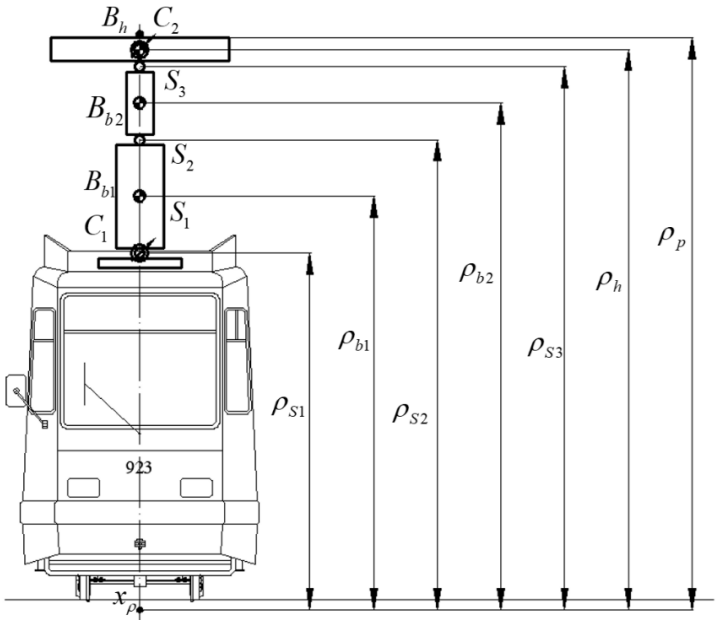


Figure 8: Distances from the current collector components to the rotation axis

Considering that the current collector joint connections are their most vulnerable areas, emphasis has been placed on the forces arising at these points. The calculation results are listed in Table 3 and compared with the calculation results obtained following the 2019 measurement.

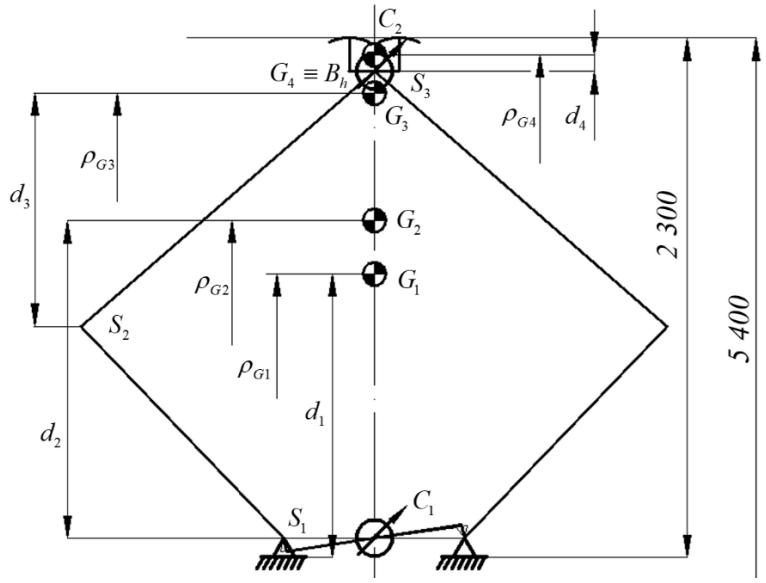


Figure 9: Characteristic points and the distances between them

5. COMPARISON WITH DATA FROM THE PREVIOUS MEASUREMENT

In the measurement conducted in 2019, the sensors were positioned on the ceiling of the passenger compartment, directly beneath the current collector base [1, 2]. Based on the obtained results, conclusions were drawn regarding the accelerations and forces acting on the various current collector components. For this purpose, the corresponding analytical relationships were derived. In 2019, the measurements were performed using two types of trams equipped with Pf-80 and KE28 current collectors. In 2025, the measurement was carried out exclusively with a T6A2-BG type tram featuring a KE28 current collector; therefore, it is appropriate to make the comparison specifically with the data for this type of current collector. In the calculations performed in [1, 2], it is assumed that the angular velocities and accelerations of the entire current collector are uniform. Based on this assumption, calculations were carried out regarding the loads in the joint connections between the current collector components. The 2025 measurement indicates that the accelerations at the current collector head are approximately 5 times greater than those at the base. Consequently, the corresponding loads within the joints are significantly higher. Table 2 presents the results of both acceleration measurements, obtained from the sensors at the base and the current collector head, respectively.

Table 2: Acceleration measurement results from the base and collector head sensors

	measurement in 2019	measurement in 2025
Acceleration values on the X-axis on the roof, ms^{-2}	0,95	2,42
Acceleration values on the X-axis on the current collector, ms^{-2}	-	12,67

The discrepancy between the acceleration values at the current collector base in the two measurements can be attributed to the different routes and, consequently, the varying track conditions. The 2019 measurement was conducted in a track section characterized predominantly by parallel rail subsidence. In contrast, for the 2025 measurement, a section with a staggered or random distribution of rail subsidence was selected. Table 3 presents the values calculated from the 2025 measurements and compares them with the mapped values computed based on the 2019 measurement.

Table 3: Values calculated from the 2025 measurements and mapped values computed following the 2019 measurement

	measurement in 2019	measurement in 2025
Bending moment at the current collector head hinge, Nm	3,62	482
Bending moment of the current collector middle joints, Nm	78,14	726
Bending moment of the current collector base joints, Nm	188,22	1297

A comparison of the data in the table indicates that the 2025 measurement yielded higher values for the bending moments in the current collector joint connections.

6. CONCLUSION

The presented results of the acceleration measurements in tram current collectors, arising from the tram track conditions, demonstrate significant values. These findings can be considered to account for the frequent failures of current collectors on trams operating on lines whose routes traverse track sections under emergency and pre-emergency conditions. The presented measurement validates the conclusions drawn in [1, 2] and provides a clearer understanding of the loads acting within tram current collectors.

REFERENCES

[1] Mihaylov E.M., Iontchev E., Miletiev R., Petkov B., „Assessing the impact of the lateral swinging of the tram to its electrical current collector", XIX Scientific - Expert Conference RAILCON'20, Ниш, Сърбия, 2020 г.

[2] Mihaylov Emil M., Emil Iontchev, Rosen Miletiev, Metodi Atanasov, Rashko Vladimirov, „Comparative analysis of the effect of lateral swinging of the tram body on different types of electrical current collectors", XI International Conference "Heavy Machinery-HM 2023", Vrnjačka Banja, 21– 24 June 2023, pp B.95-100, ISBN-978-86-82434-01-6

[3] Йончев Е., Михайлов Е.М., Владимиров Р., Милетиев Р., „Система за измерване на динамични параметри на токоприемник на трамвай", XXV Международна научна конференция "УНИТЕХ'2025" – Габрово, ТУ – Габрово, 2025 г., ISSN: 1313-230X, ISSN: 3033-1404

[4] BNO055 Intelligent 9-axis absolute orientation sensor, Bosch Sensortec, Document number BST-BNO055-DS000-18, October 2021, https://www.bosch-sensortec.com/media/boschsensortec/downloads/datasheets/bst-bno055-ds000.pdf, accessed 09/2025

[5] ESP32-WROOM-32D & ESP32-WROOM-32U, Datasheet Version 2.6, Espressif Systems (Shanghai) Co., 2025-08-05, https://espressif.com/documentation//esp32-wroom-32d_esp32-wroom-32u_datasheet_en.pdf, accessed 09/2025

[6] SparkFun IoT RedBoard - ESP32 Development Board, https://www.sparkfun.com/sparkfun-iot-redboard-esp32-development-board.html, accessed 09/2025

SESSION B: RAILWAY ENGINEERING

Measurement of distortion power and power factor in electric traction substations "Serbia railway infrastructure" when powering electric locomotives in traction and recuperative braking

Branislav Gavrilović[1], Živoslav Avramović[1], Zoran Stjepanović[1], Darko Stjepanović[1]

[1] The European University Brčko District, Tehnical Faculty, Brčko, Bosna and Hercegovina

ARTICLE INFO

* **Correspondence:** gavrilovicbranislav5@gmail.com

DOI: https://doi.org/10.46793/ET26.B05G **PAGES**: B35 – B40

ABSTRACT

The aim of this paper is to point out the need to adapt the existing power measurement methods in electric traction substations of the "Railway Infrastructure of Serbia" in accordance with the provisions of the European standards EN 50470–1 and EN 50470-3:2022. The widespread use of power converters on electric locomotives, as typical non-linear consumers, has called into question the validity of existing methods of recording railway consumption. The waveform of the traction and regenerative braking current of electric locomotives deviates from the sinusoidal waveform, due to which part of the delivered energy remains invisible, which, among other things, leads to significant losses.

To solve this problem, it is necessary to detect and quantify the power factor that takes into account the power distortion due to the power supply of electric locomotives as nonlinear consumers. This manuscript proposes and describes indirect methods of measuring electricity with single-phase intelligent electricity meters that enable not only an efficient tariff policy, but also the need for compensation of higher harmonics at the point of connection of electric traction substations to the three-phase electrical network.

KEYWORDS

Voltage and current harmonics, Traction mode, Regenerative braking.

1. INTRODUCTION

The presence of higher current harmonics in the traction system of the "Serbian Railway Infrastructure" has a very negative impact on the elements of the transmission power system. This is because these elements are designed for optimal operation at a fundamental frequency of 50Hz, higher harmonics cause additional dissipation, significantly reduce their service life and lead to numerous other problems [1, 2, 3]. A large number of studies indicate that a true insight into the power delivered from the three-phase electrical network to the electric traction system of Serbian railways cannot be obtained by measuring only active and reactive power [4]. The nonlinear character of the consumers and the accompanying harmonic distortions of currents and voltages in the electric traction system, among other problems, have generated additional losses in the transmission network to the electric locomotives. Because of these losses, it is necessary to devise a reliable and at the same time economically justified method for registering the value of harmonic distortions.

Nonlinear harmonic distortions of current and voltage cause the appearance of a new power component that is not registered by the existing meters in traction substations. Namely, in addition to the active (P) and reactive power components (Q) that are measured, the distortion power (D_B) is not measured in traction substations. By measuring the distortion power (D_B) or by measuring the apparent power (S) and power factor both in the traction mode and in the recuperative braking mode of the trains, the presence of higher voltage and current harmonics at the point of connection of electric traction substations to the three-phase power grid is taken into account. In this way, the level of negative impact of the electric traction system as a pronounced nonlinear load on the three-phase power system can be measured in real time, all with the aim of taking appropriate protective measures.

In the following, we will describe the method measuring of: effective value of two-channel current and single-phase voltage, active power and power factor in accordance with the provisions of the European standards EN 50470–1 and EN

50470-3:2022, which are suitable for traction and regenerative braking of trains on electrified lines of the "Railway Infrastructure of Serbia".

2. APPARENT POWER COMPONENTS IN ELECTRIC TRACTION SUBSTATIONS FOR NONLINEAR LOAD CASE

According to the established tariff policy, "Railway Infrastructure of Serbia" only pays for active/reactive energy used. Installed electric meters in electric traction substations are not able to measure the energy that would eventually be accepted during recuperative braking of trains and transferred to the three-phase electrical network. Due to this fact, trains with the possibility of recuperating electricity still do not use this possibility, which reduces their energy efficiency.

The appearance of electric traction vehicles with recuperative braking requires the installation of various power converters based on semiconductors (diodes, thyristors, transistors). This fact created additional problems for the portable power grids that supply the single-phase electric traction system and electric locomotives with the possibility of recuperative braking. Part of the power that occurs due to the presence of higher harmonics, which is delivered or received by electric locomotives, is invisible from the point of view of the transmission power network if standard electro-mechanical or electronic meters are used for consumption registration. In the mentioned conditions, due to the presence of higher harmonics, the following relationship applies:

$$S^2 > P^2 + Q^2 \tag{1}$$

where: S- apparent power, P- active power, Q- reactive power.

Given that the following applies to simple periodic systems: $S^2=P^2+Q^2$, it is clear that the difference in the value of the apparent power originates from the presence of higher harmonics, that is, it is a consequence of voltage and current distortion in the electric traction system. Due to the presence of higher harmonics in the electric traction system, the distortion power must also be introduced so that the following corrected expression for the apparent power is valid:

$$S^2 = P^2 + Q^2 + D_B^2 \tag{2}$$

where D_B is the distortion power. The essence of this equation is that in the absence of harmonics $D_B=0$, $S^2 = P^2 + Q^2$ holds. It is obvious that this equality becomes a special case of applying equation (2). From equation (2) it follows that the distortion power is:

$$D_B = \sqrt{S^2 - P^2 - Q^2} \tag{3}$$

The addition of the D_B vector representing the harmonic distortion, as shown in Figure 1, creates a new power triangle for a true power factor with harmonics.

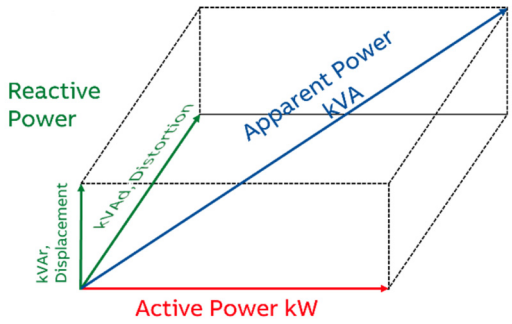


Figure 1: Power triangle for a power factor which includes harmonics

kW - Active Power is the power that performs useful work.
kVA - Apparent Power. Vectorial summation of kVAR and kW.
kVA_R - Reactive Power.
kVAd – Reactive Power Distortion.

The power factor ($\cos\varphi$) under non-linear load conditions is:

$$P_F = \frac{P}{S} = \frac{P}{\sqrt{S^2 - Q^2 - D_B^2}} \tag{4}$$

Under linear load conditions ($D_B=0$), the power factor is:

$$P_F = \cos\varphi = \frac{P}{S} = \frac{P}{\sqrt{S^2 - Q^2}} \tag{5}$$

where is φ - the angle between the active and apparent power vectors.

Power factor and harmonics are two important concepts in electrical engineering, often discussed in the context of power quality and efficiency in electrical systems. Power factor defined in this way accounts for the harmonics thus reducing the total harmonic content contributing to the efficiency of the electrical system.

By measuring the power factor based on expression (4) and taking it into consideration when creating the account (penalization of significant non-linear consumers such as the electric traction system of the "Infrastructure of the Railways of Serbia"), the transmission power network as a distributor of electricity is able to significantly reduce the level of losses in the system and to reduce other problems that can cause the propagation of harmonics in the transmission network.

3. METHODS OF MEASUREMENT OF APPARENT POWER COMPONENTS IN ELECTRIC TRACTION SUBSTATIONS

The existing measuring equipment in electric traction substations provides a display of the consumption of active and reactive power/energy from the power grid, and when powering electric locomotives in the traction mode, i.e.:

Existing measuring equipment in electric traction substations allows the display of active and reactive power/energy from the power grid when powering electric locomotives in traction mode, i.e.: +P(kW), +W (kWh), +Q(kVAr), Wr(kVArh). However, with existing meters, measurement of active and reactive power/energy in electric traction substations (-P(kW), -W (kWh), -Q(kVAr), -Wr(kVArh)) is not possible due to the derived solution during regenerative braking of electric locomotives.

For the above reasons, the "Serbia Railway Infrastructure" must require new measuring equipment and methods, as well as improving the metrological characteristics of existing measuring equipment. This measuring equipment, in addition to being able to display the direction of electrical energy flow, must also provide reliable measurement of power factor under nonlinear load conditions of the electric traction system, which will indicate the efficiency of energy use and equipment performance. This includes assessing energy consumption patterns and identifying power factors under nonlinear load conditions of the electric traction system. The following text describes intelligent power factor measurement under nonlinear load in electric traction substations of the "Infrastructure of Serbian Railways" when powering electric locomotives in traction and regenerative braking modes.

3.1. Hardware Implementation of the Measurement Sistem

Regardless of the large number of electronic meters on the market today, the measurement data and software that are characteristic of these meters still do not allow the measurement of distortion power and power factor in conditions of two-way flow of electricity. Due to the stated reason, it can be said that only the described methodology and power factor measurement enables the necessary measurement in the "Serbian Railway Infrastructure". In addition, measuring the power factor P_F in electric traction substations in the mode of electric traction and recuperative braking of locomotives can directly indicate the efficiency of the use of electrical energy and the optimization of the performance of equipment in the electric traction system.

The hardware implementation of the proposed measuring system is shown in Figure 2. The measuring system is actually an adapted version of the existing electronic meters currently on the market. In this particular case, an electronic meter of single-phase electronic type EWGE11X (product of EWG from Niš, R. Serbia, www.ewg.co.rs) was used. This meter corresponds to European standards EN 50470-1 and EN 50470-3, as well as international standards IEC 62052-11, IEC 62053-21 and IEC62053-23.

EMG11X meter microcontrollers integrate:

- Analog inputs (analogue front end - AFE)
- Independent digital metrology unit (computation engine - CE)
- 8051-compatible microprocessor (MPU)
- LCD drivers
- RAM and Flash memory
- Real time clock (RTC)
- Greater number of multifunctional I/O pins.

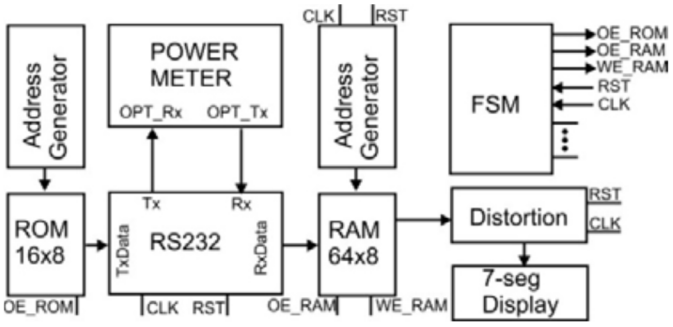


Figure 2: Block diagram of realized system

Figure 3 shows a block diagram of the two main microcontrollers used to implement the EWGE11X electronic meter.

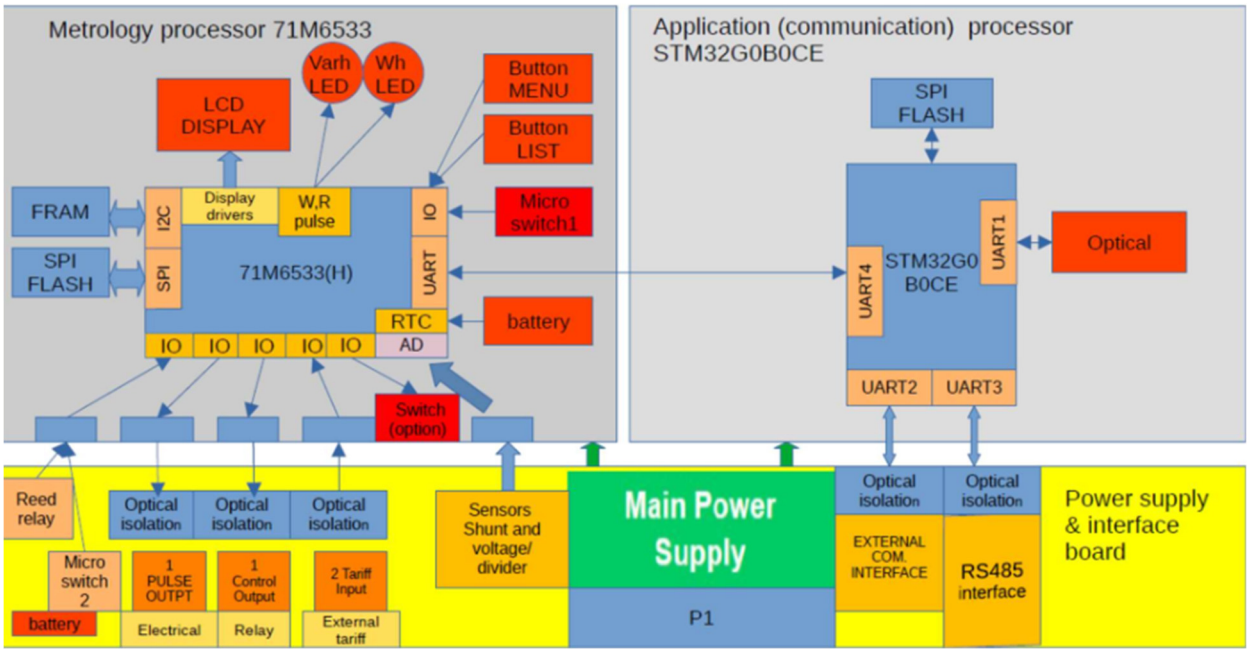


Figure 3: Microcontrollers of the electronic meter EMG11X - block diagram

The analog input functions as a data acquisition system controlled by a microprocessor unit (MPU). The basic blocks of the analog input are: input multiplexer, delta-sigma A/D converter, FIR filter and reference voltage block. Input analog signals are multiplexed before sampling using an A/D converter. The samples from the A/D converter are passed through the FIR filter and the filtering results are stored in the XRAM memory in the chip itself, where they can be accessed by the metrology unit (CE) or the microprocessor (MPU).

The metrological unit (CE) is a dedicated signal processor that performs the following metrological calculations:

- Multiplication of the instantaneous current value by the instantaneous voltage value to obtain the energy per sample, when multiplied by a constant sampling time.
- Gain and offset compensation.
- Sampling delay compensation on all channels (due to multiplexing of analog signals).
- 90° phase shift for reactive power Q calculation.
- Frequency measurement.
- Summation and calculation of effective values (RMS) of voltage and current.
- Calculation of fundamental and harmonics of active, reactive and apparent power.
- Calculation of fundamental and harmonics of current and voltage.
- Input signal frequency tracking (frequency and phase shift)
- Input signal amplitude tracking (delay detection)
- Temperature measurement and compensation.

Block diagram of the proposed meter for calculating the power factor under nonlinear load is shown in Figure 4. Namely, after multiplication, the AccSb block sums the squares of the active (P), reactive (Q) and distortion powers (D_B), and the square root of their sum is divided by the active power (P) (equation 4).

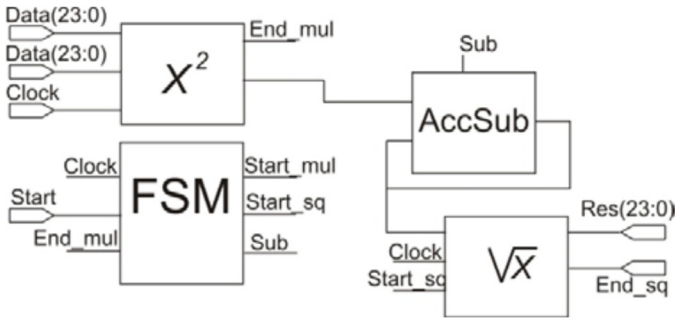


Figure 4: Block diagram for calculating the power factor under nonlinear load

The proposed meter is capable of recording all measured data in real time and transmitting it to a central location for billing purposes. In addition to providing accurate billing information, these meters can provide a number of other benefits. Namely, given the possibility of measuring the distortion power D_B and the power factor P_F, which takes this power into account, it is possible to measure the delivered power/energy, which has not been possible in electric traction substations so far, both in electric traction mode and during regenerative braking of electric locomotives. The described pawer meters were verified in the LEDA laboratory of the Faculty of Electronics, University of Niš, Republic of Serbia [20].

4. TARIFF SYSTEM

The described and proposed single-phase meter, which could be used in the "Railway Infrastructure of Serbia", allows for reliable determination of the level of vulnerability of the electric traction system with existing electric locomotives, as well as the three-phase power grid to which the 110/25kV electric traction substations are connected.

"Serbia Railway Infrastructure" has an interest in having electricity consumption in electric traction substations controlled by meters that register distortion power and power factor. Namely, the current practice is that the distributor covers losses by increasing the price of energy. In this way, losses caused by increased power/energy distortion are paid by all users, regardless of the nature of the consumers they use. Users who cannot afford expensive, energy-efficient devices will have increased bills as they cover the losses caused by energy-efficient devices! Also, those more conscious users who are willing to buy efficient devices with expensive filters to remove nonlinear distortions will also pay the costs of unscrupulous consumers. Therefore, there is a mutual interest (of distributors and consumers) to implement electronic meters that register both the energy consumption of the distortion and the power factor of that distortion.

In order to complete the system for registering electricity consumption burdened by nonlinear consumers, it is necessary to improve the tariff system. It is fair to discourage consumers from introducing harmonics into the electricity grid in an uncontrolled manner. For this reason, there is growing support for the need to introduce some allowable limit value for distortion power D_B , or power factor P_F, which would not be taken into account when creating the final bill for consumed electricity. At the same time, this means that the measured value of the distortion power D_B, i.e. the power factor P_F, should be reduced by that amount. We propose a correction that can be expressed using the form:

$$D_p = D_B - \gamma S \tag{6}$$

where D_B and γS represent the value of the measured distortion power and apparent power, respectively, and γ denotes a constant to be defined. The decision on the value of the constant γ that will correct the measured values would be made by railway experts in agreement with electricity distributors. The authors of this manuscript propose that the value of the coefficient γ be 0.1. The aforementioned value was obtained after considering the permissible value of voltage/current harmonics at the connection point prescribed by the IEEE 519-2022 standard.This standard was considered a reference, instead of IEC61000-3-2:2019, which is mainly applied in Europe, for a very pragmatic reason. Namely, IEC61000-3-2:2019 regulates the characteristics of electronic devices, while the IEEE 519-2022 standard defines the permissible values of voltage/current harmonics at the point where the consumer is connected to the three-phase electrical network, i.e. at the point of connection of the electronic meter. It is important to note that these standards are compatible, so compliance with the IEC61000-3-2:2019 standard also guarantees compliance with the requirements prescribed by the IEEE 519-2022 standard.

5. CONCLUSION

The presence of higher current harmonics in the traction system of the "Railway Infrastructure of Serbia" has an extremely unfavorable effect on the elements of the three-phase electrical system. Since they are typically designed to operate optimally at a fundamental frequency of 50Hz, higher harmonics cause additional dissipation, significantly shorten equipment life, and lead to numerous other problems.

Modern electric traction drives are designed in accordance with the requirements of increasing energy efficiency, and therefore very often a part of the energy from electric locomotives can be returned to the three-phase power grid via contact network and electric traction substations. Apart from the economic aspect of reducing consumption, electric locomotives with regenerative braking introduce non-linear effects associated with the injection of currents that are extremely rich in higher harmonics. That is why it is important to quantify the non-linear component that is introduced into the power grid via electric traction substations.

Due to all of the above, the research results indicate the need to change the electricity charging policy, i.e. the need to compensate for higher harmonics at the connection point of traction substations of the Serbian Railway Infrastructure. This intervention is not cheap, and its optimal implementation requires serious techno-economic analysis. Due to all of the above, it is proposed to bill electricity using electronic meters with the described measurement method, which allow, under conditions of highly nonlinear load, precise measurement of various electrical quantities, including positive and negative distortion power and power factor, which have not been measured in traction substations so far.

REFERENCES

[1] SRPS EN 50470-1:2009/A1:2020: “Electricity metering equipment (a.c.) - Part 1: General requirements, tests and test conditions - Metering equipment (class indexes A, B and C)”, (2020)

[2] SRPS EN 50470-3:2009/A1:2020: “Electricity metering equipment (a.c.) - Part 3: Particular requirements - Static meters for active energy (class indexes A, B and C), (2020)

[3] SRPS EN IEC 62052-11:2022/A12:2024: “Electricity metering equipment - General requirements, tests and test conditions - Part 11: Metering equipment”,(2024)

[4] SRPS EN IEC 62053-21:2022: “Electricity metering equipment - Particular requirements - Part 21: Static meters for AC active energy (classes 0,5, 1 and 2)”, (2022)

[5] SRPS EN IEC 62053-23:2022: “Electricity metering equipment - Particular requirements - Part 23: Static meters for reactive energy (classes 2 and 3), (2022)

[6] SRPS EN 50470-3:2022: Electricity metering equipment - Part 3: Particular requirements - Static meters for AC active energy (class indexes A, B and C), (2022)

[7] SRPS EN 50470-1:2009/A1:2020: Electricity metering equipment (a.c.) - Part 1: General requirements, tests and test conditions - Metering equipment (class indexes A, B and C), (2020)

[8] Milenko Marinko, Marjan Urekar:”Development of a system for powering smart electricity meters”, Proceedings, LXVII ETRAN Conference, ML3.6 - Page 1 of 5, East Sarajevo, BiH, (2023)

[9] Integral Energy Power Quality Centre: “Technical note No. 3 - Harmonic Distortion in the Electric Supply System,” University of Wollongong, http://goo.gl/8QpwdO, (2000)

[10] B Gavrilović, Z Pavlović, V Radičević, M Stojanović: „Determining the Power Supply Quality of the Diode Locomotive in the Electric Traction System“, Modelling, 2024 - mdpi.com., (2024)

[11] B Gavrilovic, VA Baboshin: „Simulations of the operation of the fast light innovative regional train from “Serbian Railways” in traction and electric braking mode“, Mechanical Engineering Advances, 2024 - ojs.acad-pub.com., (2024)

[12] M Ali and Z Arifin 2021 J. Phys.: Conf. Ser. 2111 012038, Journal of Physics: Conference Series, Volume 2111, 4th International Conference on Electrical, Electronics, Informatics, and Vocational Education (ICE-ELINVO 2021), Yogyakarta, Indonesia, (2021)

[13] D. Stevanović , P. Petković, “Harmonics in Power System Problems and Solutions (in Serbian*)”, Proc. Of XII International Scientific Symposium INFOTEH®- JAHORINA 2013, Jahorina, Bosnia and Herzegovina, pp. 203-208, (2013)

[14] Dejan Stevanović, Predrag Petković and Volker Zerbe: “Single phase system for detection of harmonic pollution sources at power grid”, Proceedings of the 5th Small Systems Simulation Symposium 2014, Niš, Serbia, (2014).

SESSION B: RAILWAY ENGINEERING

Comparative analysis of methods for the investigation of brake wear particles emission of railway vehicles

Jasna Glišović[1*], Danijela Miloradović[1], Jovanka Lukić[1], Nenad Miloradović[2], Saša Vasiljević[3]

[1] University of Kragujevac, Faculty of Engineering/Department of Motor Vehicles and Motors, Kragujevac, Serbia

[2] University of Kragujevac, Faculty of Engineering/Department of Mechanical Structures and Mechanization, Kragujevac, Serbia

[3] Academy of Professional Studies Šumadija Department in Kragujevac, Kragujevac, Serbia

ARTICLE INFO

[*] **Correspondence:** jaca@kg.ac.rs

DOI: https://doi.org/10.46793/ET26.B06G **PAGES**: B41 – B49

ABSTRACT

The braking systems of railway vehicles are a key element for the safe and efficient operation of trains. Although according to official reports, the railway sector contributes less than 1% to total PM emissions in Europe, PM concentrations in enclosed spaces of underground railways are of great concern due to the confined nature of these spaces. Numerous studies have identified brake wear as the primary source of these emissions, as particles are generated during the braking process, under conditions of intense heating and wear in the contact zone between brake pads and discs. The aim of this paper is to critically analyze previous experimental studies to arrive at the most efficient and appropriate methodology that would provide knowledge about the mechanisms of particle generation caused by wear of brake pair elements in order to find ways to control the number and size distribution of particles in the air. In this regard, the advantages and disadvantages of various field tests with full-size brakes, laboratory tests on real brakes and reduced-scale brakes, as well as laboratory tests with a pin-on-disc machine were analyzed.

KEYWORDS

Railway vehicles, brakes, particle matter, methodology.

1. INTRODUCTION

In most cases, the transport of passengers and goods by rail is considered the most environmentally friendly mode of transport. With the increasing use of electrically powered motors in railway vehicles, exhaust emissions are no longer a main concern and attention is shifting to particulate matter from wear of brakes and others materials, as well as re-suspension from the ground. The task of the brakes is to maintain or reduce speed, that is, to stop vehicles in motion within a prescribed stopping distance, as well as to secure vehicles against self-movement.

The regulation on Brakes and Braking of Trains and Vehicles prescribes the types of brakes on vehicles and their parts, inscriptions and markings of brake devices, the method of braking trains and vehicles, the composition of trains with regard to the type of brakes, checking the correctness of brakes, calculating the brake power of a train and securing it against self-propulsion [1].

Mechanical friction brakes are the most important component necessary to achieve reliable, safe train traffic. Trains are generally considered a green energy and environmentally friendly transportation system, but airborne particulate emissions affect the accuracy of this picture. Metro systems around the world have been found to exhibit worrying levels of particulate pollution, exceeding those found in above-ground urban environments. Although human exposure in metro stations is relatively short, the intensity and the large number of people affected, including vulnerable groups such as the elderly or asthmatics, make the potential toxicity of particulates and emission reductions an urgent subject of research.

It must be noted that particulate emissions from rail transport have been much less studied than from road transport. Progress in legislation, toxicity studies or even mitigation alternatives in road transport can be used as a basis for future studies of rail transport emissions. Unlike road vehicles, a similar methodology is lacking for non-exhaust emissions from

rail transport. Abbasi et al. [2, 3] proposed a method for measuring the airborne wear particle emission rate (AWPER) from wheel-rail contacts and brake materials. This method can be used by manufacturers to demonstrate the competitive advantage of their products.

Studies of particle concentrations in metro systems around the world have attracted considerable attention in the past few decades. In 2011, Abbasi et al. [2] found that particle concentrations in metro systems in Stockholm, London and Budapest reached values of 300-1000 μg/m^3, which significantly exceeded the maximum limit of 50 μg/m^3 per day set by the EU for urban transport. By analyzing the materials present in the studied PM emissions from rail transport, about 84% can be attributed to mechanical friction processes, such as friction between the rail, the wheel, and the brake disc and pad. In situations of lower speed, such as approaching a metro station or in emergency situations, the mechanical braking system is activated and the brake system becomes one of the primary sources of airborne wear particles. Studies on brake wear emissions are still scarce, especially for disc brakes of high-speed trains. Available studies indicate that factors that have a significant influence on brake wear emissions include material composition, brake pressure, initial speed and contact temperature. Higher train speeds resulted in an increase in the total and maximum particle number concentrations, but also led to an increase in the particle number concentration for the peak diameter in the size distribution analysis. Higher speeds can also be associated with increased contact temperatures due to increased frictional heat and the thermal and chemical processes involved. The high temperatures resulting from increased disc speed and braking force can facilitate the formation of nanoparticles. Technical specifications of the vehicle, such as axle loads, technical specifications of the lifting table, material properties of the wheels and brakes, braking system and braking scheme, are important for particulate matter emissions. But in addition to factors related to the railway vehicles, emissions are also influenced by railway infrastructure conditions, such as technical specifications of the railway, track gauge, track substructure and track superstructure, track layout, switches and type of power transmission [4].

There are only few studies investigating the emission of wear particles from railway above ground. The aim of this paper is to perform a qualitative analysis and select the most optimal methodology to investigate the influence of railway braking conditions, such as contact pressure and disc temperature, in particular in terms of the range of emitted particles, their shape, size and composition. Knowledge of the nature of these particles (size, shape and composition) and their emission in real time is necessary for the development of efficient numerical models that can successfully predict the spread of particles near their origin. Based on the measurements, conditions would be obtained for the correct selection of friction conditions that give lower particle emission for a given dissipated kinetic energy and given rules for the injection properties of particles in a CFD model for studying their spread in a confined space, such as a metro station. In order to assess the influence of individual parameters on the nature of emitted PM, an experiment will be performed on a pair of railway braking materials, varying one parameter at a time, while maintaining the equivalent dissipated energy.

2. RAILWAY BRAKING SYSTEM

2.1. Railway Brakes

In the past, trains were equipped almost exclusively with mechanical brakes. Today, on modern high-speed trains, the braking system is usually controlled by a computer that allows variable distribution of braking power between different subsystems. The aforementioned braking subsystems usually include electro-dynamic and mechanical (friction) brakes. The application of newer technologies in railway vehicles and track infrastructure has also led to the introduction of inductive braking systems. Research into new linear induction motor techniques has been accompanied by increased weight. An overview of the different categories of brakes for rail vehicles is shown in Figure 1.

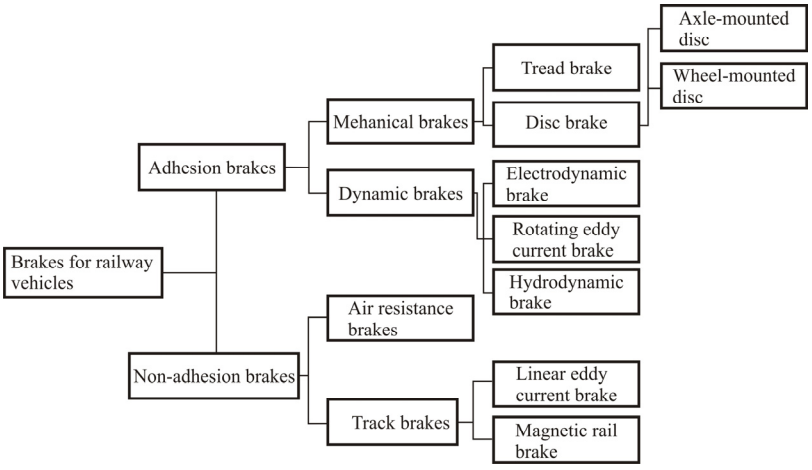


Figure 1: Different types of brakes on railway vehicles

Due to the frictional force developed between the brake pad and the brake disc, mechanical (friction) disc brakes are subject to a significant degree of wear, in contrast to electro-dynamic brakes which are much less susceptible. However, as the efficiency of electro-dynamic brakes depends on the speed of the railway vehicle, mechanical brakes are still

necessary for additional braking, especially at lower speeds. Usually below 60 km/h, friction brakes are activated to provide additional braking, and at speeds below 20 km/h, mechanical brakes take over the entire braking force. Mechanical brakes are also critical in the event of electro-dynamic brake failure or if it is necessary to reduce the speed of the train quickly, such as in emergency braking.

The two main types of friction braking systems are tread brakes (block brakes) and disc brakes. Tread brakes (Figure 2c) are still widely used on most freight trains in the EU, and disc brakes are fitted to many passenger trains, especially those operating at high speeds. The friction coefficient, and thus the efficiency of a disc brake, is very stable and almost independent of speed, and they exhibit good thermal properties and robustness. This type of brake can be mounted directly on the wheel or on the axle. Figure 2a shows a high-speed train, which uses axle-mounted disc brakes. Here, each axle is equipped with four brake discs and two pairs of brake pads on each side. A photograph of a disc brake pad mounted on a train wheel is shown in Figure 2b.

Brakes for high-speed trains must demonstrate superior properties and ensure safety, which means that they must withstand maximum temperatures during operation. This includes a high and stable coefficient of friction (COF), high thermal conductivity, as well as resistance to heat and wear, low absorption of water, oil or brake fluid, low aggressiveness towards the brake disc, high comfort properties, as well as reduced particle emissions and noise minimization. Organic brake pads have poor thermal stability and can therefore only be used when temperatures do not exceed 300°C. Since temperatures of up to 900°C can occur at the disc/pad interface during sudden braking of high-speed trains, sintered metal matrix composites are a suitable choice for this application because they can withstand temperatures of over 1000°C. These friction materials are obtained by powder metallurgy, i.e. they are pressed and sintered from a uniform mixture consisting mainly of metal powders as a matrix, matrix reinforcement materials, friction enhancers, and lubricants. Common matrix and matrix reinforcement materials are iron and copper, while metal and non-metal oxides or carbides are used as friction enhancers. To improve wear resistance, a lubricant such as molybdenum disulfide (MoS_2) is often used [5, 6].

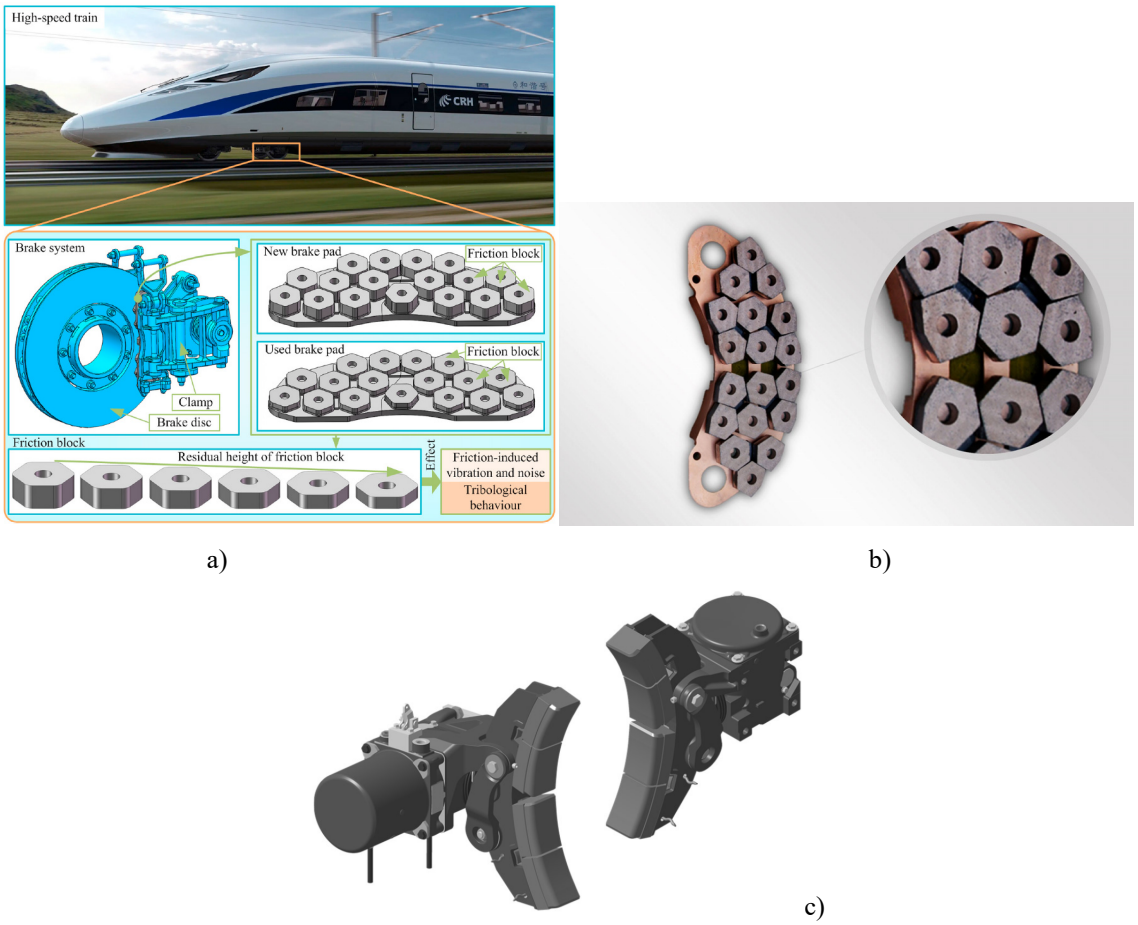


Figure 2: High-speed train disc and tread brake pad [7]

Research has proven that the friction and wear behavior of brake pad materials is primarily influenced by material characteristics, braking conditions, environmental conditions, surface conditions, and structural parameters. Train brake discs wear mainly due to the sliding contact between the pads and the disc. Abrasive and adhesive wear are the two primary wear mechanisms associated with disc brakes. The formation of brake wear particles is a result of abrasive and adhesive wear with oxidative processes. Two more specific wear modes that often occur in brake pads are fiber delamination, as well as fragmentary wear, such as induced pitting.

2.2. Brake Wear Particle Emission

Particulate matter can have harmful effects on the environment, climate, and human health, but the potential health risks of PM pollution were not widely investigated until the early 2000s. Results from many studies have indicated a link between exposure to PM particles and adverse health effects, but it has been established that health effects depend not only on particle concentration, but also on particle composition, as well as size, shape, and morphology. Larger particles are usually trapped and filtered in the nose and throat, while particles with a diameter of less than 10 μm can reach the lower respiratory tract. Ultrafine particles can further enter the bloodstream by penetrating the respiratory system, making them even more harmful than larger particles, as seen in Figure 3.

The travel distance and time until the particles settle on the ground due to gravity are dependent on the size and density of the particles. While coarse particles settle within hours, fine particles can remain suspended in the air for multiple weeks. About 90 % of PM is made up of fine and ultrafine particles under typical conditions. Non-exhaust emissions are a result of wear processes such as from the brakes, tires, clutch, engine and road surface, corrosion of vehicle parts, street signs and barriers as well as the re-suspension of settled particles on the road.

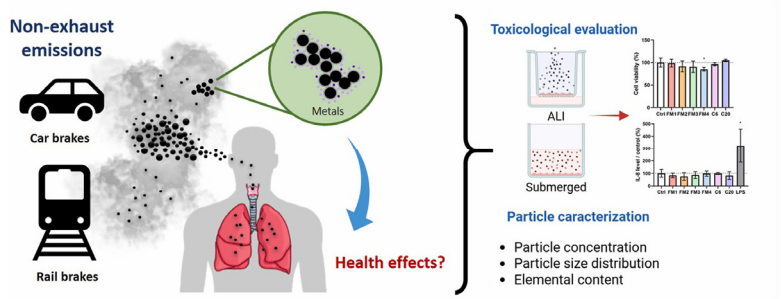


Figure 3: Human respiratory tract and deposition location of particles of different sizes.

2.3. Wear Particle Emissions in Rail Traffic

The data show that as of 2021, around 129,200 tonnes of $PM_{2.5}$ emissions and 201,400 tonnes of PM_{10} emissions in the EU are linked to transport as an industry, of which 171,300 tonnes of PM_{10} and 105,500 tonnes of $PM_{2.5}$ are from road transport alone. Rail vehicles contribute 38% of non-road PM_{10} emissions and 25% of $PM_{2.5}$ emissions. Strict regulations in the EU have contributed to a 55% reduction in annual $PM_{2.5}$ emissions and a 40% reduction in annual PM_{10} emissions between 2010 and 2019. However, 96% of urban residents are still exposed to levels of fine PM particles above the latest WHO (World Health Organization) guidelines. With stricter control of exhaust emissions, the relative share and importance of non-exhaust emissions are increasing. It has been found that up to 50% of non-exhaust PM_{10} emissions from road traffic are caused by disc brake wear. In car disc brakes, the braking torque on the rotor is generated by the friction force between the rotor and the brake pads. Since the rotor and the wheel are connected, the car will slow down due to the friction between the tire and the road. A similar situation applies to trains, where the braking force is transmitted by the contact between the wheel and the rail. During the friction braking process, wear occurs on the brake pads as well as on the rotor due to sliding contact. These brake wear particles can either deposit on the surface or be dispersed into the air. They have the potential to harm the environment and human health, as previously established. Studies have shown that it can be assumed that 35-50% of brake wear particles are in the air. Furthermore, brake wear has been found to be one of the main sources of metal particles in the air in cities.

Different types of brakes are widely used in railway transport, such as disc brakes and tread brakes. Fridell et al. [8, 9] observed in their research characteristic peaks in particle number concentration during deceleration phases, which, however, were not present in every braking process. The reason for this is the type of brake used in the braking process. Modern passenger trains are equipped with mechanical brakes as well as regenerative brakes (electric). Peaks in particle number concentration only occur if mechanical brakes are used for deceleration. Furthermore, the category of mechanical brakes must be divided into block brakes, which are widely used for freight trains, and disc brakes, which represent the standard braking system for passenger trains. While a wheel is usually equipped with one brake disc, axles can be equipped with two, three or even more brake discs.

The Euro 7 standard extends the regulatory requirements for the type-approval of braking systems and tyres with regard to particulate emissions and the abrasion process that occurs. The regulation applies to tyres of classes C1 (passenger cars and commercial vehicles), C2 (light commercial vehicles) and C3 (heavy commercial vehicles), in accordance with UN Regulation 117. As a result, the Euro 7 standard includes a requirement to limit the maximum emissions of PM_{10} particles from brake pads and tyres. Generally speaking, for electric cars the limit will be 3 mg/km, for combustion passenger cars, hybrids and hydrogen cars 7 mg/km, and for large vans and delivery vehicles 11 mg/km. The Euro 7 emission standard will first apply to the type-approval of new models of light vehicles and their braking systems 2.5 years after the entry into force of the regulation. A year later, all newly registered vehicles will have to meet the new Euro 7 requirements. Euro 7 for heavy goods vehicles will apply to new vehicle models 4 years after entry into force, and after 5 years to all new vehicles [10]. However, emission level limits have so far been based solely on the mass of particles per unit volume

of air, without considering particle number, composition and toxicity. The research community therefore suggests that this be taken into account for future limit values and assessment.

3. METHODS TO ANALYZE BRAKE WEAR EMISSIONS

The lack of standardized sampling and measurement methods for brake wear particle emissions in existing studies further complicates the comparison of results and conclusions. A major problem is that most studies are limited to automobile disc brakes, with very few examining other modes of transport, such as rail transport [11]. In order to evaluate braking performance and the influence of applied brake material, as well as to simulate different braking scenarios, a number of braking systems can be used, from small-scale ones, such as pin-on-disc tribometers, to component testing on a brake dynamometer, and even full-scale field tests (Figure 4). Compared to brake dynamometers, their main advantages are the small dimensions of the test platform and lower operating costs. Brake dynamometers with disc brakes are mainly used for homologation testing and evaluation of damage mechanisms of various disc brake systems. However, in recent years, these devices have been used more frequently to generate brake wear particles due to their ability to replicate vehicle and kinetic energy data under controlled laboratory conditions that are very close to those in the real world.

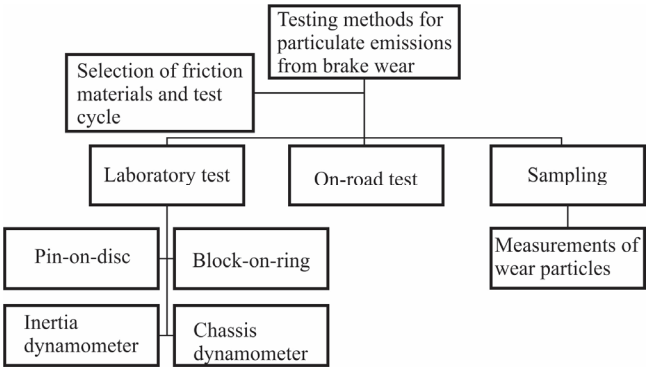


Figure 4: Overview of different laboratory and on-road testing methods for particulate emissions from brake wear

The pin-on-disc test is a robust method for studying the formation of airborne particles, based on the similarity of morphology and size distribution, with the prerequisite that the same contact conditions are used in laboratory tests as in field tests. Pin-on-disc tribometers are often used to generate brake wear particles in air under controlled operating parameters (contact pressure, sliding speed and friction material temperature) [12, 13, 14].

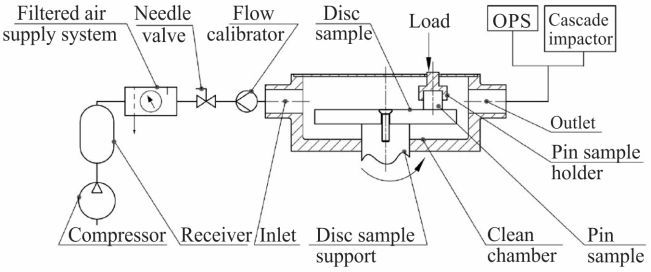


Figure 5: Schematic of the experimental set-up based on a pin-on-disc tribometer [12]

One of the few published on-board measurements was conducted by Fridell et al. [8, 9], who reported particle characteristics and emission factors for wear particles from a running Regina train. They investigated particles size in 7 intervals and aimed to investigate 27 elements among collected particles. The curve negotiation effects were not traceable in his work as they used one sampling point that was located between the two coaches. The main objective of field test conducted by Abbasi et al. [3] was to study the characteristics of particles generated by disc brakes on a moving train using a measurement system installed on the train. Two points were defined for sampling airborne particles, one close to the disc brake pad-rotor contact and the other below the frame, not close to the mechanical brake or wheel-rail contact; the number and size distribution of the detected particles were recorded and evaluated under different conditions (Figure 6). The difference in the distance from the sampling point to the main sources of particles resulted in different results. The morphology and elemental composition of the collected particles were also studied. Different realistic operating conditions were taken into account, such as train acceleration, train deceleration or train movement through a curve. This type of research could pave the way for the development of significantly simpler and cheaper sensors that can be mass-produced and used in a wide range of applications, e.g. where there are potential operational disruptions (e.g. due to faulty braking systems), thereby reducing particulate emissions to an absolute minimum.

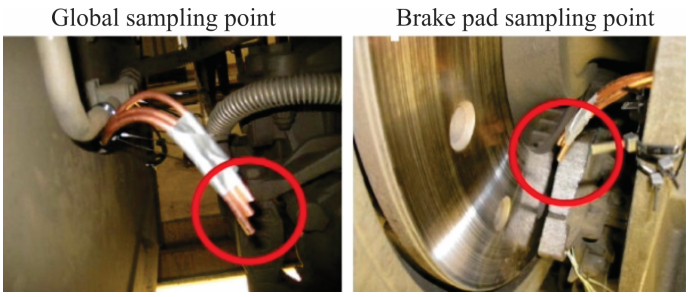


Figure 6: Particle measurement devices, along with connecting tubes arrangement [3]

To study the emission of wear particles from subway trains, it is necessary to measure both the particle concentration and the air flow. A common method for this is to use a tunnel because the tunnel design provides a given volume and cross-section that facilitate air flow estimates (Figure 7). The particle concentration is usually measured at the tunnel exit with a sampling frequency of the order of 1 Hz, together with two-dimensional wind flow measurements. Two pairs of diode laser/photodetectors placed on the side of the track can be used to measure the speed and length of the test trains. To assess the emission from individual trains, the particles emitted in the tunnel must be ventilated between each train passage, and therefore it is necessary to choose a tunnel where the train frequency is low, from only one to a few trains per hour. The tunnel is ventilated by both the flow caused by passing trains and the ambient wind, but choosing a tunnel whose length is the same as the length of the train facilitates wind-induced ventilation. The small cross-sectional area facilitates train-induced ventilation, but probably limits wind-induced ventilation. The use of a stationary particle measurement method allows for the estimation of the emission coefficient at the tunnel exit, where there is a well-defined measurement volume, and allows for the comparison of emission levels between passenger and freight train types [8].

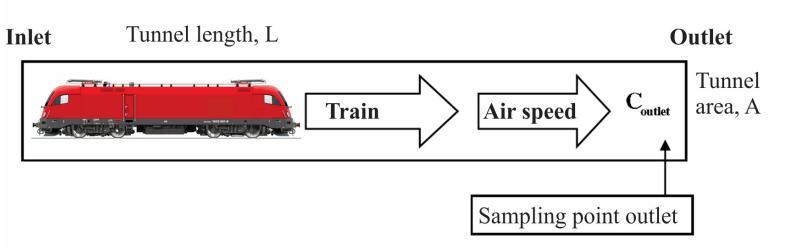


Figure 7: Tunnel measurements of particle emissions [8]

A tribological test in the laboratory can be carried out on an inertial dynamometer type test bench to simulate the braking performance of a high-speed train and to examine the influence of the residual height of the friction block on the vibration and noise characteristics caused by friction [7]. A customized test bench for simulating the braking performance of a high-speed train is shown in Figure 8. Test bench of this type are often also made in a reduced scale of 1:2 to simulate the friction, load and speed ratios of a real braking system. This test device can be used to test the friction of brakes under different conditions and operating modes (traction braking, emergency braking, etc.), as well as to simulate the tribological behavior and particle emissions of the braking system. The samples of discs and blocks used in this study must be reduced to an appropriate scale.

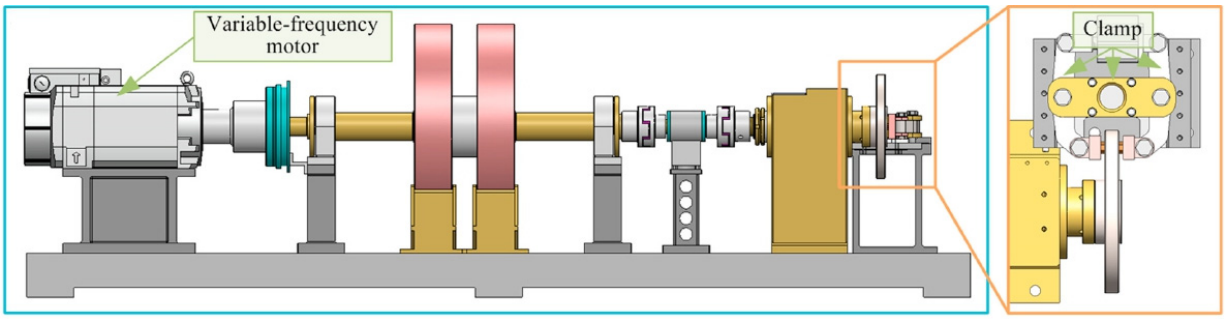


Figure 8: The test device to analyze high-speed train braking performance:
control console schematic diagram of test device [7]

The newly built test bench for railway brakes (Figure 9) was put into operation in 2023 at the Graz University of Technology. This test bench allows the testing of disc brake systems as well as tread brakes (block brakes). The kinetic energy of the vehicle is simulated using an electric mass simulation. The drive motor accelerates the drive train of the test bench to any speed in the range from 0 to 3,000 rpm. This also allows the testing of brakes on high-speed trains, which operate at a maximum speed of 500 km/h [15, 16, 17].

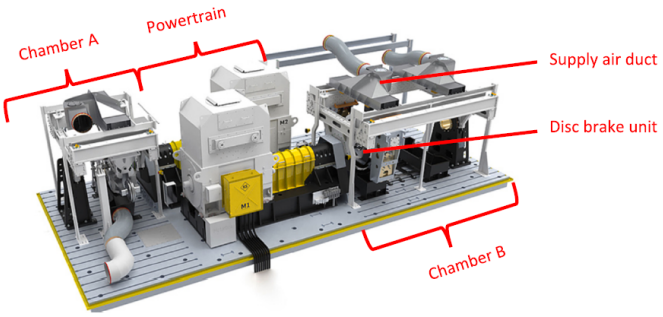


Figure 9: Scheme of the brake test bench at Graz University of Technology [15, 16]

4. DESIGN OF EXPERIMENTS

Two approaches are used for particle characterization: a particle collection approach for off-line analysis and an instantaneous measurement approach with on-line particle size distribution measurement.

The aim of future research is to analyze the influence of railway braking conditions, such as contact pressure and disc temperature, in particular with regard to the range of emitted particles, their shape and composition. Knowledge of the nature of these particles (size, shape and composition) and their emission in real time is necessary for the development of numerical fluid models that can correctly predict their propagation near the zone of their generation. An important aspect of the research is the selection of friction conditions that give lower particle emission for a given dissipated kinetic energy and the given rules for the particle injection properties in the CFD model to study their propagation in a confined space, such as a metro station. The assessment of the influence of these parameters on the nature of emitted PM particles will be carried out on a different railway braking materials, varying one parameter at a time, while maintaining the equivalent dissipated energy.

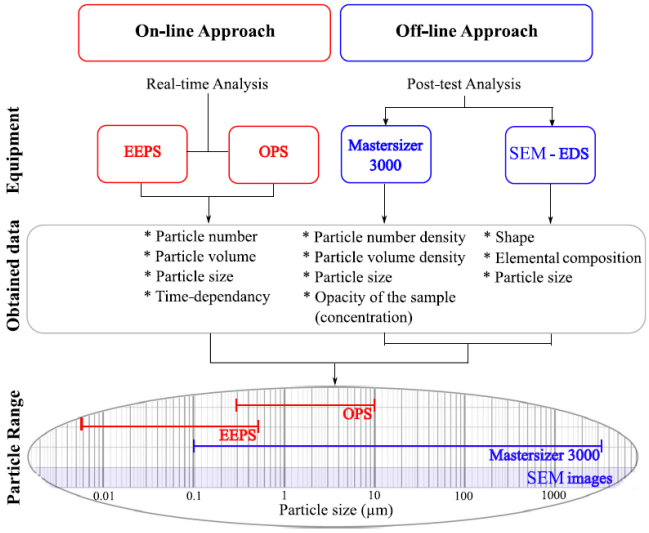


Figure 10: Description of the experimental approach [18]

4.1. On-Line Measurements of Particle Emissions

It is important to recognize that this type of experiment only provides qualitative insight into how sliding speed and contact pressure can affect wear, friction, and airborne particle emissions. The quantitative results cannot be translated into real-world scenarios, as factors such as the detection range and calibration of particle sampling instruments and the difference in contact area compared to actual disc brakes affected these results. In addition, many studies have found that the brake temperature has a significant effect on the generation of brake wear particles. Therefore, many experimental setups include thermocouples or infrared cameras to record the temperature at the contact surface [5].

The advantage of inertial brake dynamometers is that the device tests real brakes that are also used on road vehicles under controlled conditions. Such tests belong to the group of tests that are closest to the actual braking process on a vehicle. An inertial brake dynamometer (BrakeDyno 2020) was developed at the Faculty of Engineering of the University of Kragujevac. In order to simulate the kinetic energy of one quarter of the vehicle, an inertial mass is applied, which is driven by an electric motor, and at the end of the shaft is the rotor itself, i.e. disc brakes. Changing the number of revolutions of the flywheel enables the selection of the value of the kinetic energy, thus changing the simulated initial braking speed and the load of one quarter of the vehicle. The brake is activated using pneumatic and hydraulic mechanisms. The entire braking process, starting speed, pressure and load of one quarter of the vehicle is controlled

electronically. BrakeDyno 2020 contains a series of sensors that can be used to monitor various output parameters (brake pad temperature, braking torque, braking time, ...). Bearing in mind that the inertial dynamometer has been developed so that adaptations can be made for various measurements and research, Figure 11 shows the appearance of BrakeDyno 2020, which has been modified in order to measure the concentration of particles that result from the wear of friction pairs of brakes.

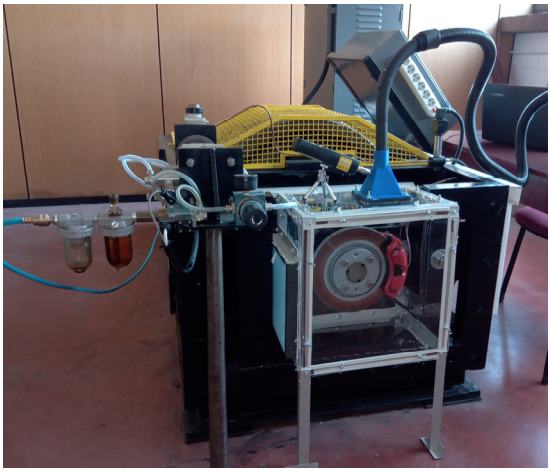


Figure 11: BrakeDyno 2020 in the process of adaptation for particle measurement

The directly measured variables are: angular velocity of the disc, number of revolutions of the disc, braking time, brake pressure, i.e. pressure in the brake installation, both in the hydraulic and pneumatic parts of the brake installation, braking torque, brake pad temperature, ambient temperature, air humidity, $PM_{2.5}$ and PM_{10} particle concentrations.

The variables determined by recalculation are: simulated vehicle speed, kinetic energy, deceleration, coefficient of friction between friction pairs. Such research, which would require modification of the test bench for railway vehicle brakes, could:

- Identify the factors influencing the concentration of particles generated by the wear of brake friction pairs;
- Determine whether factors (parameters) can differently affect the concentration of particles by changing their intensity, depending on the brake pad used;
- Identify whether the applied brake pads also have an impact on the concentration of particles in terms of their composition and design characteristics;
- Analyze the influence of individual factors on the resulting concentration of PM_{10} and $PM_{2.5}$ particles individually;
- Determine the influence of individual factors on the relationship between the concentrations of PM_{10} and $PM_{2.5}$ particles individually.

4.2. Off-Line Analysis of Particles

In the post-test characterization approach, collection experiments are conducted to further characterize the generated particles and compare different size analysis methods. These results provide information about the nature and shape of the large particles.

5. CONCLUSIONS

As we have not yet reached definitive conclusions regarding the toxicity of railway brake wear particles, it is essential to pursue a preventive approach aimed at reducing particle emissions. This is particularly important given evidence from previous research suggesting that brake wear particle components, such as iron, can induce oxidative stress in cells. As trains continue to rely on mechanical brakes without the widespread adoption of alternative technologies such as inductive braking, it is crucial to develop strategies to reduce particle emissions. These efforts are essential to minimize potential adverse health and environmental impacts.

A test that provides results similar to real conditions is the use of an inertial brake dynamometer, where the same friction elements as those on a real vehicle are used. Since the measurement is performed in laboratory conditions, it is much easier to isolate the emissions resulting from brake wear. The inertial brake dynamometer is usually modified by installing a housing in which the brake assembly is located. In laboratory conditions, it is much easier to set and control the braking factors, but also to identify other parameters that have an impact on the resulting particle emissions, such as the temperature between the friction pairs, relative humidity, ambient temperature, etc.

ACKNOWLEDGEMENTS

Paper is financially supported by The Ministry of Science, Technological Development and Innovation of the Republic of Serbia, Contract No. 451-03-65/2024-03/200107.

REFERENCES

[1] Pravilnik o kočnicama i kočenju vozova i vozila, ("Sl. glasnik RS", br. 68/2021)

[2] S. Abbasi, J. Wahlström, L. Olander, et al. “A study of airborne wear particles generated from organic railway brake pads and brake discs”, Wear, Vol. 273(1), pp. 93-99, https://doi.org/10.1016/j.wear.2011.04.013, (2011)

[3] S. Abbasi, L. Olander, C. Larsson, U. Olofsson, A. Jansson, U. Sellgren, “A field test study of airborne wear particles from a running regional train”, Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, Vol. 226(1), pp.95-109, https://doi.org/10.1177/0954409711408774, (2012)

[4] S. Abbasi, A. Jansson, U. Sellgren & U. Olofsson, “Particle Emissions From Rail Traffic: A Literature Review”, Critical Reviews in Environmental Science and Technology, Vol. 43(23), pp. 2511–2544, https://doi.org/10.1080/10643389.2012.685348, (2013)

[5] L. Tanzeglock, “Particle Emissions of Train Brakes”, Master Thesis, Division of Production and Materials Engineering, Lund University, Lund (Sweden), (2024)

[6] Y. Lyu, L. Tanzeglock, Q. Zhang et al. “Friction, wear and particle emissions from copper-based train brakes”, Proceedings 22st Nordic Seminar on Railway Technology, KTH Royal Institute of Technology, Stockholm (Sweden), 18-19 June 2024, (2024)

[7] Z.Y. Xiang, J.K. Zhang, S.L. Xie, et al. “Friction-induced vibration and noise characteristics, and interface tribological behavior during high-speed train braking: The effect of the residual height of the brake pad friction block”, Wear, Vol. 516–517, pp. 204619, https://doi.org/10.1016/j.wear.2023.204619, (2023)

[8] E. Fridell, A. Björk, M. Ferm, A. Ekberg, “On-Board Measurements of Particulate Matter Emissions from a Passenger Train”, Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, Vol. 225(1), pp. 99-106, https://doi.org/10.1177/09544097JRRT407, (2011)

[9] E. Fridell, M. Ferm, A. Ekberg, “Emissions of particulate matters from railways – Emission factors and condition monitoring”, Transportation Research Part D: Transport and Environment, Vol. 15(4), pp. 240-245, https://doi.org/10.1016/j.trd.2010.02.006, (2010)

[10] Z. Stępień, “Vehicle related non exhaust particle emissions – Euro 7 requirements”, Combustion Engines. Vol. 199(4), pp.15-29, https://doi.org/10.19206/CE-190606, (2024)

[11] A. T. Juárez-Facio, M. Introna, A. M. Montes, et al. “In vitro toxicity of car and train brake wear emissions using Air-Liquid Interface and submerged exposure models”, Atmospheric Environment: X, Vol. 29, p. 100426, https://doi.org/10.1016/j.aeaoa.2026.100426, (2026)

[12] Y. Lyu, Q. Zhang, L. Tanzeglock, et al. “Mapping of friction, wear and particle emissions from high-speed train brakes”, Wear, Vol. 572–573, p. 206027, https://doi.org/10.1016/j.wear.2025.206027, (2025)

[13] Y. Tsybrii, I. Zglobicka, M. Kuciej, et al. “Airborne wear particle emission from train brake friction materials with different contents of steel and copper fibres”, Wear, Vol. 504–505, p. 204424, https://doi.org/10.1016/j.wear.2022.204424, (2022)

[14] Y. Tsybrii, O. Nosko, I. Zglobicka, M. Kuciej, “Emission and properties of airborne wear particles from train brake friction materials based on novolac phenolic resins and butadiene rubbers”, Wear, Vol. 546–547, p. 205332, https://doi.org/10.1016/j.wear.2024.205332, (2024)

[15] D. Fruhwirt, P. Sturm, T. Nöst, et al. “PM emissions from railways – Results of tests on a wheel-rail test bench”, Transportation Research Part D: Transport and Environment, Vol. 122, p. 103858, https://doi.org/10.1016/j.trd.2023.103858, (2023)

[16] D. Fruhwirt, T. Nöst, P. Leonhardt, et al., “Characterization of emissions from axle-mounted rail disc brakes”, Transportation Research Part D: Transport and Environment, Vol. 130, p. 104181, https://doi.org/10.1016/j.trd.2024.104181, (2024)

[17] D. Fruhwirt, T. Nöst, P. Leonhardt, et al. “Emission characteristics of railway tread brakes”, Atmospheric Environment: X, Vol. 25, p. 100320, https://doi.org/10.1016/j.aeaoa.2025.100320, (2025)

[18] C. Octau, D. Méresse, M. Watremez, J. Schiffler, M. Lippert, et al. “Characterization of particulate matter emissions in urban train braking - An investigation of braking conditions influence on a reduced-scale device”, Environmental Science and Pollution Research, Vol. 27(15), pp.18615-18631, https://doi.org/10.1007/s11356-020-08337-8, (2020)

SESSION B: RAILWAY ENGINEERING

Influence of signal-to-noise ratio on the accuracy of experimental determination of wheel-rail interaction forces

Milan Bižić[1*], Dragan Petrović[1]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Kraljevo, Serbia

ARTICLE INFO

[*] **Correspondence:** bizic.m@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.B07B **PAGES**: B51 – B57

ABSTRACT

This paper analyzes the influence of the signal-to-noise ratio (SNR) on the accuracy of the experimental determination of wheel-rail interaction parameters – vertical force Q, lateral force Y, and contact point position y_{tk}. The research was conducted on an instrumented wheelset developed from a standard wheelset, with strain gauges optimally positioned and connected into Wheatstone bridges. Experimental tests were carried out on a specialized test stand within the Laboratory for Railway Engineering and Structural Testing at the Faculty of Mechanical and Civil Engineering in Kraljevo, University of Kragujevac. The noise level in the electronic components of the measurement system and its influence on the final measurement error were analyzed based on the results of corresponding static and dynamic measurements. Using appropriate statistical analysis, it was determined that the noise level originating from the Wheatstone bridges and associated electronics causes significant dispersion of output signals, especially at lower force intensities. The results show that under lower loads (characteristic of empty vehicles) the SNR is less favorable, leading to an increase in the relative measurement errors of the aforementioned parameters. Conversely, under higher loads and critical values of the Y/Q ratio, which are of interest for assessing vehicle running stability according to relevant international standards, the SNR is significantly more favorable and measurement errors decrease. The obtained results confirm that the SNR is a key factor in the reliability of the measurement system for the experimental determination of wheel-rail interaction parameters. One of the directions for further improvement in measurement accuracy is the application of modern fiber-optic sensors, which have significantly higher resistance to electromagnetic interference and enable a more favorable SNR, thus opening a promising path toward a new generation of instrumented wheelsets with even higher measurement accuracy.

KEYWORDS

Signal-to-noise ratio, SNR, Wheel-rail interaction, Wheel-rail contact forces, Wheel-rail contact point, Instrumented wheelset, Derailment coefficient, Strain gauges.

1. INTRODUCTION

Wheel-rail interaction represents one of the most complex and most studied phenomena in the dynamics of railway vehicles [1, 2]. On an extremely small contact surface, measuring only 1.5 to 5 cm², very intense forces are generated, which have a key influence on running stability, safety, comfort, as well as on the wear and fatigue of railway vehicle and track components. Among the most important parameters for assessing the dynamic behavior of railway vehicles, according to international standards UIC 518 and EN 14363, are: the vertical wheel-rail contact force Q, the lateral wheel-rail contact force Y, their ratio Y/Q, which is a direct indicator of derailment risk, and the position of the wheel-rail contact point y_{tk} [3, 4].

In order to determine these parameters as accurately as possible, the most reliable method is experimental testing using instrumented wheelsets. Modern instrumented wheelset systems, such as the one presented in this paper, are essentially standard commercial wheelsets with monoblock wheels onto which strain gauges are glued. The strain gauges are connected in appropriate Wheatstone bridge configurations and placed at pre-determined optimal radial distances in order to maximize sensitivity to the forces being measured [5–7]. Although continuous efforts are being made to improve the design, calibration and signal processing algorithms, these measurements are still characterized by significant

measurement uncertainty, and the study of errors and their causes therefore remains one of the key development directions in this field [8].

From the aspect of measurement accuracy, the signal-to-noise ratio (SNR) has been identified as the dominant influencing parameter, and its increase is directly related to the possibility of increasing measurement accuracy [9]. Recent research has also shown that errors arising from contact point movement and wheel rotation represent a significant source of problems [10–12]. However, the influence of noise on the signal, particularly pronounced under low loads (characteristic of empty wagons), remains one of the greatest challenges.

In accordance with the foregoing, the main objective of this paper is to examine the influence of the SNR on the accuracy of determining the wheel-rail interaction parameters (Q, Y, Y/Q and y_{tk}), based on a detailed analysis of experimental data obtained using an experimental setup consisting of an instrumented wheelset and a test stand. The test results have shown that the experimental setup used enables the determination of the vertical force Q, the lateral force Y and their ratio Y/Q with a relative error of less than 10%, and the contact point position y_{tk} with an error of less than 15%. Of particular importance is the comparative analysis of measurement accuracy under low and high loads, as well as the indication of the practical implications of the obtained results. This quantitative assessment provides the basis for drawing final conclusions with a special emphasis on future research directions in the field of developing instrumented wheelsets.

2. FORCES IN WHEEL-RAIL INTERACTION

When negotiating curves, there is a constant risk of the leading wheel flange climbing onto the outer rail head, i.e., of the railway vehicle derailing from the track. The key indicators of this risk, i.e., the key influencing parameters on which the running stability of railway vehicles in curves depends, are the forces in the wheel-rail interaction. Knowing these forces precisely allows very reliable conclusions to be drawn about the running stability of railway vehicles in curves, based on the results of research into the derailment mechanism.

In practical analyses of the running stability of railway vehicles, the most commonly used parameters are the vertical forces Q and the lateral forces Y, which are derived from the actual wheel-rail interaction forces – the normal forces N and the lateral tangential forces F_η (Fig. 1).

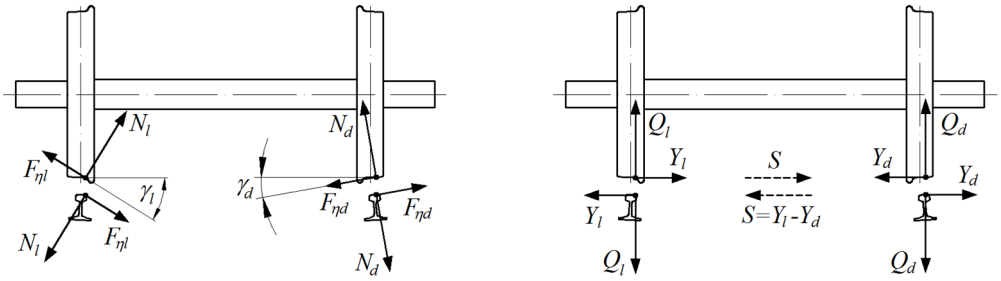


Figure 1: Vertical forces Q and lateral forces Y derived from normal forces N and lateral tangential forces F_η in wheel-rail interaction

The lateral force Y acts parallel to the track plane, while the vertical force Q acts normal to the track plane, so that the values of these forces are obtained by projecting the actual forces onto the said planes. Accordingly, the vertical and lateral forces in the interaction of the left (l) and right (d) wheel and rail are:

$$Q_l = N_l \cos\gamma_l + F_{\eta l} \sin\gamma_l \tag{1}$$

$$Y_l = N_l \sin\gamma_l - F_{\eta l} \cos\gamma_l \tag{2}$$

$$Q_d = N_d \cos\gamma_d - F_{\eta d} \sin\gamma_d \tag{3}$$

$$Y_d = N_d \sin\gamma_d + F_{\eta d} \cos\gamma_d \tag{4}$$

The resulting lateral force S in the wheel-rail interaction is:

$$S = Y_l - Y_d \tag{5}$$

The total vertical force Q in the wheel-rail interaction during motion is variable and consists of static, quasi-static, dynamic, and additional components. The static component is constantly present during motion and practically represents the static load Q_0, which includes the dead weight of the railway vehicle and the weight of the cargo. In determining it, it is assumed that the total weight of the vehicle and cargo is evenly distributed across all wheels. The quasi-static component appears only when railway vehicles negotiate curves, in cases where the running speed differs from the equilibrium speed. In that case, the magnitude of uncompensated lateral acceleration and the height of the vehicle's center of mass dictate the redistribution of the total vertical load between the wheels of the wheelsets. The dynamic component is present to a greater or lesser extent during the motion and arises from numerous disturbances originating from track irregularities and damage, irregularities and damage of wheelsets, running instability, traction, braking, etc. The additional component most often arises from uneven load distribution and asymmetry, i.e., errors in the assembly and adjustment of certain structural

elements of railway vehicles (most often suspension elements). The total resulting lateral force S during motion is also variable and consists of quasi-static, dynamic, and additional components. The quasi-static component appears when railway vehicles negotiate curves and depends on the uncompensated lateral acceleration and the distribution of forces between the wheelsets in the bogie. The dynamic and additional components arise due to a large number of disturbances, similarly to the case of the vertical force Q.

The study of the derailment mechanism by flange climbing and the assessment of running stability of railway vehicles in curves are directly related to the ratio of lateral and vertical forces in wheel-rail interaction. To determine the limit values of the Y/Q ratio, the Nadal equation is most commonly used.

Due to the under-radial steering of the wheelset and the existence of a certain angle of attack of the wheel against the outer rail, a lateral tangential force F_η appears, which tends to lift the wheel flange onto the rail head (Fig. 2). In general, this force equals the product of the normal force N acting in the contact surface and the coefficient of friction μ between the wheel and the rail. Therefore, if the wheel flange is pressed more strongly against the rail head (greater normal force N) or if the coefficient of friction μ is higher, the tangential force F_η that tries to raise the wheel flange onto the rail head is also greater.

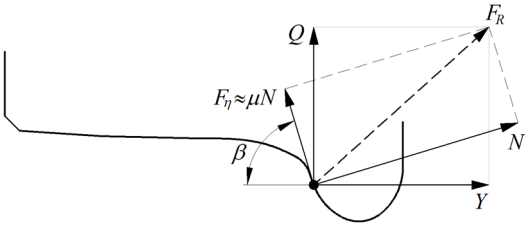


Figure 2: Forces in the initial phase of wheel flange climbing onto the rail head

In analysis of this problem, Nadal assumed that the angle of attack of the wheel against the rail has rather large values ($\psi \geq 0.5°$), so that the contact angle γ is approximately equal to the wheel flange angle β, and the tangential force F_η acts almost vertically. According to Nadal, the wheel flange begins to climb onto the rail head if the vertical component of the normal force N and the tangential force F_η is equal to or greater than the wheel load Q. Based on Fig. 2 and the previous expressions, the lateral and vertical components of the normal force N and the tangential force F_η in the wheel-rail interaction are:

$$Y = N \sin\beta - \mu N \cos\beta \tag{6}$$

$$Q = N \cos\beta + \mu N \sin\beta \tag{7}$$

Based on the previous expressions, the Nadal equation can be defined, which enables the determination of the critical ratio of lateral to vertical force Y/Q at which the process of wheel flange climbing onto the rail head begins, using the known coefficient of friction μ and the wheel flange angle β:

$$\frac{Y}{Q} = \frac{\tan\beta - \mu}{1 + \mu \tan\beta} \tag{8}$$

In the international standards UIC 518 and EN 14363, limits, i.e., limit values of the ratio of vertical and lateral forces in wheel-rail interaction, are defined, which must be satisfied when assessing the dynamic behavior quality of railway vehicles. In this case, the absolute maximum values of the ratio of vertical or lateral force are not relevant for comparison with the defined limit values; rather, following appropriate statistical processing, the 99.85 percentile of the signal filtered with a low-pass filter having a cutoff frequency of 20 Hz is taken into account:

$$\left(\frac{Y}{Q}\right)_{20\text{Hz, 2m, mean, 99.85\%}} \leq \left(\frac{Y}{Q}\right)_{doz} = 0.8 \;\; - \text{for curves of radius} \;\; R_k \geq 250 \;\; m \tag{9}$$

In experimental tests, i.e., prototype testing of railway vehicles, the aforementioned Y/Q ratio is directly recorded using instrumented wheelsets, with one of the greatest problems being its contamination due to the influence of noise.

3. EXPERIMENTAL SETUP

For the tests, an instrumented wheelset was used, based on a standard commercial wheelset intended for freight railway vehicles (Fig. 3). The wheelset has monoblock wheels with an S1002 profile, a nominal rolling diameter of 920 mm, and a static load per wheel of 112.5 kN. The main components of the measurement system are connected to each other via a wireless Ethernet network, which enables communication between them. These components include: the instrumented wheelset equipped with strain gauges placed on both wheels; the electronic-computer unit for receiving and storing measurement signals during experimental testing (housed in the measuring wagon during track tests); and the computer module for processing and displaying experimental results (which can be located outside the measuring wagon during track tests).

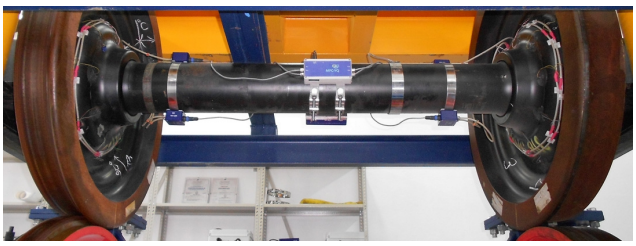


Figure 3: Instrumented wheelset[4,9]

Each wheel of the instrumented wheelset contains glued strain gauges at four radial distances from the wheel center, namely: ρ=360 mm (outer side of the disc) – for measuring the vertical force Q, ρ=168 mm (inner side of the disc) – for measuring the lateral force Y, ρ=262 mm (inner and outer side of the disc) – for measuring the contact point position y_{tk} [6, 9]. At each radial distance, 8 strain gauges are placed, arranged at an angle of 45° and connected in a full Wheatstone bridge configuration. Signals from the measuring bridges are collected in the appropriate computer unit via a telemetry measurement system, while the angular position of the axle Φ_0 is also measured using biaxial accelerometers.

Experimental tests were carried out on a special test stand within the Laboratory for Railway Engineering and Structural Testing at the Faculty of Mechanical and Civil Engineering in Kraljevo, University of Kragujevac (Fig. 4). The test stand enables experimental testing and calibration of instrumented wheelsets with a load capacity of 22.5 t for normal track gauge, i.e., independent application of vertical force Q and lateral force Y under laboratory conditions. The main parts of the test stand are: 1–lower wheelset, 2–horizontally movable carrier of the lower wheelset, 3–upper wheelset (wheelset being tested-calibrated), 4–vertically movable carrier of the upper wheelset, 5–supporting structure, 6–hydraulic systems for applying vertical and lateral forces, 7–transducers for recording the applied values of vertical and lateral forces, 8–drive electric motor with gearbox and cardan shaft, 9–safety systems, and 10–control module.

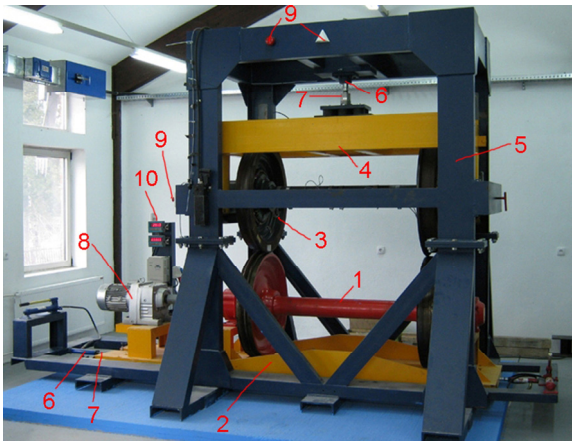


Figure 4: Special test stand for testing and calibration of instrumented wheelsets [4, 9]

The lower wheelset of the test stand has its left wheel machined to simulate the rail head, while the upper wheelset rests on rolling bearings on the right side, ensuring free rotation. The vertical force is applied by a hydraulic cylinder acting on the vertically movable carrier of the upper wheelset. The lateral force is applied by a horizontal hydraulic cylinder acting on the horizontally movable carrier of the lower wheelset. The values of the applied forces are measured by force transducers, with the maximum applicable force values being Q=225 kN and Y=100 kN. The rotational motion of the tested wheelset is achieved indirectly via the lower wheelset, which is driven by an electric motor with a gearbox and a cardan shaft. The drive electric motor has a power of 7.5 kW, and the rotational speed is regulated by a frequency converter located on the control module.

4. SIGNAL-TO-NOISE RATIO AND MEASUREMENT ERROR

In the first phase of experimental tests, static measurements were carried out primarily to verify the linearity of the system. The measurements were performed for three characteristic positions of the contact point between the wheel and the rail simulation, as well as for eight angular positions of the wheel. The formed diagrams of the dependence of the wheel-rail interaction forces applied on the test stand and the signals from the measuring bridges showed a trend of linear dependence between them, thus confirming the linearity of the system. As expected, a certain measurement uncertainty was observed, i.e., a certain degree of dispersion of the measurement results relative to the ideal dependence between the signal and the force, in which all discrete signal values would lie on a single straight line. Typical examples are shown in Figure 5.

The main cause of this problem is noise in the electronic components of the measurement system – in this case, the force transducers and the measuring bridges with strain gauges. It was observed that the influence of signal dispersion is particularly pronounced in cases where the measuring bridges have lower sensitivity to the applied loads, i.e., when the

SNR has lower values. Therefore, it is quite clear that maximizing the SNR represents a key problem in the development of measurement systems for instrumented wheelsets used for the experimental determination of forces in wheel-rail interaction.

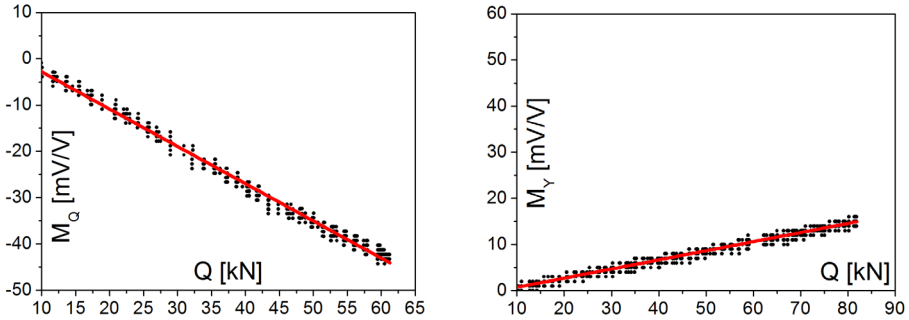


Figure 5: Dispersion of measurement results of signals from the measuring bridges for measuring *Q* and *Y* forces

In the next phase, dynamic measurements were carried out with the aim of testing the measurement accuracy of the instrumented wheelset. The accuracy was tested for several different load cases and combinations, i.e., known wheel-rail interaction parameters applied on the test stand, which were to be reconstructed by separating the mixed signals from the instrumented wheelset. Emphasis was placed on two characteristic test cases: in the first, the loads on the test stand approximately corresponded to the wheel loads for an empty freight wagon, and in the second, for a loaded freight wagon. By applying the method for separating mixed signals from the measuring bridges and the corresponding algorithm, the values of the applied parameters in the wheel-rail simulation interaction were reconstructed [4, 9]. Comparative diagrams of the applied and reconstructed values are shown in Figs. 6–8.

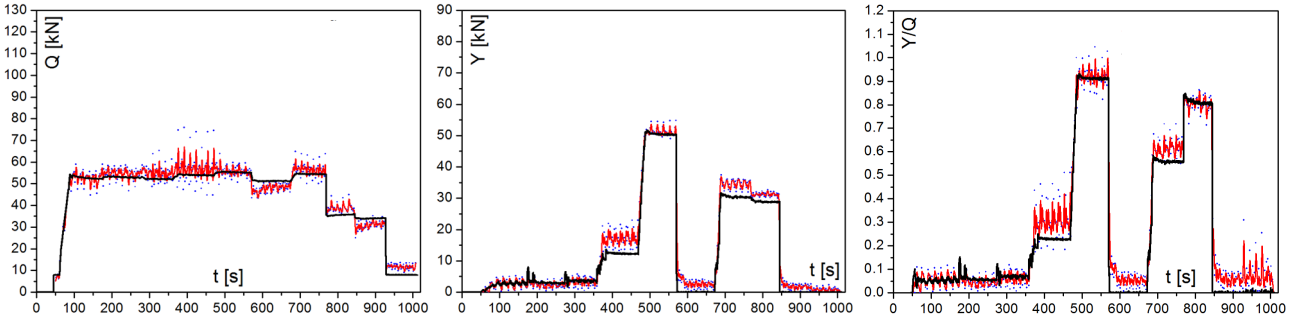


Figure 6: Comparative diagrams of the applied and reconstructed values of *Y*, *Q*, and *Y*/*Q* for the first test case [9]

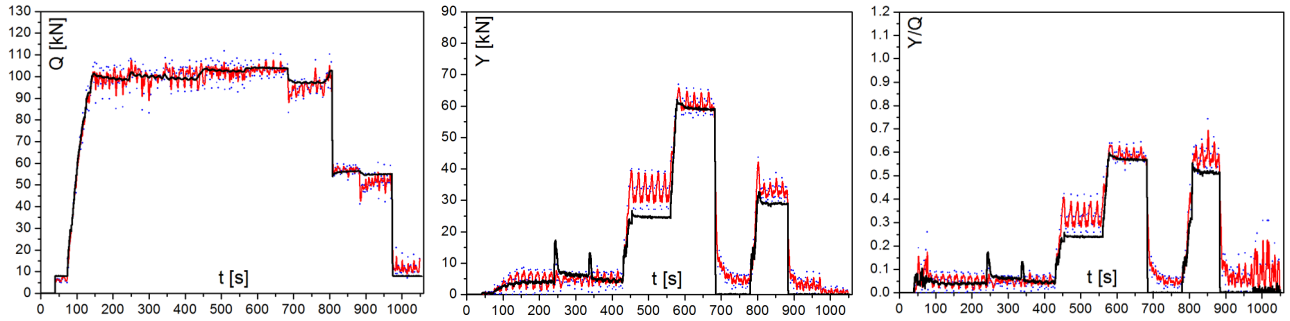


Figure 7: Comparative diagrams of the applied and reconstructed values of *Y*, *Q*, and *Y*/*Q* for the second test case [9]

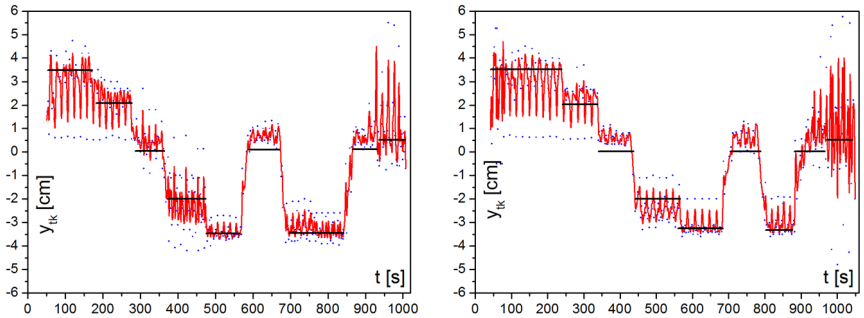


Figure 8: Comparative diagrams of the applied and reconstructed values of y_{tk} for the first test case (left) and the second test case (right) [9]

A very high degree of agreement can be observed between the applied values (black lines in the diagrams) and the reconstructed values (red lines in the diagrams), which confirms the high measurement accuracy of the given instrumented wheelset. The obtained results showed that for angular positions with lower signal values from the measuring bridges,

i.e., with a less favorable SNR, a higher degree of dispersion of the wheel-rail interaction parameters obtained by inverse identification, i.e., by reconstruction of the values applied on the test stand, is characteristic (blue points in the diagrams). In angular positions where the sensitivity of the measuring bridges to the applied loads is higher, the degree of dispersion of the reconstructed wheel-rail interaction parameters is significantly smaller. Precisely due to the dispersion caused by the influence of noise, and for easier visualization, the obtained results of the reconstructed wheel-rail interaction parameters are shown using B-splines (red lines in the previous diagrams).

Based on the detailed statistical analyses conducted, it was estimated that the given instrumented wheelset enables the measurement of the vertical force Q, the lateral force Y, and their ratio Y/Q in wheel-rail interaction with errors within limits of less than 10% [4, 9]. Furthermore, it was estimated that the measurement error of the contact point position y_{tk} is within limits of less than 15%. It is particularly important to emphasize that during flange contact (with significant values of the Y/Q ratio or Y force), the previously estimated measurement errors decrease, which is extremely important for the reliability of experimental testing of derailment safety of railway vehicles according to international standards.

For a more detailed analysis of measurement uncertainty, three measurements were performed in static mode with no load, during which output signals from the measuring bridges and signals from the force transducers, i.e., the alpha-numeric displays, were recorded. All three measurements lasted approximately 300 s, thereby sampling a sufficient amount of experimental data for valid statistical analysis. Based on the measurement results, diagrams of the signal changes from the four measuring bridges and of the forces in the wheel-rail simulation interaction were formed as a function of time. In addition, diagrams (histograms) of the statistical distribution of the recorded signals were created, showing the mean values and standard deviations. As an example, the diagram of the signal change from the measuring bridge for measuring the Q force, its statistical distribution, mean value and standard deviation for Measurement 1 are shown in Figure 9.

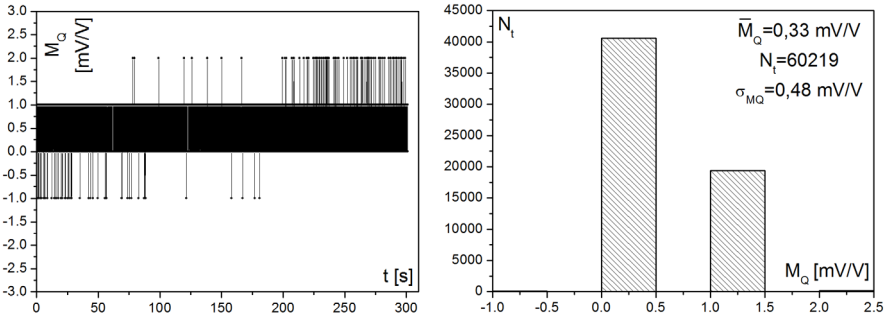


Figure 9: Diagram of the signal from the measuring bridge for measuring the Q force, statistical distribution, mean value and standard deviation for Measurement 1

Analysis of the results obtained from the formed diagrams showed that the degree of dispersion of the recorded force values in the wheel-rail simulation interaction is not large and, in the most unfavorable case, ranges from –2 to 2 mV/V. The recorded values of the vertical force Q scatter within the range of 0–15 N, and the lateral force Y within the range of 0–30 N. Based on this, it can be concluded that noise originating from the force transducers does not have a significant influence on the measurement uncertainty during calibration of the measurement system on the test stand. Therefore, it is evident that the present measurement uncertainty arises exclusively as a consequence of noise in the electronic components related to the operation of the measuring bridges. If the noise level in the electronic components related to the operation of the measuring bridges were reduced, the aforementioned dispersion intervals of the output signals would also decrease. This would consequently result in a significantly lower degree of dispersion of the wheel-rail interaction parameters obtained at the output of the inverse identification algorithm. The ultimate outcome would be a smaller measurement error of the unknown wheel-rail interaction parameters.

It should be noted that the previous analysis was performed on raw signals, without any filtering and statistical processing. In practical applications where filtering and statistical processing would be implemented, the measurement errors are expected to be even smaller than the reported values.

5. CONCLUSION

This paper analyzed the influence of the signal-to-noise ratio (SNR) on the accuracy of determining wheel-rail interaction parameters – vertical force Q, lateral force Y, ratio Y/Q, and contact point position y_{tk} – using an instrumented wheelset based on strain gauges. Static measurements confirmed the linear dependence between the applied forces and the output signals from the measuring bridges, but dispersion was observed due to noise in the electronic components, particularly pronounced at low SNR values (characteristic of empty vehicles). Dynamic measurements for two load levels (empty wagon and loaded wagon) showed that higher forces produce higher output signals, a more favorable SNR, and significantly lower relative measurement errors. Statistical data processing estimated that the given instrumented wheelset enables measurement of the parameters Q, Y and Y/Q with relative errors of less than 10%, and the contact point position y_{tk} with an error of less than 15%. During flange contact (high Y/Q or Y values), the errors are even smaller, which is extremely important for assessing derailment risk according to international standards.

Measurements at zero load confirmed that the dominant noise source originates from the measuring bridges, while the force transducers have a negligible influence. Due to the complexity of the measurement setup and the large number of influencing parameters, absolutely accurate determination of the error is not possible, but the statistical estimate provided in this paper represents a reliable basis. In addition to SNR, measurement accuracy is also influenced by a number of other parameters – the precision of strain gauge application, deviations of their nominal resistance, material inhomogeneity of the wheel, geometric imperfections, etc. Also, the stiffness of the wheel is extremely high, especially in the vertical direction, which results in relatively small output signals from the measuring bridges, even when the aforementioned problems are solved in an optimal manner. Since the SNR cannot be arbitrarily increased by using classical strain gauges, one of the directions for future research should be directed toward the application of fiber-optic sensors, which offer greater resistance to electromagnetic interference and a significantly more favorable SNR, thus enabling a new generation of instrumented wheelsets with even higher measurement accuracy.

ACKNOWLEDGEMENTS

The authors are grateful to the Ministry of Science, Technological Development and Innovation of the Republic of Serbia for support (contract no. 451–03–34/2026–03/200108).

REFERENCES

[1] S.D. Iwnicki, Handbook of Railway Vehicle Dynamics, CRC Press, Taylor & Francis Group, Boca Raton, (2006)

[2] A.H. Wickens, Fundamentals of Rail Vehicle Dynamics – Guidance and Stability, Swets & Zeitlinger Publishers, Lisse, (2003)

[3] X. Jin, Evaluation and analysis approach of wheel-rail contact force measurements through a high-speed instrumented wheelset and related considerations, Vehicle System Dynamics, Volume 58, Issue 8, pp. 1189-1211, https://doi.org/10.1080/00423114.2019.1612073, (2020)

[4] M. Bizic, Research of influential parameters in wheel-rail interaction on running stability of railway vehicles, Doctoral dissertation, University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Serbia, (2015), (in Serbian)

[5] V.R. Bagheri, P.H. Tehrani, D. Younesian, Optimal strain gauge placement in instrumented wheelset for measuring wheel-rail contact forces, International Journal of Precision Engineering and Manufacturing, Volume 18, pp. 1519–1527, https://doi.org/10.1007/s12541-017-0180-7, (2017)

[6] M. Bižić, D. Petrović, Determination of optimal locations of strain gauges on instrumented wheelsets, The XVIII International Scientific-Expert Conference on Railways – RAILCON 2018, 11-12 October 2018, Niš, Serbia, pp. 137-140, (2018)

[7] M. Bižić, D. Petrović, Determination of optimal layout, number and way of connection of strain gauges on instrumented railway wheelsets, The X International Conference HEAVY MACHINERY – HM 2021, 23-25 June 2021, Vrnjačka Banja, Serbia, pp. E17-E22, (2021)

[8] O. Ozbek, A new approach for calibration of an instrumented wheelset, Thesis text, Politecnico di Milano, (2013)

[9] M.B. Bižić, D.Z. Petrović, M.C. Tomić, Z.V. Djinović, Development of method for experimental determination of wheel–rail contact forces and contact point position by using instrumented wheelset, Measurement Science and Technology, 28 (7), 25 p., https://doi.org/10.1088/1361-6501/aa666f, (2017)

[10] D. Li, L. Huang, H. Dai, J. Wang, X. Ma, A parametric approach for enhancing the accuracy of instrumented wheelset: A case study, Measurement, Volume 242, Part D, 116133, https://doi.org/10.1016/j.measurement.2024.116133, (2025)

[11] E. Gomez, J.G. Giménez, A. Alonso, Method for the reduction of measurement errors associated to the wheel rotation in railway dynamometric wheelsets, Mechanical Systems and Signal Processing, Volume 25, Issue 8, p. 3062-3077, https://doi.org/10.1016/j.ymssp.2011.05.006, (2011)

[12] S. Bionda, E. Di Gialleonardo, G. Cazzulani, F. Braghin, A new method for the improvement of the accuracy of the estimation of wheel-rail contact forces using instrumented wheelsets, Dynamics of Vehicles on Roads and Tracks, Vol. 2, CRC Press, Taylor & Francis Group, London, https://doi.org/10.1201/9781351057189, (2017)

SESSION B: RAILWAY ENGINEERING

Systems for precision control of microclimate in world rail transport: Technological structure and intelligent management

Valentina Zochinova[1], Venelin Pavlov[1*]

[1] Todor Kableshkov University of Transport, Sofia, Bulgaria

ARTICLE INFO

[*] **Correspondence:** vpavlov@vtu.bg

DOI: https://doi.org/10.46793/ET26.B08Z **PAGES**: B59 – B63

ABSTRACT

The issue of ensuring a suitable microclimate in railway carriages and analyses of the key parameters determining the quality of the environment in the carriages – temperature, relative humidity, air flow rate, carbon dioxide (CO2) levels are considered in the publication. The detailed hierarchy of systems is described, including a real-time data sensor network, centralized logic control controllers and actuators (frequency-controlled compressors, heating modules and recirculation valves). Special attention is paid to the transition to intelligent management technologies. Mechanisms for automated optimization of energy consumption through adaptive control algorithms and remote monitoring have been considered. On the importance of preventive diagnostics and the implementation of environmentally friendly refrigerants as key trends for achieving sustainable and energy- efficient railway transport, and recommendations are made regarding the implementation of these systems in practice and their importance for safety and energy efficiency in transport conclusions has been formulated.

KEYWORDS

HVAC systems, Automated control, Energy efficiency, Microclimate, Railway safety, Railway transport, Sensor networks.

1. INTRODUCTION

In modern railway transport, ensuring an optimal microclimate is a critical factor for passenger comfort, equipment safety and energy efficiency. This article examines the technological structure of the new generation of HVAC (heating, ventilation and air conditioning) systems integrated into rolling stock around the world. The subject of analysis is the transition to intelligent control through artificial intelligence (AI), the use of environmentally friendly refrigerants and strategies for optimizing energy consumption.

Global urbanization and the drive to decarbonize transport are making railways the preferred mode of mass transportation. However, high-speed trains and urban metros face the challenge of maintaining constant air parameters under extreme external conditions and fluctuating passenger traffic. Traditional air conditioning systems no longer meet modern sustainability standards, which necessitates the implementation of precise and intelligent control technologies. The development of modern climate control systems is essential to enhance passenger comfort and energy efficiency in rail transport. The integration of advanced predictive analytics and artificial intelligence into climate control systems can further improve both passenger comfort and energy efficiency. For example, the use of real-time weather data and occupancy patterns allows for dynamic adjustments in heating, ventilation and air conditioning (HVAC) systems, ensuring optimal conditions regardless of external fluctuations. Furthermore, the implementation of smart grid technology can facilitate better energy management, reducing operating costs while maintaining high levels of comfort in train compartments. As the demand for sustainable transport solutions grows, these innovations not only promise to enhance the travel experience, but also align with wider environmental goals, highlighting the importance of smart systems in the future of rail transport. Importance of microclimate in rail transport.

As the rail industry continues to evolve, the importance of climate management extends beyond passenger comfort to maintaining safety and operational efficiency. For example, fluctuations in temperature and humidity can lead to skewed oscillations or equipment failures, posing a risk to both infrastructure and passengers. This integration of predictive analytics with real-time monitoring systems can not only improve the passenger experience, but also proactively address

potential hazards, ensuring the reliability of rail services. Furthermore, the adoption of advanced climate control technologies, such as those that use artificial neural networks to optimally manage HVAC, can be significant in mitigating energy consumption while maintaining required comfort levels. This holistic approach highlights the need for a comprehensive strategy that prioritizes both environmental sustainability and operational resilience in relation to climate variability.

2. TECHNOLOGICAL STRUCTURE OF MODERN HVAC SYSTEMS

The rail industry prioritizes passenger comfort, and it is crucial to consider the psychological effects of microclimate conditions on passengers. Studies have shown that factors such as temperature, humidity and air quality can significantly enhance passengers' mood and satisfaction, which in turn influences their overall experience and the likelihood of them choosing rail over other modes of transport. By implementing advanced air quality monitoring systems that assess not only temperature but also particles and allergens, rail operators can create a more pleasant environment that affects relaxation and well-being. Furthermore, integrating user feedback mechanisms could allow for real-time adjustments based on passenger preferences, further improving the travel experience while ensuring that energy consumption remains efficient and sustainable. This multifaceted approach not only addresses immediate comfort issues but also long-term citizen loyalty among passengers, ultimately resulting in the growth of the rail sector amidst increasing competition.

Conventional HVAC systems often treat temperature control as the primary objective, but thermal comfort is a more complex, subjective concept influenced by multiple factors. Temperature, air velocity, humidity and infrared radiation are key parameters that determine the heat exchange between a person and the environment. The study proposes a logic controller (FLC) that uses the Predicted Mean Voice (PMV) index to control indoor temperature. The PMV index aims to achieve a neutral thermal sensation (PMV = 0) with a low percentage of dissatisfied persons (PPD).

FLC is well suited for thermal comfort control as it can deal with the ambiguity and subjectivity inherent in thermal sensation by treating comfort ranges. The FLC adjusts the air temperature setpoint based on the estimated PMV level and the required comfort range, taking into account various contributing comfort variables.

PMV Index and Control Actions in a Logic Controller in Railway Transport

The Predicted Mean Voice (PMV) index plays a central role in determining the control actions of a Logic Controller (FLC) for thermal comfort management. The FLC uses the PMV to estimate the thermal sensation in the rooms and subsequently adjust the heating, ventilation and air conditioning (HVAC) system to maintain comfort and optimize energy consumption.

PMV Assessment and Comparison

FLC continuously assesses the PMV level, which is an index of thermal sensation that predicts the average thermal sensation rating on a standard scale for a large group of people. This assessment is based on environmental variables such as room air temperature, relative humidity, average radiant temperature and air velocity, together with human factors such as activity level and thermal resistance of clothing.

The calculated PMV level is then compared to the required comfort range. The goal is to achieve a neutral thermal sensation, typically when the PMV is equal to zero, with a low percentage of dissatisfied persons (PPD). For example, a PPD of 5% is assessed when the PMV is zero and the target indoor temperature is set relative to this point.

Determination of control actions

Based on the comparison between the estimated PMV and the desired comfort range, FLC determines the required control. If the PMV level falls outside the comfort range, the controller provides the air temperature setpoint that the HVAC system should use to create indoor thermal comfort .

The FLC processes clear inputs, such as the PMV value, through a fuzzy interface, converting them into linguistic variables (e.g., "COLD", "COOL", "NEUT", "WARM", "HOT") using membership functions .

Fuzzy control rules, which are conditional statements, then define the relationship between these linguistic input variables and the corresponding output control actions. These rules are often derived from experiential knowledge or models of plant behavior . For example, a rule might state: "If PMV is "COLD", then Vp is "FO" (where Vp is the output variable and FO is a fuzzy output set) .

The inference engine applies these rules to derive a fuzzy output, which is then converted into a clear control signal (e.g. for a heating valve) through a process of fuzzification, typically using the center of gravity (COG) method. This clear signal dictates the actual setting of the HVAC system.

Advantages of PMV-based FLC

The use of PMV within FLC allows for a more nuanced and robust control strategy compared to traditional temperature PID controllers. It addresses the subjective and vague nature of thermal comfort by treating comfort ranges as fuzzy variables.

This approach allows the controller to maintain excellent tracking of the comfort reference state and achieve significant energy savings (e.g. 20% reduction in energy consumption over a typical 7-day winter period). FLC also demonstrates greater robustness and general applicability across different building types, without requiring specific re-tuning for each application, unlike PID controllers. In essence, the PMV index acts as a basic feedback for the FLC, guiding the system to make intelligent, comfort-driven adjustments to the indoor environment, thereby increasing occupant satisfaction and energy efficiency. Improving microclimate systems in rail transport is key to ensuring safety, comfort and energy efficiency.

The architecture of microclimate systems in the rail sector consists of three main levels:

- Actuators: Inverter-controlled compressors, multi-speed fans and electronic thermostatic valves (RTVs) that allow smooth power regulation.
- Sensor network: Sensors for measuring temperature, relative humidity, carbon dioxide (CO2) levels and pressure sensors (critical when passing through high-speed tunnels).
- Control module: Local controllers (PLCs) integrated into the overall train control system (TCMS – Train Control and Management System).

3. REGULATORY FRAMEWORK AND INTERNATIONAL STANDARDS

The design, testing and operation of HVAC systems in global rail transport are subject to strict international standards that guarantee safety and thermodynamic comfort:

- EN 14750:2024 [1]: Basic European standard defining the parameters of thermal comfort (temperature, humidity, air flow velocity) and the methods for their testing in urban and regional rolling stock.
- EN 13129 [2]: Standard for mainline and high-speed trains regulating the air comfort in passenger compartments under extreme external conditions.
- UIC 553 / UIC 553-1 [3]: Code of the International Union of Railways defining ventilation rates and ventilation programs for passenger carriages.
- ISO 19659 [4]: A series of international standards for railway applications unifying the terminology and requirements for energy efficiency of HVAC installations.
- Regulation (EU) 517/2014 [5]: Regulatory act stimulating the phase-out of fluorinated greenhouse gases (HFCs) and the transition to natural agents (R744).

4. INTELLIGENT CONTROL AND ALGORITHMS

The key to precise microclimate is software logic. Static settings are replaced by adaptive algorithms:

- Occupancy-based Control: Using data from CCTV cameras or weight sensors on the carriages, the system calculates the exact number of passengers. Ventilation dynamically adjusts the flow of fresh air based on CO2 levels, preventing air stagnation and saving energy when the carriage is half-empty.
- Predictive Control: Intelligent systems use GPS data and weather forecasts. If the train is approaching a tunnel or an area with extreme temperatures, the HVAC system adjusts parameters in advance to compensate for sudden changes in pressure and heat load.
- Machine Learning and Diagnostics (CBM): By collecting data in cloud structures, predictive maintenance algorithms are applied, which reduces breakdowns along the way.

5. REFRIGERANT BENCHMARKING / ENERGY EFFICIENCY BENCHMARKING

An important structural component in modern transport is the transition to environmentally friendly fluids under the pressure of regulatory frameworks [5]. Instead of traditional freons, world leaders in the industry are implementing natural refrigerants such as carbon dioxide (R744), whose global warming potential (GWP) is equal to 1 [6]. Table 1 provides a comparative technical and environmental analysis of the most widely used fluids.

Comparison 1: Global Warming Factor (GWP) of Refrigerants

The transition to R744 (CO_2) eliminates the risk of greenhouse gas emissions from possible leaks from air conditioning installations during their life cycle.

Table 1: Comparison of technical and environmental parameters of refrigerants.

Refrigerant	Fluid type	GWP potential	Working pressure	Ecological status
R407C	Synthetic (HFC)	1774	Low / Medium	In process of replacement
R134a	Synthetic (HFC)	1430	Low	Limited by regulations
R513A	Synthetic blend	631	Average	Transitional solution
R744	Natural fluid	1	Natural fluid	Sustainable / Future standard [6]

Comparison 2: Annual energy consumption by control type

Traditional systems operate on the "On/Off" principle or maintain constant flow rates. The implementation of inverter precise control and management based on occupancy (number of passengers and CO2 levels) dramatically optimizes consumption.

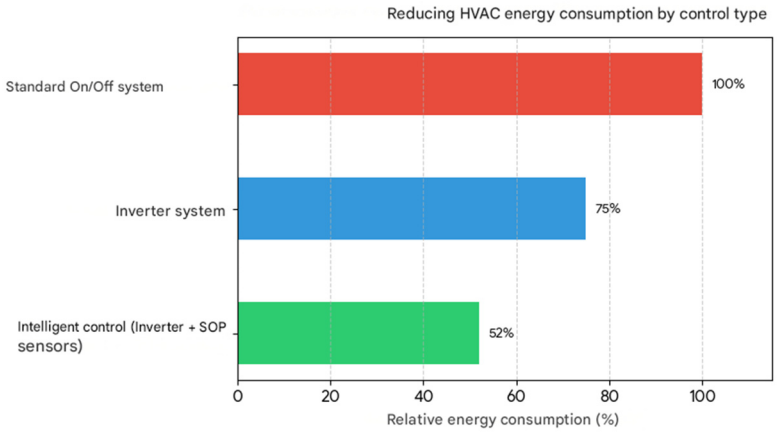


Figure 1:

Comparison 3: Coefficient of Performance (COP) in heat pump mode

At low outdoor temperatures, heat pump systems with R744 show significant superiority in heating mode over conventional combustion or electric heaters, saving up to 78% energy in winter conditions.

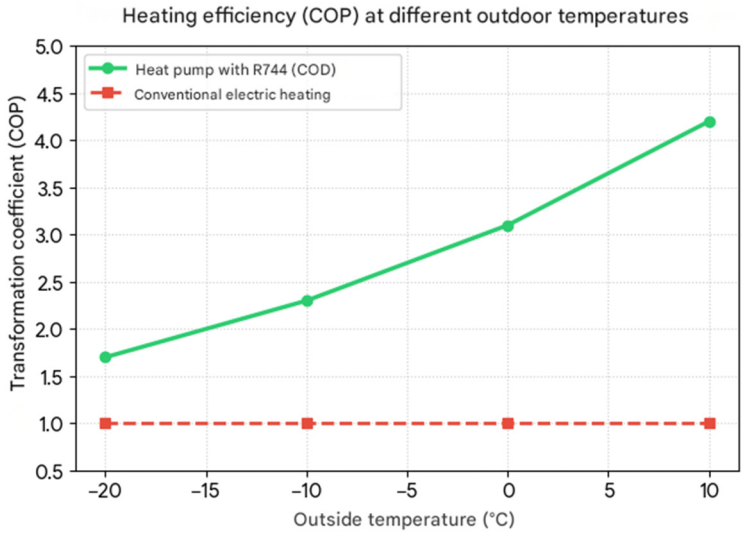


Figure 2:

6. ENERGY EFFICIENCY AND HEAT PUMP TECHNOLOGIES

HVAC systems are among the largest consumers of electricity in a train (up to 30% of the total auxiliary power). Modern configurations use reversible heat pumps with R744, which provide both cooling in summer and efficient heating in winter. They show high Coefficient of Performance (COP) at negative temperatures and utilize waste heat from traction inverters, saving up to 30% of the total energy compared to conventional on/off systems [6].

7. CONCLUSION

The transformation of rail climate systems is moving from purely mechanical solutions to software-defined, environmentally friendly and intelligent ecosystems. The implementation of precise control not only ensures first-class passenger comfort, but also drastically reduces the carbon footprint and operating costs of the rail sector globally.

REFERENCES

[1] European Committee for Standardization (CEN). (2024). EN 14750:2024 - Railway applications - Air conditioning for urban and suburban rolling stock. Brussels: CEN.

[2] European Committee for Standardization (CEN). (2009). EN 13129:2009 - Railway applications - Air conditioning for main line rolling stock - Comfort parameters and type tests. Brussels: CEN.

[3] International Union of Railways (UIC). (2005). UIC Leaflet 553: Ventilation, heating and air conditioning in passenger carriages - Standard tests. Paris: UIC.

[4] International Organization for Standardization (ISO). (2021). ISO 19659-1:2021 - Railway applications — Heating, ventilation and air conditioning systems for rolling stock — Part 1: Terms and definitions. Geneva: ISO.

[5] European Parliament and Council. (2014). Regulation (EU) No 517/2014 on fluorinated greenhouse gases and repealing Regulation (EC) No 842/2006 (F-Gas Regulation). Official Journal of the European Union.

[6] Liebherr-Transportation Systems & ÖBB. (2020). Project eco2jet: Evaluation of eco-friendly air conditioning systems with CO2 (R744) refrigerant for passenger rail vehicles. Vienna: Joint Industry Report.

SESSION B: RAILWAY ENGINEERING

Methods of increasing the wear resistance with repairs of BDZ freight wagon bogies

Venelin Pavlov[1*]

[1] Todor Kableshkov University of Transport, Sofia, Bulgaria

ARTICLE INFO

[*] **Correspondence:** vpavlov@vtu.bg

DOI: https://doi.org/10.46793/ET26.B09P **PAGES**: B65 – B68

ABSTRACT

The increase in the time between various repairs of freight wagons determines the urgency of the problem of improving the wear resistance and service life for the restoration of supporting nodes and parts of the bogie for freight wagons due to excessive wear of the friction working surfaces. Most of the freight wagon bogie parts can be restored by subsequent processing. According to the existing technology and methods, the quality of freight wagon bogie repair does not provide the bogie mileage of specified kilometres necessary between repairs or in inter-repair periods. In this regard, to study the impact of surface restoration quality parameters on the wear process of parts in operation and after the repair is a priority task. The problem of providing methodology for various repairs of freight wagon bogies aims at increasing the wear resistance of bogie parts, which is a complex task and includes the use of modern methods to predict wear and technological warranty of the resource of repair parts.

KEYWORDS

Bogies, Freight wagons, Wear resistance, Repair, Hardfacing, Heat treatment, Railway transport.

1. INTRODUCTION

Railway transport represents a strategic sector for the Bulgarian economy. The reliability of freight wagons is directly linked to the technical condition of their running gear, particularly the bogies. The bogie is a primary structural component that absorbs dynamic loads during movement, ensuring the stability and safety of the wagon.

During operation, the bogie components are subjected to significant mechanical, vibrational, and contact loads, which lead to wear, material fatigue, and the initiation of cracks. Side frames, axle box openings, the suspension system, and wheelsets are particularly critical. Research indicates that defects exceeding 1 mm in size can sharply reduce the fatigue strength of bogie frames [1].

2. MAIN CAUSES OF BOGIE WEAR

The wear of freight wagon bogies is caused by several primary factors:

- High dynamic loads;
- Friction between contact surfaces;
- Insufficient or improper lubrication;
- Corrosion processes;
- Material fatigue;
- Low quality of welded joints;
- Operation on railway tracks in poor technical condition.

Cyclic loads are particularly hazardous, as they induce the propagation of fatigue cracks in load-bearing structures. Structural analyses of freight bogie frames demonstrate that stress concentration in the areas surrounding weldments is a major factor in structural failure [1].

The most widely used bogie type for freight wagons in BDZ (Bulgarian State Railways) is the Y25 type, distributed as the BT 6 two-axle bogie. The BT 6 is a Bulgarian analogue and is technically equivalent to the European Y25, specifically the Y25cs modification.

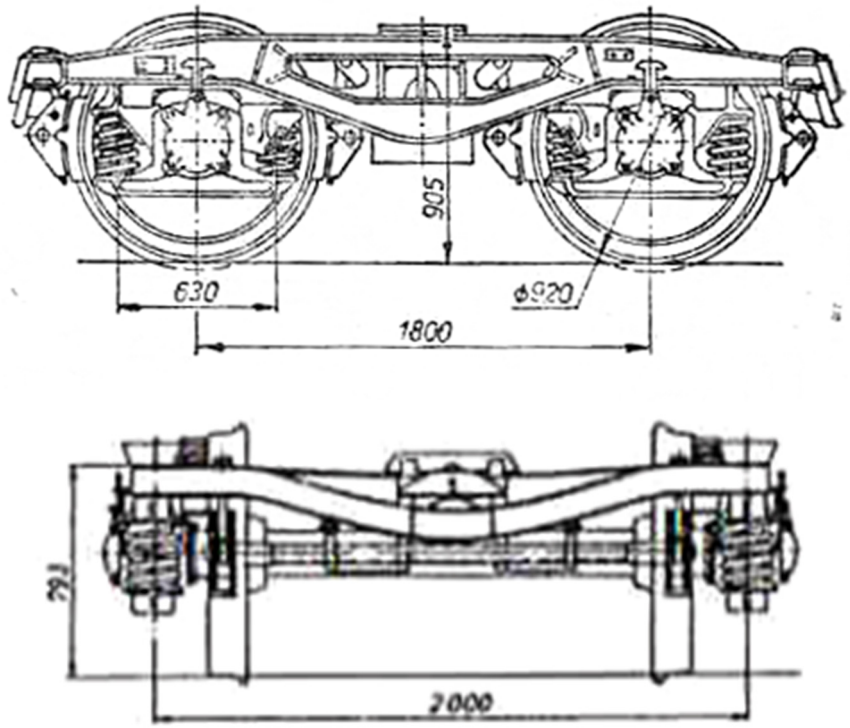


Figure 1: General view of the "Y 25Cs" (BT–6) bogie

During operation, bogies are subjected to intensive static and dynamic loads, resulting in frictional wear in the contact zones. Traditional repair methods involving standard weld hardfacing often lead to stress concentrators and the rapid recurrence of defects [2].

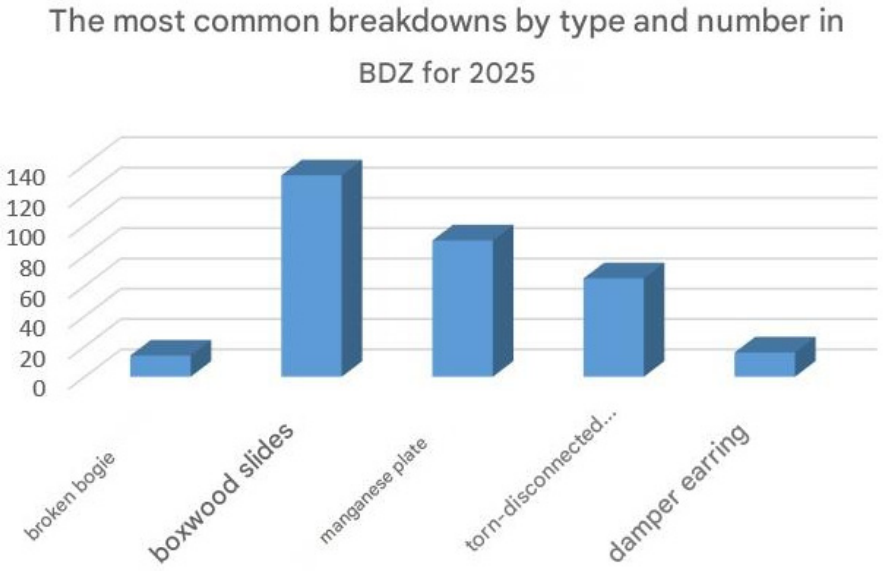


Figure 1: Most common defects in freight wagon bogies

3. METHODS FOR INCREASING WEAR RESISTANCE

3.1. Heat Treatment

Heat treatment is utilized to enhance the mechanical properties of steel components. Quenching and tempering are applied to increase:

- Hardness;
- Fatigue resistance;

- Wear resistance;
- Surface layer strength.

In bogie repair, localized induction hardening is frequently applied to highly loaded surfaces, such as axlebox openings and support elements.

3.2. Hardfacing and Component Restoration

Weld hardfacing is one of the most widespread methods for repairing worn components. Electric arc or plasma hardfacing is employed to restore the dimensions of worn surfaces.

The advantages of this method include:

- Extension of operational lifespan;
- Reduction of expenditures on new components;
- Increase in surface layer hardness.

Modern technologies utilize alloyed electrodes with high chromium, nickel, and molybdenum content, which provide superior resistance to abrasive wear.

3.3. Application of Wear-Resistant Coatings

The application of protective coatings significantly reduces the wear rate. The following are applied:

- Chromium coatings;
- Carbide layers;
- Cermet coatings;
- Polymer composite materials.

These technologies enhance resistance to friction and corrosion, particularly when operating in aggressive environments.

3.4. Non-Destructive Testing

Modern diagnostics play a key role in accident prevention. Non-destructive testing (NDT) methods include:

- Ultrasonic testing;
- Magnetic particle testing;
- Eddy current testing;
- Liquid penetrant testing.

These methods enable the detection of microcracks and internal defects at an early stage. NDT methods represent a fundamental approach to extending wagon service life and enhancing operational safety [3].

3.5. Robotic Welding

In recent years, robotic welding systems have been implemented in railway mechanical engineering. They provide:

- Higher weld seam quality;
- Minimization of human error;
- Reduced residual stresses;
- Higher reproducibility of the repair process.

Studies on robotic welding cells for bogie frames indicate an improvement in operational reliability and a reduction in defect probability [4].

4. MODERN ENGINEERING APPROACHES

4.1. High-Strength Steels

The utilization of low-alloy high-strength steels allows for:

- Reduction of the bogie tare weight;
- Increase in load-carrying capacity;
- Improvement of fatigue resistance.

4.2. Computer Modeling

The Finite Element Method (FEM) optimizes the repair of freight bogies by simulating stresses and wear within friction zones (side sliders, center pivot, axleboxes). The implementation of FEM in the repair cycle is based on fundamental mathematical contact stiffness matrices (5) and proceeds through the following six sequential stages:

- Scanning and CAD Defect Detection: 3D laser scanning of the actual geometry of the worn frame jaws to create a volumetric model in compliance with the nominal dimensions specified in the UIC 510-1 regulations [6].

- Discretization (Mesh Generation): Utilizing fine volumetric SOLID elements (hexahedrons), which provide greater accuracy per unit of computational time compared to tetrahedrons—especially when simulating bending and linear stresses in contact friction zones—ensuring high computational accuracy and smooth mesh transitions toward non-critical zones [5].
- Definition of Boundary Conditions and Loads: Applying forces in accordance with BDS EN 13749 [7], which defines the methodology to be followed to achieve a satisfactory design for bogie frames and includes design procedures, evaluation methods, verification, and manufacturing quality requirements. It is limited to the structural requirements of bogie frames, including pads and axlebox housings. For the purposes of this document, these terms encompass all functional applications, such as damper brackets.
- Contact FEM Analysis: Defining frictional contact with a realistic friction coefficient (\(\mu = 0.15 - 0.4\)). Computing Von Mises equivalent stresses and contact pressure distribution.
- Simulation of Repair Models: Conducting comparative analysis within the FEM environment between standard hardfacing, installation of manganese liners, and the use of non-metallic composites based on Archard's mathematical wear criterion [8].
- Optimization: Selection of a technology that ensures maximum mitigation of contact stress peaks.

5. CONCLUSION

Enhancing wear resistance during the repair of freight wagon bogies is a critical factor for the reliable operation of railway transport in Bulgaria. The best results are achieved through the combined application of heat treatment, hardfacing, wear-resistant coatings, robotic welding, and modern diagnostic methods.

The integration of FEM simulations into railway repair practices enables a scientifically substantiated transition from classical welding to the utilization of modern materials and methodologies.

The implementation of modern technologies in the maintenance and repair operations of the Bulgarian State Railways will result in an extended operational lifespan of the wagons, enhanced traffic safety, and reduced financial maintenance expenditures.

REFERENCES

[1] Lagerev, I.A. Fatigue strength analysis of a freight car bogie's side frame. DOAJ, 2017.

[2] Nachev, G., & Dimitrov, B. (2018). Investigation of the stress state of Y25 type freight bogie frames using FEM. Scientific Proceedings of TU-Sofia. (Originally in Bulgarian)

[3] Sapronova, S. et al. Analysis of methods for the extension of wagon service life. (tst.duit.in.ua).

[4] Edrev, St., Topalova, M. Research work about the trouble-free operation of a robotic cell for the welding of the bogie frame on a freight wagon. (Stume Journals).

[5] Zienkiewicz, O. C., Taylor, R. L., & Zhu, J. Z. (2005). The Finite Element Method: Its Basis and Fundamentals. Elsevier Butterworth-Heinemann.

[6] UIC 510-1: Traînée et roulement - Bogies - Éléments constitutifs (Railway bogies - Structural elements and technical requirements).

[7] BDS EN 13749: Railway applications - Wheelsets and bogies - Method of specifying the structural requirements of bogie frames.

[8] Archard, J. F. (1953). Contact and Rubbing of Flat Surfaces. Journal of Applied Physics, 24(8), 981-988.

[9] Nikolov, D., Nenov, N., Yosifova, D.: RFID Electronic Sensor System for Rolling Stock Recognition in Motion, 41st International Spring Seminar on Electronics Technology (ISSE) 2018, pp. 1-5, Zlatibor, Serbia, 2018, https://doi.org/10.1109/ISSE.2018.8443653.

[10] Nikolov V.: Modeling and strength-deformation analysis of a defected locomotive axle. ISSN 1312-3823 (print) ISSN 2367-6620 (online) Vol. 20(3/3), 2022. Scientific Journal mtc-aj.com article No. 2240, mtc-aj.com.

SESSION B: RAILWAY ENGINEERING

Impact of automated systems on reducing the human factor in rail transport

Irena Petrova[1*]

[1] Todor Kableshkov University of Transport, Faculty of Transport Management, Department of Technology and Organization of Transport, Sofia, Republic of Bulgaria

ARTICLE INFO

*** Correspondence:** ipetrova@vtu.bg

DOI: https://doi.org/10.46793/ET26.B10P **PAGES:** B69 – B74

ABSTRACT

In the last 5 years, the role of rail transport has grown a lot within the European Union. It is the most environmentally friendly and safest transport, transports a huge amount of cargo and passengers over short and long distances, and employs thousands of people. There is a shortage of trained personnel in all European countries. The human factor is still leading in all activities in railway transport. This leads to wrong ones that are the result of accumulated fatigue, distraction, delayed reactions, incorrect reading of signals, etc. Digitalization in rail transport aims to increase its safety and reliability. This article reviews the impact of automated systems to reduce human error. The advantages and disadvantages of modern technologies used for this purpose are analyzed. Different types of process automations are presented - semi-automatic and automatic. The challenges accompanying the implementation process are considered, such as high cost, infrastructure modernization and cybersecurity. Automated systems are not aimed at completely replacing people, but at supporting and controlling processes. The future of rail transport is impossible without the deployment of intelligent and automated systems.

KEYWORDS

Railway transport, Human factor, Automation, ERTMS, ETCS, ATO, Safety, Cybersecurity.

1. INTRODUCTION

In the context of increasing mobility and the need for safe, sustainable and environmentally friendly transport, the role of railways in passenger and freight transport is increasing. High capacity, energy efficiency and minimal environmental impact define rail transport as key. For the effective functioning and public trust of rail transport, it is important to ensure its reliability and safety. Digitalization in rail transport is a process in which modern technologies support, replace and fully realize the connection between people, systems and rolling stock. It includes intelligent signaling, continuous communication, automated infrastructure monitoring, maintenance and real-time data integration between all dependent actors – trains, infrastructure and control centers [1]. (Commission E. , 2026)

Through the digitalization of railway transport, the issue of the shortage of physical infrastructure can be solved by increasing the capacity of the existing ones. Some operational types of human error are significantly reduced, while at the same time emerging new risks such as cyber threats, problems with monitoring and verification of systems. Automation affects the frequency of operational human errors occurring in routine tasks. Automated systems are very reliable in performing repetitive operations and in reducing risk in high-risk activities. On the other hand, digitalization is changing the type of human errors that result from the use of automated systems [2]. (Vanderhaegen, 2021, 189–192)

2. HUMAN FACTOR IN RAIL TRANSPORT

The human factor in railway transport includes knowledge, experience, attention, communication skills, decisions under pressure. The operation of the rail system depends not only on technique and procedures, but also on the way in which employees apply them in practice [3]. (EUAF, 2016). The human factor is one of the most significant elements influencing railway safety. Most accidents occur not due to a single technical error, but as a result of errors in monitoring, communication, decision, or reaction. Rail traffic is high-speed, highly coordinated, and dependent on strict adherence to rules and

signals, and any mistake can have serious consequences [4]. (Human-Centered Systems Approach in a Social-Technical System, 2019)

Human errors in rail transport are divided into three main groups – operational, cognitive and organizational. The first group includes incorrect reading of signals, inappropriate estimation of distance or speed and omission of important information from the environment. The second group includes errors that are the result of an incorrect choice of action in a situation of risk, and the third - a delayed or inaccurate operational reaction [5]. (Melissa T. Baysari, 2008)

Table 1: Types of human errors in railway transport

Types of errors	Manifestation
Operational	Incorrect submission or interpretation of signals. Errors in arrow and route settings. Increasing the permissible speed. Failure to comply with the permissible speed. Failure to follow instructions and procedures.
Cognitive	Incorrect assessment of the situation. Delayed reaction. Loss of concentration. Information overload.
Organizational	Insufficient training. Lack of clear procedures. Ineffective communication. Insufficient safety control. Insufficient risk management.

In the report, human error is seen not only as a personal omission, but as a result of the organizational environment, workload, communication, and process design. It is recommended that when analyzing incidents, look not only for "who made a mistake", but also "why it was possible to make a mistake". The aim is to build more reliable working practices and reduce risks. Clear instructions, good training, fatigue, working hours, coordination between teams and the quality of information directly affect the safety and quality of services [3].(EUAF, 2016)

Causes of Human Errors

The main causes of human errors are fatigue and monotony, stress and lack of timely information. A study was conducted on 1391 high-speed train drivers in China. Its conclusions show that fatigue and stress increase human error by worsening situational awareness. The error is not an isolated personal omission. It is the result of the interaction between man, technology, environment and organization [6].(Wenil Dong, 2024)

Human errors can be caused by faulty equipment and technology, by designers, manufacturers, operators or maintenance persons, excessive speed or load. Human errors related to skills are observed at all times, as part of a series of events that caused an accident. Such events may involve negligent inspection and maintenance and are part of organizational and system errors [7].(Zhiru Wang, 2016)

Fatigue affects the working capacity of all people; it is normal to affect railway personnel as well. Shift work, night work and the need to concentrate lead to reduced alertness, slower reactions, difficulty making decisions, increased risk of mistakes. This identifies fatigue as a factor that increases the likelihood of operational disruptions and incidents in transport systems. Fatigue affects the perception of the situation, signals, the environment, and affects problem-solving abilities [8]. (Yaser Haj Yousef, 2014)

Insufficient or incorrectly submitted information between drivers, dispatchers, signaling and technical personnel can be the cause of serious traffic management violations. Research shows that poor communication is among the most common factors in the occurrence of accidents in complex transport systems. According to the analysis of railway accidents, communication gaps are often accompanied by organizational and technical weaknesses. The lack of data on the spatial and temporal characteristics of the operational situation creates a risk, as the operator is forced to act on the basis of intuition or incomplete knowledge. It is important to note that with all the training, serious experience and good education of the operator, there may be a different way of performing a procedure [9]. (RAIB, 2015). Mistakes resulting from an incorrect decision are the result of a bad choice or wrong procedure. These errors are the result of a deliberate act or omission on the part of the operators, where the plan proves to be inappropriate for the situation [10]. (Ndaru)

Impact on Safety

Human errors have a serious impact on railway safety and are the cause of collisions, derailments, accidents at level crossings, traffic schedule problems. As a result of them, significant material damage, victims and injured occur. The main directions for ensuring the safety of the transport process include the reliability of technical means, the human factor and the legislative framework. The human factor is the most vulnerable link in the safety system, as they work in an area of uncertainty and subjective decision-making [11]. (Зоя Хубенова, 2016)

3. AUTOMATED SYSTEMS IN RAIL TRANSPORT

Automation and computers increase reliability and safety by performing some of the control and computational functions in complex technical systems. In rail transport, they support monitoring, signaling and failure, speed control, which reduces dependence on momentary human reactions. Modern automated systems do not eliminate risk, but transform it, as they create new dependencies on software, interface and data quality, which require constant human supervision and a high level of cybersecurity. Automated systems in railway transport are a complex of technical solutions that support or partially replace humans in management, control, signaling, diagnostics and support. Their use reduces subjective judgment in critical situations and ensures standardized and safe train movement [12]. (Dekker, 2001)

The European Rail Traffic Management System (ERTMS) was developed to create a single railway area and bridge the gaps between more than 30 national and incompatible systems. It can be implemented as a supplement to the existing system or as a main one when building new infrastructure. Table 2 summarizes the advantages of ERTMS [13].(ЕСП, 2017)

Table 2: Advantages of ERTMS

Advantage of ERTMS	Influence
Increased safety	Reduces the risk of collisions and speeding.
Reducing human error	Standardized procedures and automatic intervention reduce the possibility of human decisions.
Compatibility	Uniform protocols allow seamless passage between different operators and networks.
Capacity Increase	Tight intervals and automated operations.
Improved service	Improved traffic and quality management.
Safe crossing	Higher safety at level crossings.

ERTMS allows for automatic monitoring and control of speed, signaling and traffic safety, limiting the risk of human error. It includes three subsystems. ETCS for speed control and traffic protection and GSM-R for safe radio communication between the dispatcher, infrastructure and train. The third ATQ system helps trains run automatically. In this way, the need for manual protective actions and subjective decisions of staff through automatic monitoring and intervention is reduced. Automatic intervention ETCS can stop or penalize the train if the observed parameters are exceeded and thus limits the possibility of human error.

The European Train Control System (ETCS) provides cab signalling and automatic train protection. The system receives information about the movement of the train from balises, radio block centers or other communication channels. If the driver does not react when speeding or approaching danger, the system automatically triggers a braking.

Level 0 and level NTS (National Train Control) are related to the compatibility status of the borders. Level 0 describes the mode of operation in which the locomotive is equipped with ETCS but runs on an unequipped railway. At Level 0, there is no exchange of information between the train and the infrastructure via ETCS and traffic control is carried out through the national signalling system. On-board equipment monitors only the maximum permissible speed, and the driver relies on roadside signals.

In ETCS Level 1, the information is transmitted automatically via balises located along the route. Eurobalise electronic devices mounted on the tracks, which send information to the passing train about the permitted speed, the status of the signals, specific characteristics of the section, the traffic permit. The transmission of information is known as stop transmission or point data transmission. Switching to Level 1 leads to an increase in traffic safety, controls speed, allows the use of existing national signaling, prevents crossing at a prohibitory signal, lower cost than ETCS Level 2.

In ETCS Level 2, there is continuous communication using a radio connection, which allows dynamic motion control. There is a continuous exchange of data between the train and the infrastructure via radio communication. Eurobalises continue to be used, but to determine the exact location of the train on the track. ETCS Level 2 continuously monitors the speed, location, section, condition of the switches, distance to the transition train. In the event of any deviation from the permissible parameters, the system immediately warns the driver or automatically triggers the emergency braking. All this leads to a reduction in human error, increased safety and higher line throughput.

ETCS Level 3 is the carrier of the concept of a mobile unit and efficient use of infrastructure. At the transition levels, the railway is divided into block sections. One block can only be occupied by one train.

Table 3: Comparison of different levels

Level	ETCS Infrastructure	Speed control	Data transmission
Level 0	Nope, I don't	Limited	None
Level 1	Yes (close)	Yes	Tochkovo
Level 2	Yes	Yes	Continuously
Level 3	Yes	Full	Continuous and dynamic

Table 3 shows a summary of the data on how the degree of automation and control of train traffic changes at different levels of ETCS. With the transition from Level 0 to Level 3, the role of the driver gradually decreases, and the role of automated systems increases [14]. (Commission E. , 2026)

4. IMPACT OF AUTOMATION

Automation in rail transport leads to a reduction in human error, because it transfers critical functions such as speed monitoring, traffic control and part of operational solutions to systems with continuous automatic control. This is extremely important in conditions where fatigue, stress or lack of information can slow down the driver's response and increase the risk of an accident (P). Automation reduces the likelihood of accidents, since the driver's attention and reaction are not leading.

Cab signaling provides the driver with information about the permissible speed, subsequent limits, braking curve and thus reduces the interpretation of external signals in case of poor visibility, high speed or complex infrastructure. With proper implementation of automation in railways, compliance with traffic schedules can be improved, energy consumption can be reduced and operational errors can be reduced.

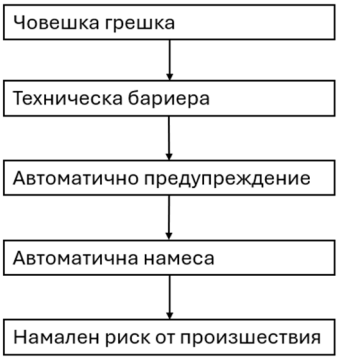


Figure 1: Impact of automation on the human factor

As a disadvantage of automation, the possibility of reducing the practical skills of staff can be considered, since it is rarely switched to manual control. In emergency situations, the operator must quickly regain knowledge, skills and make a decision, which can be difficult after a long period of passive observation. To maintain this knowledge and skills, periodic trainings, simulations and competence checks are carried out, as they are a mandatory element of safe automation.

5. CYBERSECURITY IN RAIL TRANSPORT

Cybersecurity

Digitalization is an inevitable process that also affects rail transport, bringing with it positive and negative consequences. Due to the increased number of connected systems and interactions between digital and physical components, the risk of cyberattacks increases. This necessitates the development of new approaches and architectures to improve cyber resilience, as there are problems with outdated security systems. The study shows that there is a lot of research in the field, but there is a need for a cybersecurity strategy in the railways that incorporates the new challenges. Most developments are theoretical and without the application of artificial intelligence [15]. (Ravdeep Koar, 2022) The problems, threats, and applications of artificial intelligence in cybersecurity are discussed in detail in [16] .(G. Hristov, 2023)

ERTMS works through a constant volume of data between trains, infrastructure and dispatch headquarters. Such exchanges must be protected from external attacks, disruptions and unauthorized access. The security of communication channels is very important, because even a small problem in them can affect the movement and safety of trains. The system should allow future upgrades to introduce new features, communication protocols without disrupting the operation and [17] (C).

The future development of rail transport will be related to the use of artificial intelligence, sensor systems, digital twins, predictive maintenance and autonomous train control. Through the new technologies, it will be possible to detect anomalies in real time, optimize schedules, manage energy consumption and predict possible failures. The aim is to achieve preventive safety, with the system identifying dangerous trends before they turn into incidents [18] (T).

The application of autonomous trains in railways is more complex than an automated metro. The reason is that metro systems operate in a closed and controlled environment, while rail operations include level crossings, a variety of rolling stock, open infrastructure, interaction with outsiders, weathering and higher speeds. To ensure the safety of passengers, an autonomous train must recognize obstacles, assess the condition of the infrastructure, communicate with dispatch systems and, in case of refusal, pass into a safe state. The reduction of the human factor does not mean its complete exclusion, but the redistribution of functions between man and machine. Automation takes over routine and critical defense functions, and the human remains responsible for supervision, ethical judgment, emergency management, and strategic decisions [19], [20], []. (U), (F).

Summary of the Study

The more the human factor is reduced through automation, the greater the dependence on cybersecurity increases. Table 4 summarizes the data from the survey.

Table 4: Major cyber risks

ETCS Level	Automation	Influence of the human factor	Main cyber risks
Level 0	Very low	Very high	Limited risks; Local unauthorized access to on-board systems; Lack of centralized protection.
Level 1	Low	High	Manipulation of balize data; Unauthorized access to on-board equipment; Vulnerability in data exchange between infrastructure and train.
Level 2	High	Average	Attacks on the GSM-R network; Refusal to communicate; Falsification of messages; DoS attacks; Compromise of the Radio Block Centre (RBC).
Level 3	Very high	Low	Cyber attacks on traffic management systems; Compromising train location data; Attacks on train integrity control systems; Attacks on cloud and artificial intelligence-based platforms; GPS/ GNSS spoofing.

Table 4 clearly shows the upward trend towards the level of automation from ETCS Level 0 to ETCS Level 3, the influence of the human factor is decreasing, but the dependence on digital technologies and the associated cyber risks is increasing. At Level 0, safety depends entirely on driver actions and traditionalsignalling, which is why cyber risks are limited. With the transition to Level 1, the use ofEurobalise devices begins, introducing the first potential vulnerabilities related to the exchange of digital data.

At Level 2, rail transport relies on continuous communication via GSM-R and Radio Block Centre, which significantly reduces human error, but creates opportunities for attacks on communication infrastructure and control centers.

At Level 3, cybersecurity becomes a key factor for the safe operation of railway infrastructure.

CONCLUSION

The study shows that rail automation reduces the risk of human error, but increases dependence on digital technologies and serious cyber protection is needed to ensure safe operation. The future development of ERTMS needs to be accompanied by measures to increase automation and cybersecurity solutions.

ACKNOWLEDGMENTS

The conducted research did not receive funding and there is no conflict of interest.

REFERENCES

[1] European Commission, "Predictive Maintenance for railway switches. Smart sensor networks on a machine learning analytics platform", Retrieved from https://cordis.europa.eu/: https://cordis.europa.eu/article/id/418452-system-uses-ai-and-industrial-iot-to-usher-in-new-era-of-predictive-maintenance-in-rail, (2026, 05 22)

[2] F. B. Vanderhaegen, "Human factors and automation in future railway systems", Cogn Tech Work 23, https://doi.org/10.1007/s10111-021-00670-3, (2021)

[3] EUAF. "Making the railway system work better for society" Luxembourg, https://www.era.europa.eu/system/files/2022-10/Integrating%20Human%20Factors%20in%20European%20Railways%20-%20Information%20for%20workers%20%28EN%29%20-%20BG.pdf, Publications Offi ce of the European Union, (2016)

[4] Human-Centered Systems Approach in a Social-Technical System, "Retrieved from U.S. Department ot Transportation Federal Railrood Administration", https://railroads.dot.gov/program-areas/human-factors/key-functions, (2019)

[5] T. Melissa, A.S. Baysari, ”Understanding the human factors contribution to railway accidents and incidents in Australia", Accident Analysis & Prevention, https://doi.org/10.1016/j.aap.2008.06.013, (2008)

[6] W. F. Wenil Dong, "High-speed train drivers' human error under fatigue and stress: the role of situation awareness and individual differences", Mational Library of Medicine, PMID: 38950888 https://doi.org/10.1080/00140139.2024.2369207, (2024)

[7] Z. Wang, G. Su, M. Skitmore. J. Chen, A. Chan&B. Xia, "Human Error Risk Management Methodology for Rail Crack Incidents", Urban Rail Transit, https://link.springer.com/article/10.1007/s40864-016-0032-2?utm_source=chatgpt.com, (2016)

[8] G. Peneva Yaser Haj Yousef, "Software Projects Success: Definition", Mechanics Transport Communications, ISSN 1312-3823 (print), volume 22, issue 2, article № 2632, (2014)

[9] RAIB, "RAIB: current investigations and safety digests", Retrieved from gov.uk: https://www.gov.uk/government/publications/raib-current-investigations-register/rail-accident-investigation-branch-current-investigations, (2015)

[10] W. Ndaru, "Human Error Theory – Human Factor Analysis and Clasification System MODUL 6 Human Error Theory – Human Factors Analysis and Classification System", https://www.academia.edu/37945903/Human_Error_Theory_Human_Factor_Analysis_and_Clasification_System_MODUL_6_Human_Error_Theory_Human_Factors_Analysis_and_Classification_System.

[11] Z. Hubenova, V. Gergov. "The Human Factor in Evaluating the Efficiency of Complex Human-Machine Systems in Transport", Mechanics Transport Communications, Vol. 14(3/2), ISSN 2367-6620 (online), Article No. 1393, (2016)

[12] S. Dekker, "The Field Guide to Human Error", Sweden: Linkiiping Institute of Technology, V https://www.leonardo-in-flight.nl/PDF/FieldGuide%20to%20Human%20Error.PDF, (2001)

[13] European Court of Auditors, 'A Single European Rail Traffic Management System – Will Political Choices Become a Reality?', Luxembourg: European Union, eca.europa.eu, (2017)

[14] European Commission, "ETCS Levels and Modes" Retrieved from transport.ec.europa.eu: https://transport.ec.europa.eu/transport-modes/rail/ertms/what-ertms-and-how-does-it-work/etcs-levels-and-modes_en, (2026)

[15] P. Ravdeep Koar, "A review on cybersecurity in railways", Jornal of Rail and Rapid Transit, https://doi.org/10.1177/09544097221089389, (2022)

[16] G. Hristov, A. Aleksieva-Petrova and I. Stankov, "CyberAttacks and Artificial Intelligence: A Systematic Mapping,"2023 International Scientific Conference on Computer Science (COMSCI), Sozopol, Bulgaria, pp. 1-7, https://doi.org/10.1109/COMSCI59259.2023.10315929, (2023)

[17] https://transport.ec.europa.eu/transport-modes/rail/ertms/what-are-benefits_en (accessed 12.06.2026)

[18] K. Trifonov, "Modern Dispatch Systems for Railway Traffic Management in Bulgaria: Status and Prospects", Mechanics Transport Communications, Vol. 23(2), article No 2662, ISSN 2367-6620 (online), (2025)

[19] Dimitrov D., Intelligent Transportation Management: Current Issues and Solutions ,LAP LAMBERT Academic Publishing, ISBN: 978-620-9-43886-8, (2026)

[20] D. Dimitrov and K. Lalov, "Intelligent Rail Transport Management Systems-Methodological Essence, Goals and Tasks", ENVIRONMENT. TECHNOLOGY. RESOURCES. Proceedings of the International Scientific and Practical Conference, https://doi.org/10.17770/etr2024vol2.8050, (2024)

[21] D. Dimitrov, I. Manchev, „Methodological and technical aspects of models for digital management of transport systems“, ENVIRONMENT. TECHNOLOGY. RESOURCES. Proceedings of the International Scientific and Practical Conference, https://doi.org/10.17770/etr2024vol2.8049, (2024)

SESSION B: RAILWAY ENGINEERING

An electronic business model for the regulation of railway traffic and transport based on artificial intelligence

Zoran G. Pavlović[1*], Ana Savić[2] , Veljko Radičević[1], Rade Vasiljević[1]

[1] ATAAS, Belgrade, Departments School of Railroad Transport, Belgrade, Serbia

[2] ATAAS, Belgrade, Department of Electrical Engineering and Computer Science, Belgrade, Serbia

ARTICLE INFO

[*] **Correspondence:** zoran.pavlovic@vzs.edu.rs

DOI: https://doi.org/10.46793/ET26.B11P **PAGES**: B75 – B80

ABSTRACT

The accelerated development of railway systems is based on modernization of infrastructure, interoperability of signaling and telecommunication systems, and digital transformation of operational processes. Modern railway traffic management requires efficient coordination between infrastructure systems, dispatch centers, rolling stock, and information technologies. Due to the complexity of railway operations, human errors may occur during traffic regulation and transport management. This paper presents an electronic business model for railway traffic and transport regulation based on artificial intelligence and advanced information and communication technologies. The proposed model integrates internal databases, intelligent algorithms, and digital communication systems into a unified platform for sustainable and safe railway traffic management. Particular attention is focused on the application of AI systems in decision-making processes, predictive analytics, automation, and operational safety. The paper also analyzes the advantages and limitations of AI implementation in railway systems and discusses future development trends in digital railway transport. Research results indicate that AI-based electronic business models can significantly improve operational efficiency, interoperability, safety, and sustainability in railway traffic regulation.

KEYWORDS

Railway traffic and transport, artificial intelligence, interoperability, electronic business, safety, sustainability

1. INTRODUCTION

Railway transport represents one of the most important components of sustainable transport systems. Modern railway infrastructure is increasingly based on advanced digital technologies, intelligent control systems, and interoperable communication networks. The modernization of railway systems requires the integration of signaling devices, telecommunication infrastructure, dispatch centers, electric traction systems, and rolling stock into a unified digital environment [1].

The development of Industry 4.0 technologies and digital transformation has significantly influenced railway traffic management. Artificial Intelligence (AI), Internet of Things (IoT), cloud computing, and big data analytics have become essential technologies in railway transport systems [2]. These technologies enable real-time monitoring, predictive maintenance, automated traffic regulation, and intelligent decision-making.

Modern railway systems require interoperability between different devices and infrastructures. Interoperability implies the ability of railway systems and subsystems to exchange information and operate without additional adjustments [3]. However, due to the complexity of railway operations and the large number of activities involved, human errors and operational interruptions may occur.

The application of artificial intelligence enables automation and optimization of many operational processes in railway transport. AI systems can analyze large volumes of operational data, predict failures, optimize schedules, and support dispatchers in traffic regulation [4]. The aim of this paper is to present an electronic business model for railway traffic and transport regulation based on artificial intelligence and digital technologies. Advanced technologies that can be applied in rail traffic and transport are shown in Table 1.

Table 1: Modern AI Technologies in Railway Systems

Technology	Application in Railway Systems	Main Advantages
Artificial Intelligence	Traffic regulation and predictive analytics	Increased efficiency and safety
Internet of Things (IoT)	Monitoring railway infrastructure	Real-time data collection
Big Data Analytics	Analysis of operational data	Improved decision-making
Cloud Computing	Data storage and communication	Centralized information management
Digital Twins	Simulation of railway operations	Predictive maintenance
Machine Learning	Failure prediction	Reduction of operational disruptions
Intelligent Dispatch Systems	Automated train control	Improved interoperability
Computer Vision	Monitoring railway crossings and stations	Increased operational safety

2. DIGITAL TRANSFORMATION OF RAILWAY SYSTEMS

Digital transformation in railway transport involves the integration of intelligent technologies into operational and business processes. Railway companies increasingly rely on advanced information systems to improve operational efficiency, safety, and sustainability [5].

Modern railway systems generate large volumes of data through signaling systems, onboard sensors, dispatch centers, and communication networks. These data can be processed using AI algorithms to improve traffic management and optimize operational processes.

The digital transformation of railway systems includes:

- intelligent traffic management;
- predictive maintenance;
- automated dispatching systems;
- digital communication platforms;
- real-time monitoring systems;
- electronic business operations.

One of the most important aspects of digital transformation is the development of intelligent transport systems (ITS). ITS technologies enable integration between railway infrastructure, trains, passengers, and operational management centers [6].

Artificial intelligence plays a key role in modern railway systems because it allows automation of repetitive tasks and supports operational decision-making. AI systems can detect anomalies in infrastructure conditions, predict equipment failures, and optimize train movement.

The European railway sector increasingly promotes digital railway strategies based on interoperability and automation. The implementation of the European Rail Traffic Management System (ERTMS) represents one of the most important examples of digital railway integration [7].

3. ARTIFICIAL INTELLIGENCE IN RAILWAY TRAFFIC REGULATION

Artificial intelligence systems are increasingly applied in railway traffic regulation and operational management. AI algorithms analyze operational data and support dispatchers in making optimal decisions in real time [8].

The main applications of AI in railway traffic regulation include:

- train scheduling optimization;
- predictive maintenance;
- intelligent signaling systems;
- automated traffic control;
- risk assessment and safety management;
- energy consumption optimization.

AI systems process information from multiple sources simultaneously, including sensors, dispatch centers, GPS systems, and communication networks. Based on these data, intelligent systems can identify potential operational risks and propose corrective actions.

Machine learning algorithms are particularly important because they can identify patterns in operational data and predict future events. For example, AI systems can predict delays, equipment failures, or traffic congestion based on historical and real-time data [9].

Another important application is predictive maintenance. Traditional maintenance approaches are based on predefined schedules, while AI-based predictive maintenance uses real-time data analysis to determine the actual condition of infrastructure and rolling stock.

AI technologies also improve passenger information systems and communication processes. Intelligent systems can automatically provide information about train schedules, delays, and alternative transport solutions.

Despite numerous advantages, the implementation of AI in railway systems also presents challenges:

- cybersecurity risks;
- reliability of AI algorithms;
- high implementation costs;
- data privacy concerns;
- interoperability between different systems.

Therefore, railway companies must establish appropriate standards and regulatory frameworks for AI implementation.

4. ELECTRONIC BUSINESS MODEL IN RAILWAY TRANSPORT

Electronic business (e-business) represents the application of digital technologies and information systems in business operations. In railway transport, e-business models support communication, operational management, logistics, ticketing systems, and infrastructure monitoring.

The proposed AI-based electronic business model consists of the following components:

- Centralized database;
- AI analytical module;
- Communication infrastructure;
- Intelligent dispatch center;
- User interface;
- Security and monitoring system.

The centralized database stores operational information related to trains, infrastructure, traffic schedules, maintenance activities, and passenger services. The AI analytical module processes operational data and generates recommendations for traffic regulation.

The intelligent dispatch center represents the core of the system. It integrates operational data from multiple sources and enables real-time monitoring of railway traffic.

The model also supports interoperability between railway companies and infrastructure managers. This interoperability enables efficient information exchange and coordination between different railway systems.

Electronic business models contribute to:

- reduction of operational costs;
- increased safety;
- improved communication;
- better passenger services;
- optimization of logistics processes;
- sustainable transport development.

Cloud computing technologies further improve scalability and accessibility of railway information systems. Railway operators can access operational information from different locations and coordinate activities more efficiently.

4.1. Case Study: AI-Based Predictive Railway Traffic Regulation Using ERTMS and Intelligent Dispatch Systems

To demonstrate the practical applicability of the proposed electronic business model, a case study based on AI-supported railway traffic regulation within an ERTMS environment is presented. The case study analyzes the implementation of artificial intelligence technologies for predictive maintenance, intelligent dispatching, and operational optimization in modern railway systems [7], [8].

The analyzed model is based on the integration of:

- ERTMS signaling systems;
- IoT infrastructure sensors;
- centralized cloud databases;
- AI analytical modules;
- intelligent dispatch centers;
- real-time communication systems [1], [5].

The operational architecture of the system is shown through the interaction between railway infrastructure, dispatch centers, and AI analytical platforms. Sensors installed on railway infrastructure continuously collect operational data related to:

- rail temperature;
- vibration levels;
- train speed;
- braking systems;
- signaling equipment status;
- energy consumption [6].

The collected data are transmitted through communication networks to a centralized database where machine learning algorithms analyze operational conditions in real time [2], [8].

- The AI system performs:
- anomaly detection;
- delay prediction;
- predictive maintenance;
- automatic route optimization;
- operational risk assessment [4], [9].

A simplified operational workflow is presented below:

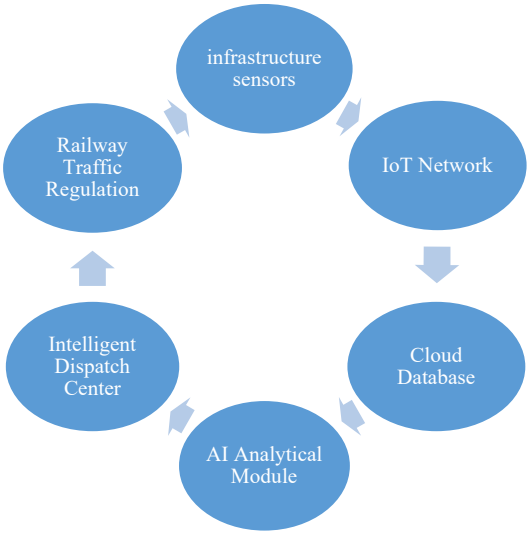


Figure 1: Operational workflow

The case study compares operational performance before and after AI implementation.

Table 2: Operational Performance Before and After AI Implementation

Operational Indicator	Traditional System	AI-Based System
Average train delay	12.4 min	4.8 min
Infrastructure failures per year	18	7
Energy consumption	100%	83%
Dispatcher reaction time	5-8 min	1-2 min
Predictive maintenance capability	Limited	Advanced
Operational safety level	Medium	High

The obtained results indicate that AI-based systems significantly improve operational efficiency and safety. Predictive analytics enables early detection of infrastructure failures and reduces operational disruptions [4], [9].

One important aspect of the case study is predictive maintenance. Traditional maintenance systems are based on fixed maintenance intervals, while AI-supported maintenance continuously evaluates the actual condition of infrastructure components. Machine learning algorithms process historical and real-time operational data to identify potential failures before they occur [9].

In addition, intelligent dispatch systems improve interoperability between railway operators and infrastructure managers. AI-supported traffic regulation enables dynamic optimization of train routes and minimizes operational conflicts in real time [3], [7].

The implementation of digital twins further improves operational management. Virtual simulation models allow railway operators to test different traffic scenarios and optimize operational processes before applying changes to real railway systems [5], [11].

Despite the advantages, several implementation challenges were identified [10]:

- high initial investment costs;
- cybersecurity risks;
- interoperability issues between legacy systems;
- need for employee digital training;
- regulatory and legal limitations.

However, long-term operational benefits significantly outweigh implementation challenges. The application of AI technologies contributes to [1], [10], [11].

- improved punctuality;
- increased infrastructure reliability;
- reduced operational costs;
- better passenger services;
- sustainable railway transport development.

The presented case study confirms that AI-based electronic business models can significantly improve railway traffic regulation and transport management within modern digital railway systems.

5. SAFETY AND SUSTAINABILITY ASPECTS

Safety represents one of the most important priorities in railway transport. AI technologies significantly contribute to improving operational safety through real-time monitoring, predictive analytics, and automated decision-making [10].

AI systems can identify abnormal operational conditions and issue warnings before accidents occur. Computer vision systems are increasingly used for monitoring railway crossings, stations, and railway infrastructure.

Another important aspect is sustainability. Railway transport is considered one of the most environmentally friendly transport modes. Digital technologies and AI systems contribute to sustainable railway operations through:

- optimization of energy consumption;
- reduction of operational disruptions;
- efficient traffic management;
- reduction of greenhouse gas emissions.

AI-based optimization algorithms can reduce unnecessary train stops and improve energy efficiency. Intelligent traffic management systems also reduce congestion and improve punctuality.

However, sustainability also requires long-term investments in digital infrastructure and employee training. Railway companies must continuously improve digital competencies and technological capabilities.

6. FUTURE DEVELOPMENT TRENDS

Future railway systems will increasingly rely on autonomous technologies, intelligent communication networks, and advanced AI systems. The concept of “smart railways” involves full integration of digital technologies into operational and business processes [11].

Future trends include:

- autonomous trains;
- digital twins of railway infrastructure;
- 5G communication systems;

- AI-driven predictive analytics;
- blockchain-based logistics systems;
- intelligent passenger services.

Digital twins will enable simulation and optimization of railway operations in virtual environments. This technology will significantly improve predictive maintenance and infrastructure management.

Autonomous railway systems are also expected to become more common in the future. AI technologies will enable trains to operate with minimal human intervention while maintaining high safety standards.

Cybersecurity will remain a critical issue due to the increasing dependence on digital technologies. Railway companies will need advanced cybersecurity strategies to protect operational data and communication systems.

7. CONCLUSION

Artificial intelligence and digital technologies are transforming railway traffic regulation and transport management. Modern railway systems increasingly depend on intelligent information systems, interoperability, and automation.

The electronic business model presented in this paper demonstrates how AI technologies can improve operational efficiency, safety, sustainability, and interoperability in railway systems. Intelligent dispatch systems, predictive maintenance, and automated traffic management significantly reduce operational risks and improve service quality.

Despite the advantages of AI implementation, challenges related to cybersecurity, interoperability, and ethical issues remain important. Successful digital transformation requires continuous investment in digital infrastructure, employee education, and regulatory development.

Future railway systems will increasingly integrate AI technologies, autonomous operations, and intelligent communication systems, contributing to safer, more efficient, and sustainable railway transport.

REFERENCES

[1] European Commission, Digital Transformation of Railways in Europe, Brussels, Belgium, 2024. https://transport.ec.europa.eu/document/download/774e79c9-1ece-4514-8f16-a2b98049c82e_en, (2024)

[2] Hadj-Mabrouk, H.: A literature review on the applications of artificial intelligence to European rail transport safety. IET Intell. Transp. Syst. 1–34 (2024). https://doi.org/10.1049/itr2.12587, (2024)

[3] ERA, European Railway Interoperability Report, European Union Agency for Railways, https://www.era.europa.eu/system/files/2023-07/Annual%20overview%20for%20Interoperability%20-%202023.pdf, (2023)

[4] Jun Liu, Gehui Liu, Yu Wang, Wanqiu Zhang. Artificial-intelligent-powered safety and efficiency improvement for controlling and scheduling in integrated railway systems. High-speed Railway, 2024, 2(3): 172-179 https://doi.org/10.1016/j.hspr.2024.06.002, (2024)

[5] UIC, Digital Rail Report 2025, International Union of Railways, Paris, France, 2025. https://uic.org/com/enews/article/uic-publishes-the-latest-global-rail-sustainability-report-on-world-sustainable, (2025)

[6] Pencheva, E.; Atanasov, I.; Trifonov, V. Towards Intelligent, Programmable, and Open Railway Networks. Appl. Sci., 2022, 12, 4062. https://doi.org/10.3390/app12084062, (2022)

[7] European Commission, ERTMS Deployment Strategy 2025, Brussels, Belgium, 2025. https://transport.ec.europa.eu/document/download/774e79c9-1ece-4514-8f16-a2b98049c82e_en, (2025)

[8] Usman, A.; Ali, T.M.; Irawan, C. Machine Learning Approaches for Real-Time Traffic Density Estimation and Public Transport Optimization.Eng.Proc., 107, 117, https://doi.org/10.3390/engproc2025107117, (2025)

[9] M. Z. Shaikh, S. Ali, S. Ali, E. N. Baro, Y. A. Baloch and B. S. Chowdhry, "Predictive Maintenance in Urban Railway Systems Using Machine Learning Models," 2024 Global Conference on Wireless and Optical Technologies (GCWOT), Malaga, Spain, 2024, pp. 1-5, https://doi.org/10.1109/GCWOT63882.2024.10805699, (2024)

[10] OECD, Artificial Intelligence and Sustainable Transport, OECD Publishing, Paris, 2023. https://www.oecd.org/en/topics/policy-issues/artificial-intelligence.html, (2023)

[11] Rouky, N.; Benmoussa, O.; Fri, M.; Abourraja, M.N.; Ben-Bouazza, F.-E. Artificial Intelligence Applications for Smart and Sustainable Mobility as a Service Concept: A Systematic Literature Review.Future Transp.2025, 5, 122, https://doi.org/10.3390/futuretransp5030122, (2025)

SESSION B: RAILWAY ENGINEERING

State of dispatcher systems in Bulgaria and their role in the liberalised passenger rail market

Kostadin Trifonov[1]

[1] Todor Kableshkov University of Transport, Department of Transport Technology, Organisation and Management, Sofia, Bulgaria

ARTICLE INFO

*** Correspondence:** ktrifonov@vtu.bg

DOI: https://doi.org/10.46793/ET26.B12T **PAGES:** B81 – B91

ABSTRACT

Modern dispatcher systems are a key element of railway traffic management, directly shaping punctuality, operational reliability, and passenger comfort. In a liberalised passenger rail market, where operators compete through service quality, punctuality becomes a critical determinant of competitiveness. This paper examines the current state of dispatcher systems in Bulgaria, identifying technological disparities and their impact on operations. A conceptual framework is developed to show how automation, communication quality, and safety systems influence punctuality and, consequently, perceived comfort. A comparative analysis of the Sofia-Plovdiv and Sofia-Burgas corridors illustrates how different modernisation levels affect performance. The results show that completing dispatcher system upgrades significantly improves punctuality, reduces delay propagation, and enhances service predictability. These improvements directly increase passenger comfort and strengthen operator competitiveness in an open market. The findings demonstrate that dispatcher system modernisation is not only a technical requirement but a strategic prerequisite for equal access, higher service quality, and successful liberalisation of passenger rail transport in Bulgaria.

KEYWORDS

Dispatcher systems, Punctuality, Passenger comfort, Railway operations, Liberalised rail market

1. INTRODUCTION

The liberalisation of the European passenger rail market, driven by the Fourth Railway Package (Directive 2016/2370/EU) [1], imposes stringent requirements on the quality and reliability of railway operations. Infrastructure managers are obliged to provide transparent, non-discriminatory access to all licensed operators. At the same time, service quality-measured above all by punctuality-becomes the primary competitive differentiator in an open market. Dispatcher systems occupy the pivotal position in this chain: they are the operational backbone through which traffic control decisions are made, real-time communication is sustained, and safety is enforced. Their technological state, therefore, largely determines whether the conditions for meaningful liberalisation can be met.

Bulgaria has already taken a concrete step beyond the regulatory framework: the country has liberalised its passenger rail market both de jure and de facto. A private passenger rail operator has been formally awarded the right to provide services alongside the state incumbent on domestic routes, making Bulgaria one of the few countries in Southeast Europe where effective passenger-rail competition is set to become an operational reality rather than a future regulatory objective. Although this competition will take place primarily within the framework of Public Service Obligation (PSO) contracts rather than pure open-access operations, the practical outcome remains the same: two operators will share the same infrastructure and will compete on service quality, punctuality, and passenger experience.

In February 2026, 12-year PSO contracts were signed under Regulation (EC) 1370/2007 [2]: the Western Region (~75% of train-km) was awarded to BDZh - Passenger Services, while the Northern and Southern Regions (~25% combined) were awarded to the private operator Ivkoni Express EAD, at rates of 8.99/7.63/7.84 EUR per train-km respectively [3]. Total public support over the 12 years amounts to EUR 1.8 billion for BDZh and EUR 511 million for Ivkoni Express [3]. These contracts will enter into force on 13 December 2026, and their signature constituted an explicit conditionality for Bulgaria's access to EUR 734 million in EU Recovery and Resilience Facility funding for 35 new trainsets [4].

With a new operator fully dependent on shared infrastructure once the contracts become operational, the technological state of dispatcher systems becomes a matter of direct contractual and operational urgency. The ability of the private operator to deliver competitive levels of punctuality and comfort will depend entirely on the infrastructure it does not control and must share with the incumbent, making dispatcher system performance a central determinant of the sustainability of the liberalised market that is about to emerge.

In Bulgaria, the National Railway Infrastructure Company (NRIC) manages the railway network, within which significantly different generations of signalling and interlocking installations coexist across individual stations. Modern technologies such as Centralised Traffic Control (CTC), GSM-R digital radio, and ETCS have been implemented mainly along key sections of the core TEN-T network. The most advanced deployment is along Railway Line No. 1 (Sofia-Plovdiv-Svilengrad), classified as a railway main line under the national line categorisation framework [5]: the southern section Krumovo-Svilengrad operates entirely under the Elektra-2 dispatching system, while the Sofia-Plovdiv section is undergoing full reconstruction [6].

The older analogue platforms - Dispatcher Communication System (DPS) and Unified Telegraph Communication (ETK) - remain in operation across a substantial part of the network. A study on the development of transport infrastructure in Bulgaria and the region, conducted by Dimitrov and Petrova [7], traces this uneven modernisation and shows that investments are concentrated primarily along international corridors, while large portions of the network remain only minimally upgraded.

According to the 2024-2025 Network Statement of SE NRIC [8], the total developed length of the managed network reached 6,446 km (total developed length, including all tracks and sidings; route length: 4,025 km) as of 31 December 2023, of which 3,904 km are open line with standard gauge (1,435 mm). The network comprises 294 stations, 16 block posts, 397 halts, 186 tunnels, and 995 bridges, with more than 70% electrified using a 25 kV / 50 Hz AC overhead system.

Train control is carried out through key relay-based systems (KIRS), semi-automatic and automatic block systems, and - on certain sections - centralised dispatching control. The TPMS system is deployed network-wide for real-time train monitoring and timetable management, while ARAMIS provides centralised control of the Krumovo-Dimitrovgrad-Svilengrad section from the Dispatching Centre in Plovdiv [8].

Station-level safety equipment includes KIRS, MRC, and MKC interlocking installations. Centralised control from a unified supervisory centre is available only on the sections Iliyantsi-Karlovo (Line 3, TRIS CTC) and Septemvri-Svilengrad (Line 1, Elektra-2/ARAMIS). Double-track sections total 995 km, while the narrow-gauge line Septemvri-Dobrinishte (760 mm) extends over 125 km [8].

This territorial asymmetry in infrastructure investment forms the structural context within which differences in dispatching systems must be understood. The technological heterogeneity has direct and measurable consequences for punctuality, delay propagation, and post-incident recovery. Previous research has identified the dispatching-system infrastructure as a major source of operational unreliability in the Bulgarian railway context, particularly on lines where outdated equipment remains in service [9]. The incident-risk profile is closely linked to the adequacy of maintenance, the condition of dispatching systems, and the quality of communication links [9].

The dispatcher function itself is among the most complex and cognitively demanding roles in railway operations. Modelling the work of train dispatchers has been a research concern since the early era of computerised railway management [10]. As network complexity has grown and traffic density has increased, the adequacy of dispatcher training and qualification has become an increasingly pressing issue [11]. Both the cognitive demands placed on dispatchers and the technical systems they operate have evolved substantially, culminating in the current generation of intelligent management systems that integrate real-time data, automated conflict detection, and decision-support interfaces [12], [13].

This paper analyses the current state of dispatcher systems in Bulgaria, examines their impact on operational performance across selected corridors, and discusses the implications for the liberalised passenger rail market. The structure is as follows: Section 2 reviews the European context and the relevant literature; Section 3 describes the principal technology layers of Bulgarian dispatcher systems and their deployment status; Section 4 discusses the relationship between dispatcher modernisation, punctuality, and passenger comfort; Section 5 presents a comparative corridor analysis; and Section 6 draws conclusions for infrastructure policy and liberalisation strategy.

2. LITERATURE REVIEW AND EUROPEAN CONTEXT

European experience consistently shows that the technological state of traffic management systems is a prerequisite for both high operational quality and effective rail market liberalisation. Countries that have completed or substantially advanced the deployment of ETCS, GSM-R, and CTC have recorded measurable reductions in train delays, more stable adherence to schedules, and higher passenger satisfaction. These outcomes, in turn, create the operational conditions under which open-access competition can deliver genuine consumer benefits rather than simply redistributing passengers between unreliable services. A structured study of railway passenger transport across the European Union by Stoilova et al. [14] provides empirical context for these trends, documenting the wide variation in punctuality performance and service quality across member states and identifying infrastructure modernisation as the most consistent differentiator between high- and low-performing national networks. A multi-criteria evaluation of the TEN-T Orient-East Med Corridor

by Stoilova et al. [15] confirms that network-level performance gaps in Bulgaria reflect infrastructure capability deficits of precisely the kind examined here.

Within the Bulgarian context, a foundational quantitative picture is provided by Karagyozov and Todorova [16], who conducted a systematic analysis of delays affecting both freight and passenger trains across the Bulgarian railway network. Their work identified infrastructure condition and communication system limitations as primary contributors to the accumulation of delays, findings that remain relevant as the modernisation programme proceeds unevenly. The risk dimension of passenger train delays has been further examined by Stoilova [17], whose risk analysis of delay patterns in the rail network links delay probability distributions to infrastructure and operational characteristics at the line level. This probabilistic perspective complements the delay-causation analysis in [16] and provides an additional analytical tool for quantifying the operational impact of gaps in the dispatcher system. The propagation of delays through the network has been further studied using GERT-based network modelling: Todorova [18] developed a GERT model of a technical station that quantifies passenger trainset delays as a function of station operations and their stochastic variability. This approach demonstrates that delay generation at individual operational nodes-precisely the nodes where dispatcher capability is most consequential-can be modelled and predicted with sufficient precision to support investment planning.

Simulation modelling of railway sections provides a complementary tool for evaluating the impact of infrastructure and control system parameters on operational performance. Atanasov and Todorova [19] developed a simulation model of a railway section that captures the interaction between timetable structure, infrastructure characteristics, and operational disturbances. Such models are directly applicable to assessing before-and-after scenarios for dispatcher system upgrades, providing an analytical foundation for evidence-based modernisation decisions.

Passenger comfort research quantifies the commercial stakes: Todorova and Trifonov [20] establish through Fuzzy-AHP that time-related attributes carry the highest weight in passenger assessments, while Todorova's Logit-based mode choice model [21] shows that punctuality improvements translate directly into modal shift and operator revenue.

The digitisation dimension of railway management has been examined in the broader context of workforce challenges and system design. Todorova [22] analysed the implications of digitisation in railway transport for workforce structure and competency requirements, identifying the transition to digital dispatcher systems as one of the most significant sources of skill disruption in the sector. The methodological basis for intelligent railway management systems has been laid out by Dimitrov et al. [12], who articulated the goals, tasks, and architectural logic of intelligent rail management platforms, and extended this analysis to include the methodological and technical aspects of digital transport management models [23]. A comprehensive overview of problems and solutions in intelligent transport management was provided by Dimitrov, situating the challenge of dispatcher system modernisation within the broader transition toward AI-assisted and data-driven traffic control.

Real-time management systems for railway nodes have been studied in the context of marshalling yards by Dimitrov et al. [24], who proposed a methodological approach to developing real-time yard management systems, and by Tomov et al. [25], who optimised the wagon sorting process within such systems. While focused on freight operations, the system architecture and real-time data integration principles developed in this line of research are directly transferable to passenger dispatcher system design. Station-level operational optimisation has been addressed specifically for passenger complexes by Todorova [26], whose analysis of shunting process optimisation at a passenger complex illustrates the operational efficiency gains achievable through systematic management improvement at the node level.

Finally, the infrastructure-charging dimension of liberalisation-central to ensuring equitable access-has been examined by Todorova [27], who analysed infrastructure charges in the new EU member states and identified the tension between cost-recovery requirements and the need to attract new operators. Unequal infrastructure capabilities-including disparities in dispatcher systems-compound this tension by creating differential operating conditions that no pricing mechanism can fully compensate for.

3. DISPATCHER SYSTEMS IN BULGARIA: ARCHITECTURE AND STATUS

3.1. Railway Line Numbering and Categorisation

Under the Ordinance on the Categorisation of Railway Lines in the Republic of Bulgaria (Council of Ministers Decree No. 293/2001, last amended 2022) [5], the national railway network is divided into four categories: railway magistrales (equivalent to European-corridor main lines), Category I lines (international connections and magistrale feeders), Category II lines (regional lines), and Category III lines (private/industrial sidings). The three lines examined in this paper span the two highest categories and are numbered as follows in Annex 4 of the Ordinance.

Railway Line No. 1 (railway magistrale) runs from Kalotina-West (state border with Serbia) through Sofia - Plovdiv - Dimitrovgrad to Svilengrad (state borders with Greece and Turkey). This corridor forms the Bulgarian section of the Orient/East-Med TEN-T Core Network corridor and is classified entirely as a railway magistrale under Annex 1 of the Ordinance. It is the principal Sofia-Plovdiv axis examined in this paper. The stations along the line, from Sofia to Svilengrad, include: Sofia (София) - Elin Pelin (Елин Пелин) - Kostenets (Костенец) - Septemvri (Септември) - Pazardzhik (Пазарджик) - Plovdiv (Пловдив) - Krumovo (Крумово) - Dimitrovgrad (Димитровград) - Harmanli (Харманли) - Svilengrad (Свиленград).

Railway Line No. 3 (Category I) runs from Iliyantsi (Илиянци) - Karlovo (Карлово) - Tulovo (Тулово) - Dabovo (Дъбово) - Zimnitsa (Зимница), and Karnobat (Карнобат) - Komunari (Комунари) - Sindel (Синдел) - Varna Ferry (Варна фериботна). This line provides the direct Sofia-Burgas connection via an inland route and is classified as Category I in Annex 2 of the Ordinance (section Iliyantsi-Karlovo-Tulovo). The 138 km section from Iliyantsi (Sofia area) to Karlovo is equipped with the TRIS CTC system managed by the Sofia Dispatcher Centre. Key stations along the Sofia-Burgas axis of Line No. 3 include: Iliyantsi (Илиянци) - Kurilo (Курило) - Bunovo (Буново) - Anton (Антон) - Karlovo (Карлово) - Tulovo (Тулово) - Stara Zagora (Стара Загора) - Zimnitsa (Зимница) - Karnobat (Карнобат) - Burgas (Бургас).

Railway Line No. 8 (railway magistrale) runs from Plovdiv (Пловдив) - Filipovo (Филипово) - Skutare (Скутаре), and Plovdiv Razpredelitelna-Iztok (Пловдив разпределителна-изток) - Stara Zagora (Стара Загора) - Zimnitsa (Зимница) - Karnobat (Карнобат) - Burgas (Бургас). The Plovdiv-Burgas segment of Line No. 8 is classified as a railway magistrale under Annex 1 of the Ordinance. It is the primary axis for the Sofia-Burgas corridor considered in this paper, with trains running from Sofia via Line No. 1 to Plovdiv and continuing east on Line No. 8. The full route from Sofia to Burgas using this path covers approximately 402 km. This line is currently under modernisation by the INTERFLO CTC system (Alstom), with commissioning targeted for 2030.

Sofia and Burgas are served by two distinct routing options. The first is via Line No. 1 (Sofia-Plovdiv, magistrale) and Line No. 8 (Plovdiv-Burgas, magistrale), totalling approximately 402 km. The second is via Line No. 3 (Iliyantsi-Karlovo-Tulovo-Zimnitsa-Karnobat-Burgas), a Category I line, which offers a more direct inland route. Both routes share the Karnobat-Burgas final section. The existence of these two competing routes means that both infrastructure quality and dispatcher system capability on each corridor directly influence which route operators prefer and how reliably the Black Sea connection can be served, a factor of particular relevance under the 2026 PSO contracts [3].

3.2. Principal Technology Layers

Bulgarian dispatcher systems cover a wide technological spectrum, structured in a five-level hierarchy - from basic mechanical dependencies to fully digital route-computer interlocking. The hierarchy progresses as follows: (1) stations without any signalling equipment (“Bez OT”); (2) key-dependency arrangements, in which the interdependencies between points and signals are ensured through mechanical keys; (3) electromechanical systems (EMC) - a transitional generation between mechanical and relay-based interlockings; (4) route-relay systems (MRC), comprising several generations - the WSSB-1 and WSSB-2 families, the N-68 family in its three variants (basic, N-68U, N-68V, MN-70), as well as Block Route-Relay Centralisation (BMRC), considered the latest generation in this group, where a computer-based control and visualisation system is now added; (5) route-computer interlocking (MKC), which provides fully digital interlocking, compatible with CTC, with real-time visualisation and centralised route control.

Route-relay systems automate route setting but do not support full CTC functionality at network level. Above them are modern digital interlocking platforms such as Elektra-2 and TRIS, which enable full CTC integration and centralised supervision. At an even higher conceptual level stand integrated automation environments of the INTERFLO type, combining CTC, train-graph visualisation, and operational data exchange between multiple operators.

The key operational threshold is the continuous deployment of MKC along the entire corridor section. Only once full coverage is achieved can the dispatcher control the entire section from a single workstation, without relying on local staff to execute interdependent commands.

3.3. Deployed Dispatcher Centres: Characteristics and Capabilities

Bulgaria operates two fully commissioned centralised dispatcher centres and has a third under deployment (2030). The Plovdiv Dispatcher Centre runs the Elektra-2 CTC system on the Krumovo-Svilengrad section of Railway Line No. 1 (Orient/East-Med TEN-T magistrale) [5]. Built on a SIL 4-certified, triple-redundant platform and transmitting all commands via optical fibre, Elektra-2 is fully centralised on this section; the Sofia-Plovdiv section is undergoing comprehensive reconstruction with ETCS Level 2 and GSM-R [6], [28]. The Sofia Dispatcher Centre runs TRIS CTC on the 138 km Iliyantsi/Kurilo-Karlovo section of Line 3, with Automatic Route Setting (ARS), event logging, and full ERTMS compatibility; the remainder of Line 3 operates under conventional supervision. INTERFLO, developed by Alstom for the full Plovdiv-Burgas axis (Line 8, magistrale), will integrate automated train control, axle-count block working, and dual-redundant signalling networks upon commissioning. The legacy N-68 relay-based systems, whose residual certified life expires around 2032 [29], lack automated conflict detection and rely on manual telephone coordination, causing delays to propagate unchecked on mixed-technology corridors-the highest-risk operational phase [30]. The Network Statement 2024-2025 confirms that SE NRIC also operates the ARAMIS Dispatching Management System on the Krumovo-DimitrovgradSvilengrad section, enabling control from a single centre while simultaneously serving passenger information boards at stations along the route. Safety equipment (signalling) at the station level relies on KIRS interlocking or station interlocking systems combined with semi-automatic or automatic block-working between stations. Where stations are undergoing reconstruction, temporary control panels (PVU) are deployed in lieu of full interlocking, introducing transient automation gaps during transition periods. For trains not equipped with on-board ETCS hardware, operation on ETCS-equipped lines is permitted at a maximum speed of 130 km/h, preserving throughput but limiting the full benefit of train-protection automation.

3.4. Comparative Analysis: Route-Relay Interlocking vs Route-Computer Interlocking (MKC)

The principal operational distinction in Bulgaria lies between legacy route-relay interlocking (RRI, type N-68), which relies on verbal authorisation and manual route execution, and modern route-computer interlocking (MKC), which performs automated safety checks and accepts commands directly from the CTC workstation. In practice, RRI operation introduces several minutes of human-mediated delay per movement, whereas MKC enables sub-minute execution with automatic conflict detection preventing incompatible route allocations.

These two interlocking families underpin the three system layers that determine corridor-level performance. The CTC supervisory layer governs conflict resolution and traffic regulation, with markedly shorter decision cycles under full CTC operation compared to legacy supervision. The GSM-R communication layer, deployed in accordance with TSI CCS and ERA specifications, replaces the analogue DPS/ETK platforms [10], [11], [31], [32] and provides the digital bearer for ETCS. The ETCS train-protection layer, defined by TSI CCS [33], ranges from Level 1 (partial overlay using lineside signals) to Level 2 (continuous radio-based supervision without reliance on lineside signalling). The deployment status of these three layers across the network is summarised in Table 1.

ETCS Level 1 provides a safety overlay via Eurobalises alongside existing lineside signals, while Level 2 transmits movement authorities via GSM-R directly to the cab, eliminating lineside signals and enabling higher capacity. Bulgaria's current deployments operate on Baseline 2 (SRS 2.3.0d); migration to current baselines (B3R2/B4) will be addressed in the forthcoming upgrade cycle [34].

3.5. Deployment Status and Identified Gaps

Table 1 maps deployment status across the three principal corridors. The data reveal a pronounced asymmetry: Line No. 1 (TEN-T magistrale) is substantially modernised, whereas the Sofia-Burgas and Plovdiv-Burgas corridors retain legacy infrastructure and lack active EU-funded modernisation programmes.

Table 1: Dispatcher technology deployment status: Sofia-Plovdiv and Sofia-Burgas corridors; target values apply to the Sofia-Burgas corridor (INTERFLO project, completion 2030)

Indicator	Line 1: Sofia-Plovdiv-Svilengrad (magistrale, TEN-T)	Line 3: Iliyantsi-Karlovo (Cat. I)	Line 8: Plovdiv-Burgas (magistrale)	Line 8 Target (INTERFLO, 2030)
Dispatcher system	Elektra-2 CTC (Plovdiv DC) Krumovo-Svilengrad: active Sofia-Plovdiv: active reconstruction (incl. Elin Pelin-Kostenets: Transport Connectivity Programme)	TRIS CTC (Sofia DC) Iliyantsi-Karlovo: fully operational-Karlovo-Varna: station interlocking (KIRS/MRC/MKC), telephonic dispatcher orders - no unified CTC centre	Legacy RRI N-68 (telephony-based) Sofia-Plovdiv section: see Line 1. Plovdiv-Burgas: INTERFLO modernisation in progress (Alstom, target 2030)	INTERFLO CTC (Plovdiv DC)
CTC coverage	100% (Krumovo-Svilengrad) Partial (Sofia-Plovdiv, under active reconstruction)	100% (138 km)	62% (partial)	100%
GSM-R / ETCS	94% / Level 2 (Krumovo-Svilengrad)	Deploying / Level 1	47% / Level 1 partial	100% FRMCS / Level 2
Legacy DPS/ETK	Decommissioned (active) Active (Sofia-Plovdiv reconstruction; new infrastructure upon completion)	Decommissioned	Active (53-61%)	Fully replaced
Mean delay / Punctuality	8.2 min / 78.4% (Krumovo-Svilengrad)	~78% (CTC section); improved punctuality following TRIS commissioning	18.6 min / 61.7%	Target: ≤8 min / ≥85%
EU funding	CEF active (Sofia-Plovdiv reconstr.)	National/partial EU	Rehabilitation of Plovdiv- Burgas railway line, Phase 2	CEF 2021-2027 (INTERFLO project)

NRIC operational reports for 2022-2024 [35] identify five systemic challenges: partial GSM-R coverage (fallback to analogue DPS/ETK); incomplete ETCS Level 1 deployment; automation discontinuities at technology boundaries; limited legacy-CTC interoperability; and, most consequentially, manual anomaly resolution, which raises average incident response time to 31 min on legacy segments versus 12 min under CTC-the primary driver of the punctuality gap documented in Table 1.

Beyond punctuality, Todorova and Filipov [9] document that legacy dispatcher infrastructure correlates with higher incident frequency and longer recovery times. Petrova [30] further identifies the mixed-technology coexistence period as the highest-risk phase, reinforcing the case for accelerated full-corridor modernisation.

Addressing these gaps is now the subject of an active pre-feasibility programme. SE NRIC is conducting pre-project studies for ERTMS deployment across five line sections (Administrative Contract No. DPTS-1/12.03.2024), co-financed

by the Cohesion Fund under the Transport Connectivity Programme with a total project value of BGN 6,240,084.51, executed over the period 12 March 2024 to 31 August 2026 [36]. The five sections under study are: Sofia-Mezdra-Gorna Oryahovitsa-Kaspichan-Sindel (existing line); Radomir-Kulata (existing line); Karnobat-Sindel (existing line under modernisation); Ruse-Gorna Oryahovitsa-Stara Zagora (existing line); and Vidin-Medkovets-Mezdra (line under modernisation). The studies are required to be aligned with the strategic deadlines of Regulation (EU) 2024/1679 of the European Parliament and of the Council of 13 June 2024 on TEN-T guidelines: core network by 2030, extended core network by 2040, and comprehensive network by 2050. For each section, three deployment scenarios for ERTMS Level 2 will be evaluated through legal, technical, economic and financial analysis, with risk and sensitivity assessment, environmental impact procedures, and detailed cost-benefit analysis prepared as the basis for subsequent procurement and construction [36].

4. DISPATCHER MODERNISATION, PUNCTUALITY AND PASSENGER COMFORT

4.1. The Modernisation-Punctuality Relationship

Dispatcher modernisation yields non-linear gains in punctuality, with the steepest improvements occurring during the transition from legacy to intermediate automation-the position of the Sofia-Burgas corridor today. GERT-based station modelling [18] and corridor-level simulation [19] provide quantitative frameworks for predicting these gains. Empirical evidence corroborates the pattern: Karagyozov and Todorova [16] document that delay accumulation is concentrated on segments with the weakest dispatcher infrastructure, consistent with Table 1. Beyond mean delay, intelligent management systems [12], [13], [23] reduce variance in operational performance-the dimension most damaging to passenger satisfaction and modal choice.

4.2. Punctuality and Passenger Comfort

Punctuality is the dominant comfort determinant in the Bulgarian context: Fuzzy-AHP analysis [20] assigns schedule adherence 22.9% of total comfort weight-ahead of vehicle hygiene (19.4%), intermodal connections (18.95%), and staff service level (15.6%)-establishing the dispatcher system as the infrastructure element most directly linked to passenger-perceived comfort.

The Logit mode-choice model [21] confirms that unreliable punctuality drives a systematic modal shift to car and coach, particularly among commuters and business travellers-so, in a liberalised market, a new operator’s ability to attract passengers depends directly on the dispatching infrastructure it uses.

Real-time passenger information-a standard component of modern CTC platforms-partially offsets the impact of delays by reducing uncertainty and the perceived loss of control. Station-level optimisation, including improvements to shunting processes at passenger complexes [26], complements dispatcher-driven schedule adherence at the departure end of journeys.

4.3. Dispatcher Modernisation and Market Liberalisation

The Fourth Railway Package mandates non-discriminatory access to infrastructure. Yet, where dispatcher technology varies materially across corridors, operators face structurally unequal operating environments: punctuality levels achievable on a fully modernised corridor are physically unattainable on a legacy-equipped route, regardless of operator quality. Stoilova [37] documents this asymmetry for the Bulgarian context, identifying infrastructure quality differentials as a structural barrier to competitive market entry that regulatory pricing-or access-charge design calibrated for cost recovery rather than quality differentials [27]-alone cannot offset.

A further dimension concerns the equitable treatment of competing operators. Every conflict-resolution decision allocates delay between train operating companies; legacy manual dispatching lacks the formal optimisation frameworks to do so transparently, creating scope for discrimination favouring incumbent operators [38]. Modern CTC platforms with automated rescheduling incorporate formal fairness constraints [39], and their event logs (as in TRIS CTC and INTERFLO) provide verifiable records enabling post-hoc regulatory audits. Dispatcher modernisation thus serves a dual function: operational improvement and institutional infrastructure for equitable multi-operator access.

Workforce transition also requires structured management: Todorova [22] and Dimitrov and Petrova [40] favour corridor-by-corridor rollout, with Petrova and Trifonov [41] identifying an IPMA Agile-Waterfall approach as most effective for projects of this complexity [11].

The governance dimension of infrastructure management is equally significant for the success of liberalisation. Vasilev [42] identifies the effective management of public railway infrastructure as resting on three interdependent functions: access control, traffic management, and maintenance and development. Within the traffic management function, Vasilev specifically points to the deployment of ERTMS systems and modern information technologies as the primary lever for process improvement, while noting that key performance indicators (KPIs) for traffic management - including train kilometres, punctuality ratios, and track-occupation levels due to maintenance - must be monitored on a five-year horizon to support strategic governance. The same study identifies a structural weakness common to Balkan railway infrastructure managers: the absence of a modern cost-management system capable of attributing traffic-management costs at the level

of individual corridors or operational units [42]. In a multi-operator environment governed by PSO contracts, this deficiency directly impairs the transparent attribution of delay costs and infrastructure-use charges between competing operators - precisely the accountability requirement that the 2026 PSO contracts impose.

The 2026 PSO contracts add a contractual dimension to the dispatcher system's capabilities. With a single Line No. 3 service crossing two PSO lots (BDZh Western Region; Ivkoni Express Southern/Northern Regions), every dispatcher delay-allocation decision directly affects each operator's KPI performance and subsidy entitlement. Legacy N-68 segments, lacking automated event logging, cannot attribute delays between operators, as required by Regulation (EC) 1370/2007 [2]. Modern CTC platforms (TRIS CTC, INTERFLO) produce the operator-level performance records that the framework demands; legacy systems do not. Dispatcher modernisation is therefore a legal prerequisite for the equitable functioning of the new market structure.

4.4. Formal Model of Dispatcher System Performance

A compact formal model links three normalised indices-Automation Level A, Communication Quality C, Safety Integration S ($\in[0,1]$) - to two outputs: Punctuality Index P and Comfort Index K [43], [32], [39]. $P = w_1 \cdot A + w_2 \cdot C + w_3 \cdot S + \varepsilon$ (1) [w_1=0.45, w_2=0.30, w_3=0.25] $\delta = \delta_0 / (1+k \cdot A \cdot C)$ (2) [the $A \cdot C$ term captures that automation benefits require communication support] $K = \alpha \cdot P + \beta \cdot R + \gamma \cdot I$ (3) [α=0.50, β=0.35, γ=0.15; R = reliability, I = information index] Results in Table 2.

Table 2: Model output: predicted punctuality, delay propagation, and passenger comfort by corridor and modernisation state

Indicator	**Line 1 Sofia-Plovdiv-Svilengrad (magistrale, TEN-T)**	**Line 3 Sofia-Karlovo (TRIS CTC)**	**Line 8** Plovdiv-Burgas (legacy RRI/CTS)	**Line 8 Target (INTERFLO, 2030)**
Automation Level A	0.82	0.60	0.25	0.88
Communication Quality C	0.90	0.55	0.30	0.90
Safety Integration S	0.88	0.65	0.35	0.90
Predicted P - eq. (1)	0.84 + ε	0.60 + ε	0.29 + ε	0.88 + ε
Punctuality observed / target	78.4 %	~78%	61.7 %	≥85 %
Delay propagation δ - eq. (2)	0.64 (attenuates)	0.90	1.30 (amplifies)	≤0.65
Comfort Index K - eq. (3)	0.93	~0.79	0.70	≥90

Note: δ_0 = 1.5, k = 2.0; ε = 0.08 (modernised main line), 0.19 (Cat. I line), 0.44 (legacy); R and I values: Line 1 0.95/0.85, Line 3 0.82/0.70, Line 8 legacy 0.78/0.40, Line 8 target 0.95/0.85.

Three findings emerge from Table 2. First, the delay propagation differential (δ = 0.64 vs 1.30) confirms that a 10-min disruption attenuates to 6.4 min on modernised Line 1 but amplifies to 13 min on legacy Line 8 [43], [32]. Second, eq. (2) confirms automation gains require GSM-R support: automation alone yields $\delta \approx 0.99$ [31]. Third, the comfort gap (K = 0.93 vs 0.70) exceeds the raw punctuality gap by 7 pp, reflecting modernisation's multi-pathway effect on reliability and information provision [20].

5. CORRIDOR COMPARISON: SOFIA-PLOVDIV AND SOFIA-BURGAS

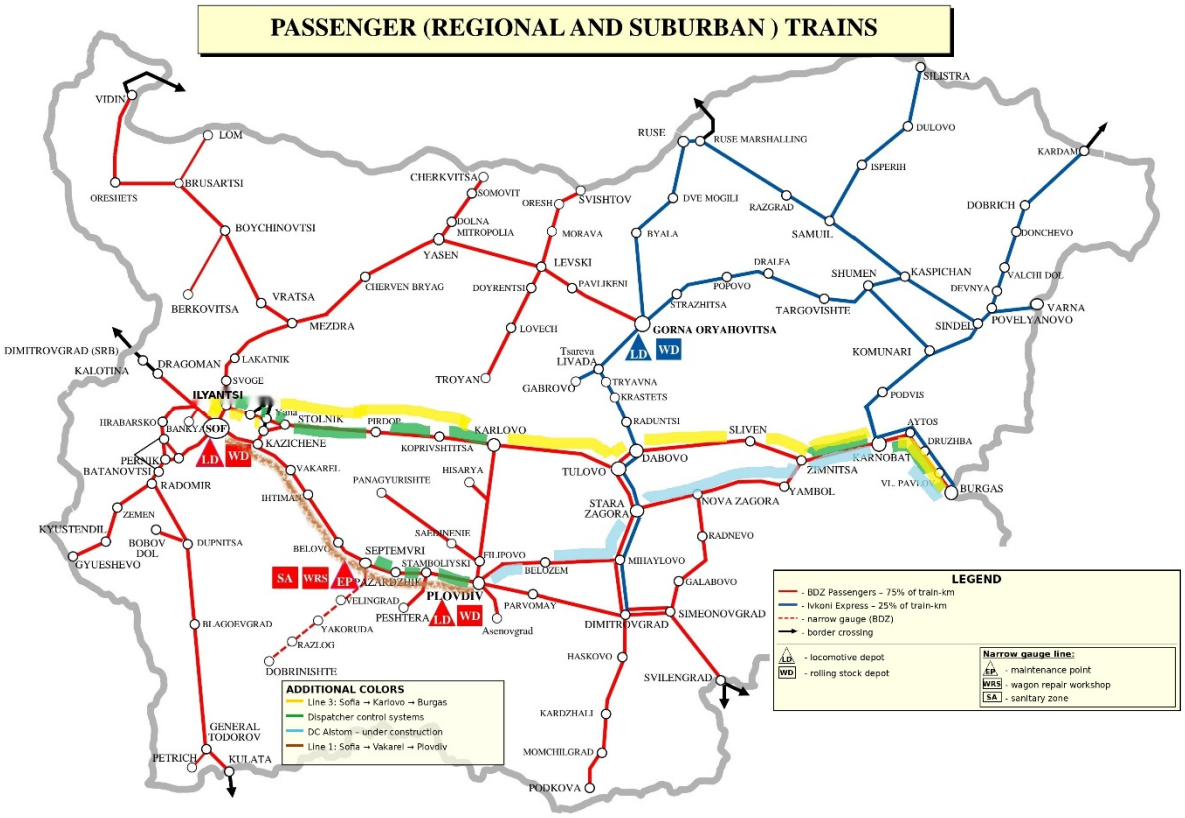


Figure 1: Passenger (regional and suburban) train network of Bulgaria: routes operated by BDŽ - Passenger Services (red) and Ivkoni Express (blue) under 12-year PSO contracts effective 13 December 2026. LD - locomotive depot; WD - rolling stock depot; EP - maintenance point; WRS - wagon repair workshop; SA - sanitary zone. Source: BDŽ-EAD network map, 2026.

5.1. Sofia-Plovdiv: Advanced Modernisation

The Sofia-Plovdiv corridor (153 km) forms part of the TEN-T Core Network and carries the highest passenger density on the Bulgarian railway network. The Septemvri-Svilengrad section is fully operational under the Elektra-2 CTC system, with 94% GSM-R coverage and ETCS Level 2 active. The Sofia-Septemvri subsection (incl. Elin Pelin-Kostenets) is under active reconstruction under the Transport Connectivity Programme; legacy DPS/ETK platforms are being progressively decommissioned. The corridor's mean delay of 8.2 min and punctuality rate of 78.4% (Table 1) reflect the best performance currently achieved on the Bulgarian network, constrained by the ongoing reconstruction phase.

The residual punctuality gap (vs the EU 80% benchmark) reflects factors beyond the dispatcher system: rolling stock reliability, dwell-time variability, and mixed passenger-freight traffic. The GERT framework [18] applies to quantifying these residual sources; the low delay propagation coefficient (Table 2) confirms that a 10-minute disruption accumulates only 14 minutes of downstream impact under CTC management.

The Elin Pelin-Kostenets section, part of the Sofia-Plovdiv corridor, is progressing as the largest railway project in Bulgaria in recent decades. Divided into three subsections, it is being implemented under the Transport Connectivity Programme with a planned completion of end-2029. Upon completion, ETCS Level 2 and GSM-R will be deployed throughout, enabling full CTC integration across the reconstructed corridor.

5.2. Sofia-Burgas: Partial Modernisation

The Sofia-Burgas corridor (402 km) is the longest and most operationally complex intercity axis in Bulgaria, connecting the capital with the Black Sea coast and carrying mixed long-distance passenger, regional, night-train, and seasonal services. CTC covers approximately 62% of the route; GSM-R is available on only 47% of the line; ETCS Level 1 is partially deployed; legacy DPS/ETK systems remain active on 53% of the corridor. The mean delay of 18.6 minutes and the punctuality rate of 61.7% reflect this technological fragmentation.

The corridor's performance is particularly sensitive to initial disruptions at the western end of the line, which propagate across multiple intermediate stations before reaching the coast. The delay propagation dynamics on legacy segments-where the absence of automated conflict detection means that a 10-minute initial disruption generates a 31-minute expected network accumulation within two propagation steps, compared to 14 minutes under CTC-explain why seasonal

peaks, when traffic density is highest, produce disproportionate deterioration in punctuality. The simulation modelling approach developed in [19] is directly applicable to this corridor and could provide the quantitative basis for prioritising specific modernisation investments.

Full modernisation-GSM-R, ETCS Level 2, SCADA-CTC integration-is projected, by analogy with comparable EU corridors, to raise punctuality by 20-25 percentage points, with corresponding gains in passenger comfort and competitive parity with the Sofia-Plovdiv axis. The delay baseline established by Karagyozov and Todorova provides the reference against which post-modernisation improvements should be assessed.

5.3. Competitive Implications

The delay propagation differential ($\delta = 0.64$ on Line 1 vs 1.30 on Sofia-Burgas, Table 2) translates into competitive asymmetry: a 10-minute disruption attenuates to 14 min on the modernised corridor but amplifies to 31 min on the legacy one. This 2.2-fold difference is a structural consequence of interlocking technology, not of traffic volume or operator quality: Line No. 1 (TEN-T magistrale) benefits from active EU cohesion funding. At the same time, the Sofia-Burgas corridor operates without an equivalent programme, imposing a punctuality ceiling that no operational excellence can overcome. The Logit choice model [21] confirms that this infrastructure gap translates directly into a passenger-attraction disadvantage for new entrants. At the same time, regional benchmarking by Stoilova [44] shows that Bulgarian performance deficits are most pronounced precisely on those dimensions linked to dispatcher system capability.

Charge design alone cannot resolve this asymmetry [27]: the fundamental requirement is infrastructure harmonisation on the Sofia-Burgas and Plovdiv-Burgas corridors, without which the Fourth Railway Package's non-discrimination framework cannot produce genuinely equal competitive conditions.

The 2026 PSO contracts sharpen this competitive asymmetry. Line No. 3 crosses all three lots: the TRIS CTC-managed Iliyantsi-Karlovo section (BDZh, Western Region) benefits from automated conflict detection and real-time tracking, while the continuation through the Southern and Northern Regions - operated by Ivkoni Express - lacks equivalent CTC coverage. Delays in legacy segments cannot be clearly attributed to either infrastructure failure or operator management, undermining the PSO contractual transparency that Regulation (EC) 1370/2007 requires [2]. Ivkoni Express thus bears the operational consequences of legacy infrastructure it neither owns nor controls. Accelerating the deployment of CTC and GSM-R on the Plovdiv-Burgas and Karnobat-Varna sections is a prerequisite for the equitable operation of the 2026 PSO contracts.

6. CONCLUSIONS

This paper has examined the current state of dispatcher systems in Bulgaria and their implications for the operational performance and liberalisation of the passenger rail market. The analysis, grounded in prior research on delay analysis, simulation modelling, passenger comfort, and intelligent management systems, leads to four principal conclusions.

First, the technological heterogeneity of Bulgarian dispatcher systems leads to measurable, significant operational disparities across corridors. Corridors equipped with modern CTC, GSM-R, and ETCS achieve substantially better punctuality and delay containment than legacy-equipped routes, confirming the findings of delay analyses and simulation studies conducted on the Bulgarian network.

Second, punctuality is the primary determinant of passenger comfort and modal choice in the Bulgarian context, as established by comfort-indicator research and modal-choice modelling. Dispatcher modernisation therefore generates direct commercial value: improved punctuality translates into higher passenger satisfaction, greater willingness to use rail, and stronger operator revenue potential.

Third, the transition to intelligent management systems is not simply a technological upgrade but a qualitative transformation of the dispatcher function. Modern systems reduce operational variance-and thus the unpredictability that passengers find most damaging-in ways that legacy platforms cannot replicate. Realising this potential requires not only hardware deployment but also the retraining and qualification of dispatcher personnel to operate new systems effectively.

Fourth, the modernisation of the dispatcher system is a strategic prerequisite for the successful liberalisation of Bulgaria's passenger rail market. The performance differential between the Sofia-Plovdiv and Sofia-Burgas corridors demonstrates that formal non-discrimination requirements cannot produce genuine competitive equality when infrastructure capabilities vary materially across routes. Completing the modernisation programme on the Sofia-Burgas and Plovdiv-Burgas corridors is therefore not merely a technical investment but an institutional prerequisite for a well-functioning open market.

Future research should validate these conclusions through post-modernisation performance monitoring as the NRIC upgrade programme progresses. It should explore the application of real-time management system architectures developed for marshalling yards and passenger complexes to the broader network-level dispatcher system design.

REFERENCES

[1] European Parliament and Council, "Directive 2016/2370/EU amending Directive 2012/34/EU as regards the opening of the market for domestic passenger transport services by rail", Official Journal of the European Union, L 352, (2016). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016L2370

[2] European Parliament and Council, "Regulation (EC) No 1370/2007 on public passenger transport services by rail and by road", Official Journal of the European Union, L 315, (2007), as amended by Regulation (EU) 2016/2338. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32007R1370

[3] Ministry of Transport and Communications, Republic of Bulgaria, "Public Service Obligation Contracts for Passenger Rail Transport 2026–2038", signed 10 February 2026, Sofia (Bulgaria), (2026). [Reported in: Transportal.bg, "MTS podpisa dogovorite za zhp usluga s BDZh i Ivkoni", 10 February 2026; BNT News, "Parvi chasten zhp prevozvach u nas", 2026; BTV Novinite, "BDZh veche ne e edinstveniyat zhp prevozvach", 2026.]

[4] Ministry of Transport and Communications, Republic of Bulgaria, "National Recovery and Resilience Plan – Transport Component: Rolling Stock Procurement Programme", Sofia (Bulgaria), (2023–2026). [Reported in: RailwayPro, "A private passenger transport operator enters the Bulgarian railway market", 2026; RailJournal, "Bulgaria awards first private passenger service contracts", 2026; RailMarket, "Bulgaria opens passenger rail markets", 2026.]

[5] Council of Ministers of the Republic of Bulgaria, "Ordinance on the Categorisation of Railway Lines in the Republic of Bulgaria Included in the Railway Infrastructure and the Closure of Individual Lines or Sections of Lines", Decree No. 293/2001, State Gazette No. 112, (2001, amended 2022). https://www.lex.bg/laws/ldoc/2135450102

[6] S. Todorov and Y. Paskalev, "Modern Trends and Innovations in the Modernisation of Railway Lines Based on Experience in the Republic of Bulgaria", Transport Problems, Vol. 19(3), (2024). https://doi.org/10.20858/tp.2023.19.3.18

[7] D. Dimitrov and I. Petrova, "Prouchvane i analiz na razvitieto na transportnata infrastruktura v Bulgaria i regiona", Proceedings of the International Scientific Conference Sustainable Development of Transport Systems, Todor Kableshkov University of Transport, Sofia (Bulgaria), (2018). https://sci.vtu.bg/proceedings/2018/

[8] SE National Railway Infrastructure Company (SE NRIC), "Network Statement 2024–2025", Version 01/11.01.2024, Valid from 15.12.2024 to 13.12.2025, Sofia (Bulgaria), (2024). https://www.rail-infra.bg/en/network-statement/

[9] M. Todorova and S. Filipov, "Otsenka na riska za vaznikvane na intsidenti po zhelyazopatnata infrastruktura", Mechanics, Transport, Communications, Vol. 20(3/1), art. ID 2318, (2022). https://mtc-aj.com/article.2318.htm

[10] D. Dimitrov, "Modelirane rabotata na vlakovia dispetcher pri upravlenie na dvizhenieto v zhelyazopatniya transport", Proceedings of the International Scientific Conference Transport 2002, Technical University of Sofia (Bulgaria), pp. 183–188, (2002). https://sci.vtu.bg/proceedings/2002/

[11] D. Dimitrov, "About the Need for Performing Training and Increasing the Qualification of Dispatchers and Movement Managers in Railway Transport", HiTech 2018, Sofia (Bulgaria), (2018). https://doi.org/10.31435/rsglobal_ijite/01062018/5677

[12] D. Dimitrov and K. Lalov , "Intelligent Rail Transport Management Systems – Methodological Essence, Goals and Tasks", Environment. Technologies. Resources. Proceedings of the International Scientific and Practical Conference, Vol. 2, pp. 83–88, Online ISSN: 2256-070X, Print ISSN: 1691-5402, (June 2024). https://doi.org/10.17770/etr2024vol2.8050

[13] D. Dimitrov, "Digitisation and Intelligent Management Systems in Transport: An Overview of Problems and Solutions", Environment. Technologies. Resources. Proceedings of the International Scientific and Practical Conference, Vol. 2, pp. 89–96, Online ISSN: 2256-070X, Print ISSN: 1691-5402, (June 2024). https://doi.org/10.17770/etr2024vol2.8048

[14] S. Stoilova and G. Pop-Andonov , "Study of Railway Passenger Transport in the European Union", Transport Means – Proceedings of the International Conference, Vol. 2018, pp. 615–622, 22nd International Scientific Conference Transport Means 2018, Kaunas (Lithuania), 3–5 October 2018. ISSN 1822-296X (print), ISSN 2351-7034 (online)

[15] S. Stoilova , N. Munier, M. Kendra and T. Skrúcaný , "Multi-Criteria Evaluation of Railway Network Performance in Countries of the TEN-T Orient–East Med Corridor", Sustainability, Vol. 12, Art. no. 1482, pp. 1–23, (2020). https://doi.org/10.3390/su12041482

[16] K. Karagyozov and M. Todorova, "Analysis of Delays of Freight and Passenger Trains on the Railway Network of Bulgaria", Scientific Forum on Innovations and Competitiveness in Transport, Sofia (Bulgaria), (2014). https://mtc-aj.com/article.962_EN.htm

[17] S. Stoilova, "Risk Analysis of Delay of Passenger Trains in the Rail Network", XVI International Scientific-expert Conference on Railways RAILCON'14, Niš (Serbia), pp. 267–270, ISBN 978-86-6055-060-8, (2014)

[18] M. Todorova, "GERT Model of a Technical Station for Determining the Passenger Trainset Delay", Problems of Transport, Vol. 14(2), (2019). https://doi.org/10.26552/pot.V14.2.2019

[19] K. Atanasov and M. Todorova, "Imitacionen model na zhelyazopaten uchastuk", Mechanics, Transport, Communications, Vol. 2020(4), art. ID 1931, (2020). https://mtc-aj.com/article.1931.htm

[20] M. Todorova and K. Trifonov, "Determining the Relative Importance of Indicators of Suburban Public Transport Comfort by the Method of Geometric Mean (Fuzzy AHP)", Environment. Technology. Resources, Vol. 4, pp. 413–417, (2025). https://doi.org/10.17770/etr2025vol4.8394

[21] M. Todorova, "Choice of Passenger Transport Mode Using Logit Model", EURO–ZEL 2015, Žilina (Slovakia), pp. 216–223, ISBN 978 80 263 0936 9, (2015). https://www.researchgate.net/publication/332557511

[22] M. Todorova, "Digitalizaciata v zhelyazopatniya transport i predizvikatelstvata pred rabotnata sila v sektora", Mechanics, Transport, Communications, Vol. 20(3/1), art. ID 2301, (2022). https://doi.org/10.26552/mtc.V20.3_1.2301

[23] D. Dimitrov, "Methodological and Technical Aspects of Models for Digital Management of Transport Systems", Environment. Technologies. Resources. Proceedings of the International Scientific and Practical Conference, Vol. 2, pp. 79–82, Online ISSN: 2256-070X, Print ISSN: 1691-5402, (June 2024). https://doi.org/10.17770/etr2024vol2.8049

[24] L. Dimitrov, S. Purgic, P. Tomov and M. Todorova, "Approach for Development of Real Time Marshalling Yard Management System", HiTech 2018, Sofia (Bulgaria), (2018). https://doi.org/10.1109/HiTech.2018.8566369

[25] P. Tomov, S. Purgic, L. Dimitrov and M. Todorova, "Optimisation of the Wagons Sorting Process for Needs of Real Time Marshalling Yard Management System", HiTech 2019, Sofia (Bulgaria), (2019). https://doi.org/10.1109/HiTech48507.2019.9128123

[26] M. Todorova, "Optimisation of Shunting Process at a Passenger Complex", Railcon '24, Niš (Serbia), (2024). https://scindeks-zbornici.ceon.rs/article.aspx?query=ISSID%26and%2665&page=9&sort=8&stype=0&lang=en

[27] M. Todorova, "Infrastructure Charges in the New EU Member States", EURO–ZEL 2015, Žilina (Slovakia), pp. 211–215, ISBN 978 80 263 0936 9, (2015). https://www.researchgate.net/publication/332574819

[28] S. Todorov, K. Kutsarova Dimitrova and Y. Paskalev, "Environmental Impact of the Modernisation of the Sofia–Plovdiv Railway Line", SGEM 2024, (2024). https://doi.org/10.5593/sgem2024/5.1/s23.95

[29] R.B. Dimitrov, "Evaluation of the Useful and Residual Life of a Class of Relay-Based Centralised Traffic Control", Mir Transporta, Vol. 19(5), (2022). https://doi.org/10.30932/1992-3252-2021-19-5-1

[30] I. Petrova, "Modern Risk Analysis in the Planning and Implementation of Modernisation Projects in Railway Infrastructure", Environment. Technology. Resources, Vol. 2, (2025). https://doi.org/10.17770/etr2025vol2.8178

[31] D. Stoyanova, "Digitalisation in Railway Transport as a Factor for Improving the Quality of the Offered Railway Service", Godishnik na UNSS, Vol. 2, (2022). https://doi.org/10.37075/yb.2022.2.12

[32] Luan, "Traffic Management Optimisation of Railway Networks", PhD Thesis, Delft University of Technology (Netherlands), (2019). https://doi.org/10.4233/UUID:4F6BF5D8-E75F-4627-AB0D-5E859F37A5FD

[33] European Commission, "Commission Implementing Regulation (EU) 2023/1694 on the Technical Specifications for Interoperability relating to the Control Command and Signalling Subsystems (TSI CCS)", Official Journal of the European Union, L 222, (2023). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202301694

[34] UNIFE – The European Rail Supply Industry Association, "ERTMS/ETCS Levels", Factsheet No. 2, 2024 Edition, Brussels (Belgium), (2024). https://www.ertms.net

[35] National Railway Infrastructure Company (NRIC), Operational Performance and Infrastructure Modernisation Report 2022–2024, Sofia (Bulgaria), (2024). https://www.rail-infra.bg/en/publications/

[36] SE National Railway Infrastructure Company (SE NRIC), "Pre-Project Studies for ERTMS Deployment – General Information", Administrative Contract No. DPTS-1/12.03.2024, Transport Connectivity Programme, Co-financed by the European Union Cohesion Fund, Sofia (Bulgaria), (2024). https://www.rail-infra.bg/

[37] S. Stoilova, "Analysis of Policies of Railway Operators Using SWOT Criteria and the SIMUS Method: A Case for the Bulgarian Railway Network", Symmetry, Vol. 13(5), p. 874, (2021). https://doi.org/10.3390/sym13050874

[38] X. Luan, F. Corman and L. Meng, "Non Discriminatory Train Dispatching in a Rail Transport Market with Multiple Competing and Collaborative Train Operating Companies", Transportation Research Part C, Vol. 80, pp. 1–22, (2017). https://doi.org/10.1016/j.trc.2017.04.011

[39] J. Högdahl and M. Bohlin, "A Combined Simulation Optimisation Approach for Robust Timetabling on Main Railway Lines", Transportation Science, Vol. 56(5), (2022). https://doi.org/10.1287/trsc.2022.1158

[40] D. Dimitrov and I. Petrova, "Strategic Planning and Development of Transport Infrastructures Based on Agile Methodology", IOP Materials Science and Engineering, Vol. 664, p. 012043, (2019). https://doi.org/10.1088/1757-899X/664/1/012043

[41] I. Petrova and K. Trifonov, "Izpolzvane na integriran podhod za upravlenie na transportni proekti", Proceedings of the International Scientific Conference, Todor Kableshkov University of Transport, Sofia (Bulgaria), (2025). https://sci.vtu.bg/proceedings/2025/

[42] V.A. Vasilev, "Sistema za upravlenie na zhelyazoptnata infrastruktura – tekhnologii i mekhanizmi" [Railway Infrastructure Management System – Technologies and Mechanisms], VTU "Todor Kableshkov" Publishing House, Sofia (Bulgaria), ISBN 978-954-12-0306-4, (2023).

[43] I.A. Hansen, "Improving Railway Punctuality by Automatic Piloting", IEEE ITSC, (2001). https://doi.org/10.1109/ITSC.2001.948761

[44] S. Stoilova, "A Multi-Criteria Assessment Approach for the Evaluation of Railway Transport in the Balkan Region", Mechanics, Transport, Communications, Vol. 17(3), art. ID 1818, (2019). https://doi.org/10.26552/mtc.V17.3.2019.a15

SESSION B: RAILWAY ENGINEERING

Study on the wear of cylindrical gears of transmission mechanisms

Vanyo Ralev[1*]

[1] Todor Kableshkov University of Transport, Faculty of Transport Management, Department of Machine Elements and Chemistry, Sofia, Bulgaria

ARTICLE INFO

[*] **Correspondence:** ralev.vanio5@gmail.com

DOI: https://doi.org/10.46793/ET26.B13R **PAGES**: B93 – B102

ABSTRACT

Gears are the main elements of transmission mechanisms such as reducers for general industrial applications, gearboxes of land transport vehicles and railway reduction gearboxes. They all are subjected to substantial dynamic loading and wear during operation. This paper analyses the materials used to produce various types of reducers. The main wear processes are examined in compliance to DIN 50320. The presented heat treatment technologies are aimed to achieve the best indicators determining the workability, contact strength and bending strength of materials. Measurements of gear profiles made of different materials and intended for a railway vehicle (electric locomotive) have been done. A formula for predicting the wear of gear profiles has been worked out in a probabilistic aspect. Optimised approaches to extend the service life of gears are suggested. It is also proposed what type of materials and technologies to choose for heat treatment of the selected reducer gears.

KEYWORDS

Reducer, Gears, Materials for gears, Gear profile, Gear wear, Electric locomotive

1. INTRODUCTION

Wear is a destructive action, with gradual removal or deformation of material on solid surfaces. The causes for wear can be mechanical or chemical. The study on wear and processes related to it is known as tribology. The examinations are focused on the types of wear phenomena observed in industries (such as abrasive wear, adhesive wear, frictional wear, fatigue wear, erosive wear and corrosive wear), their mechanisms, application of surface coating to protect the surface from industrial wear, types of surface coatings, thermal spray coating and its application in industry to protect the surface against wear [1, 2]. Detailed information on wear phenomena is presented in DIN 50320 [3] and ISO 10825 [4] that help industries to minimize the costs maintenance the parts. The mechanisms of gear wear can be generally classified into four categories [5]: abrasive wear, adhesive wear, erosive wear and corrosive wear

2. ESSENCE OF THE PROBLEM

The Bulgarian machine-building companies have long-year traditions in machine manufacturing and, more precisely, in reducer manufacturing. However, the Bulgarian reducers are not competitive under current economic conditions and with great competition with world-known companies, such as SEW, Lenze, Nord, Flender, Rossi, etc. Therefore, it is necessary to optimise certain parameters. The optimization criteria are determined based on a thorough analysis of advantages and disadvantages of competitive products made in the country and broad.

The reducer quality depends on the reliable operation of all its elements. Gears, being the main reducer element, are subjected to heavy loading, so it is necessary to ensure their load-bearing capacity and extend their life by developing new tooth shapes and, above all, choosing new materials and new manufacturing processes [6, 7]. For the last fifty years, the Bulgarian reducer manufacturing companies have used only certain grades of steel to produce gears, the most widely applied being: steel 45, 40Х, 18ХГТ, 30ХГТ, 12ХН3А, etc. [8]. These grades comply with the БДС5292 standard [9] and have the following classification:

- High-quality carbon steels, which have increased strength and ductility properties and are used for non-critical and medium-loaded gears for operation in non-corrosive environments – grade 40, 45, 50, etc.

- Structural alloy steels, which are used for operation in corrosive environments, for medium and heavy-loaded gears. The use of several alloying elements greatly improves the structural properties of steels – grades 40X, 40XH, 18XГT, 30XГT, 12XH3A, 40XM, etc.

The choice of material lies in the base of technological process development and heat treatment. It determines the durability and energy consumption of processes in manufacturing finished products. Therefore, it is necessary to analyze the factors determining these costs.

3. MATERIALS AND REQUIREMENTS FOR GEAR MANUFACTURING

The materials used for gear production are metallic and non-metallic. Metallic materials, in turn, are divided into ferrous and non-ferrous alloys [12]. Non-metallic materials, such as thermoplastics and duroplastics, are not applied in vehicle reducers. The application of forged steels is the most common in production of gears with surface or volume hardening.

Surface hardened steels (by chemical heat treatment):

- The carburizing or nitriding steels are used to obtain maximum *wear resistance* of contact surfaces.
- The cementation gears are suitable for heavy duty applications, e.g. in gearboxes, and have high resistance to wear, pitting, and fatigue. The cementation gear provides maximum surface hardness, wear resistance, and bulk impact toughness.
- The surface hardened gears (by chemical heat treatment - CHT) can withstand higher loads than the bulk hardened gears, although the latter are cheaper due to easier heat treatment.

The study presented in [10] has found that surface coating improves the segment life and reduces replacement costs. The motivation for surface innovation in the surface area is to provide practically attractive surfaces. The wide range of coatings can reduce the consumption, degradation and provide wear protection of materials. As for the main disadvantages of surface carburizing (cementation) as a hardening method, they are *the deformations of parts and their cost*.

Volume-hardened steels:

- Volume-hardened steels have higher core strength than cementation steels.
- Compared to surface hardened steels (by CHT), some volume-hardened steels reach comparable hardness values.

However, the volume-hardened parts such as gears *cannot transmit the same force load* as *cementation wheels.* With volume hardening*, the residual stresses are much more unpredictable*.

Volume-hardened steels with high hardness show *insufficient impact toughness*. To obtain compressive residual stresses of sufficient depth in the volume-hardened steel, it is necessary to implement local thermal hardening such as induction hardening or chemical heat treatment such as nitriding.

Steels must meet two groups of basic requirements: technological requirements in their production (machinability, deformability and thermal susceptibility) and operational requirements (fatigue strength, heat resistance and wear resistance) [8].

Standard ISO 6336-5 [11] defines the boundary material parameters (including requirements of quality control throughout all phases of production). This standard covers both surface and volume hardened gears and gives details about the requirements for the gears of three different quality classes.

The European standard EN 10084 [12] offers a wide variety of steel grades, both with maximum permissible sulfur content and with controlled sulphur content for easier and smoother machining.

The high-strength steels such as DIN 20NiCrMo2, DIN 17CrNiMo6 and DIN 15NiCrMo16-5 are preferred for manufacruring medium-sized and large-sized gears, while DIN 16MnCr5 steel is recommended for small-sized gears.

Having studied the western manufacturers of reducers, it was found that some of the main used materials are steels with the following designations: 20MnCr5, 31CrMoV9, 42CrMo4, etc.

The analysis in [7, 8] shows the newest materials used to manufacture gears – air-hardened steels Ovatec^{TM}277 and Ovatec^{TM} 677 (self-hardening, martensitic steels).

Table 1 presents the chemical composition of steel grades: 40X,45XH,12XH2 18XГT, 20MnCr5, 31CrMoV9, 42CrMo4,Ovate^{cT}M277andOvate^{cT}M 677.

Table 1: Chemical Composition of Steels Used to Manufacture Gears

Designation	Standard	Equivalent	C, %	Si, %	Mn, %	P, %	S, %	Cr, %	Mo, %	V, %
40X	GOST 4543	DIN 37Cr4/41	0,36-0,44	0,17-0,37	0,5-0,8	max 0,035	max 0,035	0,8-1,1	-	-
45XH	БДС6354-82	EN10083-	0,41 -	0,17 -	0,5 -	max	msx	0,45 -	-	-

		1/A1	0,49	0,37	0,8	0,035	0,035	0,75		
12XH2	БДС6354-82	EN10083-1/A1	0,09 - 0,16	0,17 - 0,37	0,3 - 0,6	max 0,035	max 0,035	0,6 - 0,9	-	-
18ХГТ	GOST 4543-71	DIN 20MnCr5	0,17-0,22	0,17-0,37	max 0,3	max 0,035	max 0,035	1-1,3	-	-
20MnCr5	EN10084	GOST 18ХГТ	0,17-0,22	max 0,4	1,1-1,4	max 0,025	max 0,035	1-1,3	-	-
31CrMoV 9	EN10085	GOST 30Н3МФ	0,27-0,34	max 0,4	0,4-0,7	max 0,025	max 0,035	2,3-2,7	0,15-0,25	0,1-0,2
42CrMo4	EN10083-3	GOST 40НГМ	0,38-0,45	max 0,4	0,6-0,9	max 0,025	max 0,035	0,9-1,2	0,15-0,3	
Ovatec^{TM} 277	EN16CrMnNiMo9-5-2F	-	0,155	0,15	1,3	0,02	0,015	2,2	0,5	0,2
Ovatec^{TM} 677	DIN 100Cr6	-	0,675	1,525	1,45	0,025	0,015	1,1	0,25	0,20

The presence of alloying elements improves the material features in the following way:

- Cr (X-chromium) – improves the mechanical properties, wear resistance, acid resistance and heat resistance of steel.
- Mo (M-molybdenum) – increases the strength properties at high temperatures.
- Mn (G-manganese) – if more than 1%, it increases the strength properties of steel. Being up to10-15% in austenitic steels, it improves impact toughness and wears resistance.

The main characteristics of specified steels are:

- Structural alloy steel 40X is used in heavy-duty components (shafts, axles, gears, pistons, rods, crankshafts and camshafts, rings, spindles, pins, slats, gears, bolts, half-axles, bushings, screws, etc.). Chromium steel 40X is also used in forged, cold and hot stamped components, and steel products are used to manufacture pipes, tanks, etc. Steel 40X is difficult to weld. The EU equivalents are 37Cr4, 37Cr4KD, 41Cr4, 41Cr4KD, 41CrS, complying with GOST 4543 standard [13].
- Steel 45XH is a structural alloy steel (substitute: 40XH) used to manufacture crankshafts, gears, couplings, bolts, etc. It complies with GOST 4543 standard.
- Steel 12XH2 is a structural alloy steel (substitutes: 20XHP, 20ХГНР, 12XH3A, 18ХГТ, 20ХГР) and is used to manufacture shafts, gears and other parts that require high strength, ductility and high surface hardness. It complies with GOST 4543 standard. According to the EU, 15NiCr6 is the equivalent of steel 12XH2.
- Steel 18ХГШ is applied to produce improved and cement able parts with a responsible purpose, which require increased strength and viscosity of the core, as well as high surface hardness, operating under the influence of shock loads. The steel is a high-quality structural chromium-manganese steel and meets the standards GOST 4543-71 and ТУ 14-1-3238-81.
- Steel grade 20MnCr5 is a structural alloy steel, subjected to surface hardening and meeting the requirements of standards ISO 683-3 / EN 10084 [12, 14].
- Grade 42CrMo4 is alloy steel, subjected to quenching and tempering. Rolled steel grade 42CrMo4 is popular in mechanical engineering, used to produce responsible high-strength parts of compressors, turbines, working elements of heavy ground and underground equipment, parts for agricultural machinery produced in compliance with the requirements of DIN EN ISO 683-2 standard [15] (previously - EN 10083-3 [16]). Steel grade 42CrMo4 has low weld ability due to its high sensitivity to cracking.
- The Ovatec^{TM} family of steels are designed to provide a strong, easy heat-treatable, with high performance alternatives to the low alloy steels, which nowadays are used for component manufacturing. There are two grades available: Ovatec^{TM}277 and Ovatec^{TM}677.

- Ovatec^{TM}277 is a steel that, after cementing , can be cooled directly from the cementation hardening temperature, avoiding quenching baths and reducing quench deformation and subsequent necessity of machining. Ovatec^{TM}277 is of standard quality for bearings but 277T is of the premium bearing grade. Ovatec^{TM}277 is a self-hardening (martensitic) steel used for applications with high requirements and is designed for surface hardening (by chemical heat treatment - CHT). It is hardened and easily tempered due to its composition and can replace large part of steels for various engineering applications. The new grade of steel Ovatec^{TM}277 [17] is produced according to an innovative metallurgical technology and is widely used by European companies. Ovatec^{TM}277 steel is used for responsible gears (gears for wind turbines) [7, 8].

- Ovatec^{TM}677 steel is intended for bearings and gears for powerful transmissions. Due to the reduced content of supplied inclusions and fine distribution of oxide inclusions, Ovatec^{TM}677 is steel with excellent fatigue strength (especially suitable for gears). By using gas or air hardening, it is possible to reduce deformations, hence: to reduce production costs and improve lab our safety and environment. With its high carbon content, OvatecTM ^{TM}677 enables manufacturing of parts using technologies that are highly competitive to those for production of surface hardened parts (by CHT)

and able to achieve comparable mechanical features (e.g. high hardness comparable to that of surface hardened steels – by CTT) and especially fatigue resistance.

When comparing the chemical composition, it is noted that some steels used according to standards EN, DIN and GOST correspond completely to each other (18 ХГТ and 20MnCr5), while others use identical alloying elements (chromium, manganese, molybdenum, etc.), with a minimal difference in percentage content. Particular attention could be paid to steels Ovatec^{TM}277 and Ovatec^{TM}677, which are self-hardening (martensitic) steels, used for high-load applications.

The chemical composition of steels determines their mechanical performance and the type of used heat treatments. The indicators, which are important for mechanical strength, are: boundary of yield strength (Re, MPa), tensile strength (Rm, MPa), and relative elongation (A, %). [7, 8].

Table 2 presents the parameters used forcomparison.

Table 2: Mechanical characteristics of steels

№	Designation	Re, MPa	Rm, MPa	A,%	Impact toughness, J/cm	HB
1	40Х	784	980	10	59	217
2	45ХН	640	785	13	59	248-293
3	12ХН2	590	780	12	69 - 88	229 - 249
4	18ХГТ	833	980	9	79	240
5	20MnCr5	690	min800	9	min20	217
6	31CrMoV9	800	1000-1200	10	25-40	248
7	42CrMo4	650	800-1100	12-13	63	241
8	Ovatec^{TM}277	850	1200	15		190
9	Ovatec^{TM}677	1450	1550	20	63	220

Ovatec™677 steel shows lower fatigue strength. Additional surface treatment is necessary to increase fatigue life in order to be comparable to cementation steels.

The data in Table 2 (without OvatecTM677 steel) makes possible to establish that there are no major differences in the mechanical strength of materials used. Their mechanical properties are close and it cannot be said that certain grades are much better or recommendable for gear manufacturing.

4. HEAT TREATMENT OF STEELS FOR GEARS

Heat treatment (number, type and duration) determines up to 30% of the cost of manufactured gears. Therefore, it is very important to analyze the differences in technologies, used equipment and peculiarities of heat treatment of different steel grades. [17, 19, 20, 21].

Steels with grades 1 ÷ 7 in Table 1 are heat treated at appropriate temperature and conditions in compliance to БДС EN 10825 [22] and pr БДС EN ISO 683-3:2022/prA1:2024 [14].

The heat treatment peculiarity of steel grade Ovatec^{TM}277 is that it is completely hardened and acquires a purely martensitic structure when cooled in air, which reduces deformations due to hardening compared to traditional surface-hardened steels. This allows saving funds from technological operations to remove deformations of hardening. The use of hardening environments such as oils and salts is avoided, thus improving safety and reducing environmental pollution. Carburizing time can be reduced compared to that of traditional surface-hardened steels, since steel Ovatec^{TM}277 has an increased carbon content in the base material.

Ovatec^{TM}677 can be hardened directly at hot deformation temperatures, which results in new achievements with cold machining, opening up the possibility to create completely new technologies for manufacturing different parts.

Comparing the heat treatment modes, the following important features can be noted:

- the operating temperatures in heating modes of different steel grades overlap and there are none with values that require special furnaces;
- the hardness of gear teeth working surfaces varies in the range of HRC 52-64. High hardness is a major factor for greater efficiency and contact strength of teeth and less wear;
- the steel grade 18ХГТ is distinguished by very good mechanical performance and hardness of HRC 55-63, but the heat treatment technology requires several cycles to complete the finished part (normalization, carburization, hardening, tempering). This determines a longer cycle of heat treatment and, accordingly, greater energy consumption of the process;
- the type of heat treatment is of great importance. Surface-hardened gears can withstand higher loads than volume-hardened ones, although the latter are cheaper due to easier heat treatment. The main disadvantages of surface carburization (carburization) as a hardening method are the part deformations and the cost. Deformations are controlled as much as possible during heat treatment and finally corrected by fine grinding. The higher product cost is justified from the view point of obtaining a high power-to-weight ratio and high durability;

- attention should be paid to the indicated advantages of Ovatec^{TM}277 and Ovatec^{TM}677.

5. MEASURING THE WEAR OF GEAR WHEELS IN TRANSMISSION MECHANISM OF ELECTRIC LOCOMOTIVES IN BULGARIA

The transmission mechanism of electric locomotives of series 43/44 and 42/45 has low operational reliability. That problem has been existing for a long time and is related to certain damage to electric locomotives and implementation of their repairs. The damage is mainly reduced to: damage in the cardan coupling, gears, reducer lubrication, noise, vibrations, etc.

Vibrations have an exceptional influence on the occurrence of damage in reducers [22, 24, 25]. They have been studied experimentally, and it has been established that the obtained spectra are close in configuration and can be presented in a generalized case. The amplitudes located in the frequency range of about 4500 Hz belong to the reducer housing. All main types of damage in gears lead to dynamic loading on gears and bearings, and to vibrations in the reduction gear-box, respectively, as 45% of rolling bearings are not repairable [26, 27].

The largest amount of damage was found in the parts related to: the reducer box, the square, the bearings of the driving pinion and the driven large gear wheel [28]. They are the main transmission mechanism components subjected to dynamic load and wear in operation.

The results obtained are useful for assessment of risk and consequences of failures both for improving the quality of repairs and for the operation of transmission mechanism of electric locomotives.

The gears of electric locomotive reducer go out of service mainly due to wear, damage to teeth and destruction of the gear ring [29].

To classify the characteristic damage that occurs in the pinions (driving) and in the large gear wheel (driven) of the reducers, it is necessary to study the defects in three series of electric locomotives. From the observations conducted so far, according to the requirements of БДС ISO 10825 - 43/44, 45, 46 and ZFN 201, the damage of reducer gears [28] have been classified by the following features: surface damage of teeth (wear and peeling); residual deformations (nicking); surface fatigue damage (pitting); tears and cracks (hardening cracks or fatigue cracks); tooth breakage (overload fracture or fatigue fracture).

The most common surface damage to pinions and the gear large wheel is the wear of their involute profiles. The amount of this type of damage to pinions is about 3 times greater than to large gear wheels, which confirms the theoretical hypothesis of wear.

It has been experimentally determined that the linear wear of involute tooth profile-ff of a worn pinion and a large gear

wheel it is twice as high as the permissible standard values. The maximum error of tooth direction has increased 7 times for a pinion and twice for a large gear wheel. The tooth profile has worn more intensively on one side due to the geometric deviations in rolling bearings, dynamic loading, etc. The grounds for such interpretations are given by the new methods used to calculate rolling bearings, as a result of which it is possible to increase gear efficiency [30].

Another surface damage of teeth, although being less common, is spelling of a pinion [29], which has become obsolete due to peeling on the tooth surface layer. It is caused by: heat stresses, insufficient depth of the hardened surface layer or low strength of the basic metal. Damage may also occur as a result of residual deformations such as chipping. It is characterized by depressions on teeth lateral surfaces, sometimes caused by an unknown body passing in the time of meshing.

One of the most common failures occurring as a result of surface fatigue of wheelset reducer gears is the appearance of pitting. The largest pits are located near the dividing circle, while the smaller ones appear on the surface immediately next to the tooth base. The reasons for progressive pitting are complex, such as errors in the tooth profile, surface roughness and peripheral speed, which are prerequisites for its appearance.

The tears and cracks in gears are due to improper thermal treatment, overstressing or defects in the material. During the manufacturing process of gear blanks by forging, slag or other non-metallic inclusions often appear. That defect is due to improper hardening or structural changes occurring in the core-hardened layer zone of alloyed steels.

Sometimes fatigue cracks also appear, especially in the wheeset reducer large gear wheel.

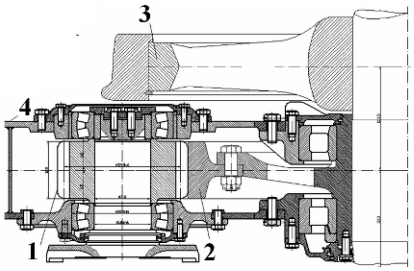
The fracture at the end of pinion and large gear wheel teeth of wheelset reducer is due to load concentration at one of the teeth ends, and sometimes it is caused by intensive pitting formation.

Besides the considered failures of reducer of electric locomotives, damage might also appear during operation as described in БДС ISO 10825 and in the company standard ZFN 201.

The change of tooth profile shape results in deterioration of reducer kinematics. In its turn, that leads to dynamic load increasing, losses due to friction and decrease of efficiency and durability of gears.

5.1. Parameters of Studied Gears

The object of study is a transmission mechanism gear of electric locomotives series 44 and 45, shown in Figure 1.



1 - pinion (driving); 2 - large gear wheel (driven); 3 - wheel axle; 4 – reducer box

Figure 1: The transmission mechanism gear of electric locomotives series 44 and 45

The main geometric parameters, material features and accuracy indicators of transmission mechanism gears of electric locomotives series 44 and 45 are given in Table 3.

Table 3: Parameters of the transmission mechanism gears of electric locomotives series 44 and 45

Parameters	Dimension	Gear for transmission ratio - u			
		u = 3,348		u = 3,591	
		$Z_1 = 23$	$Z_2 = 77$	$Z_1 = 20$	$Z_2 = 79$
Module (standard), m	mm	12	12	12	12
Execution module, m_v	mm	12,12	12,12	12,24242	12,24242
Profile angle, α	°	20	20	20	20
Output contour	-	Corrected +9,52	Corrected MAAG-3,3	Involute БДС3296-79	Involute БДС3296-79
Center distance, a_w	mm	606+0,05 47E14-14 Lo 84396	606+0,1 64E16-15 Lo 86116	606±0,05 47E14-2 Lo 83099/b	606 76E16-7 Lo 204256
Width of the tooth crown, b_w	mm	120	120	120	120
Diameter of the pitch circle, d_f	mm	265	887,4	224	933,72
Diameter of the dividing circle, d	mm	276	924	244,85	967,15
Diameter of the apex circle, da, d_a	mm	318,6	941,4	274,28	984
Output contour displacement coefficient, x	mm	+0,793	-0,275	+0,5	+0,5
Tolerance of tooth profile error, f_f	μm	16	34	16	34
Tolerance of tooth direction error, $F_β$	μm	21	40	21	40
Length of the common normal, W_m (number of teeth/length – tolerance)	mm	3/98,94 -0,08/-0,15	9/311,8 -0,15/-0,30	3/96,03 -0,09/-0,17	10/354,51 -0,14/-0,28
Tooth height, h	mm	26,8	27	25,14	25,14
Roundness at the base, R	mm	4	4	4	4
Accuracy level and type of meshing	-	6-6-6B БДС3296-79	7 – C БДС3296-79	6-6-6B БДС3296-79	7 – C БДС3296-79
Material	-	12ХН2 БДС6354-82	45ХН БДС6354-82	12ХН2 БДС6354-82	45ХН БДС6354-82
Tensile strength, Rm	MPa	784÷1079	802÷904	784÷1130	802÷904
Yield strength, Rs	MPa	min 539	min 657	min 538	min 657
Relative elongation, A	%	min 8	min 10	min 8	min 10
Minimum shrinkage	%	min 30	min 30	min 30	min 30
Impact toughness KCU2	J/sm^2	39	29	39	29
Surface layer hardness	HRC	56÷60	48÷52	56÷61	48÷52

5.2. Measured Parameters of the Transmission Mechanism Teeth Profile of Electric Locomotives Series 44 and 45

To assess the wear of a transmission mechanism teeth profile, several parameters are usually used, such as tooth thickness S_b, length of the common normal W_m and the hardened layer depth. The assessment of teeth profile wear is carried out by determining the corresponding tooth geometric profile, consisting of a part of involute and transition curve.

The tooth profile image construction is automatically done, using the obtained values of coordinates of the points on it, according to БДС1526-78, by the method described in [31]. The tooth profile geometric structuring and drawing is performed using splines of AutoCAD software. The profile geometry determination is based on software modeling on the principle of cutting teeth with a tooth comb using the centroid rolling method.

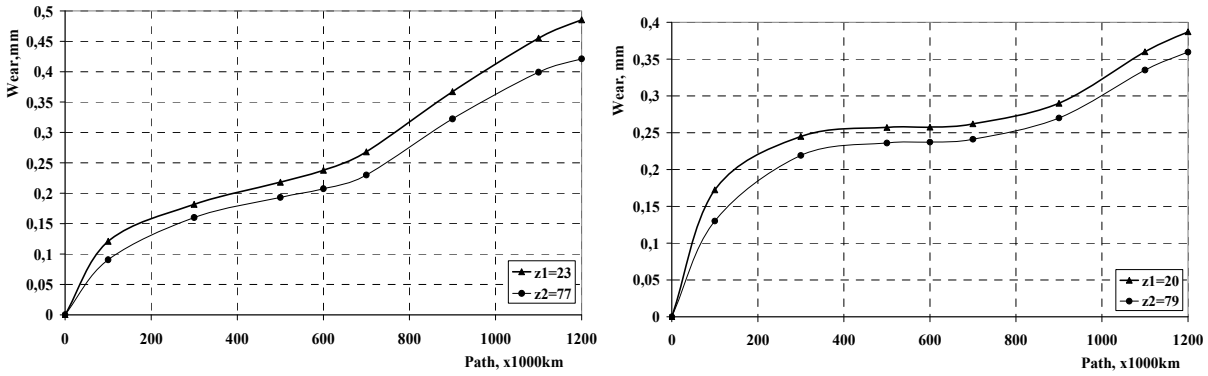
The measurements of gear wheel teeth profiles of reducers of electric locomotives series 44 and 45 were made in Sofia Locomotive Depot (Poduene Electric Depot) in the period 01.01.2024-30.06.2025. The results showed that the profile wear was more intense in the sections on the side of electric motor driving the pinion.

The methodology of obtaining a real operating gear tooth profile is based on measuring the thickness of teeth using a caliper with accuracy of 20-50 μm in 12-15 different sections along the teeth height. The deviation of thickness of the measured tooth can be determined measuring a new tooth by a caliper or calculating the thickness with a given section radius using the formula:

$$S_i = m.(d_i/d).(\pi/2 + 2\xi.tg\alpha_0) + d_i(inv\alpha_0 - inv\alpha_i) \quad (1)$$

where: m - module; d_i – current diameter, mm; d – diameter of the dividing circle; ξ - overlap coefficient; α_0 – angle of meshing; α_i – current angle of involutes.

Figure 2 shows the changeof tooth profile wear of reducer gear wheels of the studiedelectric locomotives series 44 and 45.



a) For electric locomotives series 44 b) For electric locomotives series 45

Figure 2: Change of the tooth profile wear of reduction gearbox of the studied electric locomotives

Figure 2 makes possible to draw the following conclusions:

1. The wear of the pinions and large gears of the wheel (LGW) system reducer, as a function of time and mileage, has a practically identical character with a clearly pronounced zone of initial wear during the combined operation of the wheels, a zone of low wear and a zone with a sharp increase in wear.

2. The process of slight wear of the wheel reducer of an electric locomotive series 44 begins at a mileage of 200-300 to 600 thousand km. At a mileage of more than 600 thousand km, wear increases sharply.

3. The nature of the wear of the reduction gears of an electric locomotive series 45 (Fig. 2b) allows us to conclude that the gear has a zone of very low wear at a mileage of 400-800 thousand km, after which it increases sharply.

The average 24-hour mileage of electric locomotives is 400 km and if the estimated work days in a month are 25, then the mileage per month is10,000 km. Vectors X (time in months or mileage in thousand km) and Y (wear in mm) appear as output data for the MathCad 2000 software product in which the function “submatrix” is used, to construct a polynomial regression at different degrees of the polynomial, used as a regression prediction, with the following data:

X:= data$^{<0>}$; Y := data$^{<1>}$; n:= rows(data); n:=5

Polynomial degree k:= 2

z:= regress(X, Y, k); fit:= interp(z, X, Y, x) ;

coeffs:= submatrix(z, 3, length(z) – 1,0,0)

Relative error:

$$R := \frac{\Sigma\left[\left(\text{fit}(X)-\text{mean}(Y)\right)^2\right]}{\Sigma\left[\left(Y-\text{mean}(Y)\right)^2\right]} \quad R = 0.996$$

$$i:=0\ldots n-1;\ j:=0..49;\ tx_i := \min(X) + j.\frac{\max(X)-\min(X)}{50}$$

➢ For electric locomotives series 44	➢ For electric locomotives series 45

$data := \begin{bmatrix} 100 & 0.121 \\ 300 & 0.181 \\ 600 & 0.238 \\ 900 & 0.367 \\ 1100 & 0.455 \end{bmatrix}$	$data := \begin{bmatrix} 100 & 0.172 \\ 300 & 0.244 \\ 600 & 0.258 \\ 900 & 0.290 \\ 1100 & 0.360 \end{bmatrix}$
$coeffs^T = [1.114\ 3.359 \cdot 10^{-4}\ 2.897.10^{-5}]$	$coeffs^T = [-0.044\ 5.905 \cdot 10^{-3}\ -1.913 \cdot 10^{-5}]$

The processing of statistical data on wear of gears of electric locomotives series 44 and 45 makes possible to determine the parameters of wear equation:

$$\delta = (\delta_1 + h).10^{\,t/12xA} - h \tag{2}$$

where: δ - wear, mm; δ_1 - mean wear (mileage) t; h - displacement coefficient; A - durability coefficient; t - current mileage. The parameters of equation (2) for electric locomotives series 44 and 45 are given in Table 4.

Table 4: Parameters of the equation for wearing the gears of electric locomotives series 44 and 45

Electric locomotives	Parameters of pinions (z_1 = 23/20)			Parameters of large gear wheels (z_2 = 77/79)		
	δ_1, mm	h	A	δ_1, mm	h	A
Series 44	0.72	0.3	20	0.65	-0.14	6.9
Series 45	0.008	-0.35	-10.9	0.112	0.11	18.742

Formulas (1) and (2) are used to predict the wear of tooth profile of the respective wheel with desired accuracy of results obtained.

The wear equation of pinions (z_1 = 23) of an electric locomotive series 44 is:

$$\delta = (0.72 + 0.3).10^{\,t/12.x20} - 0.3 \tag{3}$$

The wear equation of large gear wheels (z_2 = 77) of an electric locomotive series 44 is:

$$\delta = (0.65 - 0.14).10^{\,t/12x6.9} + 0.14 \tag{4}$$

The wear equation of pinions (z_1 = 20) of an electric locomotive series 45 is:

$$\delta = (0.008 - 0.35).10^{\,t/12x(-10.9)} + 0.35 \tag{5}$$

The wear equation of large gear wheels (z_2 = 79) of an electric locomotive series 45 is:

$$\delta = (0.112 + 0.11).10^{\,t/12x18.742} - 0.11 \tag{6}$$

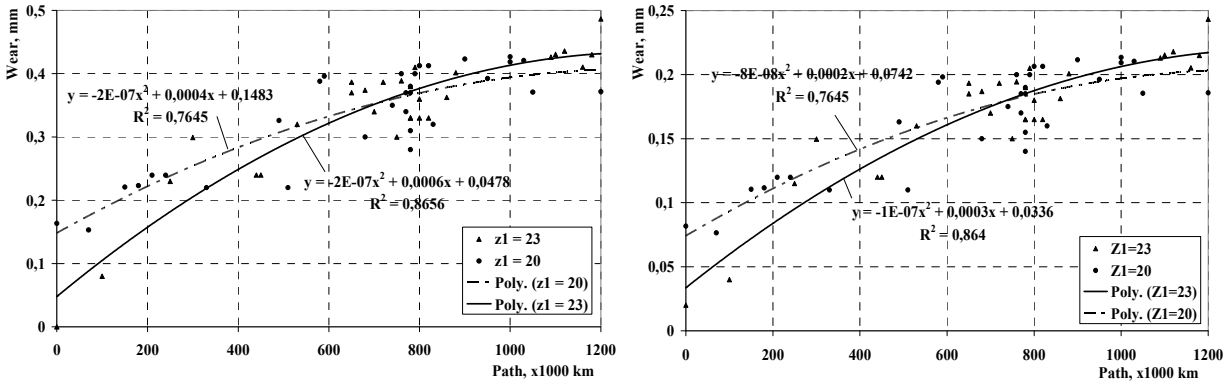
Figure 3 and 4 shows wearchangesof the reducer gear tooth profile of studied electric locomotives withusing steels 12ХН2 for pinions according toБДС6354-82 and 18ХГТ according to GOST 4543-71, and steel grade45ХН for LGW according toБДС6354-82. The dependences ofgear tooth profile wear as a function of the locomotive mileage are as follows:

- for material steel grade 12ХН2 according to БДС6354-82 (pinions) and 45ХН according to БДС6354-82 (LGW)

$$\delta = -2E-07\,p^2 + 0.0004p + 0.1483 \text{ за } z_1 = 20;\ \delta = -2E-07\,p^2 + 0.0006p + 0.0478 \text{ за } z_1 = 23 \tag{7}$$

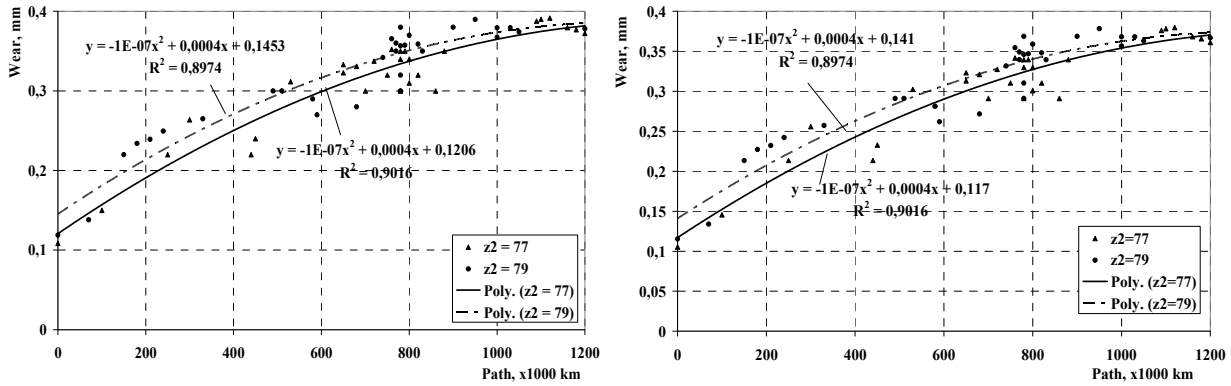
$$\delta = -1E-07\,p^2 + 0.0004p + 0.1206 \text{ за } z_2 = 77;\ \delta = -1E-07\,p^2 + 0.0004p + 0.1453 \text{ за } z_2 = 79 \tag{8}$$

where: δ - wear, mm and p – locomotive mileage, thousand km.



a) material12ХН2 for pinions. b) material 18ХГТ for pinions.

Figure 3: Change of the wear of pinions at a certain locomotive mileage



a) material 45XH for LGW and 12XH2 for pinions b) material 45XH for LGW and 18ХГТ for pinions

Figure 4: Change of the wear of LGW at a certain locomotive mileage

- for material steel grade 18ХГТ according to GOST 4543-71 (pinions) and 45XH according to БДС6354-82 (LGW)

$$\delta = -8E-08\,p^2 + 0.0002\,p + 0.0742 \;\; за\; z_1 = 20;\;\; \delta = -1E-07\,p^2 + 0.0003\,p + 0.0336 \;\; за\; z_1 = 23 \qquad (9)$$

$$\delta = -1E-07\,p^2 + 0.0004\,p + 0.117 \;\; за\; z_2 = 77;\;\; \delta = -1E-07\,p^2 + 0.0004\,p + 0.141 \;\; за\; z_2 = 79 \qquad (10)$$

From the obtained dependences (7) ÷ (10) of measured wear of gears, it has been established that the wear of pinions and LGW manufactured from steel grades 18ХГТ according to GOST 4543-71 (pinions) and 45XH according to БДС6354-82 (LGW), respectively, is two times less than the wear of wheels manufactured of steel grades 12XH2 according to БДС6354-82 (pinions) and 45XH according to БДС6354-82 (LGW). That gives grounds to use 18XHT steel for pinions. The wear of LGW from material 45XH, associated with pinions made of 18ХГТ steel, decreases by 5%. To increase the reliable operation of gears, in addition to the choice of material and heat treatment, it should be paid attention also to the choice of lubricant [32, 33].

6. CONCLUSION AND RECOMMENDATIONS

The analysis of chemical composition and mechanical features of steels, as well as of technologies for gear processing applied in various types of reducers, has shown that the Bulgarian mechanical engineering companies should keep the following guidelines for manufacturing gears:

Option 1 – to use traditional materials, as for this purpose should change technologies and part of equipment (through modernization) in order to reduce the energy consumption for heat treatments and mechanical processing;

Option 2 – to master the application of new materials and technologies.

It is a good solution to use steel grade18ХГТ as main material for gears for general machine-building reducers due to its proven qualities and mastered technologies of mechanical processing and heat treatment. It is necessary to invest in modernization of machines and installations to reduce the production costs and make products more competitive on the market. From measuring the transmission mechanism gear profile of electric locomotives operated in Bulgaria, it was established that the gears of wheelset reducers go out of use mainly due to wear, teeth damage and gear ring destruction.

The analytical dependencies for wear prediction have been obtained as a function of mileage of the respective locomotive. The derived formulas (7) ÷ (10) are recommended to apply for predicting the wear of tooth profile in a probabilistic aspect. To increase the reliability of traction reducers of electric locomotives, the following recommendations for design and production can be given:

- To develop new materials, starting from new self-hardening (martensitic) and air-hardening steels, advanced sintered steels and Nan powders.
- To develop new processes that combine advanced achievements in the field of materials, tools (as design and material), automation, thermal and mechanical processing, innovative additional processing and application of various powder metallurgy methods.

REFERENCES

[1] V. Miltenović, M. Ognjanović, “Machine elements II: power transmission elements" (in Serbian), Faculty of Mechanical Engineering – University of Niš, Niš (Serbia), (1995)

[2] В. Милтеновић, “Машински елементи – облици, прорачун примена”, Универзитет у Нишу, Машински факултет, Ниш, (2009).

[3] DIN 50320:1979-12 “Wear; Terms, Systematic Analysis of Wear Processes, Classification of Wear Phenomena”, (1979)

[4] ISO/TR 10825-2:2022 “Gears - Wear and damage to gear teeth - Part 2: Supplementary information”, (2022)

[5] М. Огнянович, “Вероятност за повреда от износване на зъбите на зъбни колела”, Сборник от 16-та Европейска конференция по счупване, Александрополис, Гърция, резюме стр. 1059-1060, статия CD-294, (2006)

[6] М. Огнянович, А. Субич, “Система за вземане на решения за проектиране на по-надеждни зъбни задвижвания”, Сборник от 3-та Международна конференция по качество и надеждност – ICQR 2002, - Университет RMIT, Мелбърн, Австралия, (2002)

[7] Ю. Димитров, В. Добрев, “Материали за изработване на зъбни колела на редуктори от общото машиностроене”, НАУЧНИ ТРУДОВЕ НА РУСЕНСКИЯ УНИВЕРСИТЕТ - 2013, том 52, серия 4, Издателски център при Русенски университет, Русе, стр. 141-147, ISSN 1311-3321 (2013),

[8] “Разработване на зъбни предавки на базата на нови материали и нови системи зъбни колела” (COLL-CT-договор 030433), Русенски университет, Русе, http://www.mech-ing.com › x-gear › Air-hardening-steel-BG, (2006)

[9] БДС5292:1986/Изменение 1:1989 “Масла смазочни AN. Технически изисквания”, (1989)

[10] B. Swain, S. Bhuyan, R. Behera, S Sanjeeb Mohapatra and A. Behera, “Tribology in Materials and Manufacturing”, https://doi.org/10.5772/intechopen.94211, (2020)

[11] ISO 6336:2019, “Calculation of load capacity of spur and helical gears”, (2019)

[12] БДС EN 10084:2000 “Case hardening steels - Technical delivery conditions”, (2000)

[13] GOST 4543 “Alloy structural steel rolled products. Technical conditions”, (1971)

[14] prБДС EN ISO 683-3:2022/prA1:2024, “Heat-treatable steels, alloy steels and free-cutting steels — Part 3: Case-hardening steels”, (2024)

[15] БДС EN ISO 683:2018 “Heat-treatable steels, alloy steels and free-cutting steels”, (2018)

[16] БДС EN 10083:2006, “Steels for quenching and tempering”, (2006)

[17] http://www.ovako.com/Global/Downloads/Product_information/Ovako_Group/EN/I Q%20steel.pdf, (2026)

[18] Ю.М. Лахтин “Термическая обработка в машиностроении, Машиностроение”, Москва, (1980).

[19] S. Kuzmanović, M. Vereš, M. Rackov, “Analysis of Characteristical Mounting Ways of Gear Wheels in Universal Gear Reducers, Budowa Maszyn i Zarzadzanie Produkcja 11 – 2009”, Zeszyty Naukowe Politechniki Poznanskiej, Wydawnictwo Politechniki Poznanskiej, Poznan, Poland, ISSN 1733-1919; стр. 81-91, (2009)

[20] S. Kuzmanović, M. Rackov, “Development Tendencies of Universal Gear Reducers", Proceedings of the 3rd International Conference Power Transmissions ‘09, Kallithea, Greece, 1-2. October 2009, ISBN 978-960-243-662-2 pp.145-148, (2009)

[21] Đ. Miltenović, “Research of thermal stability and wear of worm gear” (in Serbian), PhD Thesis, University of Banja Luka (SR BiH), (2017)

[22] EN ISO 10825 Heat-treatable steels, alloy steels and free-cutting steels - Part 2: Alloy steels for quenching and tempering, (2006)

[23] С. Кузмановић, “Универзални зупчасти редуктори са цилиндричним зупчаницима”, Универзитет у Новом Саду, Факултет техничких наука, Нови Сад, (2009).

[24] S. Ciric Kostic, M. Ognjanovic: “Excitation of Modal Vibration in Gear Housing Walls”, FME Transactions, Vol. 34(1), pp. 21-28, (2006).

[25] S. Ciric Kostic, M. Ognjanovic: “Gear Disturbance Energy Transmission through the Gear System and Frequency Spectrum”, Proceedings of the International Conference Power Transmission, Novi Sad-Serbia, pp. 167-172, (2006).

[26] M. Ognjanovic, S. Ciric Kostic: “Effects of Gear Housing Modal Behaviour at the Noise Emission”, Proceedings of the International Conference on Gears, Munich VDI-Berichte 1904-02, pp. 1767- 1772, (2005).

[27] S. Ciric Kostic, M. Ognjanovic: “The Noise Structure of Gear Transmission Units and the Role of Gearbox Walls”, FME Transactions, Vol. 35(2), pp. 105-112, (2007)

[28] М. Михалев, “Изследване на зъбния профил от предавателен механизъм на локомотиви”, Научно списание Механика Транспорт Комуникации ВТУ „Т. Каблешков”, София, БЪЛГАРИЯ том 11, брой 3, 2013 г. статия № 0886 ISSN 1312-3823, (2013)

[29] Г. Динев, “Повреди в зъбните колела от колоосни редуктори на електрически локомотиви”, Софттрейд, (2004)

[30] S. Hauri, “Gear Wear Calculation and Prediction”, https://www.powertransmission.com/gear-wear-calculation-and-prediction”, (2015)

[31] J. Pedrero, M. Pleguezuelos, J. Antona, “Load distribution model along the line of contact for involute external gears”, Mechanism and Machine Theory 45 pp.780–794, (2010)

[32] M. Vasić, M. Blagojević, M. Dragoi, “Thermal stability of lubricants in cycloidal reducers”, Engineering Today, Vol. 1(2), https://doi.org/10.5937/engtoday2202007V, UDC: 621(497.11) ISSN: 2812-9474, (2022)

[33] D. Petrovic, V. Aleksandrov, “Zeleznicka vozila. Osnove”, MF Kraljevo, ISBN 978-86-82631-66-8 p.223, (2013)

SESSION C

PRODUCTION TECHNOLOGIES, CAE AND MANAGEMENT

SESSION C: PRODUCTION TECHNOLOGIES, CAE AND MANAGEMENT

From Industry 4.0 to Industry 5.0: The transition from digitalization and cybernation to cognitization and human-machine synergy

Emine Hozdić[1,2], Elvis Hozdić[2]

[1] Kranj School Centre, Kidričeva cesta 55, Kranj, Slovenia

[2] University of Novo mesto, Faculty of Mechanical Engineering, Novo mesto, Slovenia

ARTICLE INFO

* **Correspondence:** elvis.hozdic@fs-unm.si

DOI: https://doi.org/10.46793/ET26.C01H **PAGES:** C1 – C12

ABSTRACT

The transition from Industry 4.0 to Industry 5.0 represents a fundamental shift in the evolution of manufacturing systems, moving beyond digital connectivity and automation toward cognitive, human-centric production environments. This paper examines the conceptual and structural progression from digitalization and cybernation to cognitization, highlighting how these phases redefine the interaction between technological systems and human actors.

Digitalization establishes the informational foundation of modern manufacturing through data acquisition, connectivity, and transparency. Cybernation builds upon this foundation by integrating physical and digital systems into cyber-physical production systems (CPPS), enabling adaptive, autonomous, and data-driven control. The emerging phase of cognitization extends these capabilities by incorporating learning, semantic interpretation, and context-aware reasoning into production systems, giving rise to cognitive cyber-physical production systems (C-CPPS).

Particular attention is devoted to the role of human–machine synergy, where human cognition, ethical reasoning, and contextual understanding complement machine intelligence. This integration leads to the development of Human–Cyber–Physical Systems (HCPS), in which decision-making is distributed across human and artificial agents. The paper provides a structured interpretation of this three-phase evolution and discusses its implications for the design of resilient, sustainable, and human-centric manufacturing systems.

The findings suggest that the future of industrial development will depend not only on technological advancement, but on the ability to harmonize autonomy, cognition, and human values within integrated industrial ecosystems, as envisioned in the Industry 5.0 paradigm.

KEYWORDS

Industry 4.0, Industry 5.0, digitalization, cybernation, cognitization, artificial intelligence, cognitive systems, human–machine integration, smart manufacturing, sustainability.

1. INTRODUCTION

Over the past fifteen years, manufacturing has undergone a profound transition from early digital experimentation toward increasingly autonomous, cognitive, and human-centric production systems. The initial vision of Industry 4.0 – publicly articulated around 2011 as part of Germany's high-tech strategy and subsequently adopted internationally – marked a decisive departure from isolated automation toward interconnected, data-rich industrial ecosystems built on cyber-physical systems (CPS), the Internet of Things (IoT), big-data analytics, and distributed intelligence [1-6].

The first stage of this transition, commonly referred to as digitalization, was characterized by the widespread deployment of sensing technologies, networked equipment, and digital data capture. These developments introduced real-time transparency into manufacturing environments and enabled the earliest forms of data-driven decision support. Folgado et al. [7] note that, during this period, the traditional automation pyramid began to dissolve into more flexible and distributed IIoT-oriented architectures supported by digital twins and interoperable data models.

As digital infrastructures matured, a second stage—cybernation—emerged. Within this phase, cyber-physical production systems (CPPS) [8] integrated physical processes with digital representations through bidirectional communication, adaptive control, and self-optimization. Beyond technical change, cybernation also initiated a broader reconfiguration of engineering competencies and industrial policy. As Oks et al. [9] emphasise, CPS architectures increasingly linked hardware, software, and human interfaces, thereby requiring new cognitive, cross-disciplinary and ethical capabilities in the industrial workforce.

Conceptually distinguishing digitalization and cybernation is essential for understanding the next stage of industrial evolution. As previously articulated [10], digitalization denotes the execution of operations through digital mechanisms and information flows, whereas cybernation integrates physical and digital entities into hybrid, self-regulating loops. This differentiation provides the theoretical foundation for the emerging third stage, cognitization – the embedding of learning, semantic reasoning, and context-sensitive intelligence into production systems.

Driven by workforce shortages, supply-chain volatility, sustainability imperatives and new ethical expectations, manufacturing now faces a paradigm in which production systems must not only act autonomously but also interpret their environment. In this cognitization phase, production architectures adopt cognitive capabilities such as perception, representation, learning and explainability, and humans become indispensable cognitive collaborators. Recent work on Human–Cyber–Physical Systems (HCPS) and human-machine teaming frameworks illustrates a shift from viewing humans as supervisors to recognizing them as integral participants in distributed decision processes [11,12,13].

Despite the growing body of literature on Industry 4.0 and the emerging discourse on Industry 5.0, existing research often treats digitalization, cyber-physical integration, and cognitive capabilities as separate or loosely connected developments. A systematic interpretation that integrates these stages into a coherent evolutionary framework remains limited. In particular, the transition mechanisms linking CPPS to cognitive and human-centric architectures are insufficiently formalized, and the role of human–machine synergy is frequently discussed conceptually but rarely embedded into a structured system-level model.

Against this background, the purpose of this invited lecture article is to trace the structural evolution of manufacturing system architectures across these three phases-digitalization, cybernation, and cognitization-and to illuminate the growing human–machine synergy that underpins the emerging paradigm of Industry 5.0. The central questions we address include: how have technological enablers, system structures and human roles co-evolved? What organizational transformations have accompanied these shifts? And how does the Industry 5.0 vision reshape the relationship between machine intelligence and human cognition?

Methodologically, the article presents an analytical and conceptual review of the evolution of Industry 4.0 between 2011 and 2025. Rather than focusing on empirical validation of individual technologies, the review synthesizes conceptual patterns, architectural models and structural trends in the literature. Sources include peer-reviewed publications, strategic documents of the European Commission, and the author's prior research on CPPS [8,11], cognitive manufacturing [14] and HCPS [13]. References to particular industrial sectors – such as automotive, electronics or medical technology – serve merely as illustrative examples that reflect broader tendencies rather than formal case studies.

The novelty of this contribution lies in offering a unified, three-phase interpretation of industrial evolution and introducing C-CPPS and HCPS as coherent cognitive architectures that bridge Industry 4.0 technologies with the human-centric principles articulated in the European Commission's Industry 5.0 communication [12]. By integrating historical, technological, and human-centric perspectives, the article establishes a conceptual foundation for the sections that follow, which examine: (i) the transition from digitalization to cybernation, (ii) the rise of cognitization and cognitive CPPS, and (iii) the emerging human–machine synergy at the core of Industry 5.0.

The main contributions of this paper can be summarized as follows:

- A unified three-phase model of industrial evolution (digitalization–cybernation–cognitization);
- A structured conceptual linkage between CPPS, C-CPPS and HCPS architectures;
- A system-level interpretation of human–machine synergy within Industry 5.0;
- A conceptual bridge between technological autonomy and human-centric industrial design principles.

Taken together, these elements position the proposed framework as a conceptual bridge between Industry 4.0 technologies and the emerging human-centric paradigm of Industry 5.0.

2. FROM DIGITALIZATION TO CYBERNATION: THE CORE OF INDUSTRY 4.0

2.1. Digitalization as the Foundational Phase

Digitalization constitutes the foundational phase of Industry 4.0, marking the moment when physical operations, machines and workflows became translatable into continuous digital data streams. Within Smart Manufacturing Systems (SMS) [16], this transformation enabled real-time acquisition and analysis of information originating from sensors, networked machines and human–machine interfaces. As early conceptualisations of Industry 4.0 made clear [6], digitalization is not merely the introduction of digital tools, but a structural reconfiguration of value creation based on data visibility, transparency and ubiquitous connectivity. Recent reviews, such as Folgado et al. [7], reinforce this view by demonstrating

how the traditional automation pyramid has evolved into distributed IIoT architectures supported by digital twins and interoperable data models.

At a functional level, digitalization replaces analogue signals with machine-readable data processed through software mechanisms, simulation tools and analytics platforms [17]. Three interrelated domains shape this transformation: (i) data generation via sensors and IoT networks, (ii) data transmission and storage through cloud and edge infrastructures, and (iii) data utilization through analytics, optimisation algorithms, and virtual replicas such as digital twins.

Figure 1a illustrates this digital-only configuration: all inputs, mechanisms and outputs operate within a closed digital domain characterized by unidirectional data flows (red arrows). Digitized functions are fully automated, relying on digital mechanisms embedded in software - ranging from classic algorithmic solutions to advanced AI components-that ensure the execution of computational tasks. Digital controls, expressed through rules, constraints and instructions, govern the flow of these digitized functions. Examples in manufacturing include automated material ordering, production planning, supply-chain purchasing, accounting and financial transactions [11,17].

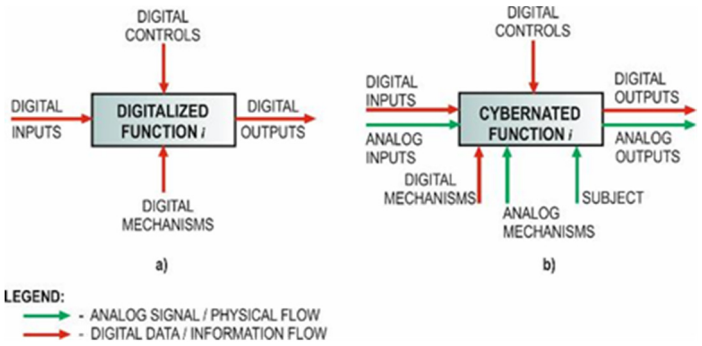


Figure 1: Digitalized and cybernated functions: a) digitalized function i – execution of digital-only mechanisms and informational flows, b) cybernated function i – integration of digital and analogue domains through feedback between digital control and physical processes [11, 17].

Digitalization thus enables the transformation of a wide range of information-centric processes in manufacturing. However, many production activities cannot be fully digitized because they involve material-transformation processes that are analogue by nature-such as machining, assembly, logistics and product testing. These processes require simultaneous management of analogue material flows and digital information flows associated with computer-based monitoring and control. Their execution depends on both physical elements (machines, actuators, sensors, operators) and computational elements (controllers, processors, databases, control programs) [11].

Consequently, today's production systems operate as hybrid analogue-digital environments in which physical and digital elements are tightly interconnected. The digital component has become increasingly advanced, enabling sophisticated monitoring, optimisation and supervisory functions across social, digital and physical domains. This hybridization establishes the structural precondition for cybernation, where digital information begins to influence, regulate and adapt physical processes in real time.

Digitalization has therefore laid the groundwork for unprecedented transparency in manufacturing systems. IoT platforms and big-data infrastructures now allow companies to observe energy consumption, machine performance, tool states and material flows at granular temporal and spatial scales. Yet transparency alone does not constitute intelligence. Digitalization provides the informational substrate, but it is cybernation that operationalises intelligence by linking digital information to physical action through bidirectional feedback loops [11,17].

2.2. Cybernation as the Integration of Intelligence and Connectivity

Cybernation marks the point at which digitalized systems evolve into intelligent, self-regulating entities. In this phase, physical processes – machines, tools and sensors – are tightly coupled with digital control architectures through feedback loops that enable adaptive, real-time decision-making. This configuration defines the essence of Cyber-Physical Production Systems (CPPS), which combine sensing, reasoning and autonomous adjustment within a unified framework.

In practical terms, CPPS architectures emerge from the convergence of IoT-enabled devices, edge computing, AI-enhanced control agents and digital twin representations. As illustrated in Figure 1b, the cybernated function extends the digitalized model by incorporating both digital and analogue channels: red arrows denote digital information flows, while green arrows represent physical or analogue signals. This dual-flow structure allows hybrid coordination between software mechanisms, physical actions and human intervention (Subject), enabling systems to respond adaptively to real-world conditions.

The general CPPS architecture (see Figure 2), following the model proposed by Hozdić et al. [17], organizes this interaction into three interconnected domains: the Physical Work System (PWS), the Cyber System (CS) and the Human Subject (HS). Communication interfaces link these domains, ensuring bidirectional exchange of information, commands and contextual feedback. This architecture represents the first structural realisation of cybernation within manufacturing.

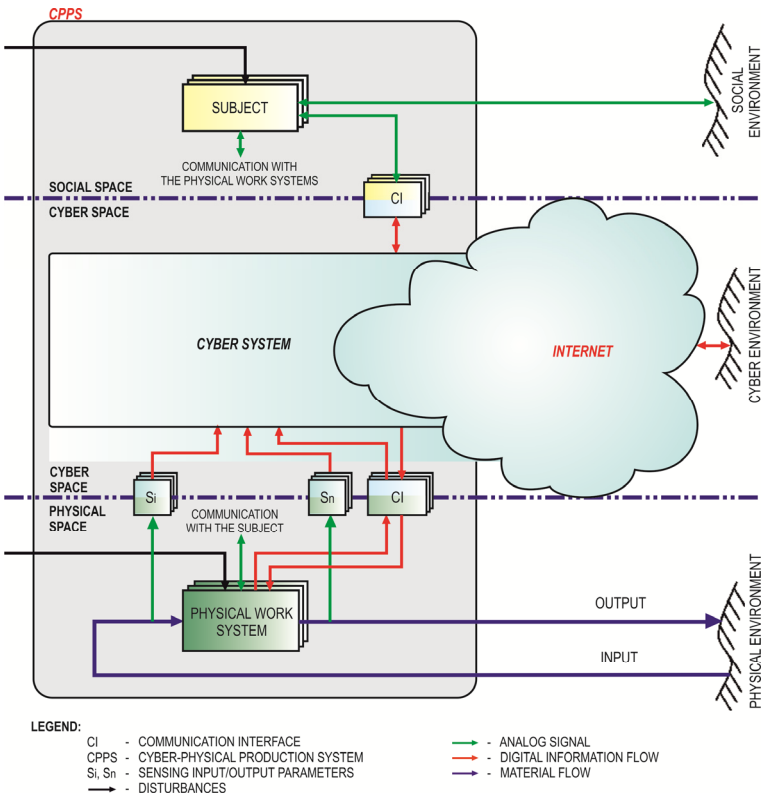


Figure 2: Generic model of a Cyber-Physical Production System (CPPS) illustrating the interaction among Physical, Cyber, and Human domains through communication interfaces (adapted from Hozdić et al. [17]).

The physical work system (PWS) inherits its structure from the Elementary Work Systems (EWS) model [18] and interacts continuously with both human actors and digital control components.

Within the social system, humans participate not only through direct collaboration with other actors but also through digitally mediated interactions that increasingly take place in cyberspace. Advances in ICT allow social processes-coordination, supervision, negotiation, task allocation-to extend beyond physical proximity and merge with cyber workflows. As a result, the three subsystems form a hybrid socio-technical environment in which physical operations, digital intelligence, and human cognition co-evolve. Figure 2 therefore provides the structural foundation for understanding how CPPS integrate physical, cyber, and social spaces as a stepping stone toward cognitization and Industry 5.0.

From a systems perspective, the CPPS architecture illustrated in Figure 2 can be interpreted as a transitional structure in which physical execution, cyber control, and human interaction are functionally coupled but not yet cognitively integrated. The absence of a dedicated learning and reasoning layer limits the system's ability to interpret context and adapt beyond predefined control logic. This structural limitation directly motivates the introduction of a cognitive layer in the subsequent evolution toward C-CPPS.

As cybernation matured, it revealed a dimension that earlier automation seldom considered explicitly: the role of context. The effectiveness of cyber-physical integration depends not only on connectivity and computational power but on the coherence, quality and interpretability of exchanged information. Cybernation thus becomes more than a technical milestone; it becomes a conceptual turning point in which decision formation, disturbance propagation and the balance of autonomy and predictability must be reconsidered. These insights naturally lead toward cognitization, where interpretation, not only control, becomes central.

2.3. Key Characteristics of Cybernated Production Systems

The core features that distinguish cybernated systems within Industry 4.0 include:

- *Connectivity:* Continuous communication among machines, devices, and platforms via industrial IoT networks ensures synchronized data flow across production layers.
- *Interoperability:* Standardized communication protocols (OPC UA, MQTT, etc.) enable seamless integration between legacy systems, robots, and cloud-based management platforms.
- *Self-monitoring and adaptability:* Embedded sensors, AI algorithms, and digital twins allow systems to detect anomalies, predict maintenance needs, and reconfigure processes autonomously.
- *Decentralized decision-making:* Instead of hierarchical control, decisions are distributed across intelligent agents, enhancing system resilience and response time.
- *Data-driven optimization:* Real-time analytics and predictive models continuously refine production parameters for maximum efficiency.

Together, these features mark a shift from traditional automation toward intelligent autonomy, where systems adapt autonomously yet remain fully transparent and controllable through digital interfaces. This evolution also redefines the human role: operators increasingly assume supervisory, interpretive and decision-support responsibilities rather than direct manual control.

Bibliometric evidence reinforces the legitimacy of distinguishing digitalization, cybernation and cognitization as successive phases. Publications on smart manufacturing grew steadily after 2011, Industry 4.0 research expanded sharply after 2012, and cognitization-related work became prominent only after 2016. These trends mirror the conceptual trajectory outlined here and indicate a broader movement within the scientific community.

2.4. The Paradox of Efficiency and Cognitive Limitation

Industry 4.0 has delivered unprecedented digital efficiency – faster processes, greater precision and global integration. Yet this technological sophistication introduces a paradox: as machines gain autonomy, the uniquely human capacities for contextual judgment, creativity and ethical reasoning risk being marginalized. Automated systems excel in computation but lack the ability to interpret intent, manage ambiguity or evaluate consequences through human-centered criteria.

This limitation underscores the growing importance of cognitive augmentation and human-in-the-loop paradigms, which form the conceptual bridge toward Industry 5.0. As articulated in the European Commission's vision [19] and recent work on human-machine teaming [13], the future of industrial intelligence lies in synergy rather than substitution: machines providing scale, repeatability and speed; humans contributing interpretation, responsibility and adaptive reasoning.

The challenge for contemporary manufacturing, therefore, is to preserve digital efficiency while expanding it with cognitive adaptability – enabling machines to learn and humans to lead. This synthesis defines the transition toward cognitization and sets the stage for fully integrated Human–Cyber–Physical Systems.

Despite the significant advancements enabled by cybernation, several limitations remain. Highly interconnected CPPS architectures often exhibit increased system complexity, reduced transparency of decision processes, and heightened vulnerability to cyber-security risks. Moreover, fully autonomous decision-making can lead to suboptimal outcomes in situations involving uncertainty, incomplete data, or conflicting objectives. These limitations highlight the need for an additional layer of interpretive and adaptive intelligence, which motivates the transition toward cognitization.

3. COGNITIZATION: THE NEXT STAGE IN THE INDUSTRIAL EVOLUTION

3.1. Definition and Conceptual Foundations

Cognitization represents the third evolutionary phase following digitalization and cybernation. While digitalization provides the informational substrate and cybernation establishes intelligent coordination between physical and digital domains, cognitization introduces learning, semantic interpretation and adaptive reasoning into production systems. In this phase, manufacturing systems no longer rely solely on predefined algorithms or static feedback loops; they gain the capability to perceive context, interpret patterns and improve performance through experience.

Cognitization thus refers to the embedding of cognitive intelligence – perception, learning and adaptive inference – into Cyber-Physical Production Systems. This development aligns with the concept of Cognitive Cyber-Physical Production Systems (C-CPPS) [11, 15], which extend traditional CPPS architectures with artificial intelligence, machine learning and semantic models. In such systems, cognition emerges from the continuous interaction among data, algorithms and human expertise, forming dynamic, self-evolving industrial ecosystems.

C-CPPS have increasingly been recognized as a central architectural element in the emerging class of Adaptive Cognitive Manufacturing Systems (ACMS) [20,21], which many authors connect with the broader vision of the fifth industrial revolution – Industry 5.0. In contrast to classical Cyber-Physical Production Systems (CPPS), these advanced systems incorporate cognitive capabilities and artificial-intelligence methods that enable greater adaptability, continuous learning, and more sophisticated forms of interaction between human actors and autonomous machines. Through this extension, C-CPPS provide the structural and functional foundations for manufacturing environments in which technical intelligence and human cognition operate as integrated, mutually reinforcing components.

This transition resonates with the European Commission's Industry 5.0 vision [19], which emphasizes human-centricity, resilience and sustainability. It also corresponds to recent formulations of Human–Cyber–Physical Systems (HCPS) [13], where humans are understood not as peripheral supervisors but as full cognitive partners in industrial decision processes.

Seen from a broader perspective, cognitization does not break with Industry 4.0 but represents its conceptual maturation. Once systems could sense, connect and adapt, questions of interpretation and contextual reasoning naturally followed. Cognitization formalizes these capabilities as central requirements for meaningful industrial autonomy.

3.2. Mechanisms Enabling Cognitization

The cognitive layer in modern manufacturing systems is enabled by a convergence of several key technologies:

- *Artificial Intelligence* (AI): Provides the analytical and decision-making capability that allows systems to infer optimal actions from complex data streams.

- *Machine Learning* (ML): Facilitates pattern recognition, predictive analytics, and continuous improvement through data-driven learning cycles.
- *Computer Vision*: Enables perceptual awareness by allowing systems to interpret visual data for quality control, defect detection, and autonomous navigation.
- *Semantic Modelling:* Structures knowledge and context relationships within manufacturing processes, enabling reasoning and interoperability among intelligent agents.
- *Adaptive Algorithms*: Dynamically modify operational parameters in real time based on feedback and learning outcomes, ensuring system robustness under changing conditions.

Together, these technologies transform CPPS into cognitive entities capable of understanding intent, anticipating variability and optimizing their behaviour beyond fixed rules. They provide the technical basis for C-CPPS, where perception, reasoning and learning are integrated into coherent, adaptive industrial architectures.

Recent developments also highlight the growing role of computer vision and perceptual AI as core enablers of cognitization in manufacturing systems. As shown by Balaska [22], vision-based systems increasingly function not merely as inspection tools but as cognitive perception modules capable of contextual interpretation, multimodal fusion and adaptive decision support. Their integration into cognitive CPPS architectures strengthens the perceptual layer by enabling machines to detect anomalies, classify patterns and anticipate deviations with human-comparable accuracy [23]. In this sense, computer-vision-driven perception becomes a foundational capability for learning-based production systems, bridging raw sensory data with higher-level cognitive inference.

While these enabling technologies provide the foundation for cognitization, their integration into coherent system architectures remains a significant challenge. The effectiveness of cognitive manufacturing systems depends not only on individual AI components but on their coordination, interpretability, and alignment with human decision-making processes. This highlights the importance of viewing cognitization as a system-level transformation rather than a purely technological upgrade.

3.3. Evolution from Reactive to Cognitive Systems

The progression from cybernated to cognitive systems marks a shift from reactive to proactive industrial intelligence. Traditional automated systems respond to deviations based on static rules; cognitive systems, by contrast, detect weak signals, predict anomalies before they occur and continuously refine their decision logic.

This evolution can be conceptualized as a shift along three dimensions:

1. From deterministic control to probabilistic reasoning, where systems evaluate uncertainty using learned models rather than fixed thresholds;
2. From static optimization to continuous self-optimization, driven by real-time data and reinforcement learning;
3. From hierarchical supervision to distributed cognition, where intelligent agents collectively learn and adapt within networked production environments.

As a consequence, manufacturing evolves from smart factories toward cognitive and socially-aware factories, where machine learning and human expertise co-develop within the same operational ecosystem.

3.4. Application Example: Cognitive Production Lines

Cognitization manifests clearly in production lines that learn from real-time data. In such environments, sensors capture performance indicators – vibration, temperature, tool wear – which ML models analyse to predict future states and optimise parameters. In processes such as injection moulding or additive manufacturing, cognitive systems autonomously adjust variables such as extrusion temperature, feed rate or cooling time based on learned correlations with output quality.

Computer-vision modules further enhance adaptability by identifying surface or geometric deviations and triggering corrective actions. Through closed-loop learning, each production cycle contributes to an evolving body of operational knowledge, supporting predictive maintenance, zero-defect manufacturing and rapid adaptation to product variation.

These principles build on earlier CPPS research by Hozdić, et al. [17], in which EWS could autonomously plan and adapt their task agendas. In cognitive environments, this logic extends into distributed learning agents that coordinate decisions across the production network. The cybernetic PPC model proposed by Hozdić et al. [17] represents an early implementation of such reasoning, where EWS evaluate alternative sequences based on workload, lead time or due-date adherence, dynamically rescheduling tasks in response to disturbances.

In practical industrial environments, these concepts are increasingly implemented in advanced manufacturing systems such as smart additive manufacturing platforms, adaptive machining systems, and digitally integrated production lines. For example, in additive manufacturing processes, cognitive systems can dynamically adjust process parameters such as layer thickness, temperature, and deposition speed based on real-time feedback and learned correlations with product quality. Such applications demonstrate how cognitization moves beyond theoretical abstraction and becomes an operational capability in modern production systems.

3.5. Toward Cognitive Cyber-Physical Production Systems (C-CPPS)

Cognitive Cyber-Physical Production Systems (C-CPPS) [15] represent the culmination of the Industry 4.0 trajectory and the conceptual threshold of Industry 5.0. Building on classical CPPS models, C-CPPS integrate three functional layers: the Physical Work System (PWS), the Cyber System (CS) and a Cognitive Layer where learning, semantic reasoning and interpretive processes occur. The human subject contributes contextual understanding, ethical judgement and domain knowledge, forming an integrated triadic architecture (see Figure 3).

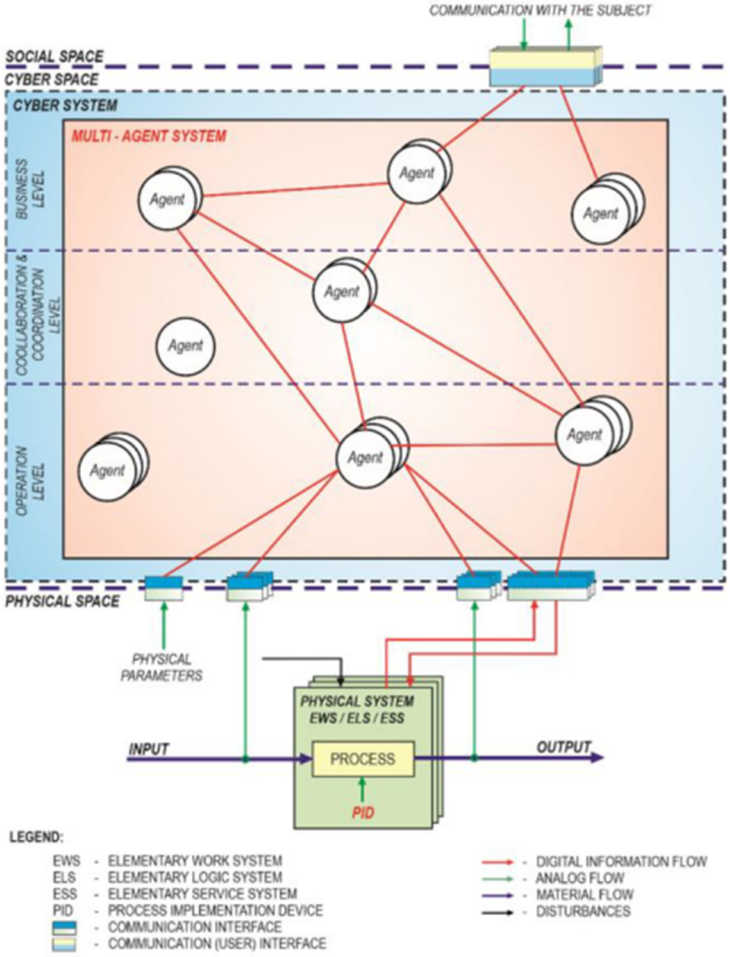


Figure 3: Conceptual architecture of a Cognitive Cyber-Physical Production System (C-CPPS) [11].

The C-CPPS extends the CPPS framework introduced by Hozdić, et al. [11, 15], which articulates the interaction between PWS, CS and the human domain via communication interfaces. In the cognitive extension, the cyber layer not only regulates but interprets, the cognitive layer learns and reasons, and the human actor exercises oversight and value-based decision-making.

Intelligence in these systems becomes distributed rather than centralized. Machines, digital agents and human operators all serve as active contributors to a shared cognitive network. Such architectures embody the human–machine synergy envisioned by Industry 5.0—where autonomy is balanced with responsibility and technological capability is aligned with human values.

Ultimately, the transition from CPPS to C-CPPS represents the material expression of cognitization: transforming automated systems into adaptive, context-aware and ethically aligned industrial ecosystems. This progression establishes the conceptual groundwork for the next industrial paradigm, in which cognition, sustainability and human creativity converge within intelligent manufacturing environments.

4. HUMAN-MACHINE SYNERGY: THE CORE OF INDUSTRY 5.0

A notable insight emerging from recent bibliometric analyses is the disproportionate attention given to the human role in manufacturing research compared to production-oriented studies. Between 2006 and 2022, over 8600 publications associated humans with manufacturing, whereas only around 1300 examined their role in production systems. This imbalance suggests a conceptual reframing: the human contribution is increasingly viewed not as an operational variable within a production flow, but as a structural component of intelligent manufacturing ecosystems. This perspective reinforces the HCPS concept developed in this article, positioning the human actor as a provider of contextual reasoning, ethical judgment and cognitive balance within autonomous and semi-autonomous environments.

4.1. From Automation to Human-in-the-Loop Paradigm

Industry 5.0 introduces a conceptual and functional shift — from the pursuit of full automation to collaborative cognition, in which humans and intelligent systems act as interdependent partners. This transition emerges not from nostalgia for traditional labor but from a clearer understanding of the limits of autonomous algorithms. Even highly advanced systems struggle when confronted with abrupt change, incomplete information or ambiguous objectives. At such boundaries, human cognition contributes irreplaceable qualities: the ability to navigate uncertainty, infer intent, weigh long-term implications and integrate ethical considerations.

Where Industry 4.0 primarily emphasised digital transformation and machine autonomy, Industry 5.0 re-introduces human creativity, cognitive flexibility and moral reasoning as essential components of intelligent production. This vision aligns with the human-in-the-loop (HITL) and human-on-the-loop (HOTL) interaction levels described by Nahavandi [23], ensuring that humans retain meaningful oversight within increasingly autonomous factories. Rather than replacing human expertise, Industry 5.0 casts humans as co-creators of industrial intelligence, shaping adaptive and responsible forms of automation.

Recent research suggests that cognitive extensions of CPPS are evolving toward holistic production architectures that dynamically coordinate cyber, physical and human agents. Englund [24] argues that future C-CPPS will require interoperability across cognitive layers, transparent reasoning interfaces and semantic models that allow distributed elements to negotiate goals, constraints and uncertainties. Such systems move beyond classical CPPS by embedding learning-based coordination mechanisms capable of restructuring workflows, reallocating resources and adjusting production logic autonomously in response to contextual variation [23]. These insights reinforce the need for C-CPPS architectures that treat cognition as a system-level property rather than a discrete algorithmic add-on.

4.2. Humans as Cognitive Partners: Toward Augmented Intelligence

Within Human–Cyber–Physical Systems (HCPS), humans act not as passive supervisors but as active cognitive partners embedded within the system's perceptual and interpretive loops. The concept of augmented intelligence reframes intelligence as a mutual enhancement: machines extend human analytical capacity, while humans contribute intuition, contextual understanding, creativity and empathy. Technical capability alone is insufficient for stable and safe operation; human judgment remains a critical corrective to both cybernetic and cognitive processes.

Augmented intelligence manifests through multimodal human–machine interfaces, digital twins, adaptive dashboards and context-aware visualizations that support intuitive interpretation of complex datasets. These tools enable operators and engineers to guide machine-learning processes by injecting domain knowledge and contextual constraints. As Bocklisch et al. [13] argue, such collaboration shapes HCPS as socio-technical systems in which knowledge circulates bidirectionally: data informs human understanding, and human interpretation reshapes algorithmic behaviour.

Complementary to this, Lou et al. [25] present a comprehensive HCPS framework that emphasizes the multi-layered role of human cognition within cyber-physical environments. Their review identifies three core dimensions—human-in-the-loop, human-on-the-loop and human-in-the-society—each representing increasing levels of human influence on technological behaviour. According to Lou et al. [25], HCPS architectures are effective only when human interpretive capacity is structurally integrated into perception, reasoning and execution layers, ensuring alignment between machine autonomy, ethical constraints and social expectations.

Effective augmented intelligence requires calibrated trust between human and machine. Governance mechanisms must define: (i) delegation thresholds, clarifying which decisions systems may take autonomously; (ii) explanatory minimums, ensuring that actions are transparent and interpretable; (iii) rollback procedures for safe recovery from erroneous or unsafe states; and (iv) model lifecycle management, including drift detection and periodic re-validation.

Such frameworks ensure accountability, minimize residual risk and enhance human acceptance of cognitive manufacturing systems.

4.3. The Role of Creativity, Ethics, and Empathy

Industry 5.0 distinguishes itself through its ethical and humanistic orientation. While machines excel at computation, optimisation and precision, only humans can attribute meaning, consider moral implications or anticipate social impact. As manufacturing environments grow more intelligent, creativity, ethics and empathy become structural – not decorative – elements of system design.

Human creativity enables innovation through abstraction and the exploration of alternative solutions beyond algorithmic optimisation. Ethical reasoning ensures that autonomous decisions align with values such as safety, fairness, sustainability and responsibility. Empathy informs the design of ergonomic workplaces, assistive robotics and socio-technical systems that respect human well-being.

These principles must be operationalised in everyday practice. Examples include progressive disclosure of information, visualisation of uncertainty, bias-checking protocols, and consent-to-act confirmations for irreversible actions. Such micro-design interventions translate ethics into operational safeguards and reinforce trust in cognitive manufacturing environments.

Ultimately, creativity, ethics and empathy determine not only how systems operate but what kinds of industrial futures are acceptable. They remind us that the trajectory of industrial intelligence depends not only on machine learning but on how thoughtfully human values are embedded into technological systems.

4.4. Human–Machine Collaboration Framework

The Human–Machine Collaboration Framework (see Figure 4) conceptualizes this synergy through three interdependent layers:

1. Perception and Awareness Layer – At the lowest level, humans and machines jointly monitor and interpret their environment. This layer integrates multisensory data streams captured through smart sensors, vision systems, wearable interfaces, digital dashboards, and augmented-reality tools. Its purpose is to build a shared situational picture in which operators perceive system states while cyber components translate raw signals into structured, context-rich information. This joint perceptual space forms the foundation for safe and coordinated behaviour.
2. Decision and Reasoning Layer – In the intermediate layer, human cognitive capabilities and machine intelligence converge. Here, artificial intelligence agents, optimization models, and rule-based algorithms exchange information with human decision-makers to evaluate alternatives, negotiate trade-offs, and co-create adaptive strategies. Unlike traditional automation hierarchies, the reasoning layer supports dynamic collaboration: humans contribute contextual interpretation, ethical judgment, and scenario framing, while cyber subsystems provide analytical depth, predictive insight, and rapid option evaluation. This shared cognitive workspace enables balanced, transparent decision-making.
3. Action and Execution Layer – The uppermost layer operationalizes decisions through coordinated activity. Robots, autonomous systems, and production equipment interact with human operators to execute tasks, adjust process parameters, and respond to deviations. Continuous feedback loops connect execution outcomes back to the perception and reasoning layers, enabling both machines and humans to learn from operational experience. This structure facilitates adaptive workflows, real-time error mitigation, and fine-grained control over safety-critical processes.

Figure 4 synthesizes these layers into a coherent HCPS architecture, highlighting the bidirectional flow of information between human cognition, cyber intelligence, and physical mechanisms. Rather than treating humans as supervisory add-ons, the model places them at the centre of the socio-technical system: influencing digital reasoning, validating autonomous actions, and shaping system behaviour through contextual and ethical insight. In this sense, the HCPS framework extends the classical CPPS model by embedding cognition, collaboration, and social intelligence into the production architecture. It represents the structural foundation of Industry 5.0, where human and machine agents jointly create resilient, adaptive, and value-aligned manufacturing systems [26].

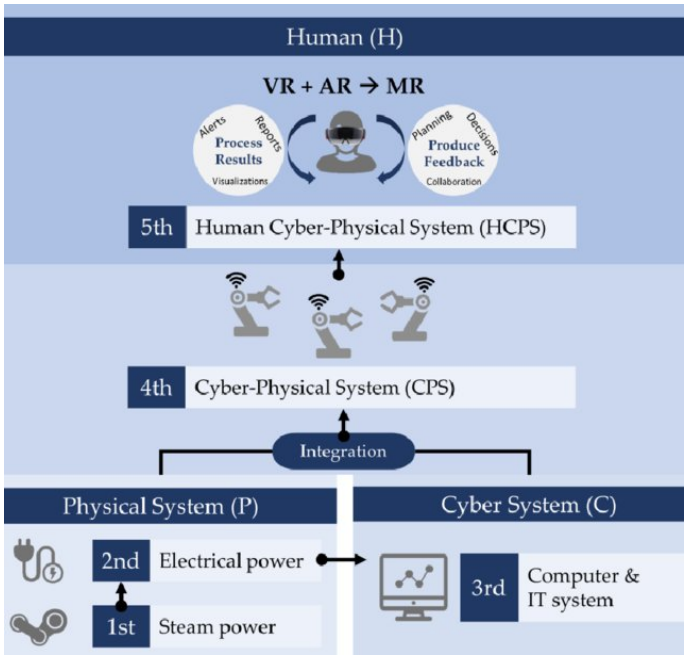


Figure 4: Conceptual model of a Human–Cyber–Physical System (HCPS) in Industry 5.0 [26]

4.5. Symbiosis and Mutual Empowerment

Human–machine synergy is ultimately defined by mutual empowerment. Technology enhances human perception, precision and predictive capability, while humans steer technological behaviour toward socially responsible and sustainable outcomes. This reciprocal relationship transforms industrial environments into collaborative intelligence systems, where ethical principles, creativity and sustainability coexist with digital efficiency.

In this emerging paradigm, industrial progress is no longer measured by automation density alone but by cognitive balance – the extent to which human values, machine intelligence and ecological responsibility operate in harmony. Industry 5.0 therefore represents a synthesis of technological excellence and human wisdom, a new phase in which innovation serves not only productivity but societal purpose.

5. CONCLUSIONS

This paper has traced the structural evolution of manufacturing systems over the past fifteen years, showing that the trajectory from digitalization to cybernation and, ultimately, to cognitization reflects not only technological progress but a deeper transformation of industrial reasoning. Digitalization established the informational backbone of modern factories through end-to-end data flows and real-time visibility. Cybernation advanced this foundation by integrating physical and digital domains into self-regulating feedback loops characteristic of Cyber-Physical Production Systems (CPPS). Cognitization now extends these architectures further, embedding learning, semantic reasoning and contextual awareness into production environments. Within this trajectory, Cognitive CPPS (C-CPPS) and Human–Cyber–Physical Systems (HCPS) emerge as the operational architectures that realize Industry 5.0's human-centric vision.

The sectoral examples examined in this work—automotive, mechanical engineering, and medical technology—illustrate this progression. Digitalization provided transparency and actionable information; cybernation enabled distributed autonomy and adaptive coordination; and cognitization reintroduced human cognition as a structural design element. In this third phase, the human role evolves from operator to supervisor and ultimately to cognitive collaborator, shaping system behaviour through judgment, creativity and ethical reasoning. The convergence of machine learning, robotics and human expertise thus produces not only higher productivity but also more responsible and resilient industrial systems.

Three managerial implications follow.

First, cognition cannot flourish atop weak digital foundations: investments in AI, robotics and autonomy must be matched by investments in data governance, semantic interoperability and lifecycle management of algorithmic components.

Second, firms must design explicit HITL/HOTL operating models to define human authority, accountability and trust calibration in autonomous environments.

Third, capability development must expand beyond technical upskilling to include systems thinking, ethics and human-factors engineering—competencies essential for acceptance, reliability and long-term value creation.

Methodologically, this article contributes a unified three-phase interpretive lens and clarifies the conceptual relationship between CPPS, C-CPPS and HCPS. Conceptually, it positions Industry 5.0 as a cognitive and ethical evolution rather than a purely technological revolution. Practically, it synthesizes cross-sectoral developments into a strategic roadmap: (i) consolidate the digital backbone; (ii) scale cyber-physical autonomy; (iii) embed cognition through human–machine teaming, trust governance and socio-technical design.

It should be noted that the conceptual nature of this study represents a limitation. The proposed framework is not validated through empirical case studies or quantitative performance indicators, which may affect its direct applicability in specific industrial contexts. Additionally, the boundaries between digitalization, cybernation, and cognitization may overlap in practice, depending on the level of technological maturity and sector-specific characteristics. Future research should therefore focus on developing measurable indicators and experimental testbeds to validate the proposed conceptual model.

Several limitations should be acknowledged. The industrial evidence referenced here draws partly on public documentation, and cross-sectoral heterogeneity constrains strict KPI comparability. Boundaries between late-cybernation and early cognitization phases may blur, especially in highly automated plants, while regulatory, labor and energy contexts differ substantially across regions. These constraints underscore the need for standardized metrics, open testbeds and safety frameworks to evaluate cognitive manufacturing maturity and ensure transferability.

These limitations open a research agenda with several urgent priorities:

- Operational metrics of cognitization (maturity models and KPIs linking learning depth, explainability, and human involvement to performance and risk);
- Benchmark datasets and testbeds for C-CPPS/HCPS under realistic disturbances;
- Trust, safety, and governance frameworks that align AI autonomy with human oversight (HITL/HOTL) and regulatory compliance;
- Socio-technical design methods that integrate ergonomics, empathy, and creativity into robotics and AI workflows;
- Sustainability-coupled optimization, uniting cognitive control with energy efficiency, circularity, and resilience.

Looking across the entire trajectory, a central insight becomes clear: technological maturity alone does not ensure industrial resilience. Organizations that minimized the human role often revealed significant vulnerabilities during periods of disruption, whereas those that cultivated structured human–machine collaboration demonstrated superior adaptability and foresight. Future competitive advantage will therefore emerge where technological intelligence and human judgment operate not in tension but in synergy.

Taken together, the evolution from digitalization to cybernation and cognitization demonstrates that industrial progress is no longer defined solely by advances in automation, but by the ability to integrate technologies into coherent socio-technical systems. Industry 4.0 delivered connectivity and autonomy; Industry 5.0 extends these foundations with cognition, ethics and sustainability. In conceptual terms:

$$\text{Industry 5.0} = \text{Autonomy} \times \text{Cognition} \times \text{Values}.$$

Enterprises that align these three dimensions—through robust digital infrastructures, interoperable CPPS, cognitive extensions and disciplined human oversight—will transform smart factories into adaptive, ethical and sustainable industrial ecosystems. Technology alone drives automation; technology guided by human cognition drives civilization. The next decade will reward those organizations that cultivate this synergy—where machines learn at scale and humans lead with purpose.

The transition from Industry 4.0 to Industry 5.0 should not be understood as a replacement of technologies, but as their conceptual evolution toward more balanced and human-centric systems. The integration of cognition, autonomy, and human values will determine the effectiveness and sustainability of future manufacturing environments. In this context, human–machine synergy emerges not as an optional feature, but as a fundamental requirement for resilient and adaptive industrial systems.

In this sense, the transition toward Industry 5.0 represents not merely a technological shift, but a redefinition of industrial intelligence as a balanced integration of machine capability and human cognition.

REFERENCES

[1] K. Schwab, “The Fourth Industrial Revolution”, World Economic Forum, Geneva (Switzerland), (2016)

[2] H. Kagermann, W. Wahlster, and J. Helbig, “Securing the Future of German Manufacturing Industry: Recommendations for Implementing the Strategic Initiative Industrie 4.0”, Final Report of the Industrie 4.0 Working Group, Acatech National Academy of Science and Engineering, Frankfurt am Main (Germany), (2013)

[3] D. Spath, S. Gerlach, M. Hämmerle, S. Schlund and T. Strölin, “Cyber-physical system for self-organised and flexible labour utilisation”, Proceedings of the 22nd International Conference on Production Research (ICPR 2013), Iguassu Falls (Brazil), 28 July–1 August 2013, (2013)

[4] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, “Internet of things: Vision, applications and research challenges”, Ad Hoc Networks, Vol. 10(7), pp. 1497–1516, https://doi.org/10.1016/j.adhoc.2012.02.016, (2012)

[5] E. Henriette, M. Feki and I. Boughzala, “The Shape of Digital Transformation: A Systematic Literature Review”, Proceedings of the 9th Mediterranean Conference on Information Systems, Samos, (Greece), 3–5 October, pp. 1–13, (2015)

[6] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann, “Industry 4.0”, Business and Information Systems Engineering, Vol. 6(4), pp. 239–242, https://doi.org/10.1007/s12599-014-0334-4, (2014)

[7] F. J. Folgado, D. Calderón, I. González, and A. J. Calderón, “Review of Industry 4.0 from the Perspective of Automation and Supervision Systems: Definitions, Architectures and Recent Trends”, Electronics, Vol. 13(4), Article 782, https://doi.org/10.3390/electronics13040782, (2024)

[8] L. Monostori, “Cyber-physical production systems: Roots, expectations and R&D challenges”, Procedia CIRP, Vol. 17, pp. 9–13, https://doi.org/10.1016/j.procir.2014.03.115, (2014)

[9] S. Oks, M. Maldonado, D. Göhringer, and T. Becker, “Cyber-Physical Systems in the Context of Industry 4.0: A Comprehensive Review”, Procedia CIRP, Vol. 107, pp. 1234–1241, https://doi.org/10.1007/s10796-022-10252-x, (2022)

[10] P. Parviainen, M. Tihinen, J. Kääriäinen, and S. Teppola, “Tackling the digitalization challenge: how to benefit from digitalization in practice”, International Journal of Information Systems and Project Management, Vol. 5(1), pp. 63–77, https://doi.org/10.12821/ijispm050104, (2017)

[11] E. Hozdić and I. Makovec, “Evolution of the Human Role in Manufacturing Systems: On the Route from Digitalization and Cybernation to Cognitization”, Applied System Innovation, Vol. 6(2), Article 49, https://doi.org/10.3390/asi6020049, (2023)

[12] M. C. Zizic, M. Mladineo, N. Gjeldum, and L. Celent, “From Industry 4.0 towards Industry 5.0: A Review and Analysis of Paradigm Shift for the People, Organization and Technology”, Energies, Vol. 15(14), Article 5221, https://doi.org/10.3390/en15145221, (2022)

[13] E.H.E. Suryadarma, P.W. Laksono, I. Priadythama, L. Herdiman, “Human Cyber Physical System in Manufacturing 4.0: An Application for Intelligent SCADA-Based Manufacturing”. In: S.N. Mohd Saffe, S.Z. Ismail, C.N. Rosyidi, M.O. Tokhi, (eds) Proceedings of the 7th Asia Pacific Conference on Manufacturing Systems and 6th International Manufacturing Engineering Conference—Volume 1. IMEC-APCOMS 2024. Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-96-4353-0_10, (2025)

[14] F. El Kalach, I. Yousif, T. Wuest, A. Sheth, and R. Harik, “Cognitive manufacturing: definition and current trends”, Journal of Intelligent Manufacturing, Vol. 36(6), pp. 3695–3715, https://doi.org/10.1007/s10845-024-02429-9, (2025)

[15] E. Hozdić and Z. Jurković, “Cognitive Cyber-Physical Production Systems: A New Concept of Manufacturing Systems on the Route to Industry 5.0”, New Technologies, Development and Application VI, Lecture Notes in Networks and Systems, Vol. 687, pp. 201–212, https://doi.org/10.1007/978-3-031-31066-9_21, (2023)

[16] J. Davis, T. Edgar, J. Porter, J. Bernaden, and M. Sarli, “Smart manufacturing, manufacturing intelligence and demand-dynamic performance”, Computers & Chemical Engineering, Vol. 47, pp. 145–156, https://doi.org/10.1016/j.compchemeng.2012.06.037, (2012)

[17] E. Hozdić, D. Kozjek, and P. Butala, “A cyber-physical approach to the management and control of manufacturing systems”, Strojniški vestnik – Journal of Mechanical Engineering, Vol. 66(1), pp. 61–70, https://doi.org/10.5545/sv-jme.2019.6156, (2020)

[18] J. Peklenik, “Fertigungskybernetik, Eine Neue Wissenschaftliche Disziplin für die Produktionstechnik”, Festvortrag Anlässlich der Verleihung des Georg-Schlesinger Preises 1988 des Landes Berlin, Berlin, (Germany), (1998).

[19] M. Breque, L. De Nul, and A. Petridis, “Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European Industry”, European Commission, Directorate-General for Research and Innovation, Brussels (Belgium), https://doi.org/10.2777/308407, (2021)

[20] H. ElMaraghy, L. Monostori, G. Schuh, and W. ElMaraghy, “Evolution and future of manufacturing systems”, CIRP Annals – Manufacturing Technology, Vol. 70(2), pp. 635–658, https://doi.org/10.1016/j.cirp.2021.05.008, (2021)

[21] H. ElMaraghy and W. ElMaraghy, “Adaptive Cognitive Manufacturing System (ACMS) – a new paradigm”, International Journal of Production Research, Vol. 60(24), pp. 7436–7449, https://doi.org/10.1080/00207543.2022.2078248, (2022)

[22] V. Balaska, A. Tserkezis, F. Konstantinidis, V. Sevetlidis, S. Symeonidis, T. Karakatsanis, and A. Gasteratos, “Machine Vision in Human-Centric Manufacturing: A Review from the Perspective of the Frozen Dough Industry”, Electronics, Vol. 14(17), Article 3361, https://doi.org/10.3390/electronics14173361, (2025)

[23] N. Vallabhaneni, P. Panneer and N. S. Kalyan Ram, "Trust-Aware Human-in-the-Loop Reinforcement Learning for Autonomous Systems," 2025 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT), Karaikal, India, pp. 1-7, doi: 10.1109/IConSCEPT66142.2025.11436874 (2025)

[24] T. Englund, J. Bruch, K. Chirumalla, and M. Ashjaei, “Towards Holistic Cyber-Physical Production Systems in Existing Production Environment: Challenges from a Case Study”, *Procedia CIRP*, Vol. 134, pp. 115–120, https://doi.org/10.1016/j.procir.2025.03.020, (2025)

[25] S. Lou, Z. Hu, Y. Zhang, Y. Feng, M. Zhou, and C. Lv, “Human-Cyber-Physical System for Industry 5.0: A Review from a Human-Centric Perspective”, IEEE Transactions on Automation Science and Engineering, Vol. 22, pp. 494–511, https://doi.org/10.1109/TASE.2024.3360476, (2025)

[26] H. B. Santoso, D. K. Baroroh, and A. Darmawan, “Future Application of Multisensory Mixed Reality in the Human Cyber-Physical System”, South African Journal of Industrial Engineering, Vol. 32(4), pp. 44–56, https://doi.org/10.7166/32-4-2551, (2021)

SESSION C: PRODUCTION TECHNOLOGIES, CAE AND MANAGEMENT

Robot-assisted lathe machine servicing

Vojislav Vujičić[1*], Ivan Milićević[1], Nenad Mijailović[1], Nedeljko Dučić[1], Veljko Aleksić[1], Marko Popović[1]

[1] University of Kragujevac, Faculty of Technical Sciences Čačak, Čačak, Serbia

ARTICLE INFO

[*] **Correspondence:** vojislav.vujicic@ftn.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.C02V **PAGES:** C13 – C18

ABSTRACT

This paper describes a simulation of an industrial robot-assisted lathe servicing process using ABB RobotStudio software. The objective of the simulation is to demonstrate how integrating a robot for material handling can improve productivity and ensure safer operation within a manufacturing environment. The first part of the paper outlines the procedure for importing 3D models of robots and the lathe into the simulation environment. Special focus is placed on defining the lathe's movable elements, such as its door, as well as a conveyor belt for moving finished parts outside the safety zone. Key operational parameters were established, including target positions for the robot (pick-up from the raw material storage, insertion into the chuck, and deposition onto the conveyor) and the robot's movement trajectory. Based on these, an optimized RAPID program was developed to control the robot and other moving components during the simulation. The final stage includes testing and visual evaluation of the entire process, confirming both the system's functionality and the robustness of the proposed automation solution.

KEYWORDS

Industrial robot, CNC lathe machine, Workpiece handling, Simulation.

1. INTRODUCTION

The rapid development of industrial automation technologies has transformed manufacturing processes across numerous industrial sectors. [1] Modern production facilities increasingly rely on automated systems to achieve higher productivity, lower operational costs, and improved product consistency. [1] Among the technologies driving this transformation, industrial robotics plays a central role due to its flexibility, precision, and capability to perform repetitive tasks continuously with minimal human intervention. [1] [2].

Industrial robots are currently employed in a wide variety of applications, including assembly, welding, packaging, palletizing, quality inspection, and machine tending. [3] Machine tending applications are particularly important because they directly influence production throughput and machine utilization. [4] In conventional manufacturing environments, machine loading and unloading operations are frequently performed manually, exposing operators to repetitive tasks, physical strain, and potential safety hazards. [4] The integration of robotic systems can eliminate these disadvantages while simultaneously improving process reliability and consistency. [3]

The implementation of industrial robots within machining environments has become increasingly common due to advancements in robot control systems, sensor technologies, and simulation software. [5] Modern offline programming platforms enable engineers to design, optimize, and validate robotic systems before physical deployment, and the use of offline programming reduces costs and risks, enables rapid tuning of robot paths, and minimizes collision states. [6] This approach reduces commissioning time, minimizes implementation risks, and lowers overall project costs. [6] By utilizing digital models of industrial equipment, engineers can identify potential problems and optimize system performance before installation in a real production environment. [6]

2. SIMULATION ENVIRONMENT AND SYSTEM COMPONENTS

The simulation was developed using ABB RobotStudio, a comprehensive software platform designed for robotic system modeling, offline programming, and virtual commissioning. The software provides tools for creating realistic workcell environments and testing robotic applications under conditions closely resembling real industrial operation.

The primary robotic component selected for this application was the ABB IRB120 industrial manipulator. This six-axis robot is characterized by compact dimensions, high repeatability, and excellent suitability for precision handling tasks. Due to its relatively small footprint and flexible motion capabilities, the robot represents an appropriate choice for machine-tending applications in compact manufacturing cells.

The robotic workcell consists of several interconnected components that collectively enable automated operation:

- ABB IRB120 industrial robot;
- Haas ST-10 CNC lathe machine;
- Two-finger robotic gripper;
- Conveyor belt system;
- Raw material storage location;
- Finished-part storage container;
- Safety enclosure and operator access zone.

To ensure realistic simulation conditions, a detailed three-dimensional model of the Haas ST-10 lathe machine was imported into RobotStudio (figure 1). Since the original machine model was available in CAD formats unsuitable for direct use within the simulation environment, it was converted into a compatible format before importation.

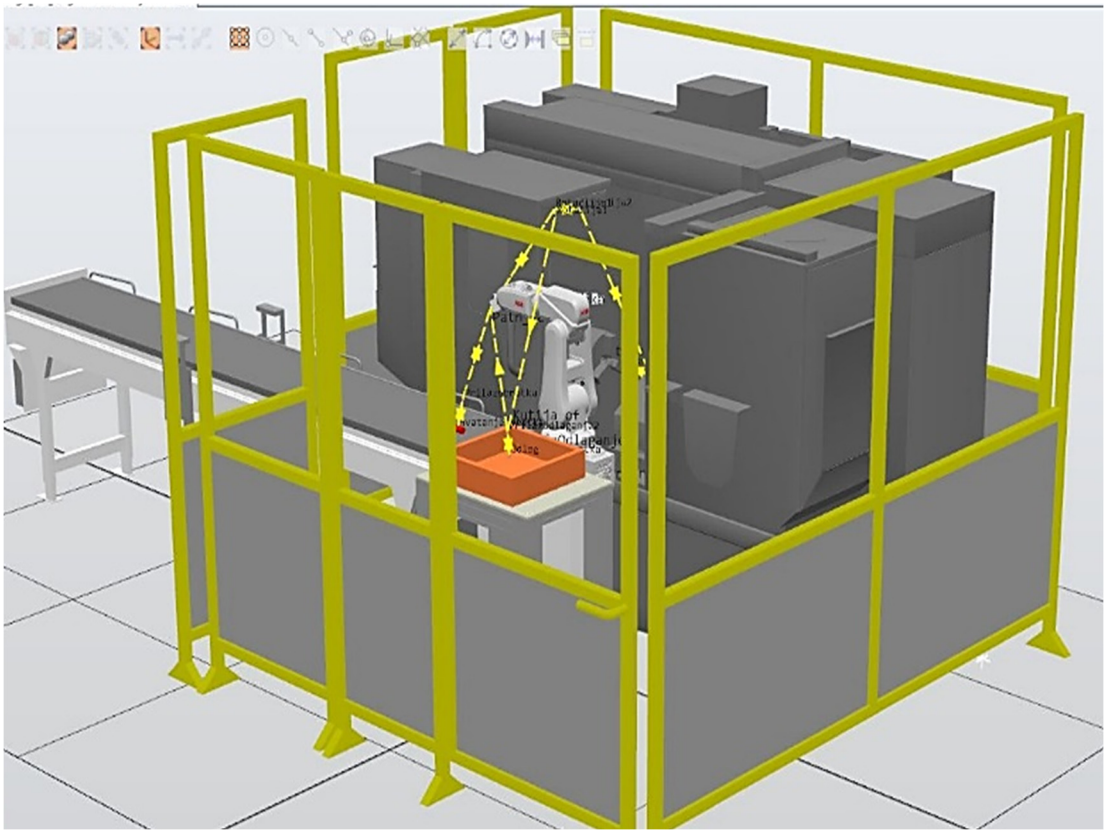


Figure 1: Robotic workcell

Modeling the lathe machine's sliding door mechanism received special attention. The opening and closing of the machine door represent a critical operation because it directly affects the interaction between the robot and the machine tool. Therefore, the door was implemented as a separate mechanical component with predefined motion constraints and operating states. A conveyor system was incorporated into the simulation to facilitate automated transportation of finished workpieces outside the robot operating area. This addition improves process continuity and reflects common industrial practices in automated machining cells. Finally, safety fencing and access gates were included to replicate realistic factory conditions and ensure compliance with industrial safety requirements. These elements also contribute to the overall realism of the virtual manufacturing environment.

3. MODELING OF THE ROBOTIC WORKCELL

After importing the required components, the next stage involved configuring and positioning the equipment within the virtual workspace. The layout design process aimed to maximize robot accessibility while minimizing unnecessary movement and ensuring collision-free operation. The lathe machine was positioned within the robot working envelope to allow efficient loading and unloading operations. The machine door mechanism was modeled as a prismatic joint with predefined open and closed positions (figure 2). Transition times between positions were configured to accurately replicate realistic machine behavior during production cycles.

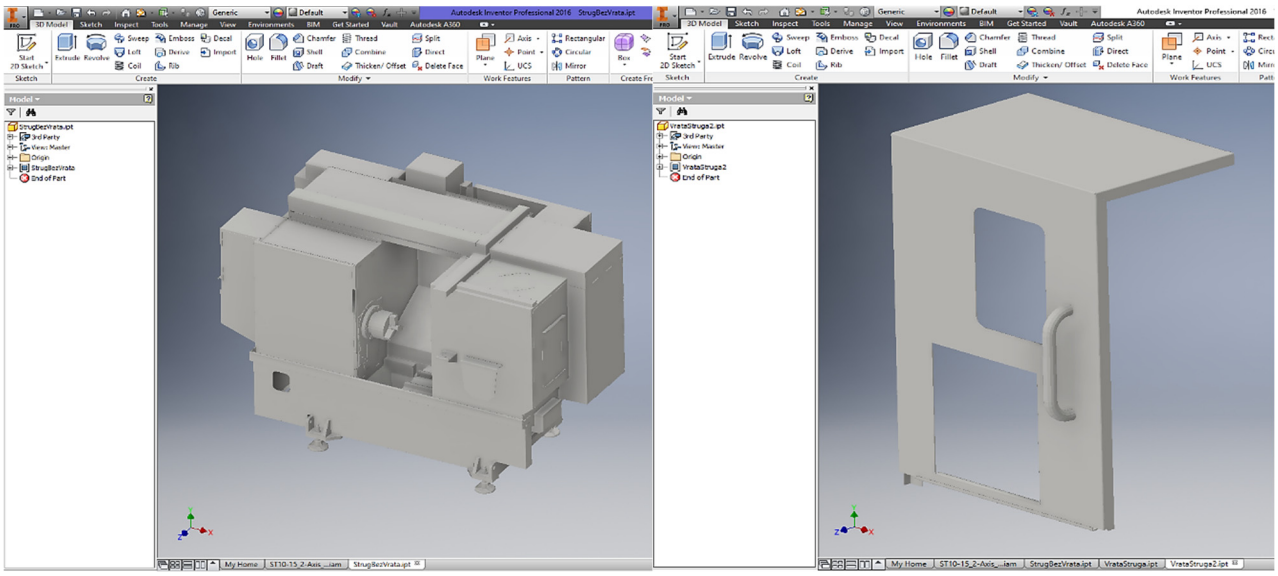


Figure 2: 3D model of lathe machine and door mechanism

The robotic gripper was modeled according to manufacturer specifications. A two-finger design was selected because of its simplicity, reliability, and suitability for handling cylindrical workpieces. The gripper mechanism was implemented with rotational joints representing the motion of the fingers during opening and closing operations shown on figure 3.

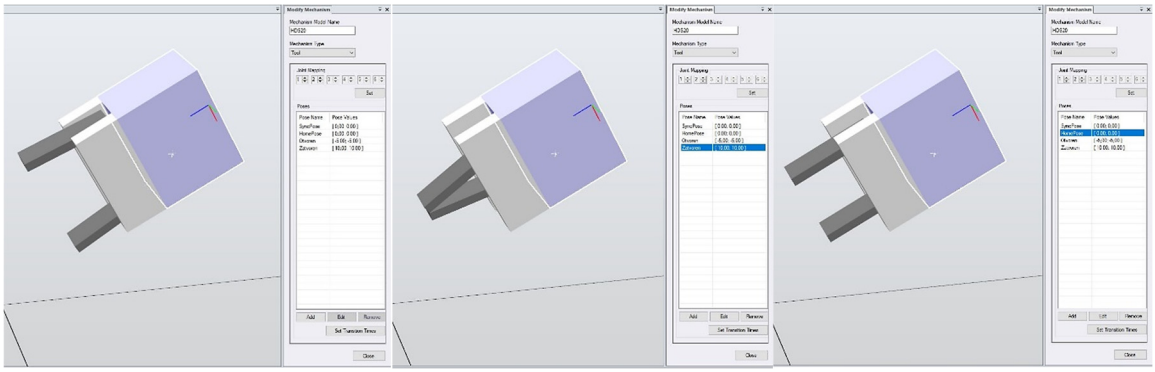


Figure 3: Robotic pneumatic gripper with rotation joints

Special attention was devoted to defining the Tool Center Point (TCP), which represents the reference point used by the robot controller during motion planning. Accurate TCP positioning is essential for ensuring precise manipulation of workpieces and repeatable robot operation. The simulation also required defining the physical properties of the gripper, including mass and center of gravity. These parameters influence robot dynamics and improve the realism of simulated motion.

Several work objects and coordinate systems were established throughout the workcell. Dedicated coordinate frames were assigned to the conveyor system, raw material location, machine chuck, and finished-part container (figure 4). The use of local coordinate systems simplifies robot programming and improves system flexibility, allowing modifications to individual components without extensive reprogramming.

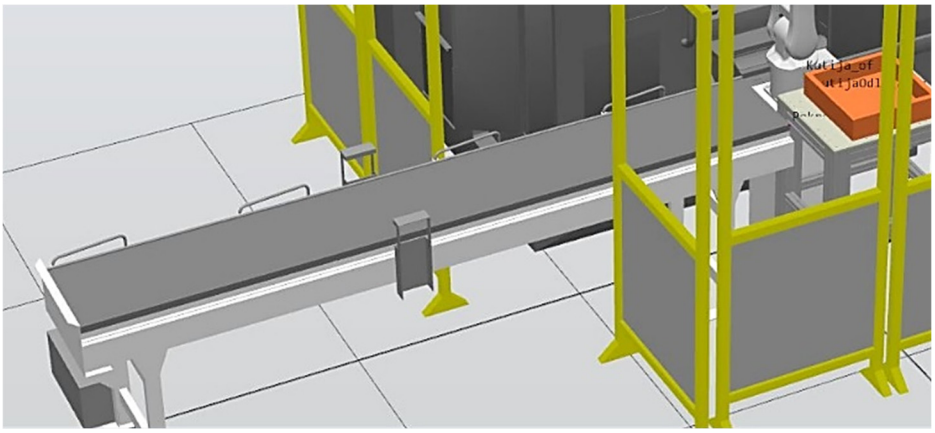


Figure 4: Conveyer belt with raw material location and finished part container

Safety considerations were incorporated during the layout design process. Robot reachability analysis and workspace visualization tools available in RobotStudio were used to verify that all required task locations were accessible without violating mechanical limitations or causing collisions.

4. ROBOT PROGRAMMING AND PROCESS SEQUENCE

Once the workcell configuration was completed, robot trajectories and control logic were developed using ABB RAPID programming language. The objective was to create a reliable and efficient operational sequence capable of servicing the lathe machine autonomously.

The first step involved defining a series of robot targets corresponding to important process locations. These targets represent specific positions and orientations that the robot must reach during operation. The main targets included:

- Home position;
- Material pickup position;
- Approach position above the raw material;
- Lathe loading position;
- Lathe unloading position;
- Finished-part placement position.

The robot motion strategy combined two different movement types. Joint-interpolated motions (MoveJ) were employed for rapid travel between distant locations because they minimize cycle time. Linear-interpolated motions (MoveL) were used whenever precise tool movement was required, particularly during insertion and extraction of workpieces from the machine chuck (figure 5).

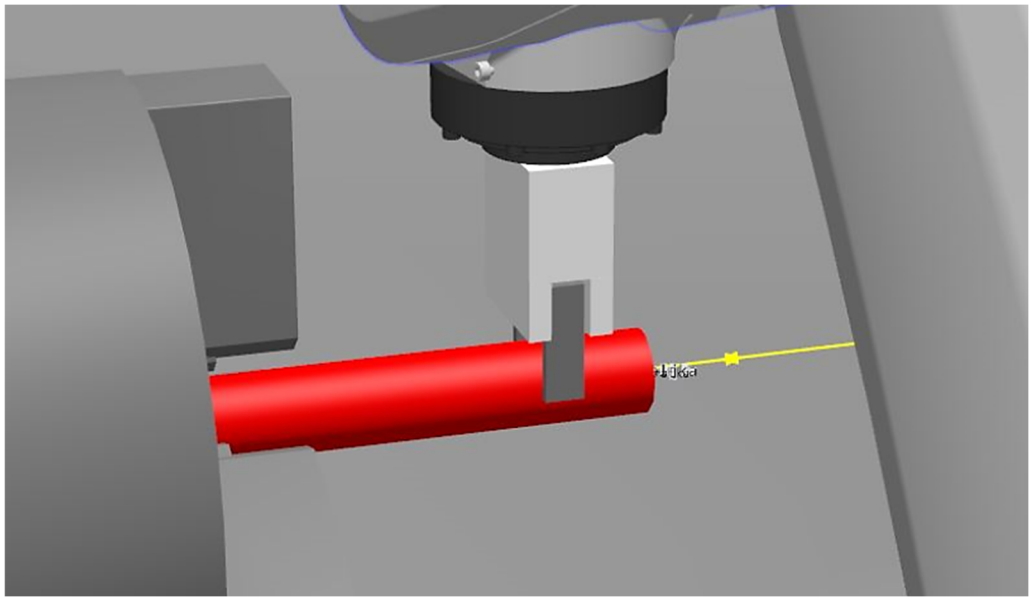


Figure 5: Insertion/extraction of workpieces from the machine

To coordinate the robot with external devices, several digital input and output signals were implemented. These signals enabled communication between the robot controller, machine tool, gripper, and conveyor system. The primary control signals were responsible for:

- Opening and closing the gripper;
- Opening and closing the machine door;
- Attaching and releasing workpieces;
- Resetting workpiece positions;
- Synchronizing robot and machine operations.

The complete production cycle begins with the robot positioned at its home location. The robot then moves toward the raw material storage area and approaches the workpiece using a controlled linear trajectory. Once the workpiece is grasped, the robot transports it toward the CNC lathe. Robot positions are shown in figure 6.

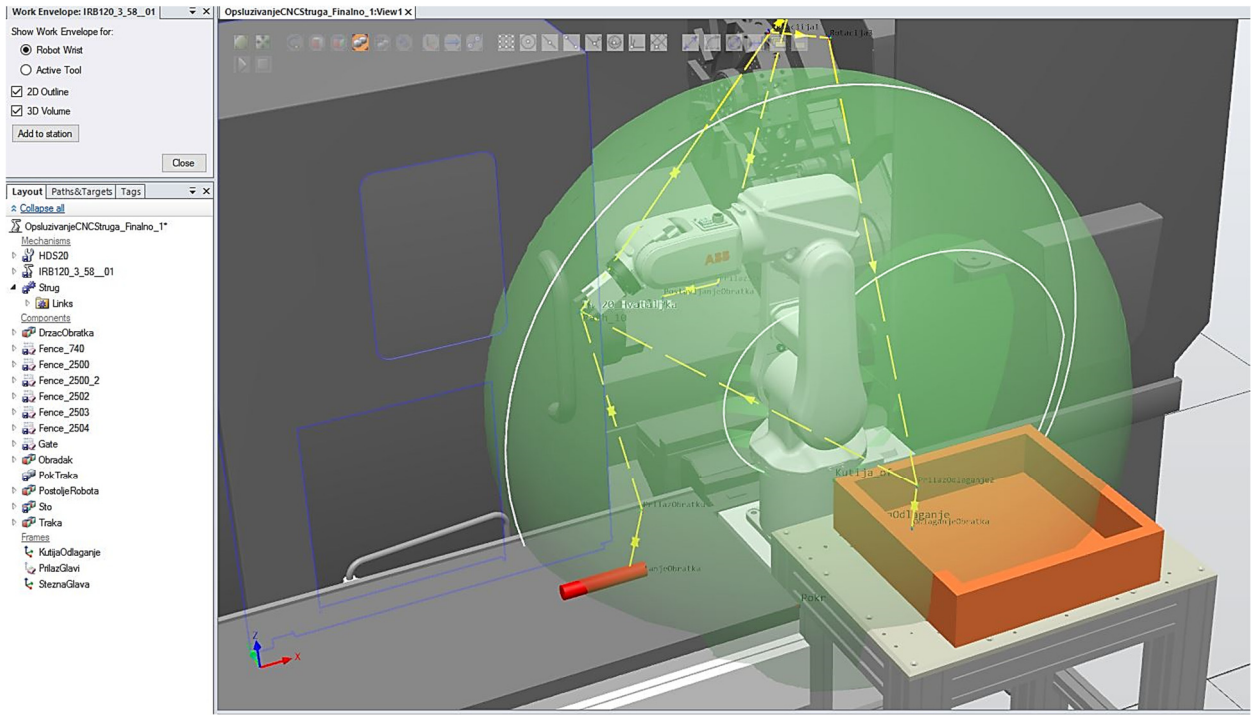


Figure 6: Robot with characteristic points and trajectories

Before loading the workpiece, the robot activates the door-opening sequence. After confirmation that the door has reached its open position, the robot inserts the workpiece into the chuck using carefully controlled linear motion. The gripper releases the part, and the robot retreats to a safe waiting position.

Following the machining operation, the door opens again, allowing the robot to remove the finished part. The workpiece is then transferred either to a conveyor system or to a designated storage container. After completing the unloading process, the robot returns to its home position and waits for the next production cycle.

5. SIMULATION RESULTS AND DISCUSSION

The developed robotic workcell was extensively tested within the RobotStudio simulation environment. Multiple simulation runs were conducted to evaluate the reliability of robot motion, verify synchronization between components, and identify potential operational issues.

The results confirmed that all programmed robot trajectories were successfully executed without collisions. The robot maintained safe distances from surrounding equipment while efficiently accessing each required process location.

The simulation demonstrated successful interaction between the robot and the machine tool throughout the complete servicing cycle. The implemented synchronization logic ensured that loading and unloading operations were performed only when the machine door was fully open and the machining process had been completed.

Workspace analysis showed that the selected robot configuration provides adequate reachability for all required tasks. Furthermore, trajectory optimization reduced unnecessary robot motion and contributed to improved cycle efficiency.

The virtual commissioning approach provided several important benefits. Potential design issues were identified and corrected before physical implementation, significantly reducing development costs and commissioning time. Additionally, the simulation environment enabled detailed evaluation of process performance without disrupting production or requiring expensive hardware modifications.

The results clearly demonstrate that robotic machine tending can provide substantial improvements in productivity, consistency, and operational safety compared with manual operation. The developed simulation therefore represents a valuable step toward practical deployment of the proposed automation solution.

6. CONCLUSION AND FUTURE WORK

This paper presented the design, programming, and simulation of a robotic machine-tending system for servicing a CNC lathe using an ABB IRB120 industrial robot. A complete virtual workcell was developed within ABB RobotStudio, including realistic models of the machine tool, robotic manipulator, gripper, conveyor system, and safety infrastructure.

The simulation successfully demonstrated the feasibility of fully automating the loading and unloading operations associated with CNC machining processes. Through careful trajectory planning, implementation of RAPID control logic, and synchronization of robot actions with machine operations, the system achieved reliable and collision-free performance.

The obtained results indicate that robotic integration can significantly improve manufacturing efficiency, reduce operator involvement in repetitive tasks, enhance workplace safety, and increase overall process consistency. Moreover, the use of simulation technology enables engineers to evaluate and optimize automation solutions before deployment, reducing both implementation risks and development costs.

Future research will focus on incorporating machine vision systems for automatic part localization, integrating advanced sensor technologies for adaptive process control, improving synchronization strategies between the robot and machining equipment, and developing a physical prototype of the workcell. Additional optimization techniques, including cycle-time reduction and intelligent scheduling algorithms, may further enhance the performance of the proposed manufacturing system.

ACKNOWLEDGEMENTS

The authors would like to express their sincere gratitude to the Ministry of Science, Technological Development and Innovation of the Republic of Serbia for its financial support in the realization of this research, based on Contract No. 451-03-34/2026-03/200132 concluded between the Faculty of Technical Sciences in Čačak and the aforementioned Ministry

REFERENCES

[1] R. A. Adebayo, N. C. Obiuto, I. C. Festus-Ikhuoria, and O. K. Olajiga, “Robotics in Manufacturing: A Review of Advances in Automation and Workforce Implications,” International Journal of Advanced Multidisciplinary Research and Studies, https://doi.org/10.62225/2583049X.2024.4.2.2549, (2024).

[2] J. Howard et al., “Industrial Robotics and the Future of Work,” American Journal of Industrial Medicine, https://doi.org/10.1002/ajim.23729, (2025).

[3] M. Soori, F. K. G. Jough, R. Dastres, and B. Arezoo, “Robotical Automation in CNC Machine Tools: A Review,” Acta Mechanica et Automatica, https://doi.org/10.2478/ama-2024-0048, (2024).

[4] Prof. C. C. Jadhav, Mr. A. Mr. R. Kale, K. Mr, M. G. Khandalkar, M. S. Ingale, and R. Bhandare, “Design and Manufacturing of Robotic Arm for Industrial Applications,” International Journal of Engineering Technology and Management Sciences, https://doi.org/10.46647/ijetms.2023.v07i04.032, (2023).

[5] R. G. Chittawadigi and D. E. Rendi, “Systematic Overview of Development of a Robot Kinematics Library for Motion Planning of Serial Robots Using Object Oriented Programming Concepts,” IEEE International Conference on Robotics and Biomimetics, https://doi.org/10.1109/robio64047.2024.10907363, (2024).

[6] M. Komák, E. Pivarčiová, and P. Herčút, “Machining of Printed Circuit Boards Using an Industrial Robot in a Simulation Environment,” Management Systems in Production Engineering, https://doi.org/10.2478/mspe-2025-0042, (2025).

SESSION C: PRODUCTION TECHNOLOGIES, CAE AND MANAGEMENT

Environmental impact analysis of laser cutting: A case study of thin stainless steel and aluminum alloy sheets

Miloš Madić[1], Milan Trifunović[1], Saša S. Nikolić[2*], Marko Milojković[2]

[1] University of Niš, Faculty of Mechanical Engineering in Niš, Niš, Serbia

[2] University of Niš, Faculty of Electronic Engineering, Niš, Serbia

ARTICLE INFO

[*] **Correspondence:** sasa.s.nikolic@elfak.ni.ac.rs

DOI: https://doi.org/10.46793/ET26.C03M **PAGES:** C19 – C24

ABSTRACT

The analysis of environmental impacts of manufacturing processes is essential for promoting cleaner and more sustainable production. The present study examined the total environmental impact in laser cutting of thin stainless steel and aluminum alloy sheets. To this aim, an integrated approach, which combines design of experiments methodology for parameter effect assessment, analytical-empirical models and life cycle assessment methods for resource consumption estimation and environmental impact evaluation, was proposed. In addition to the analysis of main laser cutting parameter effects on total environmental impact, development of regression models for predicting environmental impact, and analysis of contributions to the environmental impact, i.e., energy consumption of the laser cutting machine, assist gas consumption, and generated waste material, were also addressed. The presented approach is useful and receptive in order to quantify and predict the environmental impact of the selected laser cutting conditions directly during the process planning step.

KEYWORDS

Laser Fusion Cutting, Sustainability, Cleaner Production, Life Cycle Assessment.

1. INTRODUCTION

Alongside conventional, non-conventional technologies are being increasingly adopted in modern sheet-metal manufacturing [1]. Laser cutting technology, in particular, is widely used for contour cutting due to many advantages that bring multiple cost, quality, and efficiency benefits. Although laser cutting is a well-established technology, achieving superior performance in terms of quality, cost, and productivity remains challenging due to the complex and often nonlinear interactions among process parameters [2]. In addition, integration of an increasing number of performances is essential to maintain industrial competitiveness [3]. Therefore, process modelling is necessary to explain process behaviour, parameter effects, and their interactions, enabling effective process planning and optimization of the process. Given its complex multi-input, multi-output nature, laser cutting technology has been extensively studied with respect to various performance measures over many years. Numerous authors have recently addressed the analysis, modelling and optimization of quality characteristics, including kerf width [4,5], kerf deviation [6], kerf taper [7], surface roughness [8], dross formation [9], heat-affected zone [10], dimensional and shape accuracy [11], striations [12], microhardness [13], residual stresses [14], recast layer [15], surface morphology [16], etc.

Several studies have addressed the economic aspects [17-19], approaches for nesting and laser cutting path optimization [20-23], as well as productivity analysis of the process [24-26]. Because material processing technologies involve substantial consumption of energy, materials, and resources, the adoption of cleaner production and the promotion of environmental sustainability are essential [27]. Therefore, some studies have addressed sustainability aspects in laser cutting [28-35]. Although in comparison to conventional cutting technologies, laser cutting technology is more environmentally friendly [36], increasing awareness for sustainability highlights the importance of saving energy and material resources, reducing waste generation and minimizing environmental burden [33]. Therefore, the goal of the present study is the analysis of the effects of laser cutting parameters, which are critical to process planning, on the total environmental impact in laser cutting of thin stainless steel and aluminum alloy sheets. To this end, an integrated approach, based on

consideration of laser cutting parameter recommendation guidelines, application of design of experiments methodology and application of analytical and empirical models for resource consumption, as well as life cycle assessment (LCA) method, was developed. In addition to the analysis of main parameter effects on total environmental impact, development of regression models for predicting environmental impact, and analysis of contributions to the environmental impact, i.e., energy consumption of the laser cutting machine, assist gas consumption, and generated waste material, were also addressed.

2. APPLIED METHODOLOGY

The environmental impact of laser cutting was evaluated using the Eco-indicator 99 endpoint method [37]. It is a life cycle impact assessment (LCIA) method that aggregates environmental damage to human health, ecosystem quality, and resource depletion into a single environmental score expressed in eco-indicator points (mPts). Actually, LCA methods are the primary approach used to evaluate the environmental impacts of material cutting technologies [33].

A gate-to-gate system boundary was defined for a laser cutting operation of a high-power (4 kW) conventional CO_2 laser cutting machine (Prima Power Domino Evoluzione). Laser nitrogen cutting of stainless-steel sheet with a thickness of 1 mm was considered, while analysing energy consumption of the laser cutting machine, assist gas consumption during the process, and waste material. Process emissions and other consumables were found to have a negligible contribution [38] and were therefore excluded from the analysis.

In the estimation of these quantities, four process parameters, i.e., laser power, assist gas pressure, cutting speed and nozzle diameter, were considered. These parameters predominantly affect required laser cutting machine power for different laser output levels, assist gas consumption, melt volume and material waste generation and cycle time, which are the dominant contributors to the Eco-indicator 99 environmental impact of laser cutting. The estimation of these quantities was based on data and available models from referential literature [17,35,37-39].

In order to systematically analyse the effects of laser power, assist gas pressure, cutting speed and nozzle diameter on environmental impact, a full factorial experimental design was adopted. The levels of the input variables were determined by considering the industrial CO_2 laser cutting parameter recommendation guidelines [40]. For each laser experimental trial, i.e., a combination of laser cutting parameter values, one can calculate the contributions related to the energy consumption of the laser cutting machine, assist gas consumption during the process, and waste material. Ultimately, one can determine EI-99 single scores per unit cut length (mPts/m). The applied methodology for assessing the environmental impact in laser cutting of stainless steel and aluminum alloy sheets is given in Figure 1.

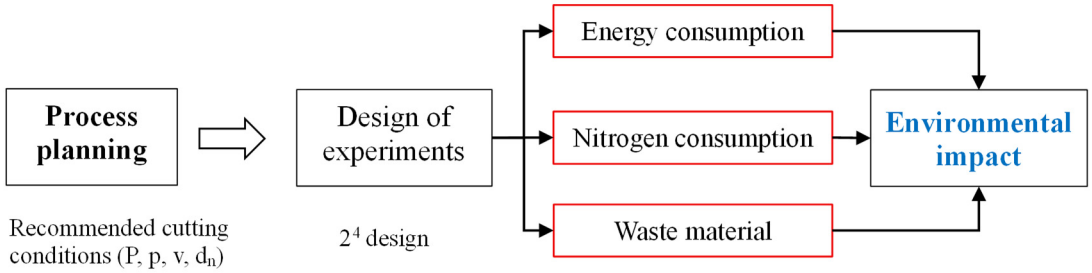
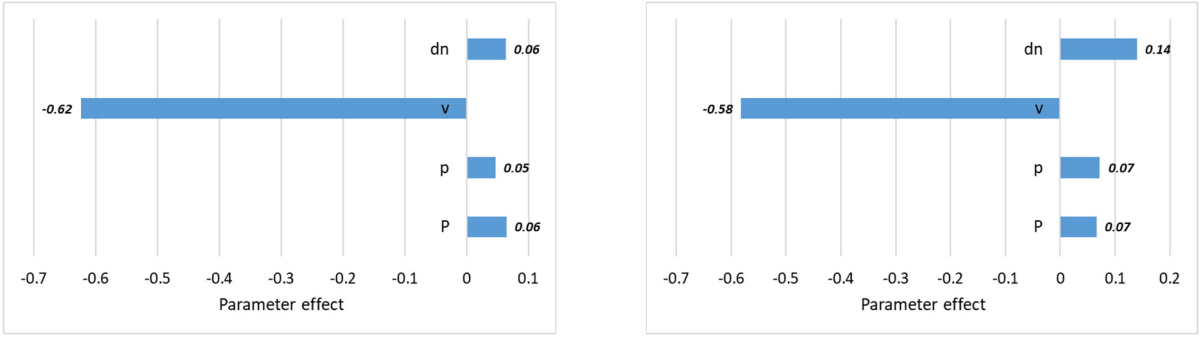


Figure 1: Applied approach for assessing the environmental impact in laser cutting

3. ANALYSIS OF RESULTS

The experimental results were analysed using standard procedures from the design of experiments methodology [41]. Main effects and interaction effects were evaluated using factorial analysis, and the significance of model terms was assessed based on their percentage contribution, calculated from the corresponding mean squares. Estimated parameter effects of laser power (P), nitrogen pressure (p), cutting speed (v) and nozzle diameter (d_n) on output (eco-indicator points per unit cut length) are given in Figure 2.



a) stainless steel b) aluminum alloy (AlMg3)

Figure 2: Estimated parameter effects on total environmental impact

As can be observed from Figure 2, cutting speed exhibits the greatest influence, indicating that environmental impact in laser nitrogen cutting of stainless steel and aluminum alloy sheets with a thickness of 1 mm is primarily governed by energy consumption per unit cut length. Nozzle diameter, laser power, and nitrogen pressure show very limited effects. While all main effects were statistically significant, all two-factor interactions were estimated to be negligible, except for the nozzle diameter and nitrogen pressure two-factorial interaction, which showed a nearly statistically significant contribution. While this suggests that the environmental impact of the laser cutting process is largely dominated by main effects, it should be noted that parameter interactions may significantly impact kerf geometry. In the present case, however, the kerf width was estimated using a simplified model that only considers sheet thickness. Laser power, nitrogen pressure, and nozzle diameter positively correlate with output, reflecting slight increases in environmental impact when these parameters are increased. On the other hand, cutting speed exhibits the strongest negative correlation, indicating that increasing cutting speed reduces the environmental impact.

Based on estimated by the output average and main effects of the considered laser cutting parameters, the regression model for predicting environmental impact (in mPts/m) in laser cutting of stainless steel and aluminum alloy can be written as:

$$EI_{score(stainless\ steel)} = 3.94 + 0.032 \cdot x_1 + 0.023 \cdot x_2 - 0.312 \cdot x_3 + 0.032 \cdot x_4 \quad (1)$$

$$EI_{score(aluminum\ alloy)} = 3.12 + 0.033 \cdot x_1 + 0.036 \cdot x_2 - 0.291 \cdot x_3 + 0.07 \cdot x_4 \quad (2)$$

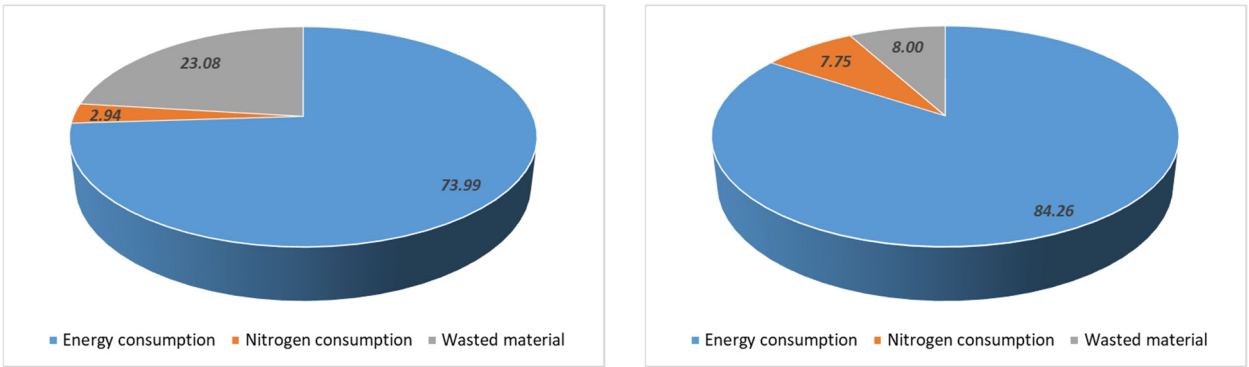
where the coded variables x_1, x_2, x_3 and x_4 represent laser power (P), nitrogen pressure (p), cutting speed (v) and nozzle diameter (d_n), respectively.

The obtained linear regression models predict the Eco-indicator 99 score as a function of laser cutting parameters. They can be used to estimate environmental impact for arbitrarily chosen laser cutting conditions, though within the studied range. In addition, they enable single and multi-objective process optimization when integrated with cost, productivity, or quality models.

The laser cutting conditions that minimize environmental impact can be easily identified. These conditions correspond to a combination of low laser power, high cutting speed, low nitrogen pressure, and small-diameter nozzles. Within experimental design points, minimal environmental impact is estimated to be 3.94 mPts/m for cutting stainless steel and 2.72 mPts/m in the case of cutting aluminum alloy. On the other hand, laser cutting conditions in which the highest amounts of energy and materials are being consumed result in an environmental impact of 4.35 mPts/m for stainless steel cutting and 3.58 mPts/m in the case of aluminum alloy. The estimated mean environmental impact for the entire experiment is 3.94 mPts/m for stainless steel cutting, while in the case of cutting aluminum alloy is 3.12 mPts/m, which is a reduction of about 20%. The lower environmental impact of the aluminum alloy is primarily due to the higher achievable cutting speed, which reduces the energy consumption per meter of cut.

It should be noted that the selection of laser cutting parameter combinations that minimize environmental impact must be done carefully and validated to ensure process quality. For instance, parameters should be verified by calculating severance energy to ensure sufficient energy input for full sheet penetration and reliable cuts with acceptable/required cut quality. This ensures that improvements in sustainability are not achieved at the expense of cutting quality performance.

The different contributions to the environmental impact, i.e., energy consumption of the laser cutting machine, assist gas consumption during the process, and waste material, are given in Figure 3. As can be seen, energy consumption by the laser cutting machine has the highest share in the total environmental impact of the laser cutting process. Due to the higher required gas flow and greater energy density for laser cutting of aluminum alloy, the relative contributions of energy consumption, nitrogen consumption, and waste material are somewhat different than those in the case of stainless steel. It should be noted that the relative contributions of energy consumption, nitrogen consumption, and waste material are not fixed and may change to a certain extent. In particular, increased gas flow requirements and higher energy power densities associated with cutting thicker sheets can alter the balance between these contributions.



a) stainless steel b) aluminum alloy (AlMg3)

Figure 3: Percentage contributions to environmental impact

The total environmental impact of one hour of productive operation using a 4 kW CO_2 laser cutting machine to cut 1 mm thick stainless steel was estimated to be 1640 mPts. This estimate, together with the trends observed in the previous analysis, is well aligned with the findings of Duflou et al. [35], who reported a total environmental impact of 1756 mPts for one hour of CO_2 laser cutting of mild steel St37-2. In that study, energy consumption accounted for approximately 66% of the total impact, followed by material waste (23%) and process gas consumption (11%), which is consistent with the dominance of energy consumption observed in the present study. Duflou et al. [34] further reported that a proper selection of the laser cutting machine, selective activation and deactivation of non-required subsystems during both productive and non-productive modes, as well as nesting optimization, can lead to tangible savings in material consumption and associated environmental impacts.

4. CONCLUSION

The present study focused on the analysis of energy consumption, assist gas consumption and generated waste material, which altogether contribute to the environmental impact in laser cutting of thin stainless steel and aluminum alloy sheets. An integrated approach, which combines the design of experiments methodology for parameter effect assessment and analytical-empirical models and LCA methods for resource consumption estimation and environmental impact evaluation, was proposed. Based on the conducted analyses, there are several main conclusions:

- Laser cutting of thin aluminum alloy sheets is more environmentally friendly, in comparison to stainless steel cutting, and this can be attributed to lower energy requirements and associated impacts per meter of cut.
- The environmental impact of both materials is strongly controlled by cutting speed.
- Process optimization for environmental performance should prioritize cutting speed, followed by nozzle diameter and laser power. It should be noted, however, that the effect of given main laser cutting parameter changes with regard to the choice of the material type, and it also changes with the change in sheet thickness, with the expectation that with an increase in sheet thickness, the effect of nozzle diameter would be pronounced.
- The absolute share of the resource consumption for both materials is similar, although due to different required gas flow, energy density levels and physical properties, relative contributions are somewhat different.
- The derived prediction models can serve as a guideline for evaluating the environmental impact of arbitrarily chosen laser cutting regimes.
- It should be noted that the obtained results may vary, given the variation intervals for the main laser cutting parameter levels, which were in the present study taken following industrial recommendations.

The presented approach is useful and receptive in order to quantify and predict the environmental impact of the selected laser cutting conditions directly during the process planning step. One limitation of the present study is that kerf width was assumed to be a function of only sheet thickness and was assumed to be the same for both materials under the same cutting conditions.

ACKNOWLEDGEMENTS

This research was financially supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-34/2026-03/200109).

This work has been supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia [Grant Number: 451-03-34/2026-03/200102].

REFERENCES

[1] T. Šajnović, D. Rodić, M. Gostimirović, M. Milošević, M. Đurđev, and D. Lukić, “Optimization of laser cutting processes in the function of surface roughness”, Annals of Faculty Engineering Hunedoara – International Journal of Engineering, Vol. 22(2), pp. 21-26, (2024)

[2] D. Rodić, M. Sekulić, B. Savković, M. Madić, and M. Trifunović, “Integration of RSM and machine learning for accurate prediction of surface roughness in laser processing”, Applied Sciences, Vol. 15(13), Article 7064, 16 pages, https://doi.org/10.3390/app15137064, (2025)

[3] M. Ramard, R. Laniel, M. Miroir, and O. Kerbrat, “Method for quantifying the criticality of laser cutting defects: influence of morphologies on design parameters”, Journal of Manufacturing and Materials Processing, Vol. 9(3), Article 34, https://doi.org/10.3390/jmmp9030091, (2025)

[4] K. Aydın and L. Uğur, “Prediction of kerf and groove widths in co$_2$ laser cutting process of pmma using experimental and machine learning methods”, Experimental Techniques, Vol. 49(5), pp. 873-886, https://doi.org/10.1007/s40799-025-00786-5, (2025)

[5] A. I. Kusuma and Y. M. Huang, “Performance comparison of machine learning models for kerf width prediction in pulsed laser cutting”, The International Journal of Advanced Manufacturing Technology, Vol. 123(7-8), pp. 2703-2718, https://doi.org/10.1007/s00170-022-10348-3, (2022)

[6] A. Sharma, P. Joshi, and K. K. Saxena, “Empirical modeling and optimization of kerf characteristics in Nd-YAG laser cutting of Al 6061-T6 sheet”, Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering, Vol. 238(2), pp. 607-618, https://doi.org/10.1177/09544089221112071, (2024)

[7] J. D. Kechagias, K. Ninikas, P. Stavropoulos, and K. Salonitis, “A generalised approach on kerf geometry prediction during CO_2 laser cut of PMMA thin plates using neural networks”, Lasers in Manufacturing and Materials Processing, Vol. 8(3), pp. 372-393, https://doi.org/10.1007/s40516-021-00152-4, (2021)

[8] E. Nas and B. Özlü, “Experimental and statistical investigation of the effect of input parameters on output parameters in laser cutting of stainless steel material”, Ironmaking & Steelmaking: Processes, Products and Applications, Online First, https://doi.org/10.1177/03019233251349871, (2025)

[9] S. F. Nabavi, M. H. Farshidianfar, A. Farshidianfar, and S. Marandi, “Dross formation modeling in the laser beam cutting process using energy-based and gas-based parameters”, The International Journal of Advanced Manufacturing Technology, Vol. 120(11-12), pp. 8169-8184, https://doi.org/10.1007/s00170-022-09019-0, (2022)

[10] A. Prajapati, G. Norkey, and G. D. Gautam, “Optimization of heat affected zone in laser cutting of Kevlar-29 fiber composite using hybrid response surface based grey wolf optimization (RSGWO) algorithm”, Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, Vol. 236(17), pp. 9622-9638, https://doi.org/10.1177/09544062221096557, (2022)

[11] K. Nadolny, M. Romanowski and P. Sutowski, “Assessing the technological quality of abrasive water jet and laser cutting processes by geometrical errors and a multiplicative indicator”, Measurement, Vol. 217, Article 113060, 16 pages, https://doi.org/10.1016/j.measurement.2023.113060, (2023)

[12] R. d. C. M. Sales-Contini, I. Amaral, R. R. Lucas, A. M. G. Pinto, R. D. S. G. Campilho, L. L. Magalhães, and F. J. G. da Silva, “Optimizing laser cutting parameters for enhanced surface quality in mild and stainless steel: an experimental study”, Engineering Reports, Vol. 7(7), Article e70283, https://doi.org/10.1002/eng2.70283, (2025)

[13] J. Selech, G. Burzynski, D. Tibebe, D. Ulbrich, and P. Banas, “Influence of laser cutting parameters on the microhardness, roughness, and microstructure of AISI 304, S355J2, and AlMg3 alloys”, Materials, Vol. 19(2), Article 240, https://doi.org/10.3390/ma19020240, (2026)

[14] M. Soori, F. K. G. Jough, and B. Arezoo, “Residual stress and surface roughness minimization in laser cutting of 304L stainless steel”, Scientia Iranica, Online First, https://doi.org/10.24200/sci.2024.62207.8240, (2024)

[15] Y. R. A. Pradana, C. H. A. Rahman, R. Afrianto, A. Aminnudin, and W. Wahono, “Recast layer and HAZ formation from laser cutting of SUS 304 plate under cutting speed variation”, AIP Conference Proceedings, Vol. 3110(1), Article 020057, https://doi.org/10.1063/5.0204931, (2024)

[16] X. Zhang, H. Zhou, B. Zhou, R. Wang, H. Han, X. Jiang, and M. Li, “Surface morphology and kerf quality during fiber laser cutting of high volume fraction SiC particles-reinforced aluminum matrix composites”, Journal of Materials Engineering and Performance, Vol. 32(13), pp. 5906-5918, https://doi.org/10.1007/s11665-022-07526-5, (2023)

[17] M. Madić, M. Radovanović, B. Nedić, and M. Gostimirović, “CO_2 laser cutting cost estimation: mathematical model and application”, International Journal of Laser Science: Fundamental Theory and Analytical Methods, Vol. 1(2), pp. 169-183, (2018)

[18] H. A. Eltawahni, M. Hagino, K. Y. Benyounis, T. Inoue, and A.G. Olabi, “Effect of CO_2 laser cutting process parameters on edge quality and operating cost of AISI316L”, Optics & Laser Technology, Vol. 44(4), pp. 1068-1082, https://doi.org/10.1016/j.optlastec.2011.10.008, (2012)

[19] M. Rimašauskas, and A. Bargelis, “Development of intelligent model for estimating manufacturing cost in sheet metalworking”, Proceedings of the 7th International DAAAM Baltic Conference Industrial Engineering, Tallinn (Estonia), 22-24 April 2010, pp. 251-256, (2010)

[20] Z. Chen, J. Lu, J. Ma, and X. Liao, “An uninterrupted cut path optimization method considering connecting bridge constraints and its integration application”, The International Journal of Advanced Manufacturing Technology, Vol. 139(3-4), pp. 1959-1977, https://doi.org/10.1007/s00170-025-15802-6, (2025)

[21] A. F. Tavaeva, and A. A. Petunin, “Optimization of technological processes for nesting parts: using a database of sheet cutting cost parameters”, Software & Systems, Vol. 38(1), pp. 150-156, https://doi.org/10.15827/0236-235X.149.150-156, (2025)

[22] B. A. Junior, G. N. de Carvalho, M. C. Santos, P. R. Pinheio, and J. W. L. Celedonio, “Evolutionary algorithms for optimization sequence of cut in the laser cutting path problem”, Applied Sciences, Vol. 13(18), Article 10133, https://doi.org/10.3390/app131810133, (2023)

[23] P. Qu and F. Du, “Improved particle swarm optimization for laser cutting path planning”, IEEE Access, Vol. 11, pp. 4574-4588, https://doi.org/10.1109/ACCESS.2023.3236006, (2023)

[24] Y. T. A. Hwang and J. M. Yang, “Laser cutting time estimate for sheet metal parts of various geometries by machine learning approach”, Proceedings of the 13th International Conference on Industrial Engineering and Operations Management, Manila, (Philippines), 7-9 March 2023, pp. 1378-1387, https://doi.org/10.46254/AN13.20230410, (2023)

[25] M. Y. Tabar, M. Hashemzadeh, and A. R. Rejani, "CO_2 laser cutting of reinforced polyester (CGFRP and GFRP) sheets: an experimental investigation into specific point energy, cutting volume efficiency and material removal rate (MRR)", Journal of Manufacturing Processes, Vol. 97, pp. 137-147, https://doi.org/10.1016/j.jmapro.2023.04.059, (2023)

[26] M. Madić, S. Mladenović, M. Gostimirović, M. Radovanović, and P. Janković, "Laser cutting optimization model with constraints: maximization of material removal rate in CO_2 laser cutting of mild steel", Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture, Vol. 234(10), pp. 1323-1332, https://doi.org/10.1177/0954405420911529, (2020)

[27] F. Le Bourhis, O. Kerbrat, L. Dembinski, J. Y. Hascoet, and P. Mognol, "Predictive model for environmental assessment in additive manufacturing process", Procedia CIRP, Vol. 15, pp. 26-31, https://doi.org/10.1016/j.procir.2014.06.031, (2014)

[28] N. Goffin, L. C. R. Jones, J.R. Tyrer, J. Ouyang, P. Mativenga, L. Li, and E. Woolley, "Industrial energy optimisation: a laser cutting case study", International Journal of Precision Engineering and Manufacturing-Green Technology, Vol. 11(3), pp. 765-779, https://doi.org/10.1007/s40684-023-00563-y, (2024)

[29] A. Krause, T. Dannerbauer, S. Wagenmann, G. Tjaden, R. Ströbel, J. Fleischer, A. Albers, and N. Bursac, "Enhancing efficiency and environmental performance of laser-cutting machine tools: an explainable machine learning approach", Procedia CIRP, Vol. 130, pp. 1674-1679, https://doi.org/10.1016/j.procir.2024.10.299, (2024)

[30] G. Sembou and G. Besseris, "Lean-and-green datacentric engineering in laser cutting: non-linear orthogonal multivariate screening using Gibbs sampling and Pareto frontier", Processes, Vol. 12(2), Article 377, https://doi.org/10.3390/pr12020377, (2024)

[31] M. d. Q. Murad, J. A. G. de Sousa, M. G. de Ataíde, G. H. N. Fernandes, L.L.T. Vaughan, and F.R. Cangue, "Selected metal cutting processes with focus on sustainability", Caderno Pedagógico, Vol. 20(9), pp. 3845-3878, https://doi.org/10.54033/cadpedv20n9-010, (2023)

[32] Y. He, H. Xie, Y. Ge, Y. Lin, Z. Yao, B. Wang, M. Jin, J. Liu, X. Chen, and Y. Sun, "Laser cutting technologies and corresponding pollution control strategy", Processes, Vol. 10(4), Article 732, https://doi.org/10.3390/pr10040732, (2022)

[33] E. G. Öztürk and Z. C. Ayturan, "Life cycle impact assessment and economic evaluation of material cutting technologies", Proceedings of the 5th International Conference on Energy and Environment: bringing together Engineering and Economics, Porto (Portugal), 2-3 June 2022, pp. 532-537, (2022)

[34] J. R. Duflou, K. Kellens, R. Renaldi, and W. Dewulf, "Environmental performance of sheet metal working processes", Key Engineering Materials, Vol. 473, pp. 21-26, https://doi.org/10.4028/www.scientific.net/KEM.473.21, (2011)

[35] J. R. Duflou, K. Kellens, T. Devoldere, W. Deprez, and W. Dewulf, "Energy related environmental impact reduction opportunities in machine design: case study of a laser cutting machine", International Journal of Sustainable Manufacturing, Vol. 2(1), pp. 80-98, https://doi.org/10.1504/IJSM.2010.031621, (2010)

[36] J. Haider and M. S. J. Hashmi, "8.02 - Health and environmental impacts in metal machining processes", Comprehensive Materials Processing, Vol. 8, Chapter 2, pp. 7-33, https://doi.org/10.1016/B978-0-08-096532-1.00804-9, (2014)

[37] M. Goedkoop and R. Spriensma, " Eco-indicator 99 manual for designers", Ministry of Housing, Spatial Planning and the Environment, The Hague (The Netherlands), (2000)

[38] K. Kellens, "Energy and resource efficient manufacturing: unit process analysis and optimisation", PhD Thesis, KU Leuven, (Belgium), (2013)

[39] M. Goedkoop, "The eco-indicator 99 methodology", Journal of Life Cycle Assessment, Japan, Vol. 3(1), pp. 32-38, https://doi.org/10.3370/lca.3.32, (2007)

[40] J. Berkmanns and M. Faerber, "Laser cutting: LASERLINE technical", Linde Gas LLC, Surrey, (England), (2008)

[41] D. C. Montgomery, "Design and analysis of experiments, Ninth Edition", John Wiley & Sons, Hoboken, NJ (USA), (2017)

SESSION C: PRODUCTION TECHNOLOGIES, CAE AND MANAGEMENT

Three-objective Pareto optimization of GMAW process parameters using a modified force balance model

Mladen Rasinac[1*], Mišo Bjelić[1], Tanja Dulović[1], Jovana Perić[1], Marina Ivanović[1]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering, Kraljevo, Serbia

ARTICLE INFO

[*] **Correspondence:** rasinac.m@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.C04R **PAGES:** C25 – C35

ABSTRACT

In this paper, a three-objective optimization model was developed for the preliminary selection of welding current and electrode wire extension length in the gas metal arc welding (GMAW) process, to achieve a droplet size characteristic for the projected spray metal transfer mode. The model is based on a modified force balance model, i.e., the MFBM approach, in which the droplet detachment condition is determined from the balance between the total detaching force and the surface tension force. A correction of the effective electrode wire neck radius, based on the balance between the pressure due to surface tension and the electromagnetic pinch pressure, was introduced to account for the tapering/necking effect at higher current values. The optimization variables are the welding current and electrode wire extension length, while the objective functions are defined to maximize the melting rate, minimize the deviation of the normalized droplet diameter from the target value, and minimize the normalized Joule/stick-out index. The obtained Pareto front indicates a clear compromise between productivity, droplet size control, and thermal loading of the electrode wire extension. A dynamic diagnostic layer was also introduced to verify the regularity of the droplet growth and detachment cycle for selected Pareto solutions.

KEYWORDS

GMAW, Metal transfer, MFBM, Pareto optimization, Projected spray transfer, Electrode wire neck radius

1. INTRODUCTION

Gas metal arc welding (GMAW) is one of the most widely used arc welding processes due to its high productivity, automation capability, and continuous feeding of filler material. The stability of this process largely depends on the mode of transfer of molten metal from the electrode wire tip to the weld pool. The metal transfer mode affects arc stability, spatter formation, weld bead geometry, penetration, and the overall quality of the welded joint [1-3]. Kim and Eagar showed that different transfer modes have different effects on arc stability, penetration, spatter, and weld bead quality, while Węglowski et al. discussed droplet size, transfer frequency, and droplet kinematics in the GMAW process.

In the GMAW process, the short-circuiting, globular, and spray metal transfer modes are most commonly considered. A particularly important region between the globular and spray modes is projected spray transfer, in which the droplets are approximately spherical, and their diameter may be approximately equal to or slightly larger than the electrode wire diameter. This mode is technologically interesting because it provides more stable metal transfer and less spatter compared with the globular mode, while at the same time representing a transitional region that may offer a more favorable compromise between transfer stability, productivity, and the thermal loading of the electrode wire compared with modes that require higher current values. However, projected spray transfer occurs within a relatively narrow range of welding parameters [1], [4].

For the preliminary selection of welding parameters, models based on the balance of forces acting on the droplet are useful. The classical static force balance model (SFBM) takes into account gravitational, electromagnetic, drag, and surface tension forces, whereas the modified force balance model, i.e., the MFBM model, additionally includes the force due to the momentum flux effect. Therefore, the MFBM model is suitable for estimating droplet size and detachment frequency in the globular and projected spray regions [4]. On the other hand, lumped-parameter dynamic models enable

additional analysis of droplet growth, oscillation, and detachment [5], while models that include tapering and necking effects indicate the importance of changes in the electrode wire tip geometry at higher currents [6].

The aim of this paper is to develop a numerically efficient three-objective optimization model for selecting the welding current and electrode wire extension length in the GMAW process. The model is based on the MFBM approach, with a pressure-balance correction of the effective electrode wire neck radius. The optimization variables are the welding current I and electrode wire extension length l_e, while the optimization is formulated as a three-objective Pareto problem. A dynamic diagnostic layer is additionally used for verification of selected Pareto solutions.

2. MODEL AND OPTIMIZATION PROCEDURE

2.1. Modified Force Balance Model

In this paper, the droplet detachment condition was described using the MFBM model. Immediately before detachment, the droplet is approximated by a spherical shape, while detachment is defined by comparing the total detaching force F_T and the surface tension force. The total force acting on the droplet is given by [4]:

$$F_T = F_g + F_{em} + F_d + F_{mf} \tag{1}$$

where F_g denotes the gravitational force, F_{em} the electromagnetic force, F_d the drag force exerted by the shielding gas or plasma, and F_{mf} the force resulting from the momentum flux effect. The droplet detachment condition can be written as:

$$F_T \geq F_s \tag{2}$$

where F_s denotes the surface tension force that retains the droplet at the electrode wire tip. The classical force balance model (SFBM) considers gravitational, electromagnetic, drag, and surface tension forces, whereas the MFBM model additionally includes the force F_{mf}. The inclusion of F_{mf} represents the main difference between the MFBM and the classical force balance model.

A schematic representation of the forces acting on the molten droplet and the corresponding lumped-parameter dynamic analogy is shown in Figure 1.

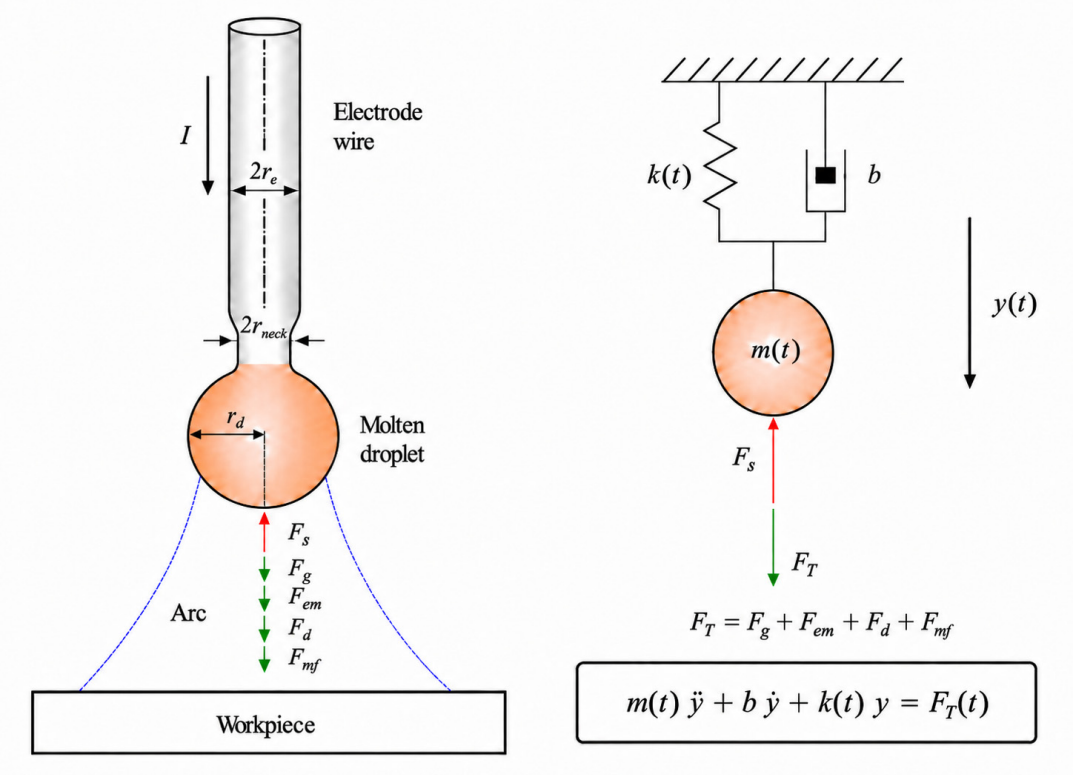


Figure 1: Forces acting on the droplet and dynamic MFBM analogy

The droplet radius is defined as:

$$r_d = \frac{D}{2} \tag{3}$$

where D is the current droplet diameter. Assuming a spherical droplet shape, the droplet mass is calculated from the expression:

$$m_d = \frac{4}{3}\pi\rho_e r_d^3 \tag{4}$$

where ρ_e is the density of the molten electrode wire metal. The gravitational force is then:

$$F_g = m_d g \tag{5}$$

where g is the gravitational acceleration.

The electromagnetic force in the MFBM model is calculated according to the expression:

$$F_{em} = \frac{\mu_0 I^2}{4\pi}\left[\ln\left(\frac{r_d \sin\theta}{r_e}\right) - \frac{1}{4} - \frac{1}{1-\cos\theta} + \frac{2}{(1-\cos\theta)^2}\ln\left(\frac{2}{1+\cos\theta}\right)\right] \tag{6}$$

where μ_0 is the magnetic permeability of vacuum, r_e is the electrode wire radius, and θ is the angle of the droplet surface region covered by the arc. In this paper, for the argon-rich shielding atmosphere, $\theta = \pi/2$ was adopted [4].

The shielding gas or plasma drag force is determined by the expression:

$$F_d = \frac{1}{2} C_d A_d \rho_p v_p^2 \tag{7}$$

where C_d is the drag coefficient, A_d is the effective droplet area exposed to the flow, ρ_p is the density of the plasma or shielding gas, and v_p is the characteristic flow velocity. The effective droplet area exposed to the flow is calculated as:

$$A_d = \pi(r_d^2 - r_e^2) \tag{8}$$

To define the momentum flux force, the following dimensionless ratio is introduced $\alpha = r_d/r_e$.

For argon-rich shielding gases, the momentum flux force is calculated as:

$$F_{mf} = \frac{\mu_0 I^2}{4\pi}\left[\frac{1}{\alpha^2} - \left(1 - \sqrt{1 - \frac{1}{\alpha^2}}\right)^2\right] \tag{9}$$

where $\alpha \geq 1$ [4].

The surface tension force in the classical MFBM model takes the form:

$$F_s = 2\pi r_e \gamma \tag{10}$$

where γ is the surface tension of the molten metal. In this paper, due to the inclusion of the tapering/necking effect, the electrode wire radius r_e in the expression for the surface tension force is replaced by the effective neck radius r_{neck}. Therefore, in the developed model, the following is used [6]:

$$F_s = 2\pi r_{neck} \gamma \tag{11}$$

This equation represents the link between the MFBM model and the introduced pressure-balance correction of the e-lectrode tip geometry. It is important to emphasize that the electromagnetic force and the momentum flux force remain in the Anzehaee/Haeri form, while the effective neck radius is used only in the surface tension force.

2.2. Pressure-Balance Correction of the Effective Neck Radius

At higher current values, a more pronounced electromagnetic action occurs in the electrode tip region, which may cause a reduction in the effective droplet neck. To take this effect into account, the effective neck radius r_{neck} is introduced. This approach is inspired by the work of Kang and Chung, in which the influence of the tapering effect and the necking region on the dynamic detachment of a droplet in the GMAW process is considered [6].

The raw value of the effective neck radius is determined from the equilibrium condition between the pressure due to surface tension and the electromagnetic pinch pressure [6]:

$$r_{neck,raw} = \frac{8\pi^2 \gamma r_e^2}{\mu_0 I^2} \tag{12}$$

Since the effective neck radius cannot be greater than the electrode wire radius, an upper bound is introduced:

$$r_{neck} \leq r_e \tag{13}$$

Additionally, for numerical stability and to avoid a physically unrealistic small neck radius, a lower bound is introduced:

$$r_{neck} \geq \eta_{min} r_e \tag{14}$$

where the following value is adopted in this paper: $\eta_{min} = 0.30$.

The final expression for the effective neck radius is:

$$r_{neck} = \max\left[\eta_{min} r_e, \min\left(r_e, \frac{8\pi^2 \gamma r_e^2}{\mu_0 I^2}\right)\right] \tag{15}$$

The critical current at which the effective neck radius begins to decrease is obtained from the condition $r_{neck,raw} = r_e$, i.e.:

$$I_{crit} = \sqrt{\frac{8\pi^2 \gamma r_e}{\mu_0}} \tag{16}$$

For the model parameters used, the following value was obtained: $I_{crit} = 221.4$ A.

In this paper, the complete Kang-Chung neck rupture criterion is not used; instead, the tapering/necking effect is introduced through a correction of the surface tension force within the MFBM detachment condition. In other words, droplet detachment is still determined according to the condition (2), while the geometrical effect of the neck is taken into account through the effective radius r_{neck}. This limitation is important because the developed model remains within the framework of a numerically simple MFBM approach and does not transition into a complete dynamic model of neck rupture.

2.3. Estimation of Droplet Diameter and Detachment Frequency

The electrode melting rate is described by the expression [4,5], [7]:

$$MR = c_1 I + c_2 \rho_{res} I^2 l_e \tag{17}$$

where MR is the volumetric melting rate, ρ_{res} is the linear electrical resistance, and c_1, c_2, are model coefficients. The first term describes the contribution associated with the current and heating in the arc region, while the second term describes the contribution of resistive, i.e., Joule heating of the electrode wire extension. Similar expressions for the melting rate and its importance in the modelling and control of the GMAW process are discussed in the works of Ersoy et al. and Thomsen [5], [7]. The mass flow rate of the molten metal is calculated as:

$$\dot{m} = \rho_e MR \tag{18}$$

The droplet diameter immediately before detachment, D_{dbd}, is determined by numerically solving the equation [4]:

$$F_T(D_{dbd}, I) - F_s(I) = 0 \tag{19}$$

The equation given above was solved as a one-dimensional root-finding problem over the droplet radius interval $r_d \in [1.001 r_e, 6 r_e]$. This equation represents the basic condition for estimating the droplet size immediately before detachment. In this equation, the total force F_T is calculated as the sum of the gravitational, electromagnetic, drag, and momentum flux forces, while the surface tension force is calculated using the effective neck radius.

After determining D_{dbd}, the droplet mass immediately before detachment is calculated as [4]:

$$m_{dbd} = \frac{4}{3}\pi \rho_e \left(\frac{D_{dbd}}{2}\right)^3 \tag{20}$$

A portion of the molten mass remains at the electrode tip after detachment. This residual mass is modelled by the expression:

$$m_{dr} = \frac{1}{2} m_{dbd} \left[1 + \frac{1}{1+\exp[-(k_4 + k_5 I)]}\right] \tag{21}$$

The mass of the detached droplet is:

$$m_{dd} = m_{dbd} - m_{dr} \tag{22}$$

Based on the mass of the detached droplet, its diameter is determined as:

$$D_{dd} = 2\left(\frac{3 m_{dd}}{4\pi \rho_e}\right)^{1/3} \tag{23}$$

The mean droplet detachment frequency is calculated from the ratio between the mass flow rate of the molten metal and the mass of the detached droplet:

$$f_d = \frac{\rho_e MR}{m_{dd}} \tag{24}$$

In this paper, the normalized droplet diameter immediately before detachment D_{dbd}/d_w is used as the main quantity for controlling the metal transfer mode, where $d_w = 2r_e$. The quantities D_{dd} and f_d are used as additional output indicators for interpreting the Pareto solutions.

2.4. Formulation of the Three-Objective Optimization Problem

The optimization variables are the welding current and the electrode wire extension length:

$$\mathbf{x} = [I, l_e] \tag{25}$$

The following ranges of the optimization variables were used in this paper:

$$180 \text{ A} \le I \le 270 \text{ A}; 18 \text{ mm} \le l_e \le 30 \text{ mm} \tag{26}$$

The first objective function refers to the maximization of the melting rate. Since the multi-objective genetic algorithm solves a minimization problem, the maximization of the melting rate is transformed into the minimization of the negative normalized value:

$$f_1^0 = -\frac{MR}{MR_{ref}} \tag{27}$$

The second objective function defines the deviation of the normalized droplet diameter immediately before detachment from the target value:

$$f_2^0 = \left[\frac{\left(\frac{D_{dbd}}{d_w}\right) - \left(\frac{D_{dbd}}{d_w}\right)_{target}}{\left(\frac{D_{dbd}}{d_w}\right)_{target}}\right]^2 \tag{28}$$

In this paper, the following target value was adopted [4]:

$$\left(\frac{D_{dbd}}{d_w}\right)_{target} = 1.20 \tag{29}$$

The third objective function is introduced as a normalized Joule/stick-out index:

$$f_3^0 = \frac{I^2 l_e}{I_{max}^2 l_{e,max}} \tag{30}$$

This quantity is dimensionless and represents an indicator of resistive heating of the electrode wire extension. It does not represent the total heat input into the base material, but rather an additional criterion used to limit an excessive increase in the welding current and electrode wire extension length.

The relative error of the normalized droplet diameter is defined as:

$$\varepsilon_D = \frac{\frac{D_{dbd}}{d_w} - 1.20}{1.20} \cdot 100\ \% \tag{31}$$

A penalty term was introduced into the optimization model for solutions that fall outside the desired projected/transitional region. First, the ratio between the momentum flux force and the electromagnetic force is defined:

$$R_F = \frac{F_{mf}}{|F_{em}|} \tag{32}$$

The penalty term can then be written as:

$$p_{proj} = w_{proj}\left[\max\left(0, R_{F,min} - R_F\right)^2 + \max\left(0, D_{min} - \frac{D_{dbd}}{d_w}\right)^2 + \max\left(0, \frac{D_{dbd}}{d_w} - D_{max}\right)^2\right] \tag{33}$$

where the following values were used in the numerical model: $R_{F,min} = 1, D_{min} = 1.00, D_{max} = 1.30$.

The active objective functions passed to the multi-objective genetic algorithm are therefore:

$$\begin{aligned} f_1 &= f_1^0 + p_{proj} \\ f_2 &= f_2^0 + p_{proj} \\ f_3 &= f_3^0 + p_{proj} \end{aligned} \tag{34}$$

The multi-objective optimization problem can be written in the form:

$$\min_{\mathbf{x}} \mathbf{f}(\mathbf{x}) = \min_{I, l_e}[f_1(I, l_e), f_2(I, l_e), f_3(I, l_e)] \tag{35}$$

with the defined bounds for I and l_e.

For each candidate pair (I, l_e), the model first calculates the effective neck radius and then determines D_{dbd} by solving (19). The obtained droplet size is used to calculate the detached droplet diameter, detachment frequency, objective functions, and penalty term. The resulting objective vector is used in the multi-objective genetic algorithm, while the dynamic layer is applied afterwards only for diagnostic verification of selected Pareto solutions.

Stricter conditions were used to identify the recommended operating window:

$$|\varepsilon_D| \leq 3\ \%;\ \frac{F_{mf}}{|F_{em}|} \geq 1;\ 1.15 \leq \frac{D_{dbd}}{d_w} \leq 1.25 \tag{36}$$

Characteristic solutions are then selected from the Pareto set: the maximum productivity solution, the target droplet solution, the minimum Joule/stick-out index solution, the three-objective compromise solution, and the recommended operating window solution.

To select the compromise solution from the Pareto set, the method of the closest point to the ideal point in the normalized objective space was used. Since all three objective functions are minimization functions, the ideal point after normalization has the coordinates $[0; 0; 0]$. For each Pareto solution i, $\min - \max$ normalization of each objective function is first performed:

$$f_{n,k}^{(i)} = \frac{f_k^{(i)} - \min(f_k)}{\max(f_k) - \min(f_k)} \tag{37}$$

The Euclidean distance from the ideal point is then calculated:

$$S_i = \sqrt{\left(f_{n,1}^{(i)}\right)^2 + \left(f_{n,2}^{(i)}\right)^2 + \left(f_{n,3}^{(i)}\right)^2} \tag{38}$$

The compromise solution is defined as:

$$i_{comp} = \arg\min_i S_i \tag{39}$$

The recommended solution within the operating window is selected using the same criterion, but only among the Pareto solutions that satisfy the conditions given in (36), therefore:

$$i_{OW} = \arg\min_{i \in OW} S_i \tag{40}$$

This procedure is not a curvature-based knee point method, but a normalized ideal-point compromise selection. However, in practical terms, it identifies a solution that represents a good balance between productivity, the target droplet size, and the Joule/stick-out index.

2.5. Dynamic Diagnostic Layer for Droplet Growth and Detachment

In addition to the static force-balance calculation, a dynamic diagnostic layer was introduced. This layer does not participate in the objective functions, but is used to verify the regularity of droplet growth and detachment for the selected Pareto solutions. The droplet is described as a variable-mass spring-damper system [5]:

$$m(t)\ddot{y}(t) + b\dot{y}(t) + k(t)y(t) = F_T(t) \tag{41}$$

where $y(t)$ is displacement of the droplet and b is linear damping coefficient.

The droplet mass increases due to the inflow of molten metal:

$$m(t+\Delta t) = m(t) + \dot{m}\Delta t \tag{42}$$

The instantaneous droplet radius is determined from the instantaneous mass:

$$r_d(t) = \left[\frac{3m(t)}{4\pi\rho_e}\right]^{1/3} \tag{43}$$

while the instantaneous droplet diameter is:

$$D_d(t) = 2r_d(t) \tag{44}$$

At each time step, the MFBM forces are recalculated using the instantaneous droplet diameter $D_d(t)$.

The effective stiffness of the system changes with the droplet mass:

$$k(t) = \max[k_{min}, k_0 + c_k(m(t) - m_0)] \tag{45}$$

The coefficient describing the change in stiffness with mass is defined as:

$$c_k = \frac{\dot{k}_{ref}}{\dot{m}_{ref}} \tag{46}$$

where the reference mass flow rate is:

$$\dot{m}_{ref} = \rho_e\left(c_1 I_{ref} + c_2 \rho_{res} I_{ref}^2 l_{e,ref}\right) \tag{47}$$

Droplet detachment is assumed to occur when the MFBM force-balance condition defined in (2) is satisfied.

After detachment, the mass is reset to the residual mass [4]:

$$m_{res} = \frac{1}{2} m(t_{det})\left[1 + \frac{1}{1+\exp[-(k_4 + k_5 I)]}\right] \tag{48}$$

The dynamic layer is used only as a diagnostic tool and not as an active optimization criterion. This use of the dynamic layer is based on the concept of modelling the droplet as a lumped-parameter system, which has been used in the literature for the analysis of oscillations, resonant behavior, and metal transfer stability [5], [8].

The main model constants and coefficients used in the numerical analysis are summarized in Table 1.

Table 1: Main model constants and coefficients used in the numerical analysis [1], [4,5]

r_e	ρ_e	g	μ_0	C_d	ρ_p	v_p	γ	c_1	c_2	ρ_{res}	k_4
0.6	7860	9.81	$4\pi \cdot 10^{-7}$	0.44	1.784	10	1.3	$3.3 \cdot 10^{-10}$	$0.78 \cdot 10^{-10}$	0.43	-0.364
mm	kg/m^3	m/s^2	H/m	–	kg/m^3	m/s	N/m	m^3/(sA)	m^3/(sΩA^2)	Ω/m	–
k_5	MR_{ref}	I_{max}	$l_{e,max}$	w_{proj}	b	k_{min}	k_0	$\dot{k}_{ref}$	m_0	I_{ref}	$l_{e,ref}$
0.0062	10^{-7}	270	30	1.0	$3 \cdot 10^{-4}$	0.01	0.2	-0.8	$5.0 \cdot 10^{-7}$	200	20
A^{-1}	m^3/s	A	mm	–	Ns/m	N/m	N/m	N/(ms)	kg	A	mm

3. RESULTS AND DISCUSSION

The optimization was performed for a welding current range of 180 to 270 A and an electrode wire extension length range of 18 to 30 mm. The target value of the normalized droplet diameter immediately before detachment was $D_{dbd}/d_w = 1.20$. The multi-objective genetic algorithm was run with a population size of 60 for 60 generations. After duplicate removal using tolerances of 0.02 A for current and 0.02 mm for electrode wire extension length, 22 Pareto-optimal solutions were retained. The obtained Pareto set shows a compromise between three requirements: increasing the melting rate, maintaining the target droplet size, and reducing the normalized Joule/stick-out index.

Figure 2 shows the Pareto front in the space of the three active objective functions. The three-dimensional distribution confirms the absence of a single dominant solution. The solutions with the highest productivity are located at higher values of welding current and electrode wire extension length, but at the same time they have a higher value of the third objective function and a lower D_{dbd}/d_w ratio. In contrast, the solutions with a lower Joule/stick-out index have lower productivity and, in some cases, produce droplets larger than the target value. This confirms that the selection of process parameters cannot be based solely on maximizing the melting rate, but must also include droplet size control and limitation of the thermal loading of the electrode wire extension.

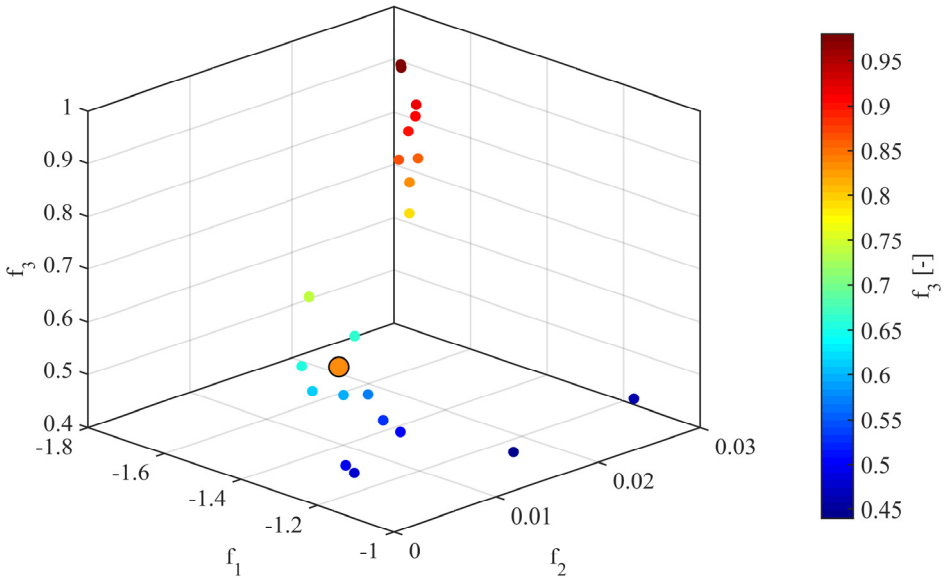


Figure 2: Pareto front in the space of three active objectives

The map of the normalized droplet diameter in the $I - l_e$ space, shown in Figure 3, indicates that the value of D_{dbd}/d_w changes predominantly with the welding current. The contour corresponding to the target value ($D_{dbd}/d_w = 1.20$) is located in the region around 231 A, while the influence of the electrode wire extension length is less pronounced. On the other hand, the melting rate map, given in Figure 4, shows that MR increases with both increasing current and increasing electrode wire extension length. Therefore, the solutions with the highest productivity are located close to the upper boundary of the analyzed space; however, such solutions deviate from the target droplet size and have an increased Joule/stick-out index.

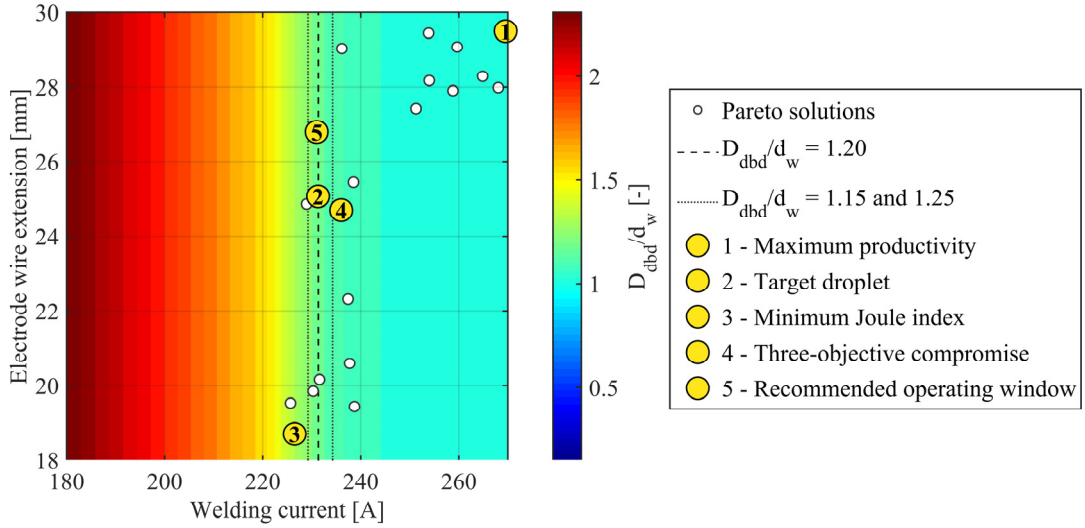


Figure 3: Map of the $I - l_e$ space with contours of D_{dbd}/d_w

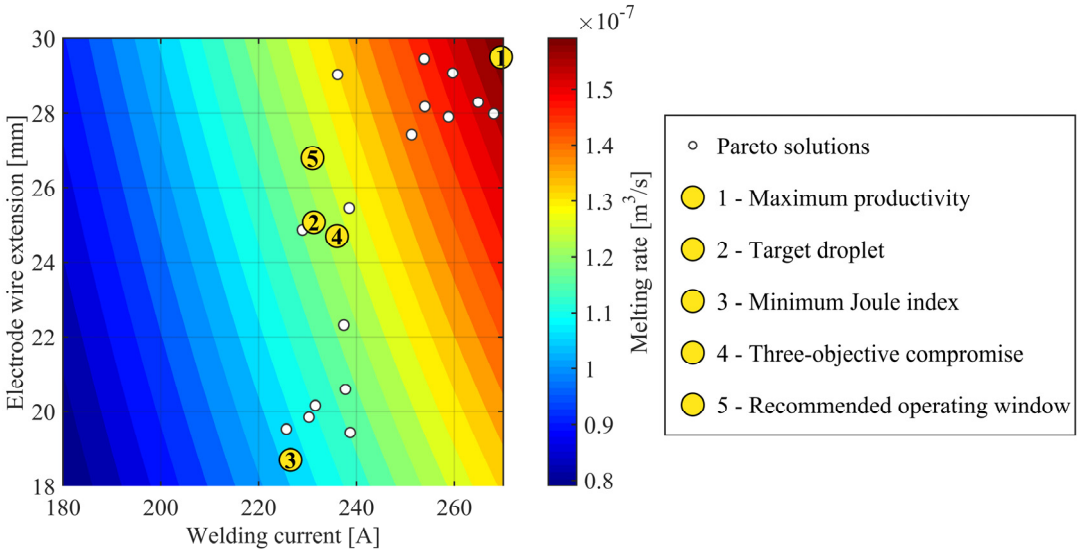


Figure 4: Map of the $I - l_e$ space with contours of melting rate

The map of the normalized Joule/stick-out index f_3, shown in Figure 5, indicates the expected increase in this quantity with increasing welding current and electrode wire extension length. This result confirms the role of the third objective function as an additional constraint that prevents the selection of solutions with excessive resistive heating of the electrode wire extension. Therefore, the maximum productivity solutions, although they provide the highest melting rate, are not necessarily the most favorable from the standpoint of the overall compromise between all three objectives.

This trend is consistent with the known behavior of metal transfer in the GMAW process: increasing the current generally reduces the droplet size and increases the transfer frequency, while the electrode wire extension length affects the melting rate and the thermal state of the electrode [1,2]. Kang and Chung further showed that, above the transition current, the tapering of the electrode tip reduces the surface-tension effect and increases the relative influence of electromagnetic forces, which supports the need to account for both droplet size and force balance in the model [6].

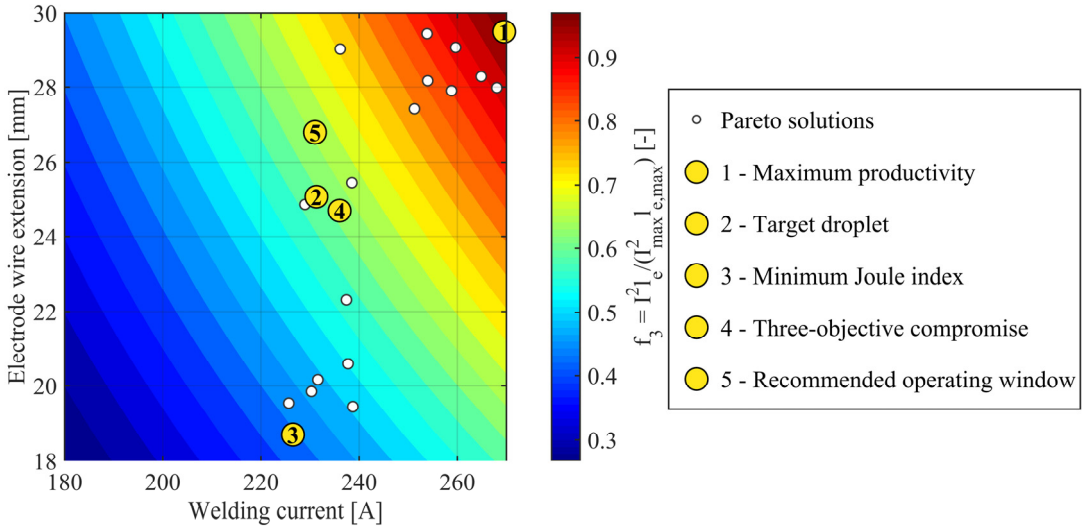


Figure 5: Map of the $I - l_e$ space with contours of f_3

Five representative solutions were selected from the Pareto set: the maximum productivity solution, the target droplet solution, the minimum Joule/stick-out index solution, the compromise solution, and the solution within the recommended operating window. Their values are presented in Table 2. The target droplet solution shows the best agreement with the prescribed value ($D_{dbd}/d_w = 1.20$), while the maximum productivity solution represents the upper limit of MR, but with a clear trade-off in droplet size and Joule/stick-out index. The solution with the minimum Joule/stick-out index has the lowest thermal loading of the electrode wire extension, but produces a droplet larger than the target value. The compromise solution represents a balanced choice in the three-objective space, while the solution within the recommended operating window maintains the droplet size close to the target value with slightly higher productivity compared with the strict target droplet solution.

Table 2: Representative Pareto solutions

Solution	I [A]	l_e [mm]	$\frac{D_{dbd}}{d_w}$ [−]	ε_D [%]	$MR \cdot 10^{-7}$ [m³/s]	f_d [Hz]	f_3 [−]
Max. productivity	269.6	29.5	1.02	−15.32	1.61	1590.8	0.98
Target droplet	231.3	25.1	1.2	−0.01	1.21	607.8	0.61
Min. Joule index	226.5	18.7	1.34	11.37	1.07	379.4	0.44
Three-obj. comprom.	236.0	24.7	1.13	−6.01	1.24	764.7	0.63
Recom. oper. window	231.0	26.8	1.21	0.50	1.24	612.0	0.65

The force contribution analysis, shown in Figure 6, indicates that the electromagnetic force (F_{em}) and the momentum flux force (F_{mf}) are dominant in the representative solutions. The gravitational force (F_g) and the drag force (F_d) have a considerably smaller contribution to the total detaching force. This force relationship confirms the suitability of the MFBM model in the analyzed region, since the inclusion of the momentum flux force represents the key difference compared with the classical force balance model. At higher currents, the influence of electromagnetic effects increases, leading to a reduction in droplet size and an increase in detachment frequency.

The recommended operating window is shown in Figure 7. This representation shows that the operating window is located within a relatively narrow current range, close to the contour corresponding to the target value of the normalized droplet diameter. Most of the high-productivity Pareto solutions are located outside the recommended operating window, because they do not simultaneously satisfy the conditions related to proximity to the target droplet size and the constraints defined through the force ratio $R = F_{mf}/|F_{em}|$. Therefore, the solution within the recommended operating window can be considered the most favorable practical choice when the priority is a stable and controlled droplet size with acceptable productivity. Figure 8 shows the selection of the compromise solution in the normalized objective space. The selected solution has the smallest Euclidean distance from the ideal point and therefore represents a balanced trade-off between productivity, droplet size, and the Joule/stick-out index, rather than an extreme solution with respect to a single objective.

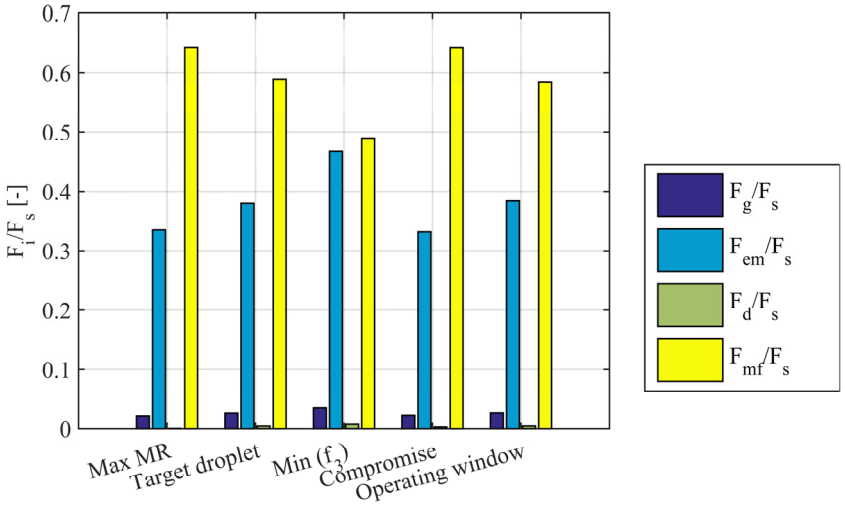


Figure 6: Force contributions for representative Pareto solutions

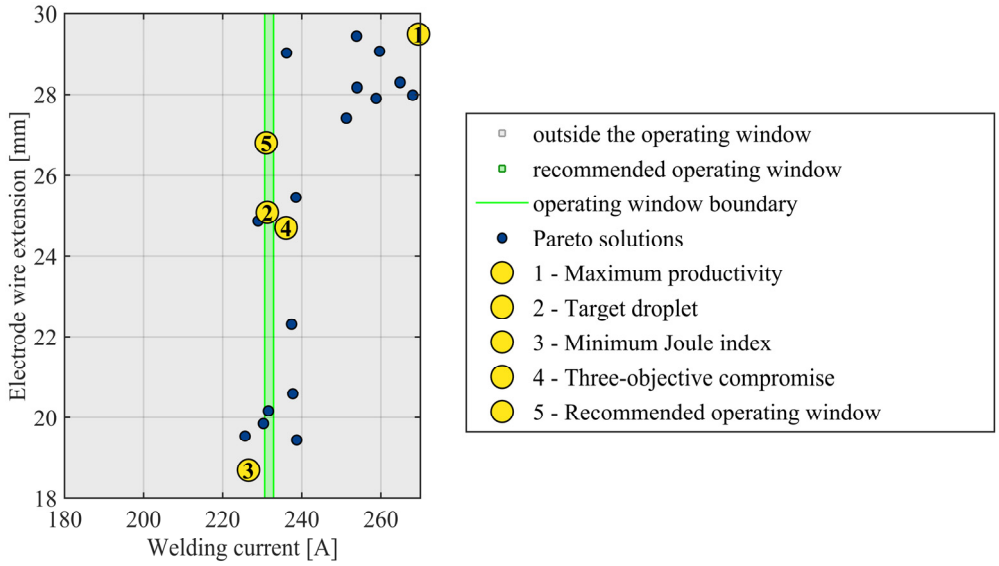


Figure 7: Recommended operating window and Pareto solutions

An additional verification of the selected solutions was performed using the dynamic diagnostic layer. The simulation was carried out for a total simulation time of $T_{sim} = 0.5$ s, with a time step of $\Delta t = 5 \cdot 10^{-5}$s as recommended in [4].

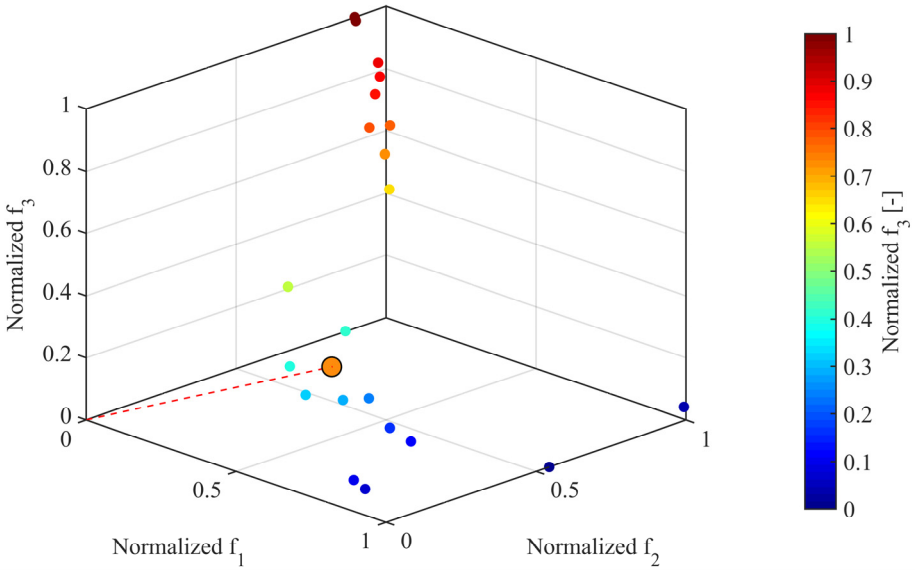


Figure 8: Compromise solution in the normalized three-objective space

The time variation of the droplet diameter for the compromise solution, shown in Figure 9, indicates the repetition of the growth and detachment cycle. After the condition defined by the MFBM criterion is reached, the droplet detaches, the mass is reset to the residual value, and a new cycle begins. The time variation of the forces for the compromise solution, shown in Figure 10, indicates the periodic approach of the total detaching force to the surface tension force, confirming the consistency of the implemented detachment condition. The repetition of the growth, detachment, and mass-reset cycles indicates regular behavior of the selected solutions within the introduced diagnostic model.

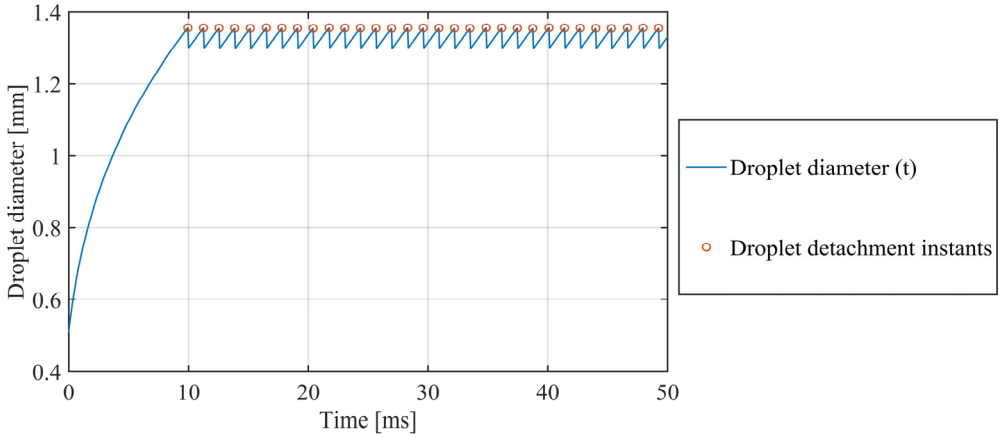


Figure 9: Time variation of the droplet diameter in the MFBM diagnostic layer

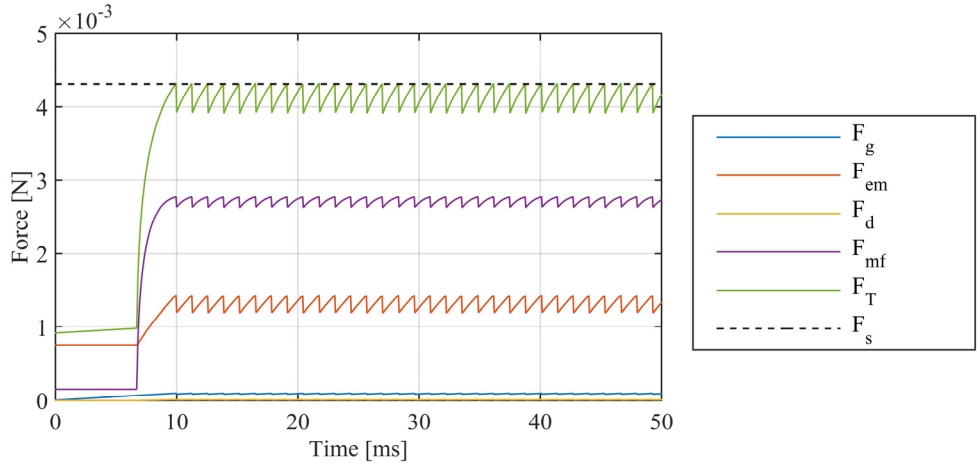


Figure 10: Time variation of forces in the MFBM diagnostic layer

Based on the obtained results, it can be concluded that the most favorable region for achieving the target droplet size is located around $I \approx 231$ A. Increasing the electrode wire extension length in this region enables an increase in the melting rate, but with a moderate increase in the Joule/stick-out index. Therefore, the recommended operating window solution at $I = 231.0$ A and $l_e = 26.8$ mm is identified as the most favorable practical solution, because it simultaneously satisfies the condition of proximity to the target droplet size, provides higher productivity than the strict target solution, and remains within an acceptable range of the third objective function.

4. CONCLUSIONS

In this paper, a three-objective optimization model was developed for the preliminary selection of welding current and electrode wire extension length in the GMAW process. The model is based on a modified force balance model (MFBM), in which droplet detachment is determined from the equilibrium condition between the total detaching force and the surface tension force. Compared with the basic MFBM model, a pressure-balance correction of the effective electrode wire neck radius was introduced to account for the influence of the tapering/necking effect at higher current values.

The optimization variables were the welding current I and the electrode wire extension length l_e, while the objective functions were defined to simultaneously maximize the melting rate, minimize the deviation of the normalized droplet diameter immediately before detachment from the target value, and minimize the normalized Joule/stick-out index. The obtained Pareto set, consisting of 22 non-duplicate solutions, shows a clear compromise between productivity, droplet size control, and thermal loading of the electrode wire extension.

The solution within the recommended operating window represents the most favorable practical choice, because it maintains the droplet size close to the target value while providing slightly higher productivity compared with the strict target droplet solution. The force contribution analysis showed that the electromagnetic force and the momentum flux force dominate in the representative solutions, while the gravitational force and the drag force are considerably smaller, confirming the suitability of the MFBM model in the analyzed projected/transitional metal transfer region.

The dynamic diagnostic layer showed regular droplet growth and detachment cycles for the selected Pareto solutions, while being used only as an additional verification tool and not as an active optimization criterion. The proposed model therefore represents a numerically efficient and physically interpretable approach for preliminary GMAW parameter selection. Future work should include experimental validation using high-speed imaging, current and voltage measurements, and comparison of the calculated droplet size and detachment frequency with experimental data.

ACKNOWLEDGEMENTS

The authors express their gratitude to the Ministry of Science, Technological Development and Innovation of the Republic of Serbia for supporting this research (Contract Nos. 451-03-33/2026-03/200108 and 451-03-34/2026-03/200108). This research contributes to the United Nations 2030 Agenda for Sustainable Development, specifically to Sustainable Development Goal 9: Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation.

REFERENCES

[1] Y-S. Kim and T. W. Eagar, “Analysis of metal transfer in gas metal arc welding”, Welding Journal, Vol. 72, pp. 269-278, (1993)

[2] M. S. Węglowski, Y. Huang, and Y. M. Zhang, “An investigation of metal transfer process in GMAW, Engineering Transactions, Vol. 56(4), pp. 345-362, https://doi.org/10.24423/ENGTRANS.189.2008, (2008)

[3] E. Uddin, U. Iqbal, N. Arif, and S. R. Shah, “Analysis of metal transfer in gas metal arc welding”, AIP Conference Proceedings, 24 July 2019, https://doi.org/10.1063/1.5114003, (2019)

[4] M. M. Anzehaee and M. Haeri, “Estimation and control of droplet size and frequency in projected spray mode of a gas metal arc welding (GMAW) process”, ISA Transactions, Vol. 50(3), pp. 409-418, https://doi.org/10.1016/j.isatra.2011.02.004, (2011)

[5] U. Ersoy, E. Kannatey-Asibu, and S. J. Hu, “Analytical modeling of metal transfer for GMAW in the globular mode”, Journal of Manufacturing Science and Engineering, Vol. 130(6), pp. 91-98, https://doi.org/10.1115/1.3006317, (2008)

[6] M. S. Kang and H. Chung, “Dynamic force balance model considering tapering effect in gas metal arc welding”, Journal of Materials Processing Technology, Vol. 257, pp. 79-87, https://doi.org/10.1016/j.jmatprotec.2018.02.029, (2018)

[7] J. S. Thomsen, “Advanced control methods for optimization of arc welding”, PhD Thesis, Aalborg University (Denmark), (2004)

[8] X. Zhang, H. Gao, and G. Zhang, “Current-independent metal transfer by utilizing droplet resonance in gas metal arc welding”, Journal of Materials Processing Technology, Vol. 279(17), https://doi.org/10.1016/j.jmatprotec.2019.116571, (2020)

SESSION C: PRODUCTION TECHNOLOGIES, CAE AND MANAGEMENT

Investigation of the influence of cutting parameters on noise emission during face turning of beech wood

Stefan Pajović[1], Vladan Grković[1], Jovana Perić[1], Aleksandra Petrović[1*], Branko Radičević[1]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Kraljevo, Republic of Serbia

ARTICLE INFO

[*] **Correspondence:** petrovic.a@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.C05P **PAGES:** C37 – C43

ABSTRACT

The paper presents the results of an experimental study of the influence of workpiece overhang length, feed rate and depth of cut on the equivalent sound pressure level during face turning of beech wood. Data analysis and mathematical model development were carried out using the Response Surface Methodology in the Design-Expert software package. The results of statistical analysis showed that the greatest influence on the equivalent sound pressure level is exerted by the depth of cut, followed by workpiece overhang, while the influence of feed rate is less pronounced. Based on the obtained data, a quadratic mathematical model was developed to predict the value of the equivalent sound pressure level. Additional experimental tests confirmed good agreement between the measured and model-predicted values. The obtained results can serve as a basis for selecting optimal process parameters in order to increase process quality and productivity during face turning of beech wood.

KEYWORDS

Face turning, Equivalent sound pressure level, Cutting parameters, Response Surface Methodology

1. INTRODUCTION

During machining processes, dynamic interactions between the machine tool, cutting tool and workpiece generate vibrations accompanied by characteristic sound emission. According to numerous studies, the main sources of vibration during turning operations on a lathe are the electric engine, the gearing between the engine and the gearbox, the existence of some worn parts inside the gearbox and the cutting process itself [1]. During the cutting process, the tool comes into contact with the material, resulting in chip removal which leads to the occurrence of vibrations [2]. Vibrations are frequently accompanied by characteristic noise pattern [1].

Several machining parameters, namely feed rate, depth of cut, cutting speed, workpiece diversity (material and hardness) and tool characteristics (tool material, cutting edge geometry, chip angle, etc.), affect sound pressure level in turning [3]. Besides the cutting parameters, the main factors that may influence the acoustic response during turning are the length of the workpiece zone which is not additionally supported or fixed, tool overhang length, the tool's active wear state, the type of tool tip and the the tool clamping method in the tool holder [1].

Some studies have reported the correlation of sound pressure level with surface roughness, accuracy of the workpiece part geometry and cutting power consumption [4]. Sound measurements are used in machine tools and in-process monitoring to detect and diagnose issues such as tool wear, tool breakage and chatter [5] - [9]. Monitoring changes in sound pressure level is proven useful for determining the optimal parameters of the cutting process and achieving maximum productivity, as it enables us to predict potential machine component failures and identify errors in the production process [10], [2]. Other studies [6] demonstrated the efficiency of using the sound generated during cutting by lathe for the adaptive metalworking control.

Mainly, the number of studies that considered sound pressure level as a criterion for machining operation is limited and focused on optimization of cutting parameters based on acoustic emissions in the field of metal processing industry. Keeping in mind economic and commercial importance and technical and technological complexity of wood processing industry, wood machining studies are not sufficiently represented, as in the case of metal processing industry.

Wood machining differs significantly from the machining of other materials. Wood is an anisotropic and heterogeneous material, so the machining responses largely depend on the direction of tool movement relative to the direction of wood grain. Studies have shown that adjusting cutting parameters can affect the surface quality of the processed wood [9].

Although numerous studies have investigated vibration and sound emission during metal cutting processes, considerably less attention has been devoted to *LAeq* analysis during wood turning operations, especially in the case of small-scale CNC machining systems and face turning configurations. Furthermore, the influence of machining parameters on *LAeq* during wood machining has not been sufficiently investigated, especially the influence of workpiece overhang length during face turning operations.

The aim of this study is to investigate the influence of feed rate, depth of cut and workpiece overhang length on *LAeq* during CNC face turning of beech wood using an HSS (high-speed steel) cutting tool. Statistical analysis and multiple linear regression modeling were employed to determine the significance of machining parameters and to develop predictive models for *LAeq* estimation. Additionally, confirmation experiments were performed to validate the developed regression models.

2. MATERIALS AND METHODS

2.1. Workpiece Material

The workpiece material used in this study is wood, which is a non-homogeneous and anisotropic material. Its mechanical and physical properties significantly depend on the grain orientation, density and moisture content. These characteristics strongly influence the cutting process, especially vibration behavior and sound emission.

The workpiece material used in this study was European beech wood (*Fagus sylvatica*), classified as a hardwood material. Beech wood is characterized by relatively high density, good machinability and relatively homogeneous structure [11]. The average density of the material is approximately 720 kg/m³ under air-dried conditions.

In this study, wooden bar blank with diameter of 8 mm is used. During the face turning operation, cutting was predominantly performed perpendicular to the grain direction.

2.2. Machine Tool and Cutting Tool

Machining experiments were performed on a UNI-DREH CNC lathe equipped with two CNC-controlled axes and stepper drives (2A, 1.8°). The machine positioning accuracy is 0.05 mm, while the maximum positioning speed is 300 mm/min. The spindle speed is constant and its value is 3333 rpm. The machine is intended for machining wood, plastics, wax and soft non-ferrous materials [12], [13]. Face turning operations were carried out using a high-speed steel (HSS) cutting tool with a shank size of 3.5 × 3.5 mm.

2.3. Measurement Equipment

Sound pressure measurements were performed using a Brüel & Kjær Type 2270 dual-channel sound level meter and analyzer. The device is a Class 1 measurement system intended for high-precision acoustic measurements in industrial and environmental applications. The analyzer supports broadband frequency measurements in the range from 4.2 Hz to 22.4 kHz using the Brüel & Kjær Type 4189 microphone [14]. The microphone was placed at a height of 1.5 m and at a distance of 1 m from the sound source.

In this study, the equivalent continuous A-weighted sound pressure level (*LAeq*) was used as the response parameter.

2.4. Experimental Procedure

Face turning operation of wooden bar blank with diameter of 8 mm was performed using an HSS cutting tool on educational CNC lathe.

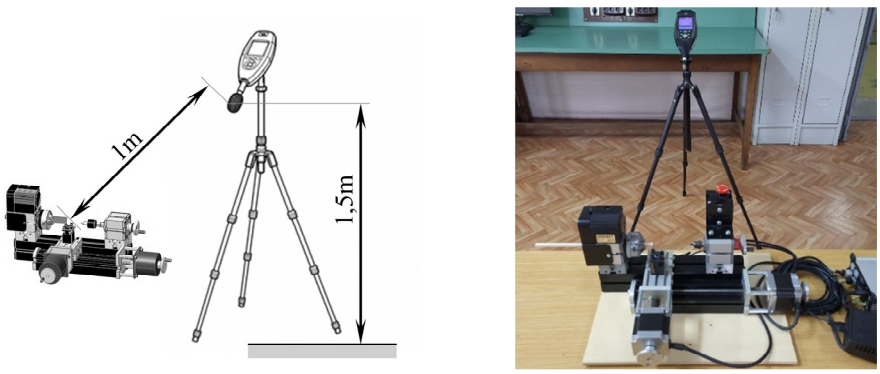


Figure 1: Experimental setup

Constant cutting parameters in this experiment were: spindle speed (3333 rpm) and workpiece diameter (8 mm). Workpiece material, tool and method of tool clamping in the tool holder also haven’t been changed.

Sound pressure measurements were performed using sound level meter during each cutting processes including time when tool is approaching to workpiece and exiting from cutting zone. Each measurement was repeated three times.

2.5. Experimental Design

Input parameters (feed rate, depth of cut and overhang length) were varied at three levels, chosen according to trial experiments and to equipment limitations, which are shown in Table 1.

Table 1: Input parameters and their levels

Factor	Symbol	Levels		
Overhang length of workpiece	*l* (mm)	15	25	35
Feed rate	*f* (mm/min)	50	100	150
Depth of cut	*a* (mm)	0,5	1	1,5

3. RESULTS AND DISCUSSION

3.1. Experimental Results

Experiments were conducted for different combinations of workpiece overhang, feed rate and depth of cut, with three levels adopted for each parameter. Each experimental combination was repeated three times to increase the reliability of the measurements and reduce the influence of random deviations. The obtained experimental *LAeq* values are shown in Table 2.

Table 2: Experimental results

Exp. no	*l* (mm)	*f* (mm/min)	*a* (mm)	*LAeq* (dB)		
				1st measurement	2nd measurement	3rd measurement
1	15	50	0,5	73	72,4	73,3
2	15	50	1	75,7	75,8	75,8
3	15	50	1,5	78	77,8	77,5
4	15	100	0,5	74,3	73,7	73,3
5	15	100	1	76,9	75,8	76,1
6	15	100	1,5	78,2	78,2	77,8
7	15	150	0,5	74,5	72,3	72,5
8	15	150	1	75,1	74,8	76,2
9	15	150	1,5	76,2	76,5	75,8
10	25	50	0,5	74,2	73,6	73,6
11	25	50	1	76,3	76	76,2
12	25	50	1,5	78,3	77,7	78,3
13	25	100	0,5	74,1	74,6	75
14	25	100	1	76,7	75,4	76,4
15	25	100	1,5	77,9	77,3	77,9
16	25	150	0,5	73,3	72,6	74,8
17	25	150	1	75,2	76,9	75
18	25	150	1,5	76,3	76	76,6
19	35	50	0,5	75,2	76,4	75,4
20	35	50	1	78,6	79,9	79,3
21	35	50	1,5	80,2	80,1	80,1
22	35	100	0,5	75,3	76,2	76,5
23	35	100	1	77,8	77,4	77,8
24	35	100	1,5	79,9	79,7	78
25	35	150	0,5	72,1	76,2	76,2
26	35	150	1	77,3	76,8	75,6
27	35	150	1,5	78,2	76,1	79

Experimental results show that the *LAeq* values are ranged from 72.1 dB to 80.2 dB. The presented data show a tendency for the value of *LAeq* to increase with an increase in the depth of cut and overhang of the workpiece. On the other hand, the influence of the feed rate on the change in *LAeq* is less significant.

3.2. Statistical Analysis and Model Development

The experimental results were processed in the Design-Expert software package using the Response Surface Methodology (RSM) [15], [16], [17]. Linear, 2FI, quadratic and cubic mathematical models were considered to describe the dependence between process parameters and *LAeq*. Their statistical indicators are given in Table 3.

Table 3: Summary statistical indicators of candidate models

Source	Std. Dev.	R^2	Adjusted R^2	Predicted R^2	PRESS	
Linear	0.8969	0.7991	0.7911	0.7759	68.18	
2FI	0.8120	0.8418	0.8288	0.8092	58.04	
Quadratic	**0.6669**	**0.8977**	**0.8845**	**0.8639**	**41.41**	**Suggested**
Cubic	0.6697	0.9071	0.8836	0.8462	46.80	Aliased

Based on the presented results, it can be seen that the quadratic model achieves the best statistical characteristics. The values of the adjusted coefficient of determination (Adj $R^2 = 0.8845$) and the predicted coefficient of determination (Pred $R^2 = 0.8639$), along with the lowest value of the Predicted Residual Error Sum of Squares parameter (PRESS = 41.41), indicate that this model most adequately describes the dependence between process parameters and *LAeq*. The cubic model was not considered due to the appearance of the aliasing effect, so the quadratic model was adopted for further analysis.

In order to assess the statistical significance of the developed model, an analysis of variance (ANOVA) was performed [18]. Analysis of results for the developed quadratic model are shown in Table 4.

Table 4: ANOVA analysis for quadratic model

Source	Sum of Squares	df	Mean Square	F-value	p-value	
Model	273.14	9	30.35	68.24	< 0.0001	significant
A-l	59.26	1	59.26	133.25	< 0.0001	
B-v	13.41	1	13.41	30.15	< 0.0001	
C-a	165.57	1	165.57	372.30	< 0.0001	
AB	2.13	1	2.13	4.79	0.0319	
AC	1.85	1	1.85	4.16	0.0452	
BC	8.80	1	8.80	19.79	< 0.0001	
A^2	9.70	1	9.70	21.82	< 0.0001	
B^2	4.21	1	4.21	9.46	0.0030	
C^2	2.97	1	2.97	6.67	0.0119	
Residual	31.13	70	0.4447			
Lack of Fit	7.28	17	0.4281	0.9511	0.5230	not significant
Pure Error	23.85	53	0.4501			
Cor Total	304.27	79				

The results of the ANOVA analysis show that the developed model is statistically significant, which is confirmed by the high value of the F-statistic (F = 68.24) and the very low p-value ($p < 0.0001$). The value of the Lack of Fit test was $p = 0.5230$, which indicates that the deviations between the experimental and model-predicted values are not statistically significant, i.e., that the model adequately describes the observed process. The analysis of individual factors showed that the greatest influence on *LAeq* is the depth of cut (F = 372.30), followed by the workpiece overhang (F = 133.25), while the feed rate has the smallest, but still statistically significant, influence (F = 30.15). Also, all interactions of the factors (AB, AC and BC) were statistically significant, which confirms the mutual dependence of the process parameters during the machining process. The basic statistical indicators of the developed mathematical model are given in Table 5.

Table 5: Statistical indicators of the developed model

Std. Dev.	0.6669	**R^2**	0.8977
Mean	76.29	**Adjusted R^2**	0.8845
C.V. %	0.8742	**Predicted R^2**	0.8639
PRESS	41.41	**Adeq Precision**	30.3647

The presented statistical indicators confirm the adequacy of the developed model. The value of the coefficient of determination ($R^2 = 0.8977$) shows that the model explains 89.77% of the total variability of the response, while the small difference between the values of Adj R^2 and Pred R^2 confirms its good predictive ability. Also, the value of the parameter Adeq Precision (30.3647) significantly exceeds the recommended minimum value of 4, which indicates a favorable signal-to-noise ratio. Based on the experimental results, a quadratic regression model was developed for predicting the *LAeq* value, shown in Equation (1).

$$LAeq = 67{,}83448 - 0{,}167001 \cdot l + 0{,}061220 \cdot f + 9{,}96358 \cdot a - 0{,}000495 \cdot l \cdot f - 0{,}046084 \cdot l \cdot a - 0{,}020106 \cdot f \cdot a + 0{,}007369 \cdot l^2 - 0{,}000194 \cdot f^2 - 1{,}62999 \cdot a^2 \quad (1)$$

where:

l – workpiece overhang length (mm),
f – feed rate (mm/min),
a – cutting depth (mm).

3.3. Model Diagnostics

As part of the diagnostics of the developed model, the normality of the residual distribution and the need for response transformation were analyzed. Based on the normal residual distribution diagram shown in Figure 2, it can be seen that the experimental points largely follow the reference line. This arrangement confirms that the residuals are approximately normally distributed and that the basic assumptions of the ANOVA analysis are met.

The Box–Cox analysis indicated that no transformation of the response variable was required while the current value of the parameter λ is within the confidence interval. The obtained results confirm that there was no need for additional response transformation before statistical analysis.

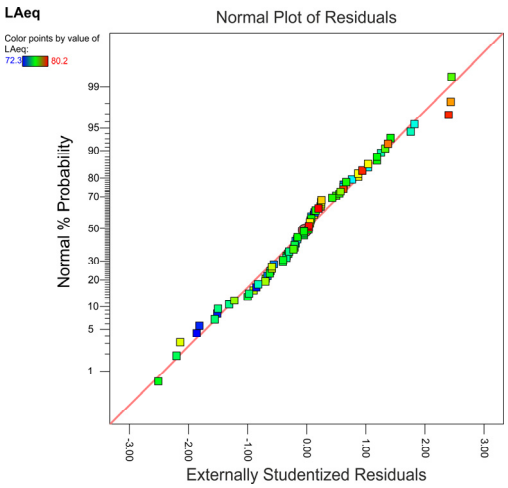


Figure 2: Normal probability plot of residuals

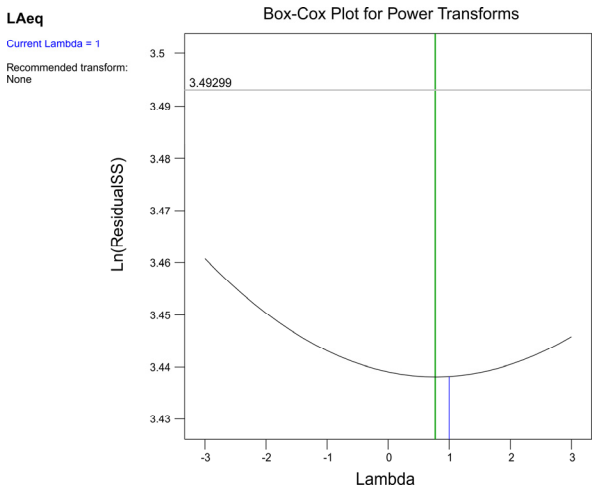


Figure 3: Box–Cox plot for response transformation

3.4. Response Surface Analysis

The effect of workpiece overhang length and feed rate on *LAeq* for different values of depth of cut was analyzed using the response surfaces shown in Figure 4.

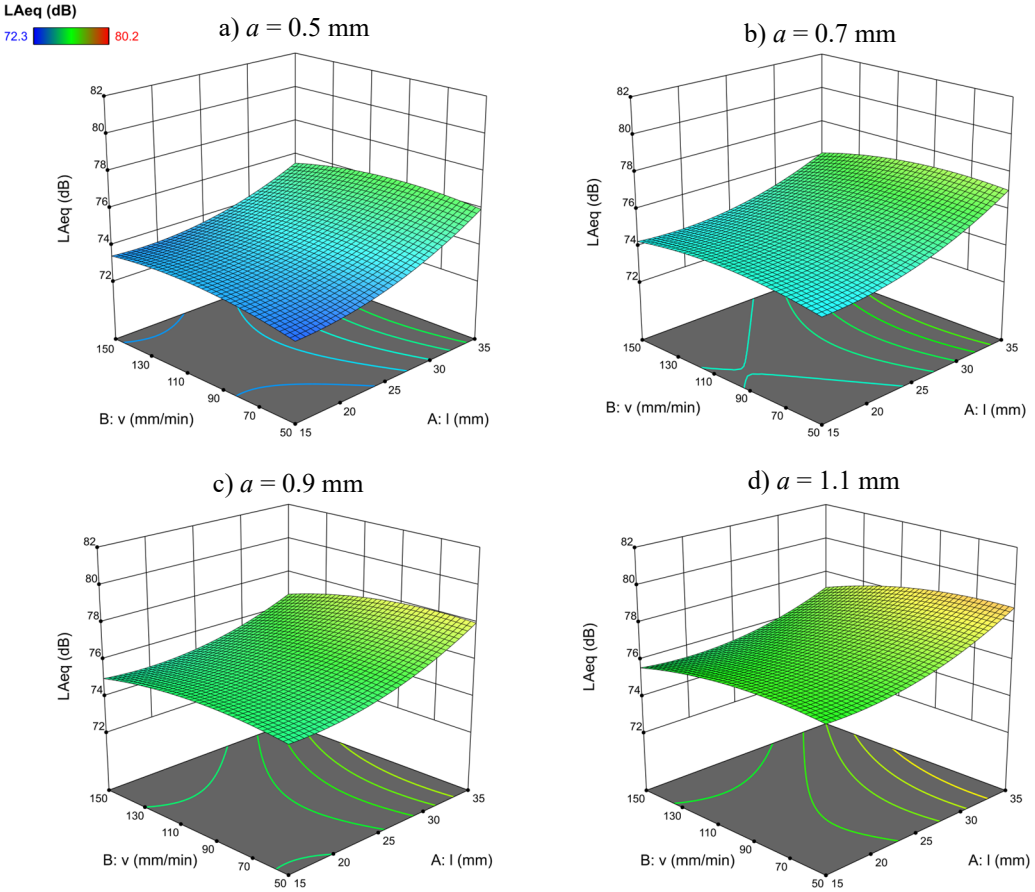


Figure 4: Response surfaces showing the influence of workpiece overhang length and feed rate on *LAeq* at different depths of cut

e) a = 1.3 mm f) a = 1.5 mm

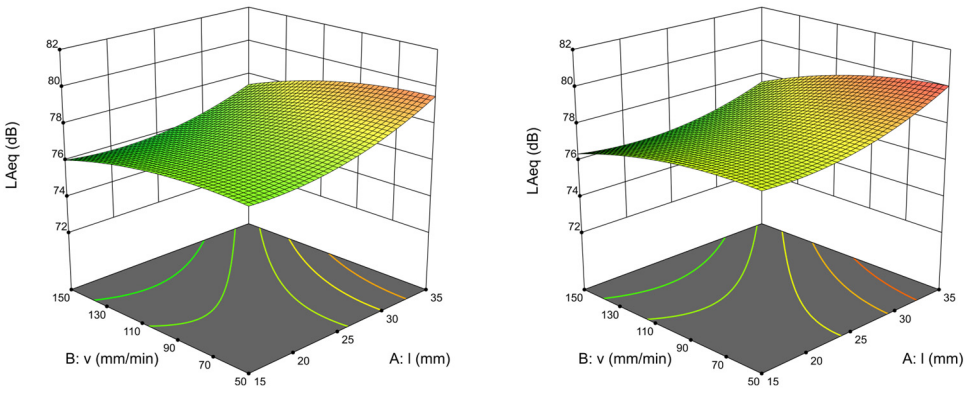


Figure 4: Response surfaces showing the influence of workpiece overhang length and feed rate on *LAeq* at different depths of cut

By analyzing the response surfaces, it can be seen that with an increase of the depth of cut, there is an increase in the value of *LAeq* for almost all combinations of workpiece overhang and feed rate. Also, increasing the workpiece overhang leads to an increase in *LAeq*, where this effect is more pronounced at larger depths of cut. On the other hand, the change in feed rate causes smaller changes in the *LAeq* value, which is in agreement with the results of the ANOVA analysis. The response surfaces additionally confirm the increasing trend of *LAeq* with increasing depth of cut and workpiece overhang length.

3.5. Model Validation

To verify the predictive ability of the developed mathematical model, additional experimental measurements were performed for three combinations of process parameters that were not used during model development. The adopted validation points and the corresponding experimentally measured *LAeq* values are shown in Table 6.

Table 6: Validation experiments and measured *LAeq* values

Exp. no	l (mm)	f (mm/min)	a (mm)	*LAeq* (dB)		
				1st measurement	2nd measurement	3rd measurement
1	30	125	1.25	76.56	77.06	76.65
2	20	60	1.40	76.55	77.71	76.46
3	17	140	0.60	74.02	73.17	73.97

The results of the validation of the mathematical model obtained in the Design-Expert software package are shown in Table 7.

Table 7: Results of the validation of the mathematical model

Analysis	**Predicted Mean (dB)**	**Predicted Median (dB)**	**Std. Dev.**	**n**	**SE Pred**	**95% PI Low (dB)**	**Data Mean (dB)**	**95% PI High (dB)**
LAeq	77.2592	77.2592	0.666874	3	0.418486	76.4246	76.7567	78.0939
LAeq	77.5980	77.5980	0.666874	3	0.423888	76.7525	76.9067	78.4434
LAeq	73.9464	73.9464	0.666874	3	0.429894	73.0890	73.7200	74.8038

Based on the presented results, it can be observed that there is a very good agreement between the model-predicted and experimentally determined *LAeq* values for all analyzed combinations of process parameters. In all cases, the experimentally determined mean values are within the 95% prediction interval, which confirms the adequacy of the developed model and its good predictive ability. The obtained results confirm that the developed model can reliably predict *LAeq* values for new combinations of process parameters within the analyzed experimental domain.

4. CONCLUSIONS

The results showed that all analyzed process parameters have a statistically significant effect on *LAeq* during face turning of beech wood. ANOVA analysis showed that the greatest impact is exerted by the depth of cut, followed by the workpiece overhang, while the impact of the feed rate is less pronounced.

Based on the experimental results, a quadratic regression model was developed to predict the *LAeq* value. Statistical analysis and validation results confirmed the adequacy of the model for describing the observed process and reliably predicting *LAeq* values within the analyzed experimental domain. The obtained results indicate that *LAeq* is sensitive to changes in machining parameters and may be considered a useful indirect indicator of machining conditions during face turning of beech wood. The developed methodology and regression model provide a basis for future research aimed at

using acoustic measurements for process monitoring, machining quality assessment and parameter optimization. The applied methodology can also be extended to future studies involving different workpiece materials, cutting tools and machining conditions.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (contract no. 451-03-34/2026-03/200108). This research contributes to the United Nations 2030 Agenda for Sustainable Development, specifically to Sustainable Development Goal 9: "Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation."

REFERENCES

[1] S.-C. Olaru, L. Slătineanu, M. Silitră, A.-M. Mihalache, and M. Coteață, "Investigation of the sound intensity level in the case of a universal lathe", MATEC Web of Conferences, Vol. 112, Article 01025, https://doi.org/10.1051/matecconf/201711201025, (2017)

[2] T. Miodragović, B. Radičević, S. Pajović, N. Kolarević, and V. Grković, "Identification of noise source based on sound intensity in vertical CNC milling machine", Proceedings of XI International Conference "Heavy Machinery – HM 2023", Vrnjačka Banja (Serbia), 21–24 June 2023, pp. C.45–C.50, (2023)

[3] A. Şahinoğlu and M. Rafighi, "Investigation of tool wear, surface roughness, sound intensity, and power consumption during hard turning of AISI 4140 steel using multilayer-coated carbide inserts", Journal of Engineering Research, Vol. 9(4B), pp. 377–395, https://doi.org/10.36909/jer.8783, (2021)

[4] A. Şahinoğlu and M. Rafighi, "Investigation of vibration, sound intensity, machine current and surface roughness values of AISI 4140 during machining on the lathe", Arabian Journal for Science and Engineering, Vol. 45(2), pp. 765–778, https://doi.org/10.1007/s13369-019-04124-x, (2020)

[5] D. Yu. Pimenov, M. K. Gupta, L. R. R. da Silva, M. Kiran, N. Khanna, and G. M. Krolczyk, "Application of measurement systems in tool condition monitoring of milling: a review of measurement science approach", Measurement, Vol. 199, Article 111503, https://doi.org/10.1016/j.measurement.2022.111503, (2022)

[6] A. Shokrani, H. Dogan, D. Burian, T. D. Nwabueze, P. Kolar, Z. Liao, A. Sadek, R. Teti, P. Wang, R. Pavel, and T. Schmitz, "Sensors for in-process and on-machine monitoring of machining operations", CIRP Journal of Manufacturing Science and Technology, Vol. 51, pp. 263–292, https://doi.org/10.1016/j.cirpj.2024.05.001, (2024)

[7] V. Nahornyi, A. Panda, J. Valíček, M. Harničárová, M. Kušnerová, I. Pandová, S. Legutko, Z. Palková, and O. Lukáč, "Adaptive control of the metalworking process with using sound generated during cutting", Research Square, https://doi.org/10.21203/rs.3.rs-986998/v1, (2021)

[8] A. Petrović, M. Rasinac, V. Grković, and M. Ivanović, "An experimental plan for noise analysis and chatter detection in milling depending on the cutting parameters", Proceedings of the 26th International Conference "Noise and Vibration", Niš (Serbia), 6–7 December 2018, pp. 41–44, (2018)

[9] M. Rasinac, B. Radičević, M. Kolarević, A. Petrović, and M. Bjelić, "Frequency analysis of noise at different milling parameters of steel", Proceedings of X International Conference "Heavy Machinery – HM 2021", Vrnjačka Banja (Serbia), 23–25 June 2021, pp. B.35–B.38, (2021)

[10] R. A. Kazeem and E. T. Akinlabi, "Experimental investigation of machine vibration rate and machine sound level in MQL turning of AISI 1525 steel employing mango oil as lubricant", International Journal on Interactive Design and Manufacturing, Vol. 19, pp. 4379–4394, https://doi.org/10.1007/s12008-024-02077-6, (2025)

[11] M. Najafian Ashrafi, H. Shaabani Asrami, Z. Vosoughi Rudgar, M. Ghorbanian Far, A. Heidari, E. Rastbod, H. Jafarzadeh, M. Salehi, E. Bari, and J. Ribera, "Comparison of physical and mechanical properties of beech and walnut wood from Iran and Georgian beech", Forests, Vol. 12(6), Article 801, https://doi.org/10.3390/f12060801, (2021)

[12] UNIMAT, "TVET – CNC – 1", Product Catalog, The Cool Tool Education, (2017)

[13] UNIMAT, "TVET – CNC – 2", Product Catalog, The Cool Tool Education, (2017)

[14] Brüel & Kjær, "Hand-held analyzers – Type 2250 and Type 2270 with Building Acoustics Software BZ-7228, Dual-channel Building Acoustics Software BZ-7229 and Qualifier Type 7830", Product Data, Brüel & Kjær, Nærum (Denmark), (2008)

[15] R. H. Myers, D. C. Montgomery, and C. M. Anderson-Cook, Response surface methodology: process and product optimization using designed experiments, John Wiley & Sons, Hoboken, NJ (USA), (2009)

[16] M. Kolarević, D. Minić, M. Rajović, V. Grković, and Zv. Petrović, "Selection of the optimal mathematical model of multiple regression in the ternary mixture experiments", IMK-14 – Research & Development in Heavy Machinery, Vol. 20(2), pp. 55–60, (2014)

[17] S. Pajović, M. Bjelić, V. Grković, and J. Perić, "Acoustic properties of standard low density polyurethane lime foam S 2535", Proceedings of 27th International Conference "Noise and Vibration", Niš (Serbia), 20–21 October 2022, pp. 137–142, (2022)

[18] D. C. Montgomery, Design and analysis of experiments, John Wiley & Sons, Hoboken, NJ (USA), (2012)

SESSION C: PRODUCTION TECHNOLOGIES, CAE AND MANAGEMENT

Investigation of cutting parameters influence on vibration during milling of HDPE

Jovana Perić[1], Mladen Rasinac[1], Stefan Pajović[1], Vladan Grković[1], Aleksandra Petrović[1*]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Kraljevo, Republic of Serbia

ARTICLE INFO

[*] **Correspondence:** petrovic.a@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.C06P **PAGES:** C45 – C52

ABSTRACT

The paper analyzes the influence of spindle speed, feed rate and axial depth of cut on vibration acceleration in feed direction during peripheral milling of high-density polyethylene. Experimental research was planned and implemented using the response surface methodology in the Design-Expert software package. Based on the experimental results, a statistical analysis was performed and a mathematical model was developed to predict vibration acceleration. The research results showed that feed rate and axial depth of cut have a dominant influence on vibration acceleration, while the influence of spindle speed is less pronounced. A reduced quadratic mathematical model was developed to describe the dependence between process parameters and vibration acceleration in feed direction. Model diagnostics and validation results confirmed its adequacy and good predictive ability. The obtained results can serve as a basis for selecting process parameters in order to reduce vibrations and improve the stability of the peripheral milling of high-density polyethylene.

KEYWORDS

Peripheral milling, Vibration acceleration, Cutting parameters, Response Surface Methodology

1. INTRODUCTION

The main sources of vibration during milling operations are the cutting process itself, the machine structure, and the tool setup. These vibrations are generally categorized into free, forced, and self-excited vibrations (chatter). Cyclic tool engagement followed by interrupted cutting and varying chip thickness leads to forced vibrations. Regenerative chatter occurs when a cutting tooth engages a wavy surface generated by the previous tooth, while friction between the tool flank and the workpiece may further contribute to self-excited vibrations. Small or long tools bend easily under cutting forces, making them highly prone to snapping or vibrating. The natural shaking frequency of the machine frame can match the cutting frequency, amplifying the vibration.

Machining parameters, above all spindle speed, feed rate, depth of cut, and radial engagement are critical drivers of vibration and chatter in milling [1]. Spindle speed determines the frequency of the cut, feed rate dictates the chip load, radial and axial depth of cut directly influence the magnitude of cutting forces, tool path determines changes of tool engagement [2] which can lead to increased vibration. Researchers [3], [4] have reported correlations between machining parameters and vibration occurrence during milling processes.

Vibrations can be detected through the measurement of acceleration, velocity or displacement [5]. Piezoelectric, capacitive, and piezoresistive accelerometers are commonly employed for vibration measurements on machine tools. Vibration measurements are often used for direct chatter detection or indirect tool wear monitoring. Several studies have reported correlations between vibration levels, surface roughness, workpiece geometric accuracy, and cutting power consumption [6], [7]. It has also been shown that vibrations affect tool life, machine service life, and machining accuracy. Pimenov et al. [8] reviewed various indirect methods for tool condition monitoring, including vibration-based measurement techniques.

Although machining vibrations have been extensively investigated in metal cutting processes, comparatively less attention has been devoted to polymer machining, particularly in the case of HDPE (High-Density Polyethylene) materials. Existing studies primarily focus on cutting forces, tool wear, or surface quality, while the influence of machining parameters on vibration behavior remains insufficiently explored. Furthermore, the potential use of vibration signals as indirect

indicators of process performance and surface quality has received limited attention. In addition, limited information is available regarding predictive modeling of vibration responses during HDPE milling under varying cutting conditions. The vibration signal acquired along the X-axis (feed direction) was selected for analysis. This direction is directly related to the process kinematics, as variations in feed rate and chip load are reflected through the cutting forces acting along the feed axis. Since the feed motion significantly influences the generation of the machined surface, vibrations measured in this direction are considered particularly representative of the machining process behavior and its response to changes in cutting conditions.

The aim of this study is to investigate the influence of spindle speed, feed rate, and axial depth of cut on vibration acceleration measured in the feed direction during CNC peripheral milling of HDPE using an high-speed steel (HSS) flat end mill. Statistical analysis and multiple linear regression modeling were employed to determine the significance of machining parameters and to develop predictive models for vibration in X-direction. Additionally, confirmation experiments were performed to validate the developed regression model.

2. MATERIALS AND METHODS

2.1. Workpiece Material

The workpiece material used in this study is HDPE, known for its exceptional strength-to-density ratio, outstanding chemical resistance, and high impact tolerance. The average density of the material is approximately 950 kg/m³, tensile strength typically ranging from 26 MPa to 38 MPa and melting point is between 120°C and 130°C [9]. HDPE has excellent machinability, rating as one of the easiest engineering plastics to cut, drill, turn, and mill. It produces minimal tool wear and creates smooth, clean surfaces when machined with sharp, high-speed tools. In this study, a box-shaped HDPE workpiece with dimensions of 50 × 50 × 35 mm was used. The peripheral milling process was performed as up-milling.

2.2. Machine Tool and Cutting Tool

Machining experiments were performed on a Unimat 4-Axis CNC Milling Machine configured as a 3-axis vertical milling machine. The machine positioning accuracy is 0.05 mm, while the maximum positioning speed is 300 mm/min. The spindle speed is adjustable between 2,000 and 15,000 rpm. The machine is intended for machining wood, plastics, wax and soft non-ferrous materials [10], [11]. Milling operations were carried out using a HSS two-flutes flat mill with a diameter of 2 mm.

2.3. Measurement Equipment

Vibrations generated during the machining process were measured using a four-channel vibration datalogger Extech VB500. The device enables simultaneous acquisition from four measurement channels and supports vibration acceleration, velocity, and displacement measurements. Measurements were performed in the frequency range from 10 Hz to 1 kHz using external piezoelectric accelerometers with magnetic mounting. The instrument accuracy is specified as ±(5% of reading + 2 digits), while acquired data are stored on an SD memory card for subsequent analysis. The device complies with the requirements of the ISO 2954 standard for mechanical vibration measurements.

2.4. Experimental Procedure

3Peripheral conventional milling was performed on polyethylene box shaped workpiece using an HSS flat mill on educational CNC milling machine. In this experiment the radial depth of cut was kept constant at 1 mm, corresponding to 50% tool engagement. The workpiece material and cutting tool also haven't been changed. The tool overhang length is set at 13 mm. Magnetic accelerometer is placed on clamping device fixed on the work table. Vibrations were measured in X direction (feed rate direction), when the tool is in full engagement with workpiece. An experimental setup and the accelerometer position is presented in Fig. 1.

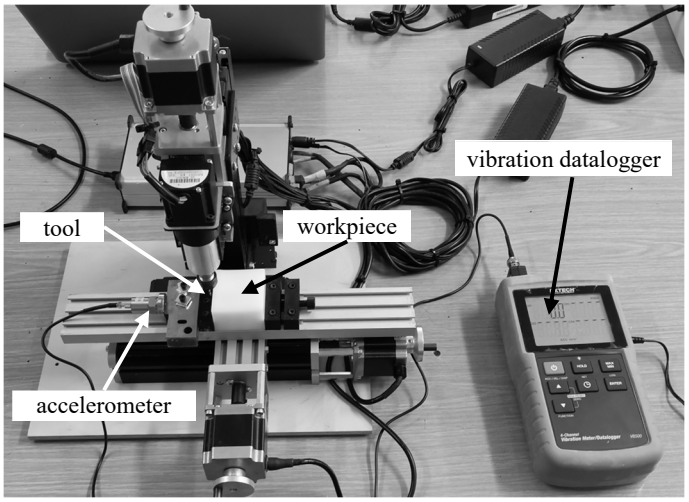


Figure 1: Experimental setup

2.5. Experimental Design

Input parameters (feed rate, spindle speed and axial depth of cut) were varied at three levels as shown in table 1. A full-factorial experimental design consisting of 27 experimental runs was employed.

Table 1: Input parameters and their levels

Factor	**Symbol**	**Levels**		
Spindle speed	*n* (rpm)	3000	4000	5000
Feed rate	*f* (mm/min)	100	200	300
Axial depth of cut	*b* (mm)	3	6	9

The vibration acceleration measured along the X-axis was considered as the response variable.

3. RESULTS AND DISCUSSION

3.1. Experimental Results

Experimental studies were conducted to investigate the influence of spindle speed (*n*), feed rate (*f*) and axial depth of cut (*b*) on vibration acceleration (*Xvib*) in feed direction during peripheral milling of HDPE. Experiments were conducted for different combinations of process parameters, with three levels adopted for each factor. The obtained experimental values of *Xvib* are shown in Table 2.

Table 2: Experimental results

Exp. no	*n* (rev/min)	*b* (mm)	*f* (mm/min)	*Xvib* (m/s^2)
1	3000	9	100	3.38
2	3000	9	200	4.34
3	3000	9	300	5.11
4	4000	9	100	2.39
5	4000	9	200	2.80
6	4000	9	300	3.90
7	5000	9	100	2.78
8	5000	9	200	2.92
9	5000	9	300	4.43
10	3000	6	100	2.86
11	3000	6	200	3.68
12	3000	6	300	4.85
13	4000	6	100	2.81
14	4000	6	200	2.96
15	4000	6	300	3.46
16	5000	6	100	2.90
17	5000	6	200	2.80
18	5000	6	300	4.10
19	3000	3	100	1.66
20	3000	3	200	1.63
21	3000	3	300	3.20
22	4000	3	100	1.22
23	4000	3	200	1.67
24	4000	3	300	3.23
25	5000	3	100	1.64
26	5000	3	200	1.89
27	5000	3	300	4.19

Based on presented experimental results, it can be seen that the *Xvib* values are ranged from 1.22 m/s² to 5.11 m/s². The results indicate an increase in the *Xvib* value with increasing feed rate and axial depth of cut. On the other hand, the influence of the spindle speed is reflected in the tendency for the *Xvib* value to decrease at higher values of this parameter.

3.2. Statistical Analysis and Model Development

Statistical analysis of experimental results was carried out in the Design-Expert software package using the Response Surface Methodology [12], [13]. Linear, 2FI, quadratic and cubic mathematical models were considered to describe the influence of process parameters on *Xvib*. An overview of their statistical indicators is given in Table 3.

,

Table 3: Summary statistical indicators of candidate models

Source	Std. Dev.	R^2	Adjusted R^2	Predicted R^2	PRESS	
Linear	0.5836	0.7220	0.6857	0.6178	10.77	
2FI	0.5668	0.7720	0.7036	0.5954	11.40	
Quadratic	**0.3454**	**0.9280**	**0.8899**	**0.8127**	**5.28**	**Suggested**
Cubic	0.2525	0.9774	0.9412	0.7676	6.55	Aliased

Based on shown statistical indicators, it can be concluded that the quadratic model most successfully describes the dependence between the process parameters and the *Xvib* value. High values of the adjusted coefficient of determination (Adj $R^2 = 0.8899$) and the predicted coefficient of determination (Pred $R^2 = 0.8127$), along with the lowest value of the Predicted Residual Error Sum of Squares parameter (PRESS = 5.28), confirm its good ability to describe and predict the response. Due to the appearance of the aliasing effect, the cubic model was not considered in further analysis, so the quadratic model was adopted for the development of the mathematical model. Analysis of variance (ANOVA) was applied to check the statistical significance of the developed model [14]. The results of this analysis are shown in Table 4.

Table 4: ANOVA analysis for reduced quadratic model

Source	Sum of Squares	df	Mean Square	F-value	p-value	
Model	26.02	7	3.72	32.81	< 0.0001	significant
A-*n*	0.5207	1	0.5207	4.60	0.0452	
B-*f*	12.20	1	12.20	107.71	< 0.0001	
C-*b*	7.62	1	7.62	67.23	< 0.0001	
AC	1.28	1	1.28	11.33	0.0032	
A^2	1.66	1	1.66	14.66	0.0011	
B^2	1.41	1	1.41	12.48	0.0022	
C^2	1.32	1	1.32	11.66	0.0029	
Residual	2.15	19	0.1133			
Cor Total	28.17	26				

The results of the ANOVA analysis show that the developed model is statistically significant, as confirmed by the high value of the F-statistic (F = 32.81) and the very low p-value ($p < 0.0001$). The analysis of individual factors showed that the greatest influence on *Xvib* is exerted by the feed rate (F = 107.71), followed by the axial depth of cut (F = 67.23), while the spindle speed has the smallest, but statistically significant, influence (F = 4.60). Also, the interaction of the AC factors was found to be statistically significant, indicating that the influence of the spindle speed depends on the selected depth of cut. The basic statistical indicators of the developed mathematical model are shown in Table 5.

Table 5: Statistical indicators of the developed model

Std. Dev.	0.3366	**R^2**	0.9236
Mean	3.07	**Adjusted R^2**	0.8954
C.V. %	10.98	**Predicted R^2**	0.8428
PRESS	4.43	**Adeq Precision**	21.6737

The statistical indicators confirm the reliability of the developed mathematical model. The value of the coefficient of determination ($R^2 = 0.9236$) shows that the model explains 92.36% of the total variability of *Xvib*. At the same time, the small difference between the values of Adj R^2 and Pred R^2 indicates a good predictive ability of the model. The value of the parameter Adeq Precision of 21.6737 exceeds the recommended minimum value of 4 by many times, which confirms an adequate signal-to-noise ratio. Based on the experimental results, a quadratic regression model for predicting *Xvib* was developed, shown in equation (1).

$$Xvib = 6.30265 - 0.003725 \cdot n - 0.011181 \cdot f + 1.27862 \cdot b - 0.000109 \cdot n \cdot b + 5.26143 \cdot 10^{-7} \cdot n^2 + 0.000049 \cdot f^2 - 0.052138 \cdot b^2 \quad (1)$$

where:

n – spindle speed (rev/min),
f – feed rate (mm/min),
b – axial depth of cut (mm).

3.3. Model Diagnostics

As part of the diagnostics of the developed model, the normality of the residual distribution and the need for response transformation were analyzed.

The normal residual distribution diagram shown in Figure 2 shows that the experimental points are located near the reference line, without significant deviations. Such a layout indicates an approximately normal distribution of the residuals and confirms the fulfillment of the basic assumptions necessary for conducting the ANOVA analysis.

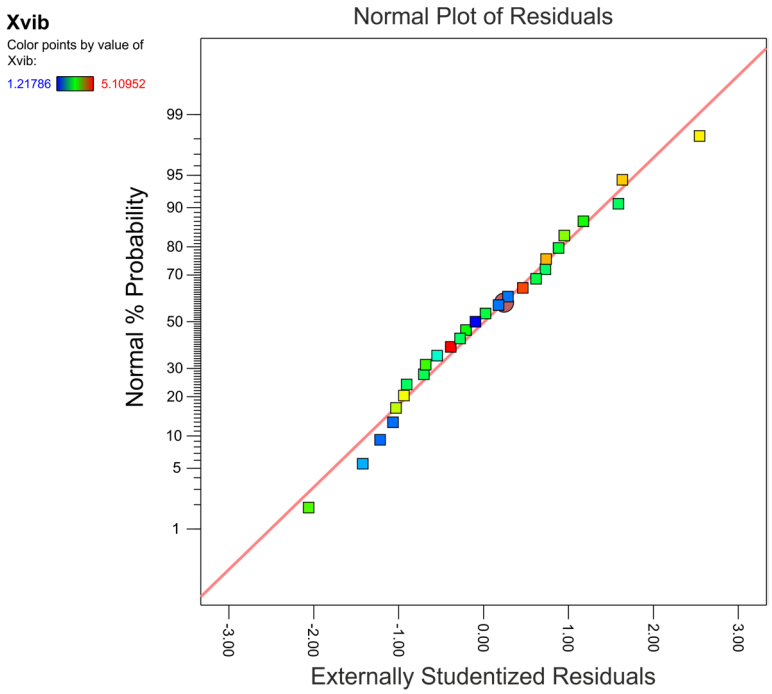


Figure 2: Normal probability plot of residuals

The Box–Cox analysis shown in Figure 3 indicated that no transformation of the response variable was required while the current value of the parameter λ is within the confidence interval. Based on the obtained results, it can be concluded that the experimental data were suitable for direct statistical processing without applying additional transformations.

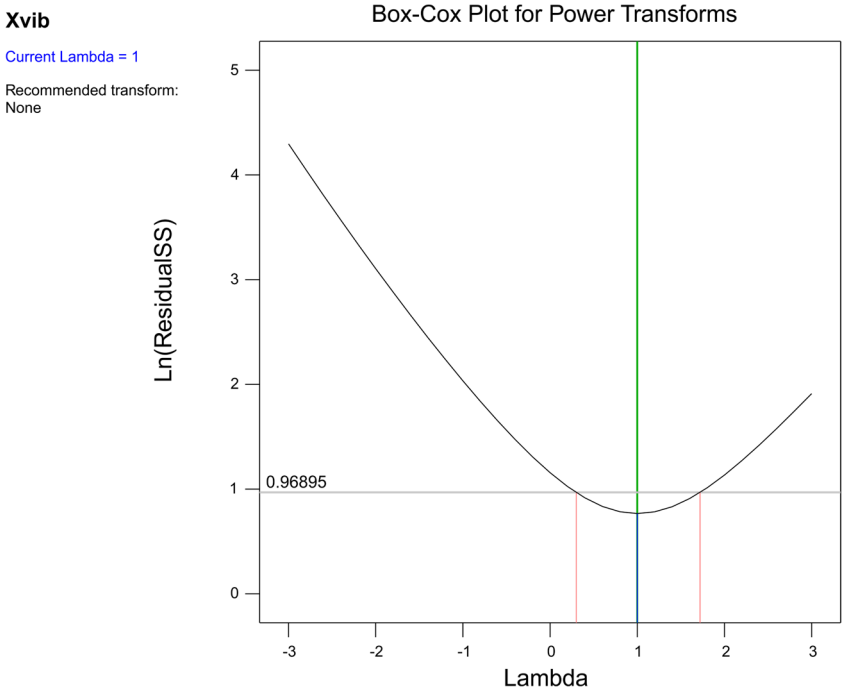


Figure 3: Box–Cox plot for response transformation

3.4. Response Surface Analysis

The mutual influence of spindle speed, feed rate and axial depth of cut on *Xvib* was analyzed using the response surfaces shown in Figure 4.

Based on the presented response surfaces, it can be seen that increasing the axial depth of cut leads to an increase in the *Xvib* value for most of the analyzed combinations of spindle speed and feed rate. At the same time, increasing the feed rate additionally intensifies *Xvib*, with this effect being more pronounced at higher axial depths of cut. In contrast, increasing the spindle speed tends to decrease the *Xvib* value, which indicates more favorable conditions for the milling process. The observed changes in the response surfaces confirm the results of the statistical analysis and indicate the dominant influence of feed rate and axial depth of cut on *Xvib*.

3.5. Model Validation

In order to assess the predictive ability of the developed mathematical model, additional experimental measurements were conducted for five combinations of process parameters that were not used during model development. The validation points and corresponding experimentally determined *Xvib* values are shown in Table 6.

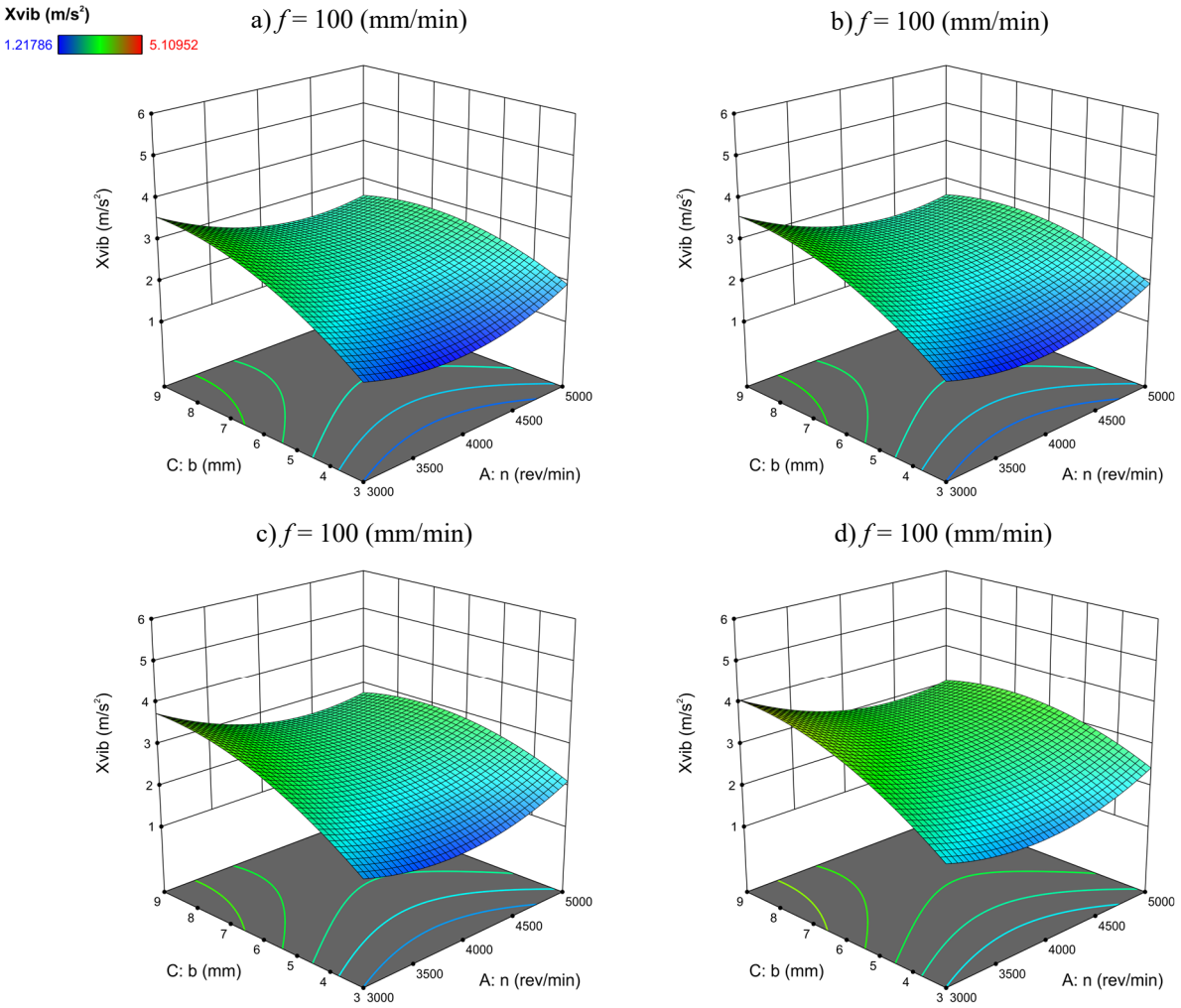
Table 6: Validation experiments and measured *Xvib* values

Exp. no	*n* (rev/min)	*f* (mm/min)	*b* (mm)	*Xvib* (m/s²)
1	3500	180	7.5	3.24
2	3800	210	6.0	2.63
3	4750	250	5.5	2.74
4	4100	110	4.0	2.31
5	3150	118	3.5	2.34

The results of the mathematical model validation obtained in the Design-Expert software package are shown in Table 7.

Table 7: Results of the validation of the mathematical model

Analysis	Predicted Mean (m/s²)	Predicted Median (m/s²)	Standard Deviation	n	SE Pred	95% PI Low (m/s²)	Data Mean (m/s²)	95% PI High (m/s²)
Xvib	3.065	3.065	0.3366	1	0.3681	2.295	3.24	3.835
Xvib	2.846	2.846	0.3366	1	0.3765	2.058	2.63	3.634
Xvib	3.325	3.325	0.3366	1	0.3662	2.558	2.74	4.092
Xvib	1.724	1.724	0.3366	1	0.3674	0.955	2.31	2.493
Xvib	1.780	1.780	0.3366	1	0.3681	1.010	2.34	2.551



e) f = 100 (mm/min) f) a = 1.5 mm

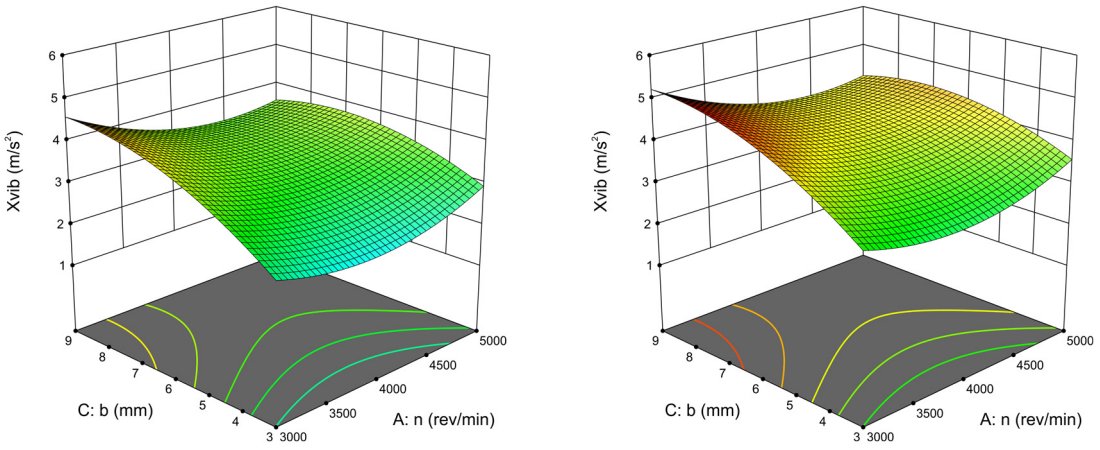


Figure 4: Response surfaces showing influence of the spindle speed and depth of cut on *Xvib* at different feed rates

By comparing the experimentally determined and model predicted values shown in Table 7, a good agreement of the results can be observed for all analyzed validation points. In all cases, the experimental values are within the corresponding 95% prediction intervals, which confirms the adequacy of the developed mathematical model. The obtained results indicate that the model has a satisfactory predictive ability and can be used to estimate the *Xvib* values for new combinations of process parameters within the analyzed experimental domain.

4. CONCLUSIONS

The conducted research showed that the spindle speed, feed rate and axial depth of cut affect *Xvib* during CNC peripheral milling of HDPE materials. Analysis of the results indicated that the feed rate and axial depth of cut have the greatest influence, while the influence of the spindle speed is less pronounced.

Based on the experimental data, a reduced square mathematical model was developed to predict the *Xvib* value. The results of statistical analysis, diagnostics and validation confirmed the adequacy of the model and its satisfactory predictive ability within the analyzed experimental domain.

The developed model can be a useful tool for estimating *Xvib* and selecting process parameters with the aim of reducing vibrations and improving the stability of the peripheral milling of HDPE materials.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (contract no. 451-03-34/2026-03/200108). This research contributes to the United Nations 2030 Agenda for Sustainable Development, specifically to Sustainable Development Goal 9: “Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation.”

REFERENCES

[1] M. Rasinac, A. Petrović, B. Radičević, V. Grković, and M. Ivanović, “Experimental analysis of noise and chatter detection in milling depending on the cutting parameters”, Acoustics and Vibration of Mechanical Structures – AVMS 2019, Springer Proceedings in Physics, Vol. 251, Springer, Cham (Switzerland), pp. 333–342, https://doi.org/10.1007/978-3-030-54136-1_33, (2021)

[2] A. Petrović, Lj. Lukić, S. Ivanović, and A. Pavlović, “Optimisation of tool path for wood machining on CNC machines”, Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, Vol. 231(1), pp. 72–87, https://doi.org/10.1177/0954406216648715, (2017)

[3] Á. Hortobágyi, P. Koleda, P. Koleda, and R. Kminiak, “Effect of milling parameters on amplitude spectrum of vibrations during milling materials based on wood”, Applied Sciences, Vol. 13(8), Article 5061, https://doi.org/10.3390/app13085061, (2023)

[4] Š. Olejárová, “Analysis of the influence of selected technological parameters on the magnitude of vibrations during material machining by milling”, Journal of Production Engineering, Vol. 25(1), pp. 15–19, (2022)

[5] A. Shokrani, H. Dogan, D. Burian, T. D. Nwabueze, P. Kolar, Z. Liao, A. Sadek, R. Teti, P. Wang, R. Pavel, and T. Schmitz, “Sensors for in-process and on-machine monitoring of machining operations”, CIRP Journal of Manufacturing Science and Technology, Vol. 51, pp. 263–292, https://doi.org/10.1016/j.cirpj.2024.05.001, (2024)

[6] A. Şahinoğlu and M. Rafighi, “Investigation of vibration, sound intensity, machine current and surface roughness values of AISI 4140 during machining on the lathe”, Arabian Journal for Science and Engineering, Vol. 45(2), pp. 765–778, https://doi.org/10.1007/s13369-019-04124-x, (2020)

[7] E. G. Plaza, P. J. Núñez López, and E. M. Beamud González, “Efficiency of vibration signal feature extraction for surface finish monitoring in CNC machining”, Journal of Manufacturing Processes, Vol. 44, pp. 145–157, https://doi.org/10.1016/j.jmapro.2019.05.046, (2019)

[8] D. Yu. Pimenov, M. K. Gupta, L. R. R. da Silva, M. Kiran, N. Khanna, and G. M. Krolczyk, “Application of measurement systems in tool condition monitoring of milling: a review of measurement science approach”, Measurement, Vol. 199, Article 111503, https://doi.org/10.1016/j.measurement.2022.111503, (2022)

[9] J. R. Araújo, W. R. Waldman, and M. A. De Paoli, “Thermal properties of high density polyethylene composites with natural fibres: coupling agent effect”, Polymer Degradation and Stability, Vol. 93(10), pp. 1770–1775, https://doi.org/10.1016/j.polymdegradstab.2008.07.021, (2008)

[10] UNIMAT, “TVET – CNC – 1”, Product Catalog, The Cool Tool Education, (2017)

[11] UNIMAT, “TVET – CNC – 2”, Product Catalog, The Cool Tool Education, (2017)

[12] R. H. Myers, D. C. Montgomery, and C. M. Anderson-Cook, “Response surface methodology: process and product optimization using designed experiments”, John Wiley & Sons, Hoboken, NJ (USA), (2009)

[13] M. Kolarević, D. Minić, M. Rajović, V. Grković, and Zv. Petrović, “Selection of the optimal mathematical model of multiple regression in the ternary mixture experiments”, IMK-14 – Research & Development in Heavy Machinery, Vol. 20(2), pp. 55–60, (2014)

[14] D. C. Montgomery, “Design and analysis of experiments”, John Wiley & Sons, Hoboken, NJ (USA), (2012)

SESSION C: PRODUCTION TECHNOLOGIES, CAE AND MANAGEMENT

Energy efficiency and carbon footprint of welded, adhesive and hybrid joining processes in the context of circular economy

Dragan Milčić[1*], Miodrag Milčić[1], Nataša Zdravković[1], Aleksija Đurić[2], Damjan Klobčar[3], Tomaž Vuherer[4]

[1] University of Niš, Faculty of Mechanical Engineering, Niš, Serbia
[2] University of East Sarajevo, Faculty of Mechanical Engineering, East Sarajevo, Bosnia and Herzegovina
[3] University of Ljubljana, Faculty of Mechanical Engineering, Ljubljana, Slovenia
[4] University of Maribor, Faculty of Mechanical Engineering, Maribor, Slovenia

ARTICLE INFO

[*] **Correspondence:** dragan.milcic@masfak.ni.ac.rs

DOI: https://doi.org/10.46793/ET26.C07M **PAGES:** C53 – C59

ABSTRACT

This paper presents a systematic comparative analysis of energy efficiency and carbon footprint of welded, adhesive, and hybrid (weld-bonding) joining processes for metallic components, within the framework of the circular economy and Extended Producer Responsibility (EPR). With global welding energy consumption estimated at over 400 TWh annually, the identification of lower-energy joining alternatives is a critical step toward industrial decarbonization aligned with Paris Agreement targets and EU climate neutrality goals for 2050. The study integrates quantitative energy balances, Life Cycle Assessment (LCA) data, and experimental results on surface preparation effects on adhesive joint performance. Results demonstrate that adhesive bonding requires up to 6000× less direct process energy than MIG/MAG welding for equivalent joints, and generates 3-50× lower CO_2 emissions per meter of joint. Hybrid joining reduces energy consumption by approximately 30% compared to pure spot welding while improving fatigue strength by 60-70%. The circular economy implications of each technology - including repairability, disassembly, and end-of-life recyclability - are examined, and a conceptual decision framework for joining technology selection is proposed, based on lifecycle performance indicators and EPR compliance.

KEYWORDS

Energy efficiency, Carbon footprint, Welding, Adhesive bonding, Hybrid joints, Circular economy

1. INTRODUCTION

The joining of metallic components is a foundational operation in modern manufacturing, with the choice of technology directly influencing mechanical performance, economic viability, and environmental sustainability. Welding, long the dominant joining method in heavy industry, automotive, and aerospace sectors, is increasingly challenged by structural adhesive bonding and hybrid joining techniques - particularly under the mounting pressure of tighter CO_2 reduction mandates, mass-optimization targets, and circular economy legislation [1, 2].

According to the International Institute of Welding (IIW), global energy consumption in welding processes is estimated at over 400 TWh per year, representing a significant share of total industrial energy use [3]. In the context of the Paris Agreement and the European Union's target of climate neutrality by 2050, the identification and adoption of energy-efficient joining technologies has become an industrial and regulatory imperative [4].

Structural adhesive bonding of metallic structures, though established since the mid-20th century, has undergone a significant renaissance in recent decades, driven by advances in high-performance epoxy and acrylic adhesive formulations and progress in surface preparation science [5, 6]. Research compiled in [7, 8] demonstrates that surface preparation methods - including mechanical abrasion, laser cleaning, and laser texturing - exert a decisive influence on adhesive joint strength, with the optimal method depending on the adherend material and adhesive system combination.

The hybrid weld-bonding approach, which combines resistance spot welding or seam welding with a continuous structural adhesive, emerged primarily from the automotive industry's need to increase body-in-white stiffness while simultaneously reducing mass and weld point count [9, 10]. This concept has demonstrated synergistic effects exceeding the sum of individual contributions, particularly with respect to fatigue strength and vibration damping [11].

Within the evolving legislative landscape of Extended Producer Responsibility (EPR) - which holds manufacturers accountable for the full lifecycle of their products, including end-of-consumer-use stages - the choice of joining technology is no longer purely a question of technical performance and production cost. It has become a strategic decision with lifecycle-wide implications for resource efficiency, repairability, disassembly, and material quality at end of life [12].

The objective of this paper is to provide a structured comparison of welded, adhesive, and hybrid joining processes across four dimensions: (i) direct process energy consumption and CO_2 emissions, (ii) thermal effects and their consequences for material integrity, (iii) stress distribution and mechanical performance, and (iv) circular economy and EPR compliance metrics - including durability, repairability, disassembly, and recyclability.

2. ENERGY CONSUMPTION AND CARBON FOOTPRINT

2.1. Welding Processes

Fusion welding processes require significant local heat input to melt the base material, with or without filler material. The Heat Input (HI) is defined as the energy deposited per unit weld length:

$$HI = \eta \cdot \frac{U \cdot I}{v} \tag{1}$$

where η is the arc-to-workpiece energy transfer efficiency (η = 0.60 - 0.95), U is the arc voltage [V], I is the welding current [A], and v is the travel speed [mm/s]. Process efficiency varies widely: MIG/MAG welding achieves 75-85%, GTAW (TIG) 50-70%, SMAW 60-75%, laser beam welding 30-50%, and electron beam welding 85-95%. Characteristic heat input values for steel range from 0.05-0.5 kJ/mm for laser welding (HAZ width 2-5 mm) to 0.8–2.5 kJ/mm for MIG/MAG (HAZ width 10-20 mm).

Beyond the direct arc energy, welding processes impose additional energy demands for the production of shielding gases (Ar, CO_2, He), pre-heating and post-weld heat treatment (PWHT) for hardenable steels, cooling systems for equipment, and post-weld mechanical operations such as grinding, straightening, and surface treatment.

2.2. Adhesive Bonding Processes

In contrast to fusion welding, adhesive bonding requires no melting of the base material. Total process energy can be expressed as:

$$E_{total} = E_{surface\ preparation} + E_{application} + E_{curing} \tag{2}$$

where: E_{total} is the total energy, $E_{surface\ preparation}$ is the surface preparation energy, $E_{application}$ is the adhesive application energy, and E_{curing} is the curing energy.

Surface preparation energy depends strongly on method: mechanical abrasion (abrasive paper) requires 5-20 kJ/m^2, atmospheric plasma treatment 10-50 kJ/m^2, laser cleaning 50-200 kJ/m^2, and laser texturing 100-500 kJ/m^2. Adhesive application energy is negligible for manual dispensing; robotic two-component mixing systems consume 0.1-0.5 kW during operation. Curing energy: room-temperature-curing epoxy systems require no external energy (cure time 30 min to 48 h), thermally-activated adhesives require heating at 80–180 °C for 20-60 min (50-200 kJ/m^2), and UV-curing systems are the most energy-efficient at 5-20 kJ/m^2 over 10-60 seconds of exposure.

A representative comparative calculation for joining two 2 mm steel plates over a 100 cm^2 bond area illustrates the magnitude of the difference: MIG/MAG welding a 400 mm seam with HI = 1.5 kJ/mm requires approximately 600 kJ total, while room-temperature-curing structural epoxy bonding of the same joint requires less than 0.1 kJ - a factor of approximately 6000.

2.3. Hybrid Weld-Bonding

Hybrid joining optimizes energy efficiency through strategic combination of the two technologies. Weld-bonding combines resistance spot welding with a continuous structural adhesive applied prior to welding. The adhesive fills the gap between adherends, while spot welds provide immediate handling strength during adhesive cure. This sequence - adhesive application, assembly, spot welding, curing - adds process steps relative to pure welding, but eliminates the need for increased weld point density. In automotive applications, replacing pure spot welding with weld-bonding (4 spot welds/dm^2 plus continuous structural adhesive vs. 8 spot welds/dm^2) reduces energy consumption from 120-180 kJ/m^2 to 100-140 kJ/m^2 - an approximate reduction of 30%. The performance gains of weld-bonding exceed the sum of individual contributions. The adhesive layer reduces peak stresses around weld spots by 40-60% (as confirmed by FEM analyses), while weld spots arrest crack propagation through the adhesive. Additionally, the viscoelastic adhesive provides vibration damping, and the sealed weld spots are protected from moisture ingress, eliminating corrosion-fatigue interaction.

2.4. Life Cycle Carbon Footprint

Complete environmental impact assessment requires Life Cycle Assessment (LCA) that encompasses upstream energy production, ancillary materials, and waste treatment. LCA studies in the automotive sector report that MIG/MAG welding generates 3.5-5 kg CO_2 per meter of joint, room-temperature-curing adhesive bonding generates 0.1-0.3 kg CO_2/m, and thermally-cured adhesive systems generate 0.5-1.2 kg CO_2/m of joint. This indicates that adhesive bonding carries a carbon footprint 3-50× lower than conventional welding - a finding of direct relevance in the context of industrial decarbonization and emerging carbon pricing mechanisms.

3. THERMAL EFFECTS AND HEAT AFFECTED ZONE

3.1. HAZ Formation in Welding

Local melting during welding generates intense thermal gradients leading to the formation of the Heat Affected Zone (HAZ) - a region in which the base material undergoes thermally induced microstructural changes without reaching melting temperature [13]. HAZ width is a function of heat input, material thermal conductivity, and joint geometry.

For low-alloy steel, MIG/MAG welding typically produces a HAZ of 10-20 mm width [14]; austenitic stainless steel (X5CrNi18-10), due to lower thermal conductivity, generates HAZ of 8-15 mm with pronounced thermal distortion [15]; Al alloys (e.g. EN AW 5754), owing to high conductivity and low melting point (630-650 °C), produce the widest HAZ at 15-30 mm and are most susceptible to dimensional distortion requiring post-weld correction [16].

Microstructural changes in the HAZ depend on temperature and cooling rate. In steels, the region adjacent to the fusion line undergoes overheating and austenitic grain growth, which - following transformation to ferrite - results in a coarse-grained structure with reduced strength and toughness [17]. Further consequences include normalization in the inter-critical zone (A_3-A_1), and partial austenitization at lower temperatures. Cooling rate governs the final microstructure: slow cooling produces ferritic-pearlitic structure; rapid cooling may generate martensite with increased hardness but reduced toughness and susceptibility to hydrogen embrittlement [18]. Tensile strength in the HAZ is typically reduced to 80-95% of base material values, while hardness may range from 70% to 150% depending on the degree of hardenability [19]. Residual tensile stresses in the weld direction can reach 300-500 MPa in low-alloy steels (70-80% of yield strength), significantly affecting fatigue performance and susceptibility to stress corrosion cracking [20].

3.2. Absence of Thermal Effects in Adhesive Bonding

The fundamental advantage of adhesive bonding lies in the complete absence of base material melting, thereby eliminating HAZ formation and all associated microstructural consequences. Adherend temperature during bonding is governed solely by curing conditions. Room-temperature-curing systems do not raise adherend temperature beyond 20-30 °C - insufficient for any metallurgical transformation. Thermally cured adhesives at 80-180 °C produce uniform, moderate heating far below the critical A_1 temperature (723 °C for steel), leaving base material properties fully intact.

The elimination of HAZ yields several practical benefits: residual stresses of thermal origin are removed (improving fatigue life and stress corrosion resistance); mechanical properties are preserved throughout the entire adherend volume (critical for precipitation-hardened alloys); and thermal distortion is essentially eliminated, reducing or eliminating the need for post-assembly dimensional correction. Residual stresses in adhesive joints arise solely from polymer shrinkage during cure (typically 2-5% volumetric contraction for epoxy systems), generating local interface stresses of only 5-15 MPa - negligible compared to welding residual stresses.

4. STRESS DISTRIBUTION AND MECHANICAL PERFORMANCE

4.1. Stress Concentration in Welded Joints

Welded joints are inherently characterized by geometric and microstructural discontinuities acting as stress concentrators. Sources include geometric notches (weld toe, root notch, surface irregularities), microstructural interfaces between base material, HAZ, and fusion zone, and internal defects (porosity, inclusions, incomplete penetration). The stress concentration factor α_k ranges from 1.5-2.5 for optimally executed joints to above 4-8 for joints with significant defects. More than 70% of fatigue failures in welded structures initiate in the HAZ/fusion zone interface region, where geometrical, microstructural, and residual stress effects combine.

4.2. Load Transfer in Adhesive Joints

Adhesive joints distribute load over a large bonded area rather than concentrating it through a small cross-section. For a single-lap joint of bonded area A under shear force F, the nominal shear stress is: $\tau = F/A$.

However, shear and peel stress concentrations occur at joint edges - the 'edge effect' analytically described in [21 and 22]. FEM analyses confirm that edge stress concentration factors α_k = 2-4 are achievable depending on geometry and adherend stiffness, with the critical failure mode being peel stress (mode I opening) at joint edges. Design strategies to mitigate edge stress concentration include optimized overlap length (L_{opt} = 15-25 × $t_{adherend}$), tapered adherend edges (reducing peel stress by 30-50%), and spew fillet formation (improving strength by 15-25%).

4.3. Fatigue Performance Comparison

For welded steel joints under axial tensile loading, fatigue limit at $N = 10^7$ cycles is typically $\Delta\sigma \approx$ 80-120 MPa (R = 0.1), representing a reduction factor of approximately 0.25-0.30 relative to static strength. For adhesive joints (high-performance structural epoxy), static shear strength is $\tau_{static} \approx$ 25-35 MPa with fatigue limit at $N = 10^6$ cycles of $\Delta\tau \approx$ 6-10 MPa (R = 0.1). Experimental data in [7] for SikaFast 580 on X5CrNi18-10 stainless steel yielded a fatigue strength of $\Delta\tau = 6.24$ MPa at 10^6 cycles.

Hybrid weld-bonding joints demonstrate exceptional fatigue performance: fatigue limit at 10^6 cycles reaches $\Delta\sigma$ = 75-95 MPa, compared to $\Delta\sigma$ = 45-55 MPa for pure spot welding and $\Delta\sigma$ = 30-40 MPa for pure adhesive bonding - improvements of 60-70% and 100-140%, respectively. This synergy is mechanistically explained by: (i) stress redistribution - the adhesive layer reduces peak stresses around weld spots by 40-60% (FEM); (ii) crack arrest - weld spots impede crack propagation through the adhesive, while adhesive prevents microcrack initiation from HAZ; (iii) damping - the viscoelastic adhesive absorbs cyclic energy; and (iv) corrosion protection - the adhesive seals weld spots from moisture ingress, eliminating corrosion-fatigue interaction.

5. SURFACE PREPARATION AND DISSIMILAR MATERIAL JOINING

5.1. Surface Preparation Effects

Experimental research reported in [7, 8] demonstrates the critical influence of surface preparation on adhesive joint performance. For X5CrNi18.10 stainless steel bonded with SikaFast 580, mechanical abrasion with P180 abrasive paper yielded the best results - a 20.8% increase in joint strength relative to unprepared specimens - with predominantly cohesive failure mode, R_a = 0.45 μm surface roughness, and controlled oxide layer chemistry. Laser cleaning proved optimal for the same material with SikaPower 880 adhesive, yielding a 17.47% strength increase. For EN AW 5754 Al alloy with SikaFast 580, P180 abrasion again provided the best performance (R_a = 1.75 μm, predominantly cohesive failure). Laser cleaning similarly proved best for EN AW 5754 with SikaPower 880.

Laser texturing, despite achieving near-zero contact angles (<10 °) and elevated oxygen content, did not improve joint strength relative to mechanical preparations. Microstructural examination revealed a redeposited layer with reduced chemical reactivity and mechanical stability. This finding has direct energy efficiency implications: the higher energy cost of laser texturing (100-500 kJ/m^2 vs. 5-20 kJ/m^2 for abrasive paper) does not translate into improved performance for the material-adhesive combinations studied.

5.2. Dissimilar Material Joining

Dissimilar material welding (e.g. steel-aluminium) is severely constrained by large melting point differentials (870 °C for Fe-Al), formation of brittle intermetallic phases (FeAl, Fe_2Al_5, $FeAl_3$) with hardness exceeding 1000 HV, and differential thermal expansion coefficients generating residual stresses near yield strength at the interface. Intermetallic layer thickness above 10 μm leads to drastic joint strength reduction. Adhesive bonding inherently eliminates all three constraints: no metallic melting means no intermetallic phase formation; the relatively low adhesive modulus (E = 2-10 GPa vs. 70-210 GPa for metals) accommodates differential thermal expansion; the adhesive layer provides electrical isolation, preventing galvanic corrosion; and the same joining process applies to any material combination.

6. CIRCULAR ECONOMY AND EPR FRAMEWORK

6.1. Conceptual Framework

Extended Producer Responsibility (EPR), as synthesized by the OECD, defines an approach whereby producers bear environmental and economic responsibilities for their products throughout the entire lifecycle, including the post-consumer phase, with the aim of improving resource efficiency and supporting circular material flows [12]. The choice of joining technology - whether a structure is to be realized by welding, adhesive bonding, or a hybrid joint (e.g. weld-bonding) - ceases to be a purely technical question of load-bearing capacity and production cost, and becomes a strategic decision with implications for the entire product lifecycle. From the standpoint of EPR and circularity, the key evaluation criteria are: durability, repairability, disassembly, recyclability, and secondary material quality.

Welded joints typically impede selective disassembly and component reuse, but allow relatively clean metal streams during recycling. Adhesive joints can improve fatigue resistance and reduce mass (thereby lowering emissions during the use phase), but may simultaneously hinder material separation and lead to waste contamination. Hybrid joints represent a compromise - they enhance structural performance and safety, but further complicate deconstruction and material recovery processes. Within this framework, joining technology selection determines lifecycle performance across all circular economy metrics. A conceptual framework linking these principles with the design and manufacture of metal structures, and highlighting the impact of joining technologies on lifecycle performance indicators and EPR compliance costs, is presented in Figure 1.

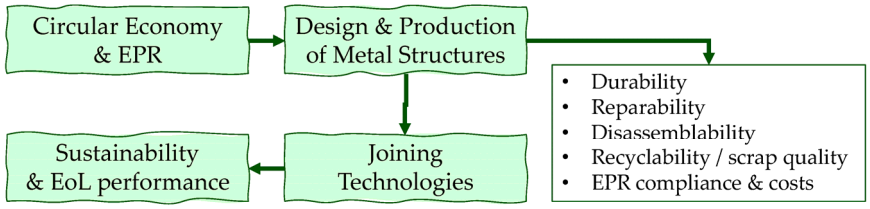


Figure 1: Conceptual framework linking circular economy and extended producer responsibility (EPR) [12]

6.2. Lifecycle Circular Economy Assessment

Welded joints generally impede selective disassembly and component reuse, as their separation requires cutting operations. The resulting metal streams are relatively clean and fully recyclable. However, welded designs are not conducive to Design for Disassembly (DfD) principles.

Adhesive joints can improve fatigue resistance and reduce structural mass - thereby indirectly reducing emissions during the use phase - but pose significant challenges for material separation at end of life. Since most structural adhesives are based on thermoset systems that do not soften under the application of heat, non-destructive disassembly is not practicable.

Hybrid joints combining welding and bonding represent a compromise between structural performance and manufacturing efficiency, while simultaneously further complicating disassembly and material recovery relative to either individual technology. Research aimed at overcoming the limitations of adhesive joints at end of life encompasses several promising directions: thermally-activated adhesives that lose load-bearing capacity above 200 °C, chemically-activated systems, dynamically covalent adhesives with reversible bond breaking, and bio-degradable formulations designed for short-lifecycle products.

6.3. EPR Cost Implications

Under EPR frameworks, producers bear financial responsibility for end-of-life management costs. Joints that complicate disassembly - such as structural adhesives - may impose higher EPR compliance costs, while designs enabling clean metal recovery at end of life can reduce these costs. Carbon pricing mechanisms additionally penalize energy-intensive processes such as conventional welding. The comprehensive evaluation framework for joining technology selection must therefore integrate: initial capital and operating costs, process energy and CO_2 emissions (direct and LCA), mechanical performance requirements, EPR compliance costs, and projected carbon price impact over product lifecycle.

7. TECHNOLOGY COMPARISON SUMMARY

The preceding analysis across energy consumption, thermal effects, mechanical performance, and circular economy metrics enables a structured comparison of the three joining technologies. Table 1 summarizes the key parameters, quantifying the trade-offs between welded, adhesive, and hybrid joints across the dimensions examined in this paper.

The data presented in Table 1 confirm that no single joining technology demonstrates unconditional superiority across all evaluated parameters. Each method exhibits a distinct performance profile that reflects its underlying physical mechanisms and process characteristics. Welding (MIG/MAG) registers the highest direct energy consumption and CO_2 emissions per unit length, alongside significant heat-affected zone widths and residual stresses. However, it retains a clear advantage in end-of-life disassembly and recycling potential, as well as lower EPR compliance risk - factors of increasing regulatory relevance in the context of the European Union's evolving product responsibility frameworks.

Table 1: Comparative overview of welded, adhesive, and hybrid joining processes

Parameter	Welding (MIG/MAG)	Adhesive Bonding	Hybrid	Best
Direct energy [kJ/m]	800–2500	0.1–200	100–140	Adhesive
CO_2 emission [kg/m]	3.5–5.0	0.1–1.2	~1.5–2.5	Adhesive
HAZ width [mm]	10–20 (steel)	None	Reduced	Adhesive
Residual stress [MPa]	300–500	5–15	Intermediate	Adhesive
Stress concentration α_k	1.5–8	1.1–4	~1.3–2.5	Adhesive
Fatigue limit [MPa]	80–120	6–10 (τ MPa)	75–95	Hybrid
Dissimilar materials	Limited	Excellent	Good	Adhesive
Disassembly/recycling	Moderate	Difficult	Difficult	Welding
EPR compliance risk	Low	Moderate–High	Moderate	Welding

Adhesive bonding presents the most favorable environmental profile in terms of energy input and carbon footprint, while also enabling the joining of dissimilar materials with minimal thermal disturbance. These characteristics make it particularly well-suited for lightweight multi-material assemblies in automotive and aerospace applications. Nevertheless, the inherent difficulty of disassembly and the limited recyclability of cured polymer adhesives represent significant constraints from a circular economy perspective, particularly as Extended Producer Responsibility legislation continues to tighten across EU member states.

Hybrid joining emerges as an intermediate solution that partially reconciles the conflicting requirements of mechanical performance and sustainability. By combining adhesive bonding with mechanical fastening or welding, hybrid joints achieve fatigue performance superior to either constituent method alone, while moderating the energy and emissions penalties associated with pure welding processes. The stress concentration factor and residual stress levels are reduced relative to welding, and dissimilar material compatibility is retained from the adhesive component. However, hybrid joints share the end-of-life disassembly challenges of adhesive bonding, limiting their circularity potential.

The matrix given in Table 2 summarizes the relative advantages of each joining technology with respect to sustainability, mechanical performance, and circular economy requirements.

Table 2: Comparative decision matrix for welded, adhesive, and hybrid joining processes

Criterion	Welding	Adhesive Bonding	Hybrid Joining
Energy Efficiency	★	★★★★★	★★★★
Carbon Footprint	★	★★★★★	★★★★
Fatigue Performance	★★★★	★★	★★★★★
Dissimilar Material Joining	★★	★★★★★	★★★★
Recyclability / End-of-Life Recovery	★★★★★	★★	★★

Rating scale: ★ = poor performance, ★★★★★ = excellent performance.

Table 2 presents a rating comparison of all three joining technologies. Adhesive bonding scores highest in energy efficiency, carbon footprint, and dissimilar material joining, making it the best choice when environmental performance is the priority. Welding scores highest in recyclability and end-of-life recovery, as metal separation and reprocessing are well-established processes. Hybrid joining performs best in fatigue resistance, which is important in automotive and industrial structures subjected to cyclic loading.

The choice of joining technology depends on the requirements of each specific application. When recyclability and end-of-life recovery are the main concern, welding is the most suitable option. When the goal is to reduce weight, join different materials, and lower emissions, adhesive bonding is preferred. When high mechanical performance under dynamic loading is required and moderate environmental impact is acceptable, hybrid joining offers the best balance.

The numerical ranges in Table 1 vary depending on material type, joint geometry, and process conditions, and should be treated as approximate values. The ratings in Table 2 agree with findings reported in the literature, though actual performance may differ based on specific material combinations, production scale, and applicable regulations.

8. CONCLUSIONS

This paper has presented a multi-dimensional comparative analysis of welded, adhesive, and hybrid joining processes from the perspectives of energy efficiency, carbon footprint, thermal and mechanical performance, and circular economy compliance. Key conclusions are:

- Adhesive bonding demonstrates clear superiority in energy efficiency - requiring 3-6000× less direct process energy than arc welding - and generates 3-50× lower CO_2 emissions per unit joint length as determined by LCA studies.
- The complete elimination of HAZ in adhesive bonding preserves base material integrity, eliminates thermally-induced residual stresses and distortion, and enables effective dissimilar material joining - advantages unattainable in fusion welding.
- Welding retains its essential position in applications requiring extreme joint strength, hermeticity, service temperatures above 200 °C, and where established industry standards and certifications are mandatory.
- Hybrid weld-bonding represents the optimum solution for mass manufacturing contexts. Fatigue strength improvements of 60-70% over pure welding, 30% energy reduction, and 20-30% stiffness gains demonstrate the synergistic potential of the combined approach.
- Surface preparation remains critical for adhesive joint performance. Mechanical abrasion (P180) is optimal for both X5CrNi18-10 and EN AW 5754 with SikaFast 580, while laser cleaning is superior with SikaPower 880. Laser texturing does not improve joint strength due to redeposited layer formation, making it energetically unjustified.
- From a circular economy and EPR perspective, the development of reversible and bio-based adhesive systems is an active research priority that may resolve the end-of-life limitations of adhesive and hybrid joining.

The energy transition and industrial decarbonization agenda strengthen the case for low-energy joining processes. However, final technology selection must holistically balance energy efficiency, mechanical performance, economic viability, EPR compliance, and end-of-life environmental impact through a lifecycle assessment approach.

REFERENCES

[1] European Commission, "A Clean Planet for All: European long-term strategic vision for a prosperous, modern, competitive and climate neutral economy", COM(2018) 773, Brussels (Belgium), (2018)

[2] R. W. Messler, "Joining of materials and structures: from pragmatic process to enabling technology", Elsevier, Amsterdam (Netherlands), (2004)

[3] International Institute of Welding (IIW), "Global energy consumption in welding processes", Technical Report, (2020)

[4] United Nations Framework Convention on Climate Change (UNFCCC), "Paris Agreement", Paris (France), (2015)

[5] E. M. Petrie, "Handbook of adhesives and sealants", 2nd Edition, McGraw-Hill, New York (USA), (2007)

[6] R. D. Adams, "Adhesive bonding: science, technology and applications", Woodhead Publishing, Cambridge (UK), (2005)

[7] N. Zdravković, "Uticaj metoda pripreme površina od nerđajućeg čelika i legure aluminijuma na mehaničke karakteristike zalepljenih spojeva", PhD Thesis, University of Niš (Serbia), (2026)

[8] N. Zdravković, D. Klobčar, D. Milčić, M. Zupančič, M. Milčić, and A. Đurić, "Influence of surface preparation of aluminum alloy AW-5754 and stainless steel X5CrNi18-10 on the properties of bonded joints", Materials, Vol. 17(11), Article 2561, https://doi.org/10.3390/ma17112561, (2024)

[9] P. Briskham, N. Blundell, L. Han, R. Hewitt, K. Young, and D. Boomer, "Comparison of self-pierce riveting, resistance spot welding and spot friction joining for aluminium automotive sheet", SAE Technical Paper 2006-01-0774, https://doi.org/10.4271/2006-01-0774, (2006)

[10] T. Sadowski and E.Z. Raczkowska, "Hybrid adhesive bonded and riveted joints - influence of rivet geometrical layout on strength of joints", Archives of Metallurgy and Materials, Vol. 57(4), pp. 1127-1135, https://doi.org/10.2478/v10172-012-0126-0, (2012)

[11] S. Maggiore, M. D. Banea, P. Stagnaro, and G. Luciano, "A review of structural adhesive joints in hybrid joining processes", Polymers, Vol. 13(22), Article 3961, https://doi.org/10.3390/polym13223961, (2021)

[12] M. Parisi and G. Di Bella, "Joining Technologies and Extended Producer Responsibility: A Review on Sustainability and End-of-Life Management of Metal Structures", Metals, Vol. 16(1), Article 49, https://doi.org/10.3390/met16010049, (2026)

[13] K. Easterling, "Introduction to the Physical Metallurgy of Welding", Butterworth-Heinemann, Oxford (UK), (1992)

[14] H. K. D. H. Bhadeshia and R. W. K. Honeycombe, "Steels: microstructure and properties", 4th Edition, Butterworth-Heinemann, Oxford (UK), (2017)

[15] J. C. Lippold and D. J. Kotecki, "Welding metallurgy and weldability of stainless steels", Wiley, Hoboken, NJ (USA), (2005)

[16] G. Mathers, "The welding of aluminium and its alloys", Woodhead Publishing, Cambridge (UK), (2002)

[17] D. A. Porter, K. E. Easterling, and M. Y. Sherif, "Phase transformations in metals and alloys", 4th Edition, CRC Press, Boca Raton (USA), (2021)

[18] N. Yurioka, H. Suzuki, S. Ohshita, and S. Saito, "Determination of necessary preheating temperature in steel welding", Welding Journal, Vol. 62, pp. 147s-153s, (1983)

[19] H. Zhang, A. Wei, X. Qiu, and J. Chen, "Microstructure and mechanical properties of resistance spot welded dissimilar thickness DP780/DP600 dual-phase steel joints", Materials & Design, Vol. 54, pp. 443-449, https://doi.org/10.1016/j.matdes.2013.08.027, (2014)

[20] P. J. Withers, "Residual stress and its role in failure", Reports on Progress in Physics, Vol. 70(12), pp. 2211-2264, https://doi.org/10.1088/0034-4885/70/12/R04, (2007)

[21] M. Goland, and E. Reissner, "The stresses in cemented joints", Journal of Applied Mechanics, Vol. 11(1), pp. A17-A27, (1944)

[22] L.J. Hart-Smith, "Adhesive-bonded single-lap joints", NASA Contractor Report CR-112236, https://ntrs.nasa.gov/citations/19740005083, (1973)

SESSION C: PRODUCTION TECHNOLOGIES, CAE AND MANAGEMENT

A fuzzy AHP-based approach for optimal transport mode selection

Nikola Petrović[1*], Vesna Jovanović[1], Jelena Mihajlović[1], Boban Nikolić[1], Saša Marković[1]

[1] University of Niš, Faculty of Mechanical Engineering, Niš, Serbia

ARTICLE INFO

[*] **Correspondence:** petrovic.nikola@masfak.ni.ac.rs

DOI: https://doi.org/10.46793/ET26.C08P **PAGES:** C61 – C66

ABSTRACT

The selection of an optimal transport mode in modern logistics systems represents a complex multi-criteria decision-making problem requiring the simultaneous consideration of economic, operational, and environmental factors. This study applies the Fuzzy Analytic Hierarchy Process (Fuzzy AHP) method to determine the weighting coefficients of evaluation criteria and rank alternative freight transport solutions. The analysis focuses on containerized freight flows, considering road, rail, and inland waterway transport as the main alternatives, evaluated according to transportation cost, delivery time, CO_2 emissions, and reliability. A key contribution of the study is reflected in the comparative analysis of three shipment scenarios with different cargo volumes (10, 100, and 500 TEU), enabling a systematic assessment of the influence of shipment size on transport mode selection. The obtained results indicate that increasing cargo volume improves the competitiveness of rail and inland waterway transport due to economies of scale, particularly in terms of cost efficiency and environmental performance, while road transport maintains an advantage primarily in delivery time. The findings confirm the applicability of MCDM approaches in sustainable transport planning and logistics decision-making.

KEYWORDS

Fuzzy AHP, Transport mode selection, MCDM, Container freight, Sustainable logistics

1. INTRODUCTION

The continuous growth of global trade and the increasing complexity of logistics systems have increased the importance of efficient and sustainable freight transport. Containerization has become one of the key elements of modern logistics because it enables standardization, easier cargo handling, and improved integration of different transport modes. Consequently, transport mode selection has become a complex decision-making problem involving multiple conflicting criteria related to economic efficiency, operational performance, and environmental sustainability.

Traditionally, transport mode selection was primarily based on transportation cost and delivery time. However, increasing environmental concerns, stricter emission reduction policies, and the growing importance of sustainable logistics have expanded the decision-making framework to include additional criteria such as CO_2 emissions, energy efficiency, and transport reliability. As a result, decision-makers are required to evaluate transport alternatives while considering uncertainty and subjective expert judgments.

Multi-Criteria Decision-Making (MCDM) methods have been widely applied in transport and logistics research to support complex decision-making processes [1–4]. Among the most frequently applied methods, the Analytic Hierarchy Process (AHP), introduced by Saaty [5], represents an effective approach for determining criterion importance and ranking alternatives. To overcome the limitations of conventional AHP related to uncertainty and vagueness in expert judgments, the Fuzzy Analytic Hierarchy Process (Fuzzy AHP) was developed through the integration of fuzzy set theory with the classical AHP approach [6,7].

Previous studies demonstrated the applicability of hybrid MCDM approaches, including AHP, fuzzy AHP, and fuzzy TOPSIS, in sustainable transport planning and multicriteria logistics evaluation problems [8–10]. In addition, recent studies confirmed the effectiveness of fuzzy MCDM methodologies in transport systems, logistics optimization, transport sustainability assessment, and engineering decision-making problems [11–17].

This paper applies the Fuzzy AHP method to the problem of transport mode selection in containerized freight transport. Three transport alternatives are considered, namely road, rail, and inland waterway transport, while the evaluation is performed according to transportation cost, delivery time, CO_2 emissions, and transport reliability. The analysis is conducted for three shipment scenarios involving 10 TEU, 100 TEU, and 500 TEU in order to investigate the influence of cargo volume on transport mode competitiveness.

The remainder of the paper is organized as follows. Section 2 presents the methodological background of the Fuzzy AHP method. Section 3 describes the case study and decision matrix construction. Section 4 presents the obtained results and discussion, while the final section provides concluding remarks and directions for future research.

2. METHODOLOGY

2.1. Fuzzy AHP Method

The Fuzzy Analytic Hierarchy Process (Fuzzy AHP) method represents an extension of the conventional Analytic Hierarchy Process (AHP) developed by Thomas L. Saaty. The method integrates fuzzy set theory into the pairwise comparison procedure in order to better represent uncertainty and vagueness in expert judgments. Unlike classical AHP, which relies on precise numerical values, Fuzzy AHP enables decision-makers to express preferences using linguistic variables and fuzzy numbers. Similar fuzzy MCDM approaches have been successfully applied in various engineering and transport decision-making problems [9,10].

The Fuzzy AHP procedure applied in this study consists of the following steps:

- Definition of alternatives and evaluation criteria;
- Construction of the fuzzy pairwise comparison matrix;
- Calculation of fuzzy synthetic values;
- Determination of fuzzy criterion weights;
- Defuzzification and normalization of weights;
- Ranking of transport alternatives.

The obtained weighting coefficients are subsequently used for evaluating the considered transport alternatives within different shipment scenarios.

2.2. Triangular Fuzzy Numbers

In this study, triangular fuzzy numbers (TFNs) are used to represent pairwise comparisons between evaluation criteria. A triangular fuzzy number is defined as:

$$\tilde{A} = (l, m, u) \tag{1}$$

where:

- l denotes the lower bound,
- m denotes the most probable value,
- u denotes the upper bound.

The linguistic importance scale applied in the analysis is presented in Table 1.

Table 1: Linguistic importance scale and triangular fuzzy numbers

Linguistic importance	Triangular fuzzy number
Equal importance	(1, 1, 1)
Moderate importance	(2, 3, 4)
Strong importance	(4, 5, 6)
Very strong importance	(6, 7, 8)
Extreme importance	(8, 9, 10)

The application of triangular fuzzy numbers enables more flexible modeling of subjective expert evaluations and reduces the limitations associated with precise numerical judgments.

2.3. Determination of Criterion Weights

The fuzzy pairwise comparison matrix is constructed based on expert evaluations regarding the relative importance of the considered criteria. The geometric mean method is then applied to calculate the fuzzy weights of criteria.
The fuzzy geometric mean value for criterion i is calculated as:

$$\tilde{r}_i = \left(\prod_{j=1}^{n} \tilde{a}_{ij}\right)^{1/n} \tag{2}$$

where:

- $\tilde{a}_{ij}$ represents the fuzzy comparison value between criteria i and j,
- n denotes the total number of criteria.

The fuzzy weights are obtained using:

$$\widetilde{w}_i = \tilde{r}_i \otimes (\sum_{i=1}^{n} \tilde{r}_i)^{-1} \quad (3)$$

After determining fuzzy weights, the defuzzification procedure is performed in order to transform fuzzy values into crisp weighting coefficients. In this study, the center of area (COA) method is applied:

$$w_i = \frac{(l_i + m_i + u_i)}{3} \quad (4)$$

Finally, the obtained weights are normalized and used for the evaluation and ranking of transport alternatives for different shipment scenarios (10 TEU, 100 TEU, and 500 TEU). The proposed Fuzzy AHP framework was applied to the transport corridor between Pančevo and Novi Sad in order to evaluate the competitiveness of road, rail, and inland waterway transport under different shipment scenarios. The obtained criterion weights were combined with the normalized performance values of transport alternatives to determine the final evaluation scores and rankings.

3. CASE STUDY

The case study focuses on the transport corridor between Pančevo and Novi Sad, Serbia, which represents an important regional freight transport route with developed road, rail, and inland waterway infrastructure. The corridor was selected due to its significant intermodal transport potential and the possibility of analyzing different transport alternatives under varying shipment volumes. The considered logistics scenario enables the comparative evaluation of economic, operational, and environmental performance of the analyzed transport modes.

The proposed Fuzzy AHP methodology is applied to the problem of transport mode selection in containerized freight transport. The analysis considers three commonly used transport alternatives, namely road transport, rail transport, and inland waterway transport. The objective of the study is to evaluate the competitiveness of these transport modes under different shipment volumes while simultaneously considering economic, operational, and environmental criteria.

3.1. Transport Alternatives

The following transport alternatives are included in the analysis:

- A_1 – Road transport, characterized by high flexibility and short delivery time, but associated with higher transportation costs and CO_2 emissions for large shipment volumes;
- A_2 – Rail transport, characterized by favorable transportation costs and reduced environmental impact, particularly for medium and large shipment sizes;
- A_3 – Inland waterway transport, characterized by high transport capacity and low environmental impact, but generally associated with longer delivery times.

3.2. Evaluation Criteria

The transport alternatives are evaluated according to four criteria frequently used in transport and logistics planning:

- C_1 – Transportation cost (€). Represents the total transportation cost for a given shipment volume. Lower values are preferred (min).
- C_2 – Delivery time (h). Represents the total transport duration from origin to destination. Lower values are preferred (min).
- C_3 – CO_2 emissions (kg). Represents the environmental impact generated during transportation. Lower values indicate better environmental performance (min).
- C_4 – Reliability (%). Represents the consistency and dependability of transport operations. Higher values are preferred (max).

3.3. Shipment Scenarios

In order to analyze the influence of cargo volume on transport mode competitiveness, three shipment scenarios are considered:

- Scenario 1 (S_1): 10 TEU
- Scenario 2 (S_2): 100 TEU
- Scenario 3 (S_3): 500 TEU

The considered scenarios enable a comparative analysis of economies of scale effects and provide insight into the sustainability and operational performance of different transport modes under varying transport demand conditions.

3.4. Decision Matrix

The decision matrix used in the analysis is presented in Table 2.

Table 2: Decision matrix for analyzed shipment scenarios

Scenario	Alternative	Cost (€)	Time (h)	CO_2 (kg)	Reliability (%)
S1 – 10 TEU	Road transport	600	2.0	1320	90

Scenario	Alternative	Cost (€)	Time (h)	CO_2 (kg)	Reliability (%)
	Rail transport	480	3.3	216	80
	Inland waterway transport	360	8.3	360	85
S2 – 100 TEU	Road transport	5200	2.5	13200	88
	Rail transport	3200	4.0	2160	84
	Inland waterway transport	2500	10.0	3600	82
S3 – 500 TEU	Road transport	25500	3.5	66000	85
	Rail transport	13500	5.0	10800	86
	Inland waterway transport	9800	12.0	18000	83

The presented values are used as the basis for the application of the Fuzzy AHP method and the comparative evaluation of transport alternatives under different shipment volumes.

4. RESULTS AND DISCUSSION

Based on expert evaluations and the application of the Fuzzy AHP method, the weighting coefficients of the considered criteria were determined (Table 3). The consistency of expert judgments was verified through the calculated consistency ratio (CR = 0.08), indicating acceptable consistency.

Table 3: Fuzzy pairwise comparison matrix of evaluation criteria

Criteria	C_1 Cost	C_2 Time	C_3 CO_2	C_4 Reliability
C_1 Cost	(1,1,1)	(2,3,4)	(1,2,3)	(4,5,6)
C_2 Time	(1/4,1/3,1/2)	(1,1,1)	(1/5,1/4,1/3)	(2,3,4)
C_3 CO_2	(1/3,1/2,1)	(3,4,5)	(1,1,1)	(4,5,6)
C_4 Reliability	(1/6,1/5,1/4)	(1/4,1/3,1/2)	(1/6,1/5,1/4)	(1,1,1)

The fuzzy pairwise comparison matrix was constructed according to the relative importance of the evaluation criteria. Transportation cost (C_1) and CO_2 emissions (C_3) were assigned higher importance because they directly reflect the economic and environmental efficiency of freight transport. The obtained normalized criterion weights are presented in Table 4.

Table 4: Normalized criterion weights obtained by Fuzzy AHP

Criterion	Description	Weight
C_1	Transportation cost	0.37
C_2	Delivery time	0.18
C_3	CO_2 emissions	0.32
C_4	Reliability	0.13

The results indicate that transportation cost and CO_2 emissions represent the most influential criteria in the transport mode selection process, confirming the increasing importance of sustainable and cost-efficient logistics systems.

After determining criterion weights, the transport alternatives were evaluated across the analyzed shipment scenarios. The obtained results are presented in Table 5.

Table 5: Final scores and ranking of transport alternatives

Scenario	Alternative	Score	Rank
S1 – 10 TEU	Rail transport	0.412	1
	Inland waterway transport	0.356	2
	Road transport	0.232	3
Scenario	Alternative	Score	Rank
S2 – 100 TEU	Rail transport	0.438	1
	Inland waterway transport	0.371	2
	Road transport	0.191	3
S3 – 500 TEU	Inland waterway transport	0.446	1
	Rail transport	0.421	2
	Road transport	0.133	3

The obtained scores indicate that rail transport achieves the best balance between economic, operational, and environmental performance for small and medium shipment volumes. However, inland waterway transport becomes the most competitive alternative for large shipment sizes due to economies of scale and lower environmental impacts. In contrast, the competitiveness of road transport decreases with increasing cargo volume because of higher transportation costs and CO_2 emissions. Figure 1 presents a graphical comparison of transport alternative scores for different shipment scenarios.

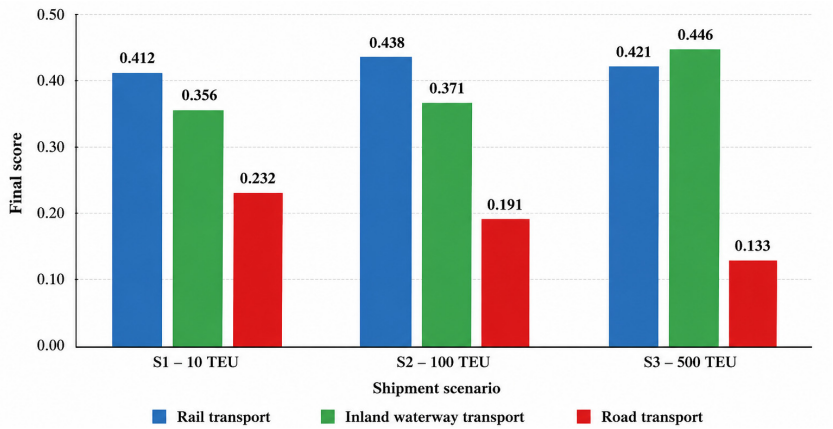


Figure 1: Comparative evaluation of transport alternatives under different shipment volumes

The graphical analysis confirms the strong influence of cargo volume on transport mode selection. Smaller shipment volumes favor road transport due to operational flexibility and shorter delivery times, while larger shipment volumes favor rail and inland waterway transport because of their superior sustainability and cost-efficiency characteristics. Similar applications of fuzzy MCDM approaches in transport and logistics systems were reported in recent studies [16,17].

The obtained findings confirm the applicability of the Fuzzy AHP method in complex transport planning problems involving uncertainty and multiple conflicting criteria.

5. CONCLUSION

This paper presented the application of the Fuzzy Analytic Hierarchy Process (Fuzzy AHP) method for transport mode selection in containerized freight transport. The proposed approach enabled the evaluation of three transport alternatives, namely road, rail, and inland waterway transport, based on multiple economic, operational, and environmental criteria including transportation cost, delivery time, CO_2 emissions, and reliability.

The analysis was conducted for three shipment scenarios involving 10 TEU, 100 TEU, and 500 TEU in order to investigate the influence of cargo volume on transport mode competitiveness. The obtained results indicated that transportation cost and CO_2 emissions represent the most influential criteria in the decision-making process. Furthermore, the results demonstrated that increasing shipment volume significantly improves the competitiveness of rail and inland waterway transport due to economies of scale and better environmental performance.

Rail transport achieved the best overall balance between economic efficiency and sustainability for small and medium shipment volumes, while inland waterway transport became the most favorable solution for large shipment sizes. In contrast, road transport-maintained advantages primarily in terms of delivery time, but showed lower competitiveness for larger cargo volumes due to higher transportation costs and CO_2 emissions.

The findings confirm that the Fuzzy AHP method represents an effective decision-support tool for sustainable transport planning under uncertainty. The proposed framework may assist logistics planners and decision-makers in selecting appropriate transport solutions while simultaneously considering operational efficiency and environmental sustainability. Future research may include additional criteria, larger transport networks, and the integration of other MCDM methods for comparative analysis.

The study is limited by the use of estimated operational data and a limited number of evaluation criteria.

ACKNOWLEDGEMENTS

This research was financially supported by the Ministry of Education, Science and Technological Development and Inovation of the Republic of Serbia (Contract No. 451-03-34/2026-03/ 200109).

REFERENCES

[1] S. Marković, N. Petrović, D. Marinković, B. Nikolić, and N. Komatina, “Environmental assessment of cruise ships and superyachts with multi-criteria evaluation of marine fuels”, Applied Sciences, Vol. 16(9), Article 4287, https://doi.org/10.3390/app16094287, (2026)

[2] N. Petrović, D. Marinković, N. Komatina, M. Runić-Ristić, T. M. Qureshi, and A. Aleksić, “Fuzzy MCDM methodology application in analysis of annual operational efficiency in passenger and freight air transport”, Mathematics, Vol. 13(21), Article 3535, https://doi.org/10.3390/math13213535, (2025)

[3] N. Petrović, V. Jovanović, D. Marinković, B. Nikolić, and S. Marković, “Logistics companies’ efficiency analysis and ranking by the DEA-Fuzzy AHP approach”, Applied Sciences, Vol. 15(17), Article 9549, https://doi.org/10.3390/app15179549, (2025)

[4] N. Mandal, P. Sarkar, and N. Petrović, “Multi-criteria decision-making for ranking renewable energy sources: a case study from the Republic of Serbia”, Journal of Engineering Management Systems Engineering, Vol. 4(2), pp. 109 –121, https://doi.org/10.56578/jemse040203, (2025)

[5] T. L. Saaty, “The analytic hierarchy process”, McGraw-Hill, New York (USA), (1980)

[6] J. J. Buckley, “Fuzzy hierarchical analysis”, Fuzzy Sets and Systems, Vol. 17(3), pp. 233-247, https://doi.org/10.1016/0165-0114(85)90090-9, (1985)

[7] D. Y. Chang, “Applications of the extent analysis method on fuzzy AHP”, European Journal of Operational Research, Vol. 95(3), pp. 649-655, https://doi.org/10.1016/0377-2217(95)00300-2, (1996)

[8] C. Macharis, J. Springael, K. De Brucker, and A. Verbeke, “PROMETHEE and AHP: The design of operational synergies in multicriteria analysis”, European Journal of Operational Research, Vol. 153(2), pp. 307-317, https://doi.org/10.1016/S0377-2217(03)00153-X, (2004)

[9] G. Büyüközkan and G. Çifçi, “A combined fuzzy AHP and fuzzy TOPSIS based strategic analysis of electronic service quality in healthcare industry”, Expert Systems with Applications, Vol. 39(3), pp. 2341-2354, https://doi.org/10.1016/j.eswa.2011.08.061, (2012)

[10] A. Awasthi, S. S. Chauhan, and H. Omrani, “Application of fuzzy TOPSIS in evaluating sustainable transportation systems”, Expert Systems with Applications, Vol. 38(10), pp. 12270-12280, https://doi.org/10.1016/j.eswa.2011.04.005, (2011)

[11] N. Komatina, N. Petrović, D. Pamučar, V. Simić, and D. Marinković, “Multi-criteria ranking of industrial presses with respect to operational performance using the square-root based evaluation method (SREM)”, Facta Universitatis, Series: Mechanical Engineering, Online First, https://doi.org/10.22190/FUME260305009K, (2026)

[12] A. Ulutaş, F. Balo, L. Sua, E. Demir, A. Topal, and V. Jakovljević, “A new integrated grey MCDM model: case of warehouse location selection”, Facta Universitatis, Series: Mechanical Engineering, Vol. 19(3), pp. 515-534, https://doi.org/10.22190/FUME210424060U, (2021)

[13] Ž. Stević, M. Baydaş, M. Kavacık, E. Ayhan, and D. Marinković, “Selection of data conversion technique via sensitivity-performance matching: ranking of small e-vans with PROBID method”, Facta Universitatis, Series: Mechanical Engineering, Vol. 22(4), pp. 643-671, https://doi.org/10.22190/FUME240305023S, (2024)

[14] D. Tadić, J. Lukić, N. Komatina, D. Marinković, and D. Pamučar, “A fuzzy decision-making approach to electric vehicle evaluation and ranking”, Technical Gazette, Vol. 32(3), pp. 1066-1075, https://doi.org/10.17559/TV-20250223002410, (2025)

[15] C. N. Wang, T. Q. Le, K.H. Chang, and T. T. Dang, “Measuring road transport sustainability using MCDM-based entropy objective weighting method”, Symmetry, Vol. 14(5), 1033, https://doi.org/10.3390/sym14051033, (2022)

[16] I. Badi, M. B. Bouraima, and C. K. Kiptum, “Evaluating the barriers to logistics outsourcing through a fuzzy multi-criteria decision-making model”, Spectrum of Mechanical Engineering and Operational Research, Vol. 3(1), pp. 65-78, https://doi.org/10.31181/smeor31202653, (2026)

[17] K. H. Gazi, A. Bera, S. P. Mondal, and A. Ghosh, “Finding humanitarian supply chain management challenges using uncertain MCDM methodology”, Spectrum of Mechanical Engineering and Operational Research, Vol. 2(1), pp. 248-279, https://doi.org/10.31181/smeor21202548, (2025)

SESSION C: PRODUCTION TECHNOLOGIES, CAE AND MANAGEMENT

Selection and analysis of key occupational safety and health indicators: A case study in the foundry industry

Aleksandar Jovičić[1], Dejan Veljković[1], Anđelija Mitrović[1*], Jelena Baralić[1]

[1] University of Kragujevac, Faculty of Technical Sciences/Department of Mechatronics, Čačak, Serbia

ARTICLE INFO

[*] **Correspondence:** andjelija.mitrovic@ftn.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.C09J **PAGES:** C67 – C75

ABSTRACT

Selection of key performance indicators (KPI) in the field of occupational safety and health requires precise definition of measurable parameters which allow efficient monitoring of the level of employee protection and evaluation of the success of implemented preventive measures. In the foundry industry, which is characterized by high temperatures, use of heavy equipment, complex technological processes and the presence of hazardous and harmful materials, continuous monitoring of situation in the field of employee health protection is a significant basis for improving working conditions, reducing professional risks and increasing the level of employee safety and health. This paper presents the identification and analysis of key indicators of occupational safety and health through a case study conducted in a foundry in Serbia. Special attention was paid to the LTIFR (Lost Time Injury Frequency Rate), SR (Severity Rate) and AFR (Accident Frequency Rate) indicators, with the analysis of their values in the period 2018-2024, which allowed the quantitative assessment of workplace injury frequency and severity. The results indicate an increased level of professional risk in the observed foundry from the case study. The analysis established that the highest risk workplaces are connected with machine molding, grinding and final treatment of castings, while arms and eyes are the most frequently injured body parts. The obtained results emphasize the importance of preventive measures, education of employees, improvement of risk assessment, more consistent use of personal protective equipment and modernization of technological processes for the purpose of improving occupational safety and health.

KEYWORDS

Occupational safety and health, Foundry industry, Workplace injuries, KPIs, LTIFR, AFR, Severity Rate, Risk assessment, Industrial safety, Occupational risk.

1. INTRODUCTION

Key performance indicators (KPI) are quantitative and qualitative indicators which serve for measuring, monitoring and managing of business results of an organization. They are instruments used by a company to monitor the degree of success in the accomplishment of its strategic goals, vision and business policy. At the same time, KPIs allow the assessment of business efficiency, identification of problematic areas and improvement of decision-making processes and, therefore, they occupy a significant place in modern management and professional practice. A large number of different indicators for measuring company performance have been developed in theory and practice. However, the main challenge of successful management relates to the selection of adequate indicators, i.e. determination of a limited number of key indicators among a great number of potential performance measures. The selected KPIs should cover those factors which can be directly influenced by the organization and which have a crucial importance for its long-term and sustainable development.

In order to assess and improve the existing business operations, it is necessary for the management to translate organizational goals into measurable categories. Various types of indicators are used in that process, where each of them measures a certain aspect of business operations and provides the management with relevant information for decision-making. KPIs allow comparison of the achieved results with the target values defined in advance, the results from previous periods, the established standards, as well as the performances of competitive organizations [1]. Since the performance indicators must be reliable, objective and informative, the selected KPIs should fulfill several basic criteria:

- to be based on relevant and verifiable data,
- to be sensitive to changes in the business environment,
- to be specific for the analyzed situation and goals of the organization,
- to be calculated based on the available data.

In addition to monitoring the efficiency of production, product quality and the level of exploitation of resources, key performance indicators in the foundry industry have a significant role in the field of occupational safety and health as well, because they allow the identification of risk factors and the analysis of frequency of workplace injuries for the purpose of improving the safety of the work environment. The foundry industry is one of the highest risk sectors from the aspect of occupational safety and health due to the complex action of technological processes and conditions of the work environment. Previous research has shown that the most frequent injuries in foundries are burns, mechanical injuries and musculoskeletal disorders, while the most serious incidents are connected with spilling of liquid metal, explosions and interventions on the equipment [2]. Besides occupational injuries, the impact of long-term exposure to harmful effects, such as silica dust, noise and heat stress, is also important, which is confirmed by the reports of Health, Safety and Environment - HSE and recent research in the field of public health [3].

Modern approaches to safety are based on monitoring key performance indicators (KPI), which enable systematic risk assessment and injury prevention. The selection of KPIs in the field of occupational safety and health is a complex process which implies precise definition of measurable parameters for monitoring the safety level in industrial systems. In the foundry industry, which is characterized by high temperatures, application of heavy industrial equipment and the presence of hazardous substances, systematic monitoring of safety indicators has a special importance for improving employee safety and reducing risks of workplace injuries. Therefore, the increasing importance is assigned to proactive indicators, such as the number of near-miss events (avoided incidents), level of compliance with maintenance procedures, exposure to harmful effects and indicators of employee behavior. Research shows that a large number of near-miss events are a reliable predictor of serious accidents, which is in accordance with classical models of industrial safety (e.g. Heinrich model, also known as the Domino Theory).

A particularly important factor is the maintenance of equipment. Papers from the field of risk assessment suggest that inadequate maintenance procedures or the non-obeyed ones are among the most frequent causes of severe and fatal injuries [1], which is confirmed by analyses of real accidents from the bases of the National Institute for Occupational Safety and Health (NIOSH) in the USA. Besides technical aspects, human factor has a key role because risk perception, the level of training and the respect of procedures directly influence the probability of injury occurrence [4].

Accordingly, efficient occupational safety and health management in foundries requires an integrated approach which combines reactive and proactive indicators. Monitoring of the injury rate, near-miss events, maintenance parameters, exposure to harmful effects and the indicators of safety behavior allows timely risk identification and directing preventive measures. Previous research indicates that such a systematic approach makes the basis for reducing the frequency and severity of injuries in the foundry industry.

This paper presents an example of identification and analysis of key occupational safety and health indicators through a case study in the foundry industry in Serbia. Special attention is dedicated to the LTIFR (Lost Time Injury Frequency Rate), SR (Severity Rate) and AF (Accident Frequency) indicators, with the presentation of their values in an eight-year period, which enables quantitative assessment of workplace injury frequency and severity. Also, the paper presents the analysis of the total number of injuries according to the degree of injury, causes of injury, then the analysis of injuries in relation to the workplace, body part which is injured, time of occurrence, and even the day of injury occurrence is analyzed.

The research methodology in the paper is based on the authors' independent interpretation of data from the foundry from the case study with the application of critical analysis, comparison and synthesis of data in the definition of findings, with the qualitative-descriptive approach in the presentation of results and derivation of conclusions. The key research question in the paper is: "Which key performance indicators most efficiently indicate the level of professional risk and the need for improving occupational safety and health measures in the foundry industry?".

Consequently, the aim of this paper is to assess the level of professional risk in the foundry industry through identification, analysis and application of KPIs in the field of occupational safety and health as well as to establish the possibilities of improving preventive measures for the purpose of reducing workplace injuries, improving working conditions and increasing the level of employee health protection.

2. KEY OCCUPATIONAL SAFETY AND HEALTH INDICATORS IN THE FOUNDRY INDUSTRY

The Fatal Incident Frequency Rate - FIFR and the Lost Time Incident Frequency Rate – LTIFR indicators stand out as the examples of reactive performance indicators, especially in the companies with pronounced risks regarding occupational safety and health. The FIFR indicator is defined as the number of fatalities per 100 million of hours worked, while the LTIFR is the number of lost-time workplace injuries per one million of hours worked. These indicators enable the assessment of the safety level in the organization and the identification of critical areas which require the improvement of occupational safety measures [5].

In this paper, special attention is dedicated to the analysis of key occupational safety and health indicators, i.e. LTIFR (Lost Time Injury Frequency Rate), SR (Severity Rate) and AFR (Accident Frequency Rate) indicators, with the presentation of their values in an eight-year period in the foundry from the case study. Quantitative assessment of both frequency and severity of workplace injuries is thus enabled, as well as comprehensive consideration of safety performances in the observed system. The analysis also includes the Fatal Accident Rate - FAR, i.e. the number of fatalities among employees and operators per 100 million working hours; however, as there were no fatalities in the observed period, this indicator was not further considered in the paper.

The Accident Frequency Rate - AFR indicator is defined as the ratio of the number of reported accidents resulting in more than one day of absence from work to the total number of hours worked [6], where the result is standardized to one million of working hours:

$$AFR = \frac{Number\ of\ accidents\ (absence>1day)\ x\ 1000000}{Total\ number\ of\ hours\ worked} \quad (1)$$

The LTIFR indicators is the frequency of lost-time injuries and it is expressed as the number of such injuries per one million of hours worked:

$$LTIFR = \frac{Total\ number\ of\ lost\text{-}time\ injuries\ x\ 1000000}{Total\ number\ of\ hours\ worked} \quad (2)$$

The Severity Rate - SR indicator is defined as the ratio of the total number of workdays lost due to injuries to the total number of hours worked, where the value is also standardized:

$$SR = \frac{Total\ number\ of\ lost\ workdays\ x\ 1000000}{Total\ number\ of\ hours\ worked} \quad (3)$$

The total number of working hours is the sum of all hours effectively worked by employees in the observed period and can be approximated using the following expression:

$$Working\ hours = number\ of\ employees\ \ x\ number\ of\ working\ days\ \ x\ 8h \quad (4)$$

The application of the abovementioned indicators allows standardized monitoring and comparison of safety levels between different time intervals and organizational units, regardless of the size of the system. Their complementarity is particularly important, taking into account that AFR and LTIFR indicate the frequency of injuries, while SR provides an insight into their consequences and seriousness. It is possible to use an integrated analysis of these indicators to identify critical points in the working process, define priorities in risk management and improve the occupational safety and health system.

3. ANALYSIS OF WORKPLACE INJURIES IN THE PERIOD 2018-2024

According to the reports of the International Labour Organization (ILO), heavy industry, including foundries, is characterized by an increased injury frequency in relation to the other industrial sectors. Similarly, the data obtained by the Bureau of Labor Statistics (BLS) indicate that the metal production sector records above-average workplace injury rates, especially regarding injuries which lead to the absence from work.

Modern approaches to safety management are based on the application of quantitative indicators, such as LTIFR (Lost Time Injury Frequency Rate), AFR (Accident Frequency Rate) and SR (Severity Rate), which allow monitoring of the frequency and severity of injuries as well as comparison of safety performances between different systems and time periods. According to the Health and Safety Executive and EU-OSHA, the LTIFR values below 5 are considered indicators of high safety performance, while the values above 20 indicate an increased risk, which is particularly pronounced in industries with a high degree of hazard, e.g. foundry production.

The analysis of workplace injuries in the foundry from the case study, in the observed period from 2018 to 2024, is based on monitoring absolute indicators (total number of injuries, injury structure and the number of lost workdays) as well as relative indicators of safety performances (AFR, LTIFR and SR). Such an approach allows comprehensive consideration of the safety situation, both from the aspect of frequency and the aspect of severity of injuries.

The total number of workplace injuries during the analyzed period shows moderate oscillations, without a clear trend of continuous rise or fall. The largest number of injuries were recorded in 2022 (19 injuries), (Table 1), while the lowest level was recorded in 2021 (13 injuries). These variations can be the consequence of changes in the scope of production, organization of work or the level of implementation of safety measures.

Table 1: Workplace injuries and key safety indicators in the foundry for the period 2018-2024

Year	Number of employees	Total injuries	Minor injuries	Severe injuries	Lost days due to injuries	AFR	LTIFR	SR
2018	156	14	14	0	161	44.87	44.87	516.3
2019	174	15	14	1	300	43.10	43.10	862.07
2020	179	17	17	0	350	47.49	47.49	977.65

Year	Number of employees	Total injuries	Minor injuries	Severe injuries	Lost days due to injuries	AFR	LTIFR	SR
2021	208	13	13	0	350	31.25	31.25	841.35
2022	202	19	19	0	154	47.03	47.03	381.19
2023	194	15	14	1	370	38.66	38.66	953.61
2024	170	16	13	3	466	47.06	47.06	1370.59

Although the AFR (Accident Frequency Rate) and LTIFR (Lost Time Injury Frequency Rate) indicators are often used as synonyms in practice, there is a certain difference between them. LTIFR refers exclusively to the injuries that have led to the absence from work, while AFR can encompass a wider range of reported injuries and accidents, depending on the methodology of organization or the standard being applied.

In this research, AFR and LTIFR have identical values, taking into account that all the reported injury cases resulted in the absence from work, which did not allow differentiation between these two indicators.

An especially important aspect of analysis is injury structure. Minor injuries are dominant in most years, which indicates the presence of frequent, but less serious incidents. However, the appearance of severe injuries in certain years (2019, 2023 and particularly 2024) considerably influences the overall safety evaluation. The year of 2024 stands out as critical, with three severe injuries, which represents the highest value in the observed period, and indicates the deterioration of working conditions or the insufficient efficiency of preventive measures.

The analysis of lost workdays additionally confirms these observations. Although the number of injuries does not show a drastic decrease, the number of lost days considerably varies, where the highest level was recorded in 2024 (466 days). This indicates increased severity of injuries, which is in accordance with the increased number of severe injuries in the same year. On the other hand, the year of 2022, despite the largest number of injuries, has a relatively low level of lost workdays (154), which indicates the dominance of minor injuries.

AFR and LTIFR indicators show relatively stable values over the mentioned period, with a slight decrease in 2021, which may indicate temporary improvement of safety conditions. However, there is a repeated increase afterwards, which indicates the lack of continuity in the implementation of safety measures. Unlike them, the SR indicator shows significant oscillations and clearly identifies critical years, 2020, 2023 and 2024.

The thorough analysis indicates that safety in the observed system cannot be adequately evaluated exclusively on the basis of the number of injuries. The combination of data about the structure of injuries and the indicators of severity shows that even a relatively stable number of injuries can be accompanied by a significant deterioration of consequences, which is particularly pronounced in 2024.

3.1. Analysis of AFR Indicators

The values of AFR indicators in the observed period range between 31 and 47, which shows a very high level of frequency of workplace injuries. According to the criteria of the International Labour Organization and the Health and Safety Executive, the values above 20 are considered high risk indicators, while the values above 30 indicate a very high-risk level characteristic of heavy industry.

Although certain oscillations are noticed, the lowest value was recorded in 2021 (31), which may indicate a temporary improvement of safety conditions. However, the absence of a continuous downward trend indicates the instability of the occupational safety and health management system and the insufficient efficiency of preventive measures.

3.2. Analysis of LTIFR Indicators

The values of LTIFR indicators completely follow AFR, as expected, because both indicators are based on the number of injuries with the absence from work. The range from 31 to 47 shows that the frequency of time-lost injuries is significantly above the recommended values of good practice.

If the international benchmark values for metal processing industry and steel industry are compared, it can be noticed that the analyzed foundry from the case study records considerably higher values of LTIFR indicators, which indicates an increased risk level and the need for improving occupational safety and health measures [7-9].

By comparing the obtained values of LTIFR indicators in the analyzed foundry from the case study, which, in the observed period, ranged from 31.25 to 47.49 injuries per million working hours, with the benchmark data from the Bureau of Labor Statistics and the American Foundry Society for foundry industry in the USA, where the injury rates range approximately from 6.0 to 8.0 cases per 100 employees, it can be concluded that the analyzed system is in the upper section of the range characteristic of metal processing industry and indicates an increased level of professional risk [10]. This suggests that there is a significant potential for improvement, especially in the field of prevention of injuries which result in the absence from work. The lack of a decreasing trend in LTIFR values during the observed period additionally confirms that safety measures neither were systematically improved nor produced long-term effects.

3.3. Analysis of SR Indicators

The SR Indicator Shows Additional Specificities of the Analyzed System. the Highest Value of SR Was Recorded in 2024 (1370.59), with a Simultaneous Increase in the Number of Severe Injuries and a Considerable Number of Lost Workdays. Such a Result Indicates That the Consequences of Injuries Became More Serious and That They Had a More Pronounced Impact on the Work Capacity of Employees and the Continuity of the Production Process. the High Values of SR Indicators Represent a Particularly Important Indicator in the Foundry Industry Because They Suggest the Possibility of Occurrence of More Serious Injuries Due to Work with Hazardous Materials and Equipment.

On the other hand, the year of 2021 shows the most favorable results from the aspect of injury frequency, which may indicate a more efficient implementation of occupational safety measures, better organization of the process or a reduced scope of production and exposure to risks. It is also interesting that the largest number of injuries were recorded in 2022 with the lowest SR value at the same time, which indicates that the injuries were less serious and that the employees' absences from work were shorter.

In comparison with modern practice in developed foundries, where special attention is dedicated to preventive action, risk assessment, training of employees and automation of high-risk operations, the analyzed results show the need for additional improvement of the occupational safety and health system. Special focus should be placed on the identification of causes of injury repetition, improvement of preventive measures, more consistent use of personal protective equipment as well as continued education of employees for the purpose of reducing the frequency and severity of workplace injuries. These results confirm that the frequency of injuries is not a sufficient safety indicator, but that it is necessary to analyze their severity.

The combined analyses of AFR, LTIFR and SR indicators shows that the system is characterized by a high level of injury frequency with significant oscillations in their severity. This confirms the need for an integrated approach to occupational safety and health management, which simultaneously covers the reduction of both frequency and severity of injuries.

The analysis of occupational safety and health was extended to the indicators presented in Table 2, which cover the review of injuries in relation to the workplace, injured body parts, time and day of injury occurrence, as well as the number of lost working hours in the foundry for the period 2018–2024.

3.4. Analysis of Workplaces from the Aspect of Workplace Injuries

The analysis of workplace injuries in the period 2018–2024 established that the largest number of injuries continuously occur at workplaces connected with machine molding, steel casting, core making and, therefore, these workplaces can be considered the highest risk segments of the production process (Figure 1). The years of 2019 and 2021 particularly stand out as the largest number of injuries were recorded in machine molding jobs in that period.

The increased injury frequency is connected with the specificities of the technological process, which include intensive manual labor, use of rotary tools, contact with metal surfaces, presence of dust, increased physical effort and repetitive operations. Additional risk factors are employee fatigue, reduced concentration and exposure to increased noise and vibrations. The continued presence of injuries recorded in the jobs of molding, core making and maintenance of plants indicates the existence of system risks within the production process. During 2023, an increase in the number of injuries in assembly jobs was recorded, which may be associated with increased work intensity and organizational load.

The analysis of injury structure shows that the most frequently injured body parts were arms and eyes, which is directly connected with the jobs of metal processing and material handling. The mentioned results suggest the need for more consistent application of personal protective equipment, especially for protection of hands and eyes. Based on the analysis conducted, it can be concluded that the key risk factors are manual labor, physical effort, working with tools and cast metal as well as the insufficient implementation of occupational safety and health measures. For the purpose of reducing the number of injuries, it is necessary to improve preventive activities through additional training of employees, intensified supervision over the use of protective equipment and improvement of the organization of work in the highest risk workplaces.

Table 2: Workplace injuries in relation to the workplace, body parts, time of injury occurrence, day of injury occurrence and lost hours in the foundry for the period 2018–2024

Year	Total injuries	Minor injuries	Severe injuries	Workplaces where injuries occurred	Injured body parts	Time of injury occurrence	Weekdays when the injury occurred
2018	14	14	0	Machine molding 2; Steel casting 5; Core making 3; Maintenance of equipment 1; Machining 1; Assembly 1; Crane operator 1	Head 5; Arms 6; Legs 2; others 1	1h–4; 2h–2; 3h–3; 4h–2; 5h–1; 6h–1; 8h–1	Monday 6; Tuesday 3; Wednesday 3; Friday 2
2019	15	14	1	Machine molding 5; Steel casting 4; Core making 1; Maintenance of equipment 2; Sand preparation 1; Pattern maker 1; Electrical engineer 1	Head 3; Arms 5; Legs 3; Eye 4	1h–4; 2h–1; 3h–2; 4h–1; 5h–2; 6h–2; 7h–1; 8h–2	Monday 3; Tuesday 1; Wednesday 2; Thursday

							2; Friday 6; Saturday 1
2020	17	17	0	Machine molding 6; Steel casting 4; Core making 1; Semi-automatic molding 1; Forklift operator 1; Maintenance of equipment 2; Melter/Furnace operator 2;	Head 1; Arms 6; Legs 2; Eye 8	1h–0; 2h–0; 3h–4; 4h–2; 5h–3; 6h–3; 7h–2; 8h–3	Monday 4; Tuesday 2; Wednesday 3; Thursday 2; Friday 4; Saturday 2
2021	13	13	0	Machine molding 10; Core making 1; Semi-automatic molding 2	Head 1; Arms 5; Eye 7	1h–2; 2h–2; 3h–2; 4h–1; 5h–1; 7h–2; 8h–3	Monday 2; Tuesday 2; Wednesday 6; Thursday 1; Friday 1; Saturday 1
2022	19	19	0	Machine molding 4; Steel casting 3; Core making 3; Melter/Furnace operator 1; Maintenance of equipment 1; Forklift operator 1; Semi-automatic molding 2; Electrical maintenance 1; Structures 1; Work on CNC laser 1; Painter 1	Head 4; Arms 11; Legs 3; Others 1	Arrival–1; 1h–2; 2h–2; 3h–4; 4h–2; 5h–3; 8h–5	Monday 3; Tuesday 2; Wednesday 6; Friday 2; Saturday 6
2023	15	14	1	Machine molding 4; Steel casting 1; Maintenance of equipment 1; Forklift operator 1; Assembly 5; Painter 1; Melter/Furnace operator 1; Core making 1	Head 1; Arms 6; Legs 3; Eye 5	1h–1; 2h–1; 3h–1; 4h–2; 5h–1; 6h–5; 7h–1; 8h–3	Monday 2; Tuesday 3; Wednesday 1; Thursday 3; Friday 3; Saturday 3
2024	16	13	3	Machine molding 1; Electrical maintenance 1; Maintenance 2; Forklift operator 1; Machining 1; Toolmaker 1; Melter/Furnace operator 1; Core making 2; Machine molding 2; Controller 1; Semi-automatic molding 2; Melter/Furnace operator 1	Head 2; Arms 6; Legs 3; Eye 5	1h–1; 2h–4; 3h–4; 4h–2; 5h–2; 6h–2; 7h–1;8h–0	Monday 3; Tuesday 3; Wednesday 3; Thursday 2; Friday 3; Saturday 2

3.5. Analysis of Injuries by Body Part

Analysis of injuries by body part shows that in the period 2018–2024 arms were most frequently injured, which indicates a high degree of manual labor and an indirect contact between employees and tools, castings and work equipment (Figure 2). There were a significant number of eye injuries since eyes were most often exposed to metal chips, dust and sparks in grinding and metal processing.

Injuries of heads and legs were less common, but their continuous presence indicates the existence of hazards related to impacts, falls and improper handling of materials in the production process. The obtained results confirm that the jobs of metal machining, grinding and molding are operations with an increased risk of mechanical injury. Therefore, it is necessary to improve the use of personal protective equipment, especially for hand and eye protection, as well as strengthen the supervision over the implementation of occupational safety and health measures.

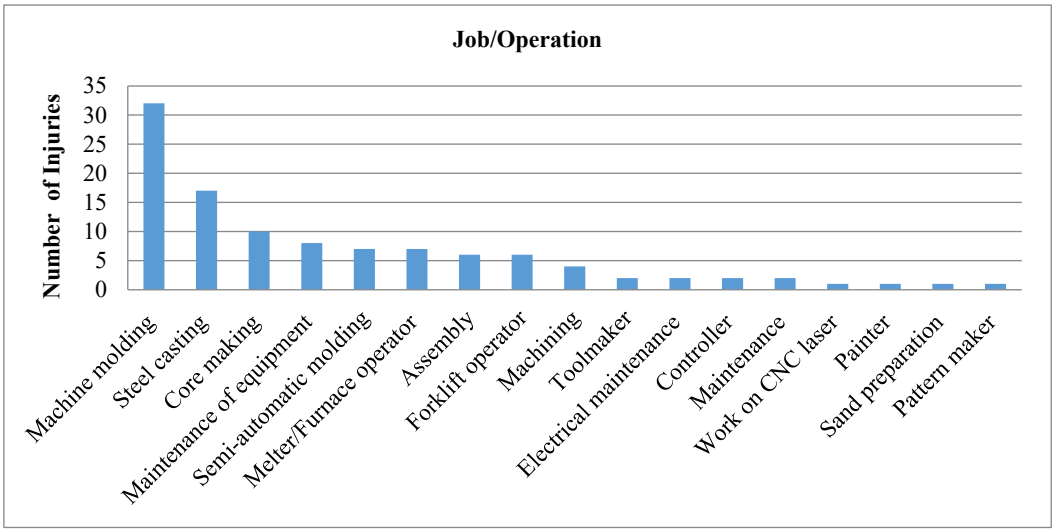


Figure 1: Total number of injuries by workplace (2018-2024)

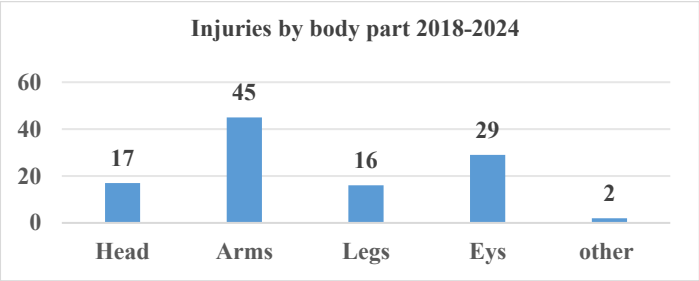


Figure 2: Injuries by body part (2018-2024)

3.6. Analysis of Injuries During the Time of Occurrence

Based on the data presented in Figure 3, it can be concluded that the distribution of injuries during the working time shows pronounced oscillations and clearly identified critical periods of injury occurrence. The highest frequency of injuries was recorded in the period which included the first three hours of work, with the peak in the third hour, which indicates an increased level of load and a possible fall in employees' concentration in the early phase of the working process. After that, there is a gradual decrease in the number of injuries in the middle of the shift, while a significant rise is again recorded at the end of the working time.

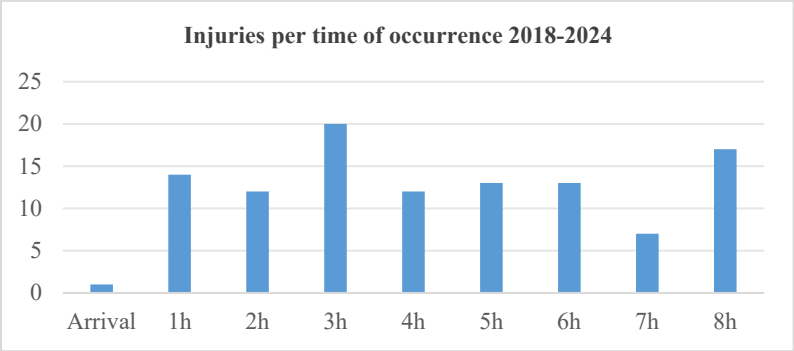


Figure 3: Injuries per time of occurrence (2018-2024)

The obtained results suggest that the factors such as fatigue, monotonous work, psychophysical load and reduced attention significantly affect the occurrence of injuries. What is particularly important is that the increase in the number of injuries occurs even at the end of the shift, which may indicate the consequences of long working hours and exhaustion of employees. The analysis of the diagram confirms the need for improvement of occupational safety and health measures, especially in the periods of increased risk. Preventive activities, e.g. organizing adequate breaks, task rotation, additional training of employees and continuous supervision of work processes, can contribute to the reduction of the number of injuries and the increased overall safety of the work environment.

3.7. Analysis of Injuries by Weekday

The analysis of injuries by weekday shows that the injury frequency is not uniformly distributed during the working week, but that there are distinct time periods of increased risk. The most critical days are Wednesday and Monday, while the lowest risk was recorded on Thursdays. The obtained results indicate a significant influence of psychophysical load, work-related fatigue and the level of employees' concentration of on the occurrence of workplace injuries.

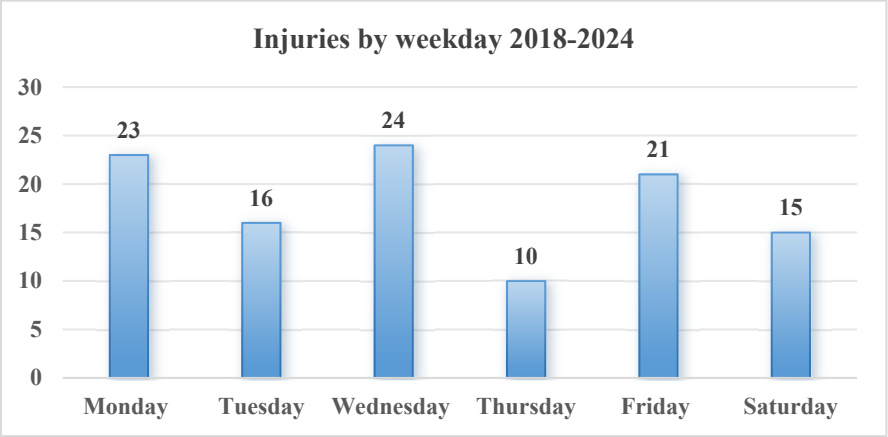


Figure 4: Injuries by weekday (2018-2024)

The results of research confirm the need for continuous improvement of preventive occupational safety measures, especially in the days of increased risk. Organization of work, adequate breaks, additional education of employees and intensified supervision of work processes can contribute to the reduction of the number of injuries and increased safety of the work environment.

4. CONCLUSION

Based on the analysis of key performance indicators regarding the occupational safety and health (AFR, LTIFR and SR) in the foundry from the case study, it can be concluded that KPI implementation and monitoring enable the efficient identification of professional risks and the assessment of the employee safety level in the foundry industry, which confirms the key research question of the paper. The analysis of values of indicators in the period 2018–2024 shows that the observed foundry is characterized by an increased level of professional risk, which is reflected in the frequency and severity of workplace injuries.

The research results indicate that the specificities of the casting process, such as high temperatures, increased physical load, presence of dust, vibrations and handling of metal materials and tools, considerably affect the safety and health of employees. The analysis of the structure of injuries established that the highest risk workplaces are connected with machine molding, grinding and final treatment of castings, while the most frequently injured body parts are hands and eyes. Also, the critical time periods of injury occurrence were identified, especially in the first hours of work, at the end of the shift as well as at the beginning and in the middle of the working week, which indicates a significant impact of fatigue, monotonous work, psychophysical load and organizational factors.

The obtained results confirm that the analysis of KPI indicators is an important tool for the assessment of success of an occupational safety and health system and the identification of areas which require the improvement of preventive action. The aim of this paper related to the assessment of the level of professional risk in the foundry industry and definition of the possibilities for improving employee protection measures has thus been achieved.

Based on the research results, it can be concluded that the existing occupational safety and health system in the analyzed foundry is not sufficiently oriented to preventive action and continuous improvement of working conditions. Therefore, it is necessary to develop an integrated approach to occupational safety and health management, which implies the improvement of risk assessment, more intensive education of employees, more consistent application of personal protective equipment, improvement of the organization of work and modernization of technological processes, with the aim to reduce the exposure of employees to hazards and harmful effects as well as to reduce the number and severity of injuries.

Further research should be directed to a more detailed analysis of the causes of injuries by applying modern methods of risk assessment, research into the influence of ergonomic and psychophysical factors on the occurrence of injuries, as well as the analysis of the connection between the organization of work and the occupational safety and health level. Special importance is assigned to the investigation of possibilities regarding the implementation of automation and digitalization in high-risk phases of production as well as the development of predictive analysis of injuries based on statistical and software tools. Also, future research can involve the comparison of safety indicators in different foundries and the evaluation of efficiency of preventive measures for the purpose of developing a sustainable and safer production process in the foundry industry.

ACKNOWLEDGEMENTS

This study was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, and these results are parts of the Grant No. 451-03-33/2026-03/ 200132, with the University of Kragujevac - Faculty of Technical Sciences Čačak.

REFERENCES

[1] A. Kokić Arsić and M. Mišić, “Health and safety and environmental key performance indicators”, Proceedings of the 8th International Quality Conference, Faculty of Engineering Sciences, University of Kragujevac, Kragujevac (Serbia), pp. 18–1–18–6, (2014)

[2] Occupational Safety and Health Administration (OSHA), “Foundries”, U.S. Department of Labor, https://www.osha.gov/foundries

[3] D. Duan, P. Leng, X. Li, G. Mao, A. Wang, and D. Zhang, “Characteristics and occupational risk assessment of occupational silica-dust and noise exposure in ferrous metal foundries in Ningbo, China,” Frontiers in Public Health, Vol. 11, Article 1049111, https://doi.org/10.3389/fpubh.2023.1049111, (2023)

[4] M. J. Jafari, F. Saghi, E. Alizadeh, and F. Zayeri, “Relationship between risk perception and occupational accidents: a study among foundry workers,” Journal of the Egyptian Public Health Association, Vol. 94, Article 24, https://doi.org/10.1186/s42506-019-0025-6, (2019)

[5] International Association of Oil & Gas Producers (IOGP), “Health performance indicators – 2020 data,” Report No. 2021s, London (UK), (2021)

[6] Health and Safety Executive (HSE), “Monitoring health and safety performance”, HSG65, London (UK), (2013).

[7] World Steel Association, “Safety and health in the steel industry – Data Report 2025”, Brussels (Belgium), https://worldsteel.org, (2025)

[8] International Marine Contractors Association (IMCA), “Safety statistics report”, London (UK), https://www.imca-int.com

[9] Safe Work Australia, “Lost time injury frequency rate (LTIFR) by industry”, Canberra (Australia), (2023), https://www.safeworkaustralia.gov.au

[10] American Foundry Society (AFS), “Occupational injury rates per 100 full time workers”, Schaumburg, Illinois (USA), https://www.afsinc.org

SESSION C: PRODUCTION TECHNOLOGIES, CAE AND MANAGEMENT

Balancing between public service obligations and market efficiency: Modern challenges in passenger rail transport

Vasko Vassilev [1*], Nina Ivanova Gergova [1]

[1] Todor Kableshkov Higher School of Transport, Faculty of Transport Management, Sofia, Bulgaria

ARTICLE INFO

[*] **Correspondence:** vvasilev@vtu.bg, ninagergova@vtu.bg

DOI: https://doi.org/10.46793/ET26.C10V **PAGES:** C77 – C81

ABSTRACT

Public service contracts (PSCs) in passenger rail are a cornerstone of sustainable mobility policy, ensuring the provision of socially necessary but commercially unviable services. However, evolving regulatory frameworks, rising operating costs, decarbonisation targets and rising service quality expectations are putting pressure on traditional contracting models. This report examines the contemporary challenges faced by PSCs in passenger rail, with a focus on the tension between public service obligations and market-driven efficiency. The study focuses on the consequences of market liberalisation and the introduction of competitive tendering under the EU's Fourth Railway Package. Key risks are identified, including information asymmetry between contracting authorities and operators. Through policy analysis, the report proposes a bargaining model that balances flexibility, accountability, and long-term investment incentives. The findings contribute to the ongoing debate on how to structure effective and sustainable public service contracts (PSCs) that can adapt to changing economic, technological and environmental conditions, while maintaining accessibility and affordability for passengers. Practical recommendations are proposed for policymakers, transport authorities and rail operators seeking to optimize public services in a rapidly changing mobility landscape.

KEYWORDS

Rail transport, Public service obligation, Sustainable mobility, Liberalization

1. INTRODUCTION

Passenger rail transport is a strategic sector that combines economic, social and environmental functions. In many countries, railway operators perform services that are not economically profitable but are essential for society. These services are defined as Public Service Obligations (PSO) and are financed through public compensation. At the same time, the processes of liberalization and opening of the railway markets in the European Union place an increasing emphasis on efficiency, competitiveness and quality of services. This creates a need to find a balance between public objectives and market principles.

The purpose of this study is to analyze the main challenges to achieving such a balance and to propose possible directions for the development of the sector.

2. THEORETICAL ASPECTS OF PUBLIC SERVICE OBLIGATIONS

The concept of public service is based on the need for certain transport services to be provided regardless of their economic profitability. The main objectives include:

- ensuring transport accessibility;
- social inclusion of the population;
- regional development;
- reducing the environmental footprint of the transport system.

Rail transport has the characteristics of a public good due to the significant external benefits it generates – reducing congestion, greenhouse gas emissions and road accidents.

Public service obligations in rail passenger transport of general interest include:

- Ensuring the necessary frequency of service.

The competent authorities must guarantee public services also in low-density areas, urban transport services outside peak hours in the late hours of the day and at weekends, regardless of demand levels.

- Assuming tariff obligations.

Public service obligations may include a regulated non-commercial tariff structure, or a limitation of tariff increases below those recommended by public service operators or at a lower rate than cost increases.

- Ensuring safety and reliability of services.
- Higher quality services and meeting the constantly changing requirements of customers.

In economic theory, public service obligations often lead to conflict between social efficiency and financial sustainability [3, 4, 5]. Operators are forced to maintain lines with low demand, which increases costs and creates a dependence on government funding.

3. EUROPEAN REGULATORY FRAMEWORK AND LIBERALIZATION OF THE RAILWAY MARKET

Over the past two decades, the European Union has pursued a consistent policy of liberalizing the railway sector. Through a series of railway packages, the aim is to create a Single European Railway Area based on competition and non-discriminatory access to infrastructure [1, 2].

Of particular importance is Regulation (EC) No 1370/2007, which regulates the award of public service obligations in passenger transport [1]. The Regulation allows for the granting of compensation for the performance of public service obligations, while at the same time setting requirements for transparency and efficiency.

The introduction of competitive procedures in the award of public contracts aims to stimulate:

- reduction of operating costs;
- improving the quality of services;
- increasing innovation;
- better use of public funds.

However, the practical application of these principles faces a number of challenges related to the specificities of the railway sector.

4. KEY CHALLENGES IN BALANCING PUBLIC SERVICE AND MARKET EFFICIENCY

Financial sustainability of public services

One of the most serious problems is the need for significant public subsidies. In many countries, ticket revenues cover only a fraction of operating costs.

Rising energy prices, inflationary pressures and the need for investments in rolling stock are increasing the financial burden on state budgets. This raises the question of the effective use of public resources and the optimization of transport services.

Competition and social responsibility

Liberalization stimulates competition, but private operators are usually interested in profitable routes with high demand. This can lead to so-called "cream skimming", where unprofitable lines remain entirely the burden of the public operator.

Therefore, regulators must develop mechanisms to ensure equality between participants and at the same time provide transport services to less populated regions.

Quality of service and passenger satisfaction

Market efficiency should not be viewed solely through the prism of financial indicators. Important criteria are:

- punctuality of trains;
- service reliability;
- accessibility for people with reduced mobility;
- information service;
- comfort and safety.

Modern public service contracts increasingly include quality indicators and mechanisms for financial incentives or sanctions for operators.

Digitalization and technological transformation

The development of digital technologies creates new opportunities for improving efficiency. Traffic management systems, electronic ticketing, big data analytics and artificial intelligence can optimize operational processes and reduce costs.

At the same time, implementing such solutions requires significant investments and institutional capacity.

Environmental requirements and sustainability

Rail transport is among the most environmentally friendly modes of transport. The European Green Deal and decarbonisation policies are strengthening its role as a tool for sustainable development.

However, there is a need for significant investments in electrification, infrastructure modernization and rolling stock renewal. These investments often cannot be realized through market mechanisms alone and require public support.

5. GOOD EUROPEAN PRACTICES

The experience of a number of European countries shows that successfully combining societal and market objectives is possible (Table 1).

Table 1: Foreign experience [6, 7, 8, 9, 10, 11]

Countries:	Czech Republic	Republic of Croatia	Kingdom of Belgium	Hungary	Slovak Republic
Organization:					
Public service obligations for long-distance and interregional transport are negotiated with	the Ministry of Transport (28.8% share of the PSO); 14 regions conclude their own PSO contracts (71.2% share of the PSO);	the Ministry of Transport	the Ministry of Transport	the Ministry of Transport; regions conclude their own PSO contracts;	the Ministry of Transport
Railway Passenger Transport Operators:					
- state-owned operator:	1	1	1 - domestic traffic only	1	1
- majority state-owned company:	-	-	-	1 - The state owns 65.6% of the company's capital, 28.3% is owned by the Republic of Austria, and the remaining 6.1% is privately owned.	-
- private operators:	7 - - with four of them having full obligations, while the other three have obligations only for a fare reduction. Private operators serve between 5% and 10% of the total passenger flow in the market.	-	3 - international transport only	-	3 - only one has a Public Service Contract and it serves 3.5% of the total passenger flow on the market.
Quality criteria:					
- reliability of operating rolling stock;	✓	✓	✓	✓	✓
- punctuality of trains - schedule compliance, train connections;	95%	95%	100%	100%	100%
- cleanliness of the rolling stock;	✓	✓	✓	✓	✓
- functioning of heating, air conditioning, opening of windows and doors, etc.;	✓	✓	✓	✓	✓
- additional amenities - free Wi-Fi access, bicycle storage, etc.	✓	✓	✓	✓	✓

Germany implements a model of regional procurement through competitive procedures, which leads to improved quality and reduced costs.

Sweden was among the first countries to liberalize its railway market, combining competition with clearly defined public obligations.

Austria demonstrates an effective model for integrating public services with national goals for sustainable transport and climate neutrality.

What successful models have in common is the presence of clear contractual relationships, stable financing, and effective control over implementation.

6. GUIDELINES FOR IMPROVING THE BALANCE BETWEEN PUBLIC SERVICE AND EFFICIENCY

To achieve a sustainable balance, it is necessary:

1. Increasing transparency in determining compensation for public services.
2. Implementation of performance and quality indicators in contracts with operators.
3. Promoting competition where it is economically justified.
4. Investments in digitalization and intelligent transport systems.
5. Better integration between rail and other modes of public transport.
6. Development of sustainable financial models for long-term support of the sector.

CONCLUSION

The balance between public service obligations and market efficiency is one of the key challenges facing modern passenger rail transport. On the one hand, society expects affordable, high-quality and environmentally sustainable transport services. On the other hand, limited public resources and liberalisation processes require operators to increase their efficiency and competitiveness.

Successful management of this balance requires a combination of an adequate regulatory framework, transparent financing mechanisms, innovation and consumer orientation. In the context of increasing demands for sustainable mobility, rail transport will continue to play a strategic role, which necessitates a constant pursuit of an optimal balance between public interest and economic efficiency.

For private railway operators to participate in the passenger market and assume public service obligations, it is appropriate to divide the railway network with railway services into segments, based on the expected revenues from public services for 5 years, guaranteeing a minimum revenue of X million euros per year for each segment. In this way, participants will be equal and will be able to participate in the tender for as many segments as they want. Then a competitive advantage will be the extent to which operators manage to minimize their costs. Another advantage may be evidence of opportunities for coordinating timetables with road or urban transport.

When assuming Public Service Obligations, it is necessary to ensure equal access for each applicant to the rolling stock and conditions for its operation and maintenance. Also, applicants must necessarily demonstrate the ability to effectively manage the life cycle of the rolling stock (operational capabilities, ongoing maintenance, repair, modernization).

The rolling stock and depots of state railway operators (freight and passenger) need to be converted into separate units (one or more) and the possibility of private investment assessed.

The analysis of practices in EU Member States (Czechia, Germany, Poland, Hungary, and Slovakia) shows the following legally defined levels of reasonable profit: 7.5% in Czechia, 6% in Germany and Poland, 4% in Hungary, and a range from 2% to 7.5% in Slovakia. In the case of public rail transport in Bulgaria, the principle of reasonable profit without a fixed percentage is applied according to the European regulatory framework, while for bus transport a profit ceiling of 5% is set by law. To achieve equality between road and rail transport, a reasonable profit level set as the maximum permissible value of 5% could be discussed.

REFERENCES

[1] Regulation (EC) No. 1370/2007 of the European Parliament and of the Council, https://eur-lex.europa.eu/eli/reg/2007/1370/oj/eng, (2007)

[2] European Commission, "Fourth Railway Package", https://transport.ec.europa.eu/transport-modes/rail/railway-packages/fourth-railway-package-2016_en, (2016)

[3] A. S. J. Smith and C. Nash, "Rail efficiency: cost research and its implications for policy", Discussion Paper No. 2014-12, International Transport Forum, OECD, https://doi.org/10.1787/5jrw1kqmbw5d-en, (2014)

[4] C. Nash, "Railway Finance in Europe", Review of Network Economics, Vol. 16(2), pp. 67–88, https://doi.org/10.1515/rne-2017-0039, (2017)

[5] J. Preston and D. Robins, "Evaluating the long term impacts of transport policy: The case of passenger rail privatisation", Research in Transportation Economics, Vol. 39(1), pp. 14–20, https://doi.org/10.1016/j.retrec.2012.05.019, (2013)

[6] Independent Regulators' Group – Rail (IRG-Rail), "Market Monitoring", https://irg-rail.eu/irg/documents/market-monitoring

[7] Ministry of Transport of the Czech Republic, "Souhrnné zprávy o závazcích veřejné služby", https://md.gov.cz/Dokumenty/Verejna-doprava/Souhrnne-zpravy-o-zavazcich-verejne-sluzby-dotaz

[8] Croatian Regulatory Authority for Network Industries (HAKOM), "Objavljeno 14. godišnje izvješće IRG-Raila o praćenju tržišta željezničkih usluga", https://www.hakom.hr/hr/objavljeno-14-godisnje-izvjesce-irg-raila-o-pracenju-trzista-zeljeznickih-usluga/12556

[9] Regulatory Body for Railway Transport (Belgium), https://www.regul.be/nl/

[10] MÁV Group, https://www.mavcsoport.hu/

[11] Ministry of Transport of the Slovak Republic, "Rail Transport", https://www.mindop.sk/ministerstvo/doprava/zeleznicna-doprava

[12] CER – Community of European Railway and Infrastructure Companies, "Annual Reports", https://www.cer.be/publications/reports

[13] OECD, "Competition Trends 2024", OECD Publishing, Paris (France), https://doi.org/10.1787/e69018f9-en, (2024)

[14] International Association of Public Transport (UITP), "Global Economic Outlook 2024: Taking the Pulse of the Public Transport Sector (UITP Statistics Brief)", UITP, https://www.uitp.org/wp-content/uploads/sites/7/2025/08/20241008_Economic-Outlook-2024_WEB.pdf, (2024)

SESSION D

AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Hybrid simulation of an inverted pendulum

Lazar Kelić[1]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering, Kraljevo, Serbia

ARTICLE INFO

* Correspondence: delboyrodney@gmail.com

DOI: https://doi.org/10.46793/ET26.D01K **PAGES:** D1 – D9

ABSTRACT

Hybrid simulation (Hardware-in-the-Loop) of an inverted pendulum is an approach that combines numerical simulation and physical experimentation, with the aim of faithfully replicating real operating conditions in the control of dynamic systems. In this paper, an experimental setup is implemented in which the system dynamics - the cart motion and the pendulum angle - are calculated on a personal computer using the Python environment and the SciPy library, while the control algorithm is executed on the Arduino microcontroller platform. This division enables the simulation and control to run on independent processors, which brings the operating conditions closer to real systems. The mathematical model of the system is based on nonlinear differential equations that describe the cart position and the pendulum deflection angle, while the microcontroller receives the state values in real time and calculates the control signal using the LQR algorithm. In this way, interaction is achieved between the software and hardware parts - Python represents the physics model, while Arduino represents the real controller.

The advantage of this approach is the possibility of analyzing delay and time response under conditions close to real ones, which is often neglected in classical simulations. Different microcontroller platforms (ATmega328P, ATmega32u4, ESP32 with FPU) make it possible to experimentally evaluate stability, latency and control-computation speed. The influence of the sampling frequency and communication channels (serial transmission) on system performance is also analyzed. Hybrid simulation represents a bridge between theoretical modeling and practical implementation, enabling safe testing of control algorithms and comparison of hardware platforms in real time.

KEYWORDS

Inverted pendulum, Hybrid simulation, LQR control, System delay, Multithreaded modeling, Real-time systems; Hardware-in-the-Loop (HIL), Arduino Leonardo, ESP32, System dynamics, Python simulation, Controller optimization, Latency, Digital control, Experimental verification.

1. INTRODUCTION

The inverted pendulum system is a classical problem in control theory because of its unstable equilibrium and nonlinear dynamics [1]. A typical implementation includes a cart with a pendulum in the upright position, while the control task is to maintain the pendulum in that position by controlling the cart position or the actuator. Modeling of such a system often begins with linearization around the upright position, after which the state equations are analyzed in detail in the form [2]

$$\dot{x} = Ax + Bu \tag{1}$$

To stabilize this linearized system, a linear-quadratic regulator (LQR) is often used, which minimizes the cost function

$$J = \int (x^\top Q x + u^\top R u)\, dt \tag{2}$$

and provides optimal feedback boundary control[3]. Previous studies have shown that LQR provides fast transients and a minimal step in simulations [4].

However, classical simulations implemented in software environments such as MATLAB/Simulink or similar tools have limitations when it comes to real execution of the control law on embedded hardware and the influence of factors such as delay, sampling and real actuators [5]. A control law designed in simulation must also take into account further processing, data loading, communication and execution on a microcontroller, which can significantly affect real-time system performance [6].

In this context, multithreading and the software architecture of embedded platforms (including interrupts, tasks and serial communication) play an important role and may be a primary causal component of latency, even greater than the computing power itself [7].

At the same time, increasing attention is being given to hybrid simulations in which the system dynamics are calculated on one processor (for example, a computer running Python), while the control logic is implemented on an independent controller (for example, Arduino) connected to the simulation framework through real-time communication. Such an approach enables a more realistic product evaluation because the simulation and control run on different platforms, allowing the influence of real hardware latencies and processing to be assessed [8]. One recent study emphasizes that bridging the gap between theory and practice requires an environment in which the simulation "lives" together with the physical platform [9].

The literature also includes studies comparing different control techniques applied to the inverted pendulum: comparisons of PID, SMC, fuzzy control and LQR show that, although advanced algorithms are attractive, in implementations with limited microcontroller resources the path to stability also requires careful study of latency and computing architecture [10]. Examples of simulations correlated with real experiments show that — even when the control law performs well in simulation — a simple microcontroller implementation can suffer degradation due to delays and process congestion [11].

Classical modeling of pendulum-dynamics simulation includes calculating position, angle, velocity and acceleration parameters, often assuming continuous time and idealized samples. However, when the system is transferred to a real platform, the time step ΔT, sampling and controller response affect performance. Studies that examined the influence of ΔT found that larger time steps can cause oscillations or even loss of stability [12]. It is particularly emphasized that microcontroller implementations must take into account deterministic execution flow and minimize variance in time response [13].

In this paper, we present an architecture that combines classical simulation of system dynamics (on a computer) and implementation of the controller on a microcontroller as an independent processor. This makes it possible to measure and optimize the simulation time step, communication latency and control cycle. In particular, we investigate how the form of the multithreaded model, interrupts and controller architecture affect the final performance of the LQR controller in a hybrid setup. The aim is to determine which device and which procedure account for latency, stability and control accuracy, and to use experimental simulation to verify whether the results correspond to theoretical expectations.

2. ANALYSIS OF LQR CONTROL COMPUTATION AND THE INFLUENCE OF DENORMALIZED NUMBERS AND NONLINEARITY

2.1. LQR on the Arduino Platform

The linear-quadratic regulator (LQR) is an optimal control method that minimizes a specified cost function based on the system state. However, in real microcontroller implementations, especially on platforms with limited computational resources, the speed of calculating the control signal has a significant influence on system stability and performance. This section provides a detailed analysis of the execution cycle for calculating LQR control, as well as the influence of denormalized numbers and system nonlinearity on execution time.

2.2. Methodology for Estimating the Number of Processor Cycles

The estimates of the number of processor cycles used in this paper are based on a combination of theoretical analysis of microcontroller architecture and empirical values obtained from the literature and documentation for the platforms used. It should be emphasized that these values do not represent absolutely exact numbers, but rather order-of-magnitude approximations that enable a qualitative analysis of the influence of computational complexity on real-time system performance.

On microcontrollers without hardware support for floating-point operations (such as ATmega328P and ATmega32u4), all arithmetic operations on the float type are implemented in software through compiler libraries. These operations include several steps: normalization, exponent alignment, mantissa manipulation and rounding, which results in a significantly larger number of instructions compared with integer arithmetic. For this reason, the number of cycles for operations such as multiplication and addition is estimated in the range from several tens to several hundreds of cycles, depending on compiler optimization and memory-access method.

In more modern microcontrollers such as the ESP32, which has a hardware floating-point unit (FPU), arithmetic operations are executed much faster, although still not in a single cycle. The actual execution time depends on the pipeline architecture, cache-memory access and possible delays caused by the operating system (FreeRTOS), which introduces variability into the execution time.

Trigonometric functions represent a special case, because in most implementations they are realized through numerical approximations or lookup tables. Therefore, their execution time is not constant, but depends on the library and the range of input values; in this paper, order-of-magnitude values (1000–2000 cycles) are used to illustrate the increase in computational complexity.

In addition to the arithmetic operations themselves, a real system is significantly affected by additional factors such as memory access, serial communication, interrupt handling and multithreading mechanisms. These factors can dominate the total execution time, which explains the experimentally obtained results in which nominally faster microcontrollers do not achieve the expected performance.

Based on the above, the cycle-count values used should be regarded as engineering approximations that enable comparison of different architectures and understanding of the influence of computational complexity on control dynamics, rather than as precise measured quantities.

2.3. Cycle-Count Estimate for Calculating the Control Signal

On microcontrollers without hardware support for floating point (such as Arduino Leonardo and Arduino Mega), all operations on real numbers are performed in software. This implies a larger number of processor cycles for each individual operation. Approximate values are:

Table 1: Estimated number of cycles

Operation	Estimated number of cycles
Multiplication (32-bit float)	~150 cycles
Addition (32-bit float)	~100 cycles
Memory read	~50 cycles
Negation (sign change)	~10 cycles

For calculating the control signal

$$u = Kx \tag{3}$$

where is K a vector of dimension 1×4, and x is a state vector of dimension 4×1, the following is required:

1. 4 multiplications: cycles $4 \times 150 = 600$
2. 3 additions: cycles $3 \times 100 = 300$
3. 4 memory reads: cycles $4 \times 50 = 200$
4. Negation of the result: 10 cycles

Total:

$$600 + 300 + 200 + 10 = 910 \text{ cycles} \approx 900 \text{ cycles (with optimization).}$$

2.4. Theoretical Assumption of Computation Time for Different Platforms

The computation speed directly depends on the processor clock frequency. The approximate execution time is:

Table 2: approximated computation time

Platform	Clock (MHz)	Computation time (μs)
Arduino Leonardo	16	900 / 8 = 56.25
Arduino Mega	16	900 / 16 = 56.25
ESP32 (80 MHz)	80	900 / 80 = 11.25
ESP32 (240 MHz)	240	900 / 240 = 3.75

The ESP32 has a hardware FPU (Floating Point Unit), which enables floating-point operations to be performed in one cycle. In this case, the total number of cycles for LQR control is approximately 20, which gives the computation time:

$$20/240 = 0.083\ \mu s.$$

2.5. Influence of Number Magnitude and Number of Decimal Places

The computation speed does not depend on the absolute magnitude of the number (for example, 0.1° or 45°), because all floating-point operations are performed in the same number of cycles. However, when denormalized numbers appear (very small values close to zero), the processor must perform additional internal normalization, which increases the computation time.

- Multiplication: 300–500 cycles (instead of 150)
- Addition: 200–400 cycles (instead of 100)
- Reading: 100 cycles (instead of 50)

Table 3: Influence of number magnitude and number of decimal places

Condition	Total cycles	Time on Arduino Mega and Leonardo (16 MHz)
Angle 0.1° (denormalized)	2210	138.125 μs
Angle 45° (normalized)	900	56.25 μs

Therefore, denormalized numbers slow down computation by approximately 2.5 times.

2.6. Influence of System Nonlinearity

The nonlinear model of the inverted pendulum requires the use of trigonometric functions and more complex expressions. An example of a nonlinear control equation is:

$$u = \frac{(M+m)\,g\sin\theta - ml\dot{\theta}^2 \sin\theta\cos\theta + u_{LQR}}{M+m(1-cos^2\theta)} \tag{4}$$

Such expressions require more computational resources. Approximate estimate:

Table 4: Estimated number of cycles for calculating complex operations

Operation	Estimated cycles
sin(θ), cos(θ)	1000–2000 cycles
Other multiplications and additions	+500 cycles
Total	~3000–5000 cycles

Converted execution time:

Table 5: Number of cycles converted into execution time on different platforms

Platform	Clock (MHz)	Time (μs)
Arduino Leonardo	16	250
Arduino Mega	16	250
ESP32 (without FPU)	80	50
ESP32 (with FPU)	240	16.7

2.7. Practical Implications and Optimization

Problems:

- Denormalized numbers increase the number of cycles and reduce precision.
- Nonlinear models drastically increase computational complexity.

Solutions:

1. Data normalization — all inputs should be scaled into a range of normalized values.
2. Fixed point (Q format) — accelerates computation up to 50 times.
3. Sine/cosine lookup tables — eliminate the need to call the sin() and cos() functions.
4. Hardware FPU — using an ESP32 or ARM Cortex-M4/7 processor provides the best computational efficiency.
5. Implementation in a timer interrupt — ensures a fixed control period without delay.

2.8. Conclusion

The analysis performed shows that:

- Denormalized numbers can slow down LQR-control computation by up to 2.5 times.
- The nonlinear system model requires up to 5 times more cycles than the linearized model.
- The best performance is achieved on the ESP32 platform with a hardware FPU unit, where the computation time is on the order of 0.083 μs.
- The use of fixed point and value normalization is the most efficient way to optimize control on microcontrollers without an FPU.

These results are of key importance in the design of a real inverted-pendulum control system, where control stability directly depends on the speed of calculating the control signal.

3. EXPERIMENT

The objective of the experiment is to determine unambiguously whether hardware can be used for inverted-pendulum control and which hardware can be used. For the purposes of the experiment, a simulation of the inverted-pendulum dynamic system was used according to the formula

$$(M+m)\ddot{x} + ml\ddot{\theta}cos\theta - ml\ddot{\theta}2sin\theta + b\dot{x} = F \tag{5}$$

and

$$l\ddot{\theta} + gsin\theta = \ddot{x}cos\theta \tag{6}$$

The simulation was written in the Python programming language. Another Python script was written to communicate between the above-mentioned Arduino platforms and the Python script on the computer. The Python script takes the system state vector and sends it via a serial connection to the Arduino platform. For each Arduino platform, a script was written that receives data from the serial port, in this case the system state vector; the LQR controller then calculates the next action and returns the control value to the Python simulation through the serial connection. The Python simulation takes that value, calculates the next state of the system and sends a new state vector to the Arduino platform.

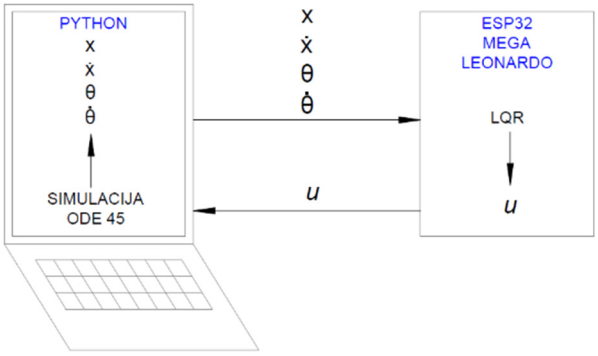


Figure 1: Communication between the Python script on the computer and the Arduino platform

3.1. Determination of LQR Gain

The parameters of the LQR regulator were determined using the standard optimal-control method for the linearized model of the inverted pendulum. Starting from the nonlinear equations of motion of the system, linearization was performed around the upright equilibrium position ($\theta \approx 0$), which produced a state-space model — equation (1), where x = [x, θ, ẋ,

$\dot{\theta}]^T$ is the state vector and u is the control signal. The A and B matrices were formed on the basis of the known system parameters (cart mass, pendulum mass, pendulum length, gravity and friction coefficients), defined in the experimental part of the paper.

The LQR gain was obtained by minimizing the quadratic cost function — equation (2), where Q and R are weighting matrices that define the trade-off between state-tracking accuracy and control energy. By selecting appropriate Q and R matrices and solving the algebraic Riccati equation in the Python environment (SciPy library), the optimal gain was calculated, equation (3), where the gain vector K was used in the microcontroller implementation. The resulting gain vector has the following form: K = [-10.0000 -136.6005 -20.5942 -27.7585], which gives the control law: u = -(-10x -136.6005θ -20.5942$\dot{x}$ -27.7585$\dot{\theta}$). The simulation is executed for up to 10 seconds. The obtained diagrams of pendulum stabilization and controller response speed are given below.

3.2. Experiment Settings

For the purposes of the experiment, the following simulation values were used:

- pendulum length l - 0.5 m,
- cart mass 200 g,
- pendulum mass 180 g
- gravitational acceleration g is 9.81 m/s2,
- friction in the pendulum shaft 0.1,
- cart sliding friction 0.1.

The script on the real hardware is executed as fast as possible, and the number of control-computation cycles in one second is limited exclusively by the hardware power of the controller itself. ΔT in the pendulum-dynamics simulation was set to 5 ms the first time and to 1 ms the second time. This experimental setup gives us the possibility to compare the classical simulation approach, where the state of the controlled system is presented with respect to the current iteration, with a more complex approach in which several independent processes are used and the state of the system is presented with respect to time regardless of the number of iterations (described in the system-delay section). In this case, it also enables testing of the simulation on real hardware while checking the capability of the hardware itself to control the given process.

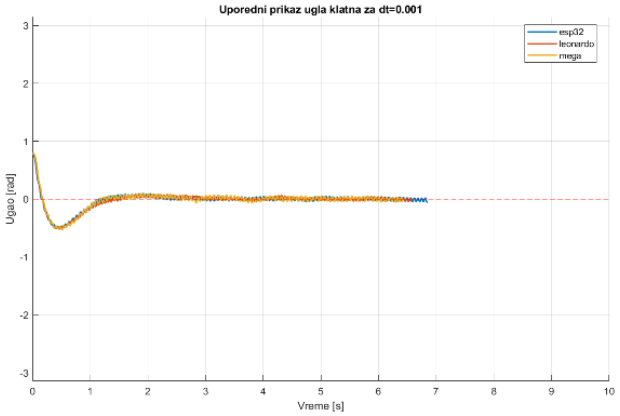


Figure 2: Comparative presentation of pendulum stabilization for ΔT = 1 ms

Figure 2 shows that all three tested controllers perform the task in a very similar way. There is slight oscillation, but the results are essentially similar, with a negligibly better result for the Leonardo, which is certainly unexpected.

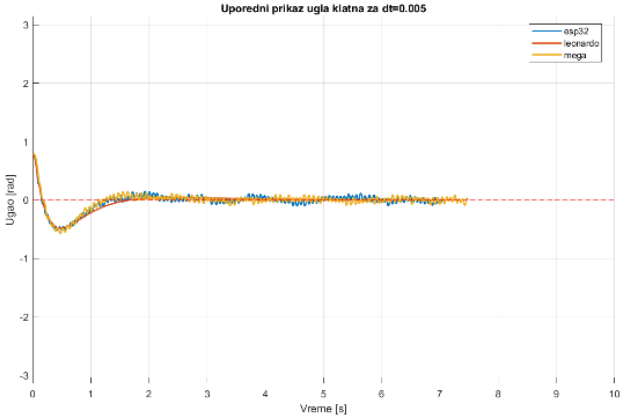


Figure 3: Comparative presentation of pendulum stabilization for ΔT = 5 ms

Figure 3 shows a larger difference in favor of the Leonardo, with a much more accurate result, while the other two controllers show increased instability when ΔT in the system-dynamics simulation is increased to 5 ms.

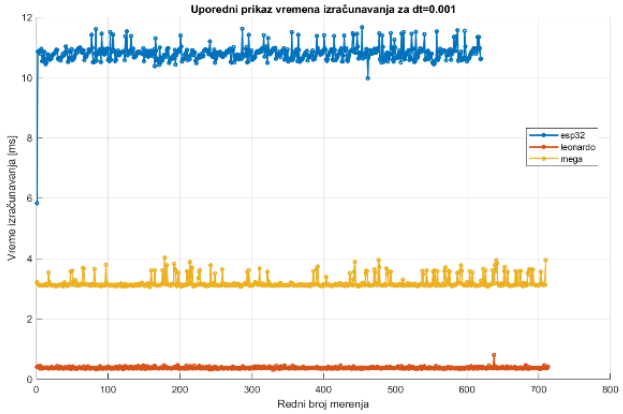


Figure 4: Comparative presentation of the computation speed of each iteration for ΔT = 1 ms

Figure 4 shows the clear dominance of the Leonardo, while the ESP32 is considerably delayed in calculating the control action, which is certainly an unexpected result.

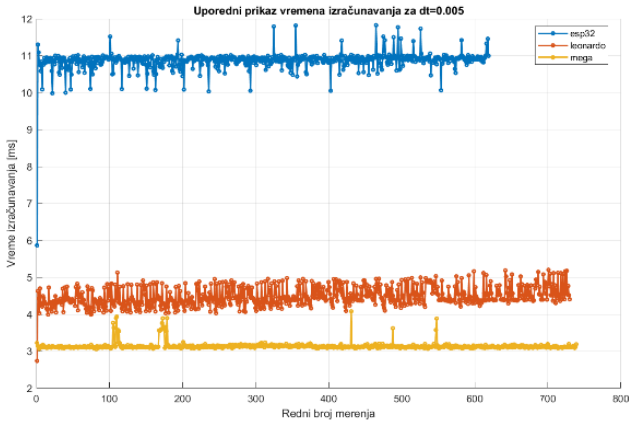


Figure 5: Comparative presentation of the computation speed of each iteration for ΔT = 5 ms

Figure 5 shows a result that differs even more from expectations. The time step of the system-dynamics simulation was increased from 1 ms to 5 ms, and the result is that the Mega is the fastest, the Leonardo is slower than the Mega, while the ESP32 still lags far behind the weaker and slower controllers.

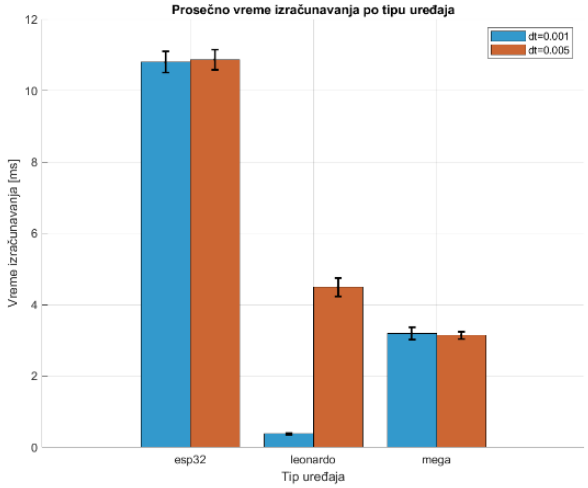


Figure 6: Average control-computation time for ΔT = 1 ms and ΔT = 5 ms

Figure 6 shows a simplified comparative presentation of the data shown in Figures 4 and 5. It can be seen that the ESP32 is by far the slowest, which is very unexpected despite all the hardware advantages it possesses. The computation speed on the Leonardo differs significantly and depends on the time step ΔT of the system-dynamics simulation, while the Mega operates in a balanced manner in both cases of the dynamic-system simulation time step ΔT.

4. INTERPRETATION OF RESULTS AND CONCLUSION

The analysis of the simulation results shown in Figures 2–6 reveals several very interesting and largely unexpected phenomena in the behavior of different controllers (Arduino Leonardo, Arduino Mega and ESP32) when calculating real-time control for the inverted-pendulum model.

Figure 2 shows that all three controllers achieve very similar results in system stabilization, with only slight oscillations and minimal deviations in the transient process. Although the ESP32 was expected to show a convincing advantage because of its significantly higher processor speed and support for hardware floating point, the best results in the initial tests were achieved by the Arduino Leonardo. The difference is small but consistent, indicating that in this operating mode the hardware potential of the ESP32 controller is probably not fully utilized, while the Leonardo, although weaker in hardware, is more stable in communication and data processing at a lower frequency level.

In Figure 3, the situation becomes more pronounced. When the time step of the system-dynamics simulation is increased to 5 ms, the Leonardo maintains good stability and control precision, while the Mega and especially the ESP32 show a noticeable increase in instability. This indicates a possible synchronization problem between the simulation and control processes. In the ESP32, due to parallel processing and the multithreaded structure, a delay may occur in serial communication or in the execution of the task that processes the pendulum data. In the Leonardo, by contrast, the entire process takes place linearly, without an internal scheduler, resulting in lower latencies and more consistent behavior over time.

Figure 4 further confirms the dominance of the Leonardo. While the ESP32 shows a pronounced response delay, the Leonardo tracks the desired signal almost perfectly. This result is completely contrary to expectations and indicates that hardware speed is not the only relevant factor. The probable cause is the different implementation of libraries and functions for floating-point arithmetic. While the ESP32 uses a 32-bit FPU that operates faster but requires multiple type conversions and data buffering within the FreeRTOS scheduler, the Leonardo executes all operations more simply and without an interrupt context, which results in lower effective delay.

Figure 5 shows another surprise: when ΔT is increased from 1 ms to 5 ms, the Mega proves to be the fastest, followed by the Leonardo, while the ESP32 remains the slowest. Here, the different architecture of memory access and timing interpretation within the simulation becomes important. The Mega has a stable timer and a deterministic execution flow, meaning that each computation is repeated in almost the same time interval. By contrast, the ESP32, which uses multiple threads and has a more complex internal organization, shows greater variability in computation time because the scheduler distributes processes depending on system priorities.

Figure 6 provides an overview of the entire situation and clearly shows that the ESP32 is the slowest, which is completely opposite to theoretical assumptions. Based on hardware specifications, the ESP32 would be expected to be several times faster than classical Arduino controllers; however, in practice it is evident that software architecture, internal task priorities and the method of communication with the external system play a key role in real time. The Leonardo, as a single-threaded microcontroller without an operating system, has an advantage in execution predictability and time consistency, which makes it more efficient in control tasks requiring minimal delay.

In general, the obtained results show that in real-time control systems, hardware power alone is not decisive; the way in which computation is integrated into the overall control process is also crucial. The ESP32, although technically superior, suffers losses due to software complexity, scheduler delays and multithreaded architecture. Arduino Leonardo and Mega, with a simple and deterministic execution flow, show better alignment with the system dynamics and lower latency.

This research emphasizes that in real-time control systems with high requirements for timing accuracy, simpler microcontrollers can often achieve better results than nominally faster devices. Future work should include analysis of the influence of arithmetic libraries, optimization of communication with the host system, and testing at larger ΔT values and with more complex control models.

REFERENCES

[1] J. de la Cruz-Alejo, H. Beatriz-Cuellar, I. C. A. Guillermo, and A. M. Ortega, “A decision-making approach on control techniques for an inverted pendulum based on, neuro-fuzzy, indirect adaptive and PID controllers”, Discover Applied Sciences, Vol. 6, p. 251, https://doi.org/10.1007/s42452-024-05921-2, (2024)

[2] U. Yıldıran, “Adaptive Control of an Inverted Pendulum by a Reinforcement Learning-based LQR Method”, Sakarya University Journal of Science, https://doi.org/10.48550/arXiv.2310.04436, (2023)

[3] E. Shala, X. Bajrami, R. Likaj, and A. Pajaziti, “Real Time Swinging Up and Stabilizing a Double Inverted Pendulum Using PID-LQR”, Strojnícky časopis - Journal of Mechanical Engineering, Vol. 73(1), pp. 159–168, https://doi.org/10.2478/scjme-2023-0013, (2023)

[4] M.-S. Phan, T.-C. Do, V.-Q. Truong, and T.-V.-A. Nguyen, “Comparative Analysis of SMC-LMI and LQR Controllers for Double Inverted Pendulum”, Journal of Measurement, Control and Automation, Vol. 4(3), pp. 1–7, https://mca-journal.org/index.php/mca/article/view/184, (2023)

[5] Z. Ben Hazem, “Study of Q-learning and deep Q-network learning control for a rotary inverted pendulum system”, Discover Applied Sciences, Vol. 6, p. 49, https://doi.org/10.1007/s42452-024-05690-y, (2024)

[6] L. Chen and Y. Zhang, “Linear-quadratic optimal control problems of state delay systems under full and partial information”, Systems & Control Letters, Vol. 176, p. 105528, https://doi.org/10.1016/j.sysconle.2023.105528, (2023)

[7] M. Khosravi, H. Azarinfar, and K. Sabzevari, “Design of infinite horizon LQR controller for discrete delay systems in satellite orbit control: a predictive controller and reduction method approach”, Heliyon, Vol. 10(2), p. e24265, https://doi.org/10.1016/j.heliyon.2024.e24265, (2024)

[8] M. Costanzo, R. Mazza, and C. Natale, “Digital Control of an Inverted Pendulum Using a Velocity-Controlled Robot”, Machines, Vol. 13(6), p. 528, https://doi.org/10.3390/machines13060528, (2025)

[9] Y. Martínez-Armero, S. López-Blandón, and E. Giraldo, “Low-cost Arduino-based Hardware-In-the-Loop Platform for Simulation and Control of Dynamic Systems”, IAENG International Journal of Computer Science, Vol. 50(4), pp. 1312–1318, https://www.iaeng.org/IJCS/issues_v50/issue_4/IJCS_50_4_18.pdf, (2023)

[10] https://www.mathworks.com/discovery/hardware-in-the-loop-hil.html

[11] A. Zhang, X. He, A. Cammi, and X. Wang, “Modelica and Arduino-based hardware-in-the-loop simulation of a nuclear-powered engineering ship”, Nuclear Engineering and Design, Vol. 429, p. 113650, https://doi.org/10.1016/j.nucengdes.2024.113650, (2024)

[12] Y.-H. Chang, F.-C. Wu, and H.-W. Lin, “Design and Implementation of ESP32-Based Edge Computing for Object Detection”, Sensors, Vol. 25(6), p. 1656, https://doi.org/10.3390/s25061656, (2025)

[13] A. Elnwegy, F. Youssef and T. A. Kern, “Study to compare different transmission procedures in the context of a robotic telemanipulation system”, Measurement: Sensors, Vol. 38, p. 101847, https://doi.org/10.1016/j.measen.2025.101847, (2025)

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Intelligent control of a UAV using a sensor glove

Simona Tosheska Radovic[1*], Dushko Stavrov[1], Gorjan Nadzinski[1], Vesna Ojleska Latkoska[1]

[1] Faculty of Electrical Engineering and Information Technologies, "Ss. Cyril and Methodius" University in Skopje, Skopje, Republic of North Macedonia

ARTICLE INFO

* Correspondence: simona.tosheska@gmail.com

DOI: https://doi.org/10.46793/ET26.D02TR **PAGES:** D11 – D22

ABSTRACT

This paper presents the development and simulation of an intelligent system for controlling a drone through hand gestures, evaluated and tested in the MATLAB/Simulink environment. First, a detailed mathematical model of the drone dynamics is built, including aerodynamic forces, motor dynamics and battery model. For precise estimation of the system states the model also incorporates the sensor noise, processed with an Unscented Kalman filter. Flight stabilization is achieved through a PID control system. The central contribution of the paper is an implementation of an intelligent control of a drone using a smart glove, equipped with flex sensors and an inertial measurement unit. The smart glove raw analog signals are processed and classified by a neural network with an 8–10–4 architecture, trained using the Levenberg-Marquardt algorithm. The network classifies the fused sensor signals into four discrete control commands, which are fed directly into the drone's flight control system, as structured reference inputs. The integrated system is verified through a simulation, where the drone successfully executes a three-dimensional trajectory, commanded exclusively by the hand gestures. The results show stable tracking of the reference signals, without oscillatory instability, confirming the practical applicability of the proposed approach.

KEYWORDS

Drone, Gesture control, Smart glove, Neural network, Kalman filter, PID controller, MATLAB/Simulink

1. INTRODUCTION

Drones, or quadcopters, have revolutionized various industries and have taken the public stage in many areas over the past few decades. The need for unmanned aerial vehicles (UAVs) with greater maneuverability and hovering capabilities compared to fixed-wing aircraft has led to a surge in research into drones. These drones have applications in diverse fields such as aerial photography, search and rescue operations, agricultural monitoring, and package delivery [1], [2]. Their popularity stems from their versatility, maneuverability, vertical takeoff and landing, and ability to access hard-to-reach areas in a safe and efficient manner [1]. However, the seemingly carefree flight of drones hides the complex engineering and control systems that enable their smooth operation.

One of the biggest challenges in drone operation is the efficient control, due to the inherent instability [3], [4]. The drone has six degrees of freedom, but only four controlled inputs, making it an underactuated system. This complexity requires highly robust and efficient control algorithms to maintain stability and achieve precise maneuvering. The process of developing an effective drone control system begins with a thorough understanding of its dynamics. In this paper, the drone dynamics are described using a mathematical model based on the physical and mathematical laws that govern drone operation.

Recent research in UAV control has increasingly moved beyond classical control methods toward intelligent, multimodal approaches, with machine learning and artificial intelligence playing a central role in advancing autonomous capabilities. [6], [7], [11], [13], [14], [15], [16] Gesture-based human-drone interaction has emerged as a particularly active research area, where control commands can be issued either through wearable sensor devices or through vision-based systems relying on cameras and computer vision algorithms, both removing the dependency on traditional remote controllers [13], [16], [22] . More advanced implementations further combine gesture recognition with obstacle avoidance and autonomous navigation, broadening the scope of intuitive human-machine interaction [25]. These developments collectively reflect a

clear trend toward intelligent, multimodal UAV control systems, which serves as the direct motivation for the research presented in this work.

The focus of this paper is the modeling of the drone dynamics and the development of a control system based on hand gesture recognition - an approach that uses the user's body as the control interface instead of a conventional remote control. For this purpose, a smart glove equipped with flex sensors on the fingers and an inertial measurement module on the hand will be realized, which in combination allow precise recording of the bending angle of each finger, the orientation of the hand, and its linear movements in real time. The collected sensor data is first denoised and normalized, and then fed into a neural network that classifies the gestures and converts them into appropriate navigation commands for the drone, such as pitch, roll, yaw, and throttle.

The entire system is implemented and verified in several interconnected stages. The first stage encompasses building a mathematical model based on the reconstructed states provided by the Kalman filter and designing a PID control system [28], [27]. The second stage is the physical realization of the smart glove containing five flex sensors [19] and IMU module [20] all integrated by Arduino Nano microcontroller. The third stage of collecting and labeling of training data for series of defined gestures is followed by training a neural network for classification, trained using the Levenberg-Marquardt algorithm [26]. The last stage is the integration of the individual stages in one fully operational system defined in MATLAB/Simulink. The end result is a fully integrated intelligent system in which each hand gesture is converted in real time into a specific control action of the aircraft, with verified stability and precision in a simulation environment.

2. SYSTEM MODELING

This section presents the system model through three interconnected parts: the dynamic model of the drone (Figure 1), the sensor noise model with estimation (Figure 3), and the controller design for stable reference tracking.

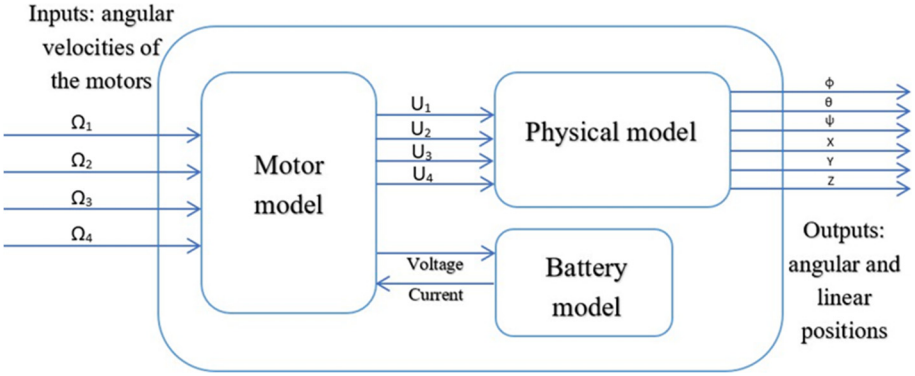


Figure 1: Block diagram of the drone dynamics model

2.1. Dynamic Drone Model

The drone has six degrees of freedom, translations along the X, Y and Z axes, as well as rotations described by the angles of roll (φ), pitch (θ) and yaw (ψ). The local NED (North-East-Down) frame is adopted as the reference coordinate system, in which the X axis points north, Y east, and Z down, normal to the XY plane (Figure 2). In this framework, a positive roll angle moves the drone to the east, a positive pitch angle to the south, and a positive yaw angle corresponds to a clockwise rotation.

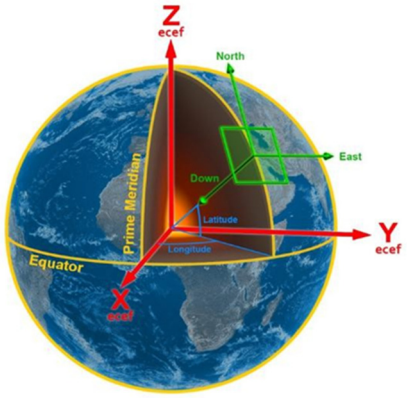


Figure 2: NED (North-East-Down) coordinate system

The model assumes that the drone is a rigid body with a symmetrical construction, that the centers of gravity and geometry coincide, and that the thrust force and drag force are proportional to the square of the angular velocity of the propeller. Aerodynamic air resistance is modeled as a force opposite to the direction of motion, proportional to speed, with a coefficient of 0.1 kg/s in each axis, a value suitable for simulation purposes [24].

2.2. Dynamics Equations

For an "X" frame configuration, the four control quantities, vertical thrust U_1, roll U_2, pitch U_3 and yaw U_4, are defined as:

$$U_1 = b \cdot \left(\Omega_1^2 + \Omega_2^2 + \Omega_3^2 + \Omega_4^2\right) \tag{1}$$

$$U_2 = b \cdot \sin\left(\frac{\pi}{4}\right) \cdot \left(\Omega_1^2 - \Omega_2^2 - \Omega_3^2 + \Omega_4^2\right) \tag{2}$$

$$U_3 = b \cdot \sin\left(\frac{\pi}{4}\right) \cdot \left(\Omega_1^2 + \Omega_2^2 - \Omega_3^2 - \Omega_4^2\right) \tag{3}$$

$$U_4 = d \cdot \left(\Omega_1^2 - \Omega_2^2 + \Omega_3^2 - \Omega_4^2\right) \tag{4}$$

where Ω_i is the angular speed of the i-th engine, b is the thrust coefficient, and d is the resistance moment coefficient [17], [24]. The resulting angular velocity of the rotor is:

$$\Omega_r = \Omega_1 - \Omega_2 + \Omega_3 + \Omega_4 \tag{5}$$

The angular accelerations $\ddot{\varphi}$, $\ddot{\theta}$ and $\ddot{\psi}$ along all three axes are obtained from the force-momentum equations, where each moment includes the gyroscopic effect of the body, the gyroscopic effect of the propeller, and the action of the actuator:

$$\ddot{\varphi} = \left(\dot{\theta}\dot{\psi}\left(I_{yy} - I_{zz}\right) + J_r\dot{\theta}\Omega_r + lU_2\right)\Big/ I_{xx} \tag{6}$$

$$\ddot{\theta} = \left(\dot{\varphi}\dot{\psi}\left(I_{zz} - I_{xx}\right) + J_r\dot{\varphi}\Omega_r + lU_3\right)\Big/ I_{yy} \tag{7}$$

$$\ddot{\psi} = \left(\dot{\theta}\dot{\varphi}\left(I_{xx} - I_{yy}\right) + U_4\right)\Big/ I_{zz} \tag{8}$$

where $\dot{\varphi}$, $\dot{\theta}$ and $\dot{\psi}$ are angular velocities, I_{xx}, I_{yy} and I_{zz} are inertia moments, J_r is rotor inertia, l is rotor axis to copter center distance.

Linear accelerations $\ddot{X}$, $\ddot{Y}$ and $\ddot{Z}$ in the NED system are defined as:

$$\ddot{X} = \left(\left(\sin\psi\sin\varphi - \cos\psi\sin\theta\cos\varphi\right) \cdot U_1 - A_x\dot{X}\right)\Big/ m \tag{9}$$

$$\ddot{Y} = \left(\left(\cos\psi\sin\varphi + \sin\psi\sin\theta\cos\varphi\right) \cdot U_1 - A_y\dot{Y}\right)\Big/ m \tag{10}$$

$$\ddot{Z} = \left(mg - \cos\theta\cos\varphi \cdot U_1 - A_z\dot{Z}\right)\Big/ m \tag{11}$$

where Ax, Ay and Az is air resistance in each axis.

It is evident from equations (9) – (11) that the linear accelerations depend on the current orientation angles. By double integration of the acceleration, the velocity and position of the drone can be obtained at all times - which forms the basis for closing the control loop.

2.3. Sensor Modeling and State Estimation

Based on the derived equations of motion (9) – (11) [17], [24], a corresponding model was realized in the MATLAB/Simulink environment. The subsystem that models the dynamics of the drone has four inputs (Ω_1–Ω_4) and six outputs (X, Y, Z, φ, θ, ψ), and internally consists of four interconnected subsystems that implement the thrust, angular and linear acceleration equations, with built-in motor and battery dynamics. To improve the interpretability of the model, the following constraints have been introduced: the Z coordinate is limited to negative values, the angular velocities of the rotor to the interval $[0, \infty)$, and the Euler orientation angles to the range $[-\pi, \pi]$.

In real conditions, the dynamic model of the drone does not accurately describe the actual behavior of the real system, due to the deliberate and necessary approximations introduced during modeling and the atmospheric conditions (wind, temperature etc.). The control approach is designed based on the feedback data provided by the Inertial Measurement Unit (IMU), which combines a gyroscope and an accelerometer. The gyroscope measures the angular velocities $\dot{\varphi}$, $\dot{\theta}$, $\dot{\psi}$, while the accelerometer measures the linear accelerations $\ddot{X}$, $\ddot{Y}$, $\ddot{Z}$ along all three axes (Figure 3). The sensor model in its general form can be represented as:

$$y = K_{bias} + LPF_{signal} + noise \tag{12}$$

where K_{bias} represents the bias factor, the LPF low-pass filter, and the last term corresponds to the measurement noise.

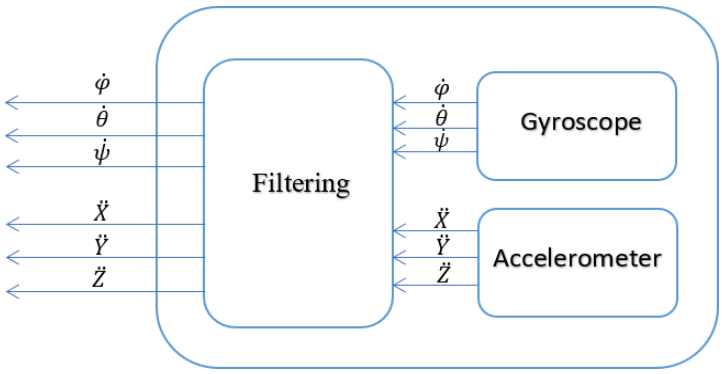


Figure 3: Block diagram of the sensor system

2.3.1. Unscented Kalman Filter

Due to the highly non-linear nature of drone dynamics, the Unscented Kalman filter (UKF) [28], [30] was chosen for state estimation. Unlike the extended Kalman filter (EKF) [29], [30], which relies on linearization via the Jacobian, the UKF uses the unscented transform (UT), a deterministic set of carefully selected sigma points to approximate the distribution of system states.

Algorithm Unscented_Kalman_filter($\mu_{t-1}, \Sigma_{t-1}, u_t, z_t$):

1: $\mathcal{X}_{t-1} = \left(\mu_{t-1}, \quad \mu_{t-1} + \gamma\sqrt{\Sigma_{t-1}}, \quad \mu_{t-1} - \gamma\sqrt{\Sigma_{t-1}}\right)$

2: $\overline{\mathcal{X}}_t^* = g(u_t, \mathcal{X}_{t-1})$

3: $\bar{\mu}_t = \sum_{i=0}^{2n} w_m^{[i]}\, \overline{\mathcal{X}}_t^{*[i]}$

4: $\bar{\Sigma}_t = \sum_{i=0}^{2n} w_c^{[i]} \left(\overline{\mathcal{X}}_t^{*[i]} - \bar{\mu}_t\right)\left(\overline{\mathcal{X}}_t^{*[i]} - \bar{\mu}_t\right)^T + R_t$

5: $\overline{\mathcal{X}}_t = \left(\bar{\mu}_t, \quad \bar{\mu}_t + \gamma\sqrt{\bar{\Sigma}_t}, \quad \bar{\mu}_t - \gamma\sqrt{\bar{\Sigma}_t}\right)$

6: $\bar{\mathcal{Z}}_t = h(\overline{\mathcal{X}}_t)$

7: $\hat{z}_t = \sum_{i=0}^{2n} w_m^{[i]}\, \bar{\mathcal{Z}}_t^{[i]}$

8: $S_t = \sum_{i=0}^{2n} w_c^{[i]} \left(\bar{\mathcal{Z}}_t^{[i]} - \hat{z}_t\right)\left(\bar{\mathcal{Z}}_t^{[i]} - \hat{z}_t\right)^T + Q_t$

9: $\Sigma_t^{x,z} = \sum_{i=0}^{2n} w_c^{[i]} \left(\overline{\mathcal{X}}_t^{[i]} - \bar{\mu}_t\right)\left(\bar{\mathcal{Z}}_t^{[i]} - \hat{z}_t\right)^T$

10: $K_t = \Sigma_t^{x,z}\, S_t^{-1}$

11: $\mu_t = \bar{\mu}_t + K_t(z_t - \hat{z}_t)$

12: $\Sigma_t = \bar{\Sigma}_t - K_t\, S_t\, K_t^T$

13: **return** μ_t, Σ_t

Figure 4: Unscented Kalman Filter algorithm [28], [30]

For a system of dimension n, $2n + 1$ sigma points are used. Each sigma point is independently propagated through the nonlinear system model, and the new predicted mean and covariance matrix are obtained as a weighted average of the transformed points. The distribution of sigma points is controlled by parameters α, β and κ: the parameter α determines the compactness of the distribution around the mean value, β takes into account the assumed error distribution ($\beta = 2$ for Gaussian distribution), while κ affects the range of sigma points in relation to the system dimension.

The key covariance matrices of the filter are: the process noise matrix Q, which reflects the uncertainty of the model; the measurement noise matrix R, which describes the reliability of the sensor; and the initial covariance matrix P_0, which defines the initial uncertainty in knowing the state of the system. Balancing these matrices achieves an optimal compromise between filter response speed and noise immunity.

2.3.2. UKF Implementation in a Simulation Environment

In the Simulink model, the UKF block receives as inputs the discretized angular velocities and linear accelerations from the sensors, along with the relevant constants of the dynamic model. The discretization step is 0.001s. Corrected values of angular and linear velocities are obtained at the output. Filtered angular velocities are directly fed into the controller, while linear velocities undergo an additional transformation of the coordinate system before further use in controller.

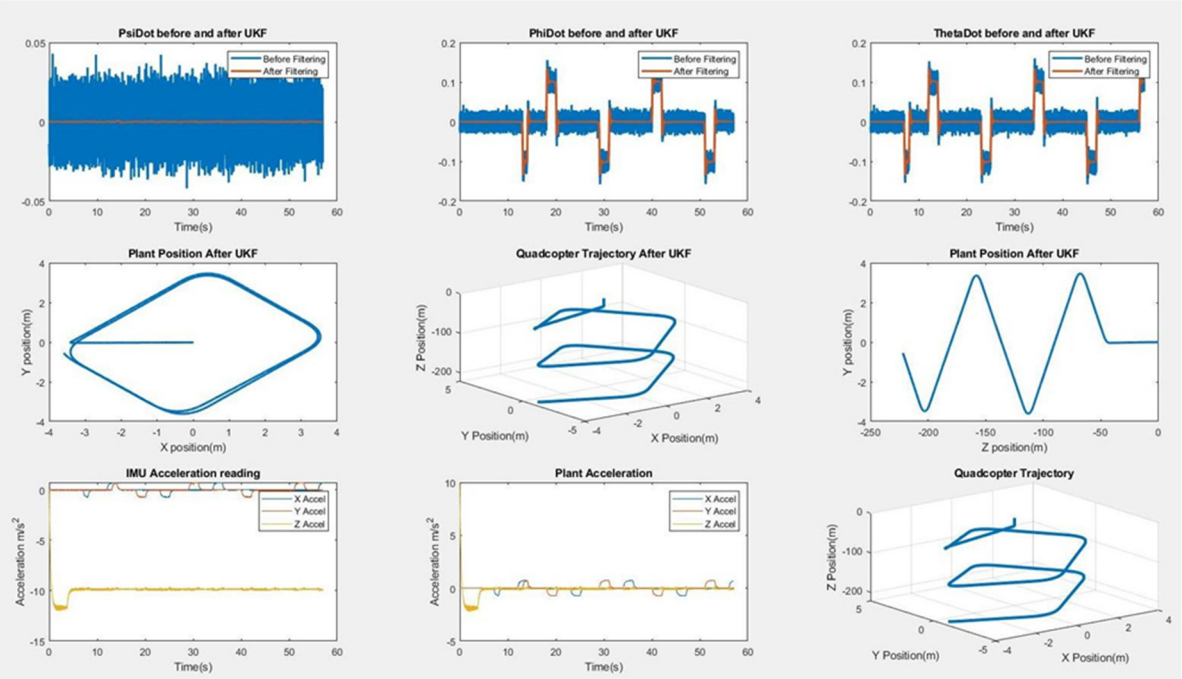


Figure 5: Simulation of a spiral movement results with using of UKF

The simulation results confirm the effectiveness of the applied approach (Figure 5). Unfiltered angular velocity signals contain significant noise that could potentially lead to instability of the control system, while UKF successfully eliminates noise and generates smooth, precise signals that closely follow the expected dynamics of the drone. The filtered orientation angles almost ideally match the reference values, which indicates a high estimation accuracy. The simulated three-dimensional trajectory, obtained at an aerodynamic drag coefficient of 0.1 kg/s, follows a circular path without significant deviations, confirming that the filter not only corrects measurement deviations but also ensures stability, predictability and robustness of the entire control system.

2.4. PID Controller

The controller is the central element of the closed-loop system whose task is to reduce the error between the reference and actual state values to zero. To stabilize the drone, a PID controller was used that simultaneously regulates four parameters: height (Z), roll angle (φ), pitch angle (θ) and yaw speed (ψ'). The control signal is formed as the sum of the proportional, integral and derivative term, where the proportional term provides general gain, the integral term eliminates the stationary error by accumulation, and the derivative term reacts to the rate of error change [27].

In order to avoid the impulse reaction of the derivative term at sudden changes of the reference signal, a PI-D structure is adopted, in which the derivative action is applied to the feedback state quantity and not to the error signal. The controller receives ten input signals, three linear accelerations, three angular velocities, and corresponding reference values and generates four motor angular velocities (Ω_1–Ω_4) as outputs.

3. INTELLIGENT CONTROL OF THE DRONE

The principles of assigning commands to the drone can be divided into automatic and manual control. Automatic control relies on built-in algorithms and sensor systems that enable stabilization and trajectory tracking without direct operator intervention. Manual control is traditionally realized through an RC controller, where the operator directly controls the height, orientation and movement of the aircraft [5], [6], [7], [8].

Recently, approaches based on the recognition of human limb movements as a natural interface for drone control have attracted increasing interest. Hand gesture recognition stands out as one of the most current areas within the human-machine interface [13], [14], [15], [16]. Hand gesture recognition systems can be divided into two categories: visual systems based on computer vision and systems based on sensor measurements. Visual systems, while not requiring the wearing of additional devices, have high computational complexity, are sensitive to lighting conditions, and can introduce latency critical for real-time applications.

As a more efficient alternative, sensor approaches based on a smart glove with resistive bending sensors and an inertial measurement module (IMU) are being proposed, which directly measure the biomechanical parameters of the hand independently of external visual conditions, with lower latency and less computing resource requirements [13], [14], [19]. The integration of gesture control with machine learning algorithms opens up the possibility of creating an intuitive and adaptive interface in which hand movements are directly mapped into control commands, thereby reducing the operator's cognitive effort and improving human-aircraft interaction [5], [10], [11], [23].

3.1. Architecture and Hardware Realization of the Smart Glove

The smart glove is a wearable interface for gestural control of the aircraft in which hand movements are recorded through a multisensory architecture and transformed into control signals. The design integrates five resistive bending sensors to monitor the position of the fingers and an IMU module to detect the orientation and angular velocity of the hand (Figure 6). The goal is to ensure stable acquisition of biomechanical parameters with sufficient resolution and sampling frequency for application in a real-time system [14], [15], [16], [19], [20].

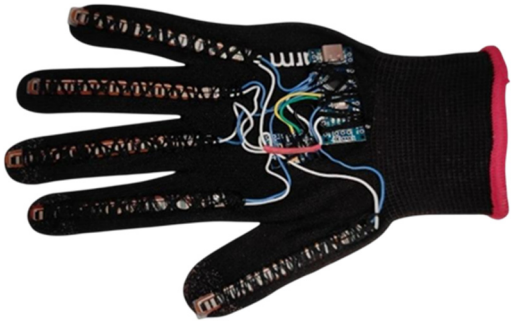


Figure 6: Smart glove

An Arduino Nano microcontroller based on an ATmega328P processor with a clock speed of 16 MHz is used for signal processing. The microcontroller reads analog values from the flex sensors via voltage dividers and communicates with the IMU module MPU6050 via the I²C protocol. The processed data is forwarded to a Simulink simulation, which serves as a virtual representation of the gesture classification processor.

Resistive flex sensors are placed on each finger of the glove. Each sensor changes resistance in proportion to the bend angle, and a voltage divider translates that change into a voltage signal that the microcontroller can process. For each sensor, two calibration values are defined, resistance at fully extended and fully flexed finger, which serve as reference points for normalizing the measured bending angle [14], [19].

The MPU6050 IMU module combines a three-axis accelerometer and a three-axis gyroscope in a single MEMS device with 16-bit analog-to-digital conversion per channel. The module provides continuous monitoring of the orientation and angular velocities of the hand in three-dimensional space, which, in combination with the bending sensor signals, enables precise detection of gestures in real time [3], [20], [22].

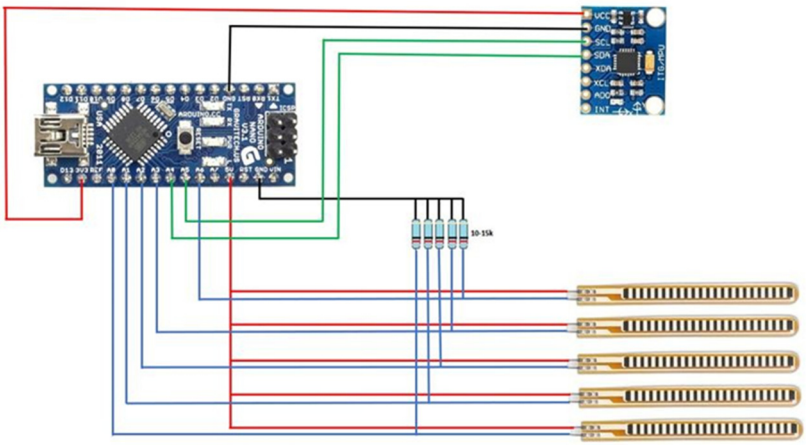


Figure 7: Electrical schematic of the sensor glove

3.2. Gesture Recognition Algorithm

The transformation of glove sensor signals into drone control commands requires an algorithm capable of modeling non-linear dependencies between the eight-dimensional measurement input into the four-dimensional control command space. The input vector consists of five normalized finger flexion values and three accelerometer components, while the control outputs correspond to the commands: roll, pitch, yaw and throttle. To realize this transformation, a multilayer feedforward neural network with an 8–10–4 architecture (8 inputs, 10 hidden and 4 output neurons), implemented in the MATLAB Neural Network Toolbox environment, was used [5], [9], [10], [11], [13].

The input layer receives eight numeric values from the glove sensor. The ten-neuron hidden layer uses a hyperbolic tangent activation function (tansig), while the four-neuron output layer uses a linear activation function (purelin), which is a standard configuration for regression tasks. The number of neurons in the hidden layer is determined empirically as a compromise between learning capacity and the risk of overfitting. Since the system performs instantaneous input-output

transformation without time dependencies, recurrent architectures are not required, and a network with one hidden layer is theoretically justified for approximating this class of nonlinear functions.

The Levenberg-Marquardt algorithm (trainlm) [26] was used for network training, which combines the advantages of gradient descent and Newton's method and is the standard and fastest algorithm for training networks with back propagation in the MATLAB environment. The error function to be minimized is the mean squared error (MSE). The data set was split randomly in the ratio 70/15/15% for training, validation and testing. The maximum number of epochs is set to 1000, and the early stopping criterion is activated after six consecutive epochs without improving the validation error, thus preventing overfitting (Figure 9 and 10).

The training dataset was formed by systematically recording sensor measurements for nine defined gestures: up, down, left, right, forward, back, left rotation, right rotation, and stop (Figure 8). Each gesture was performed multiple times with different intensities and speeds to capture natural variations in real use. Each sample represents an eight-dimensional feature vector assigned a corresponding control label, an annotation process that is a key step of supervised learning.



Figure 8: Defined hand commands (down, up, left, right, backward, forward, right rotation, left rotation, stop)

Training was completed after 29 epochs, with the best validation performance achieved at epoch 23 with a value of MSE = 0.010873. The error gradient at the end of training is 0.00003023, which indicates a good convergence towards a local minimum. The scaling parameter of the Levenberg–Marquardt algorithm (μ) reached the value of 1×10^{-15} in the last epoch, which confirms that the algorithm has entered the phase of rapid Newton descent characteristic of the vicinity of the minimum.

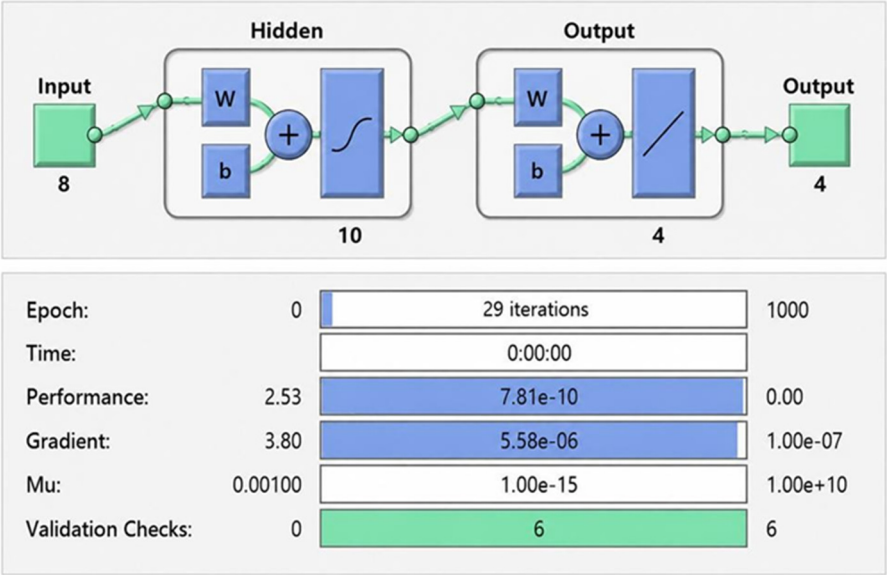


Figure 9: Neural network architecture and training parameters

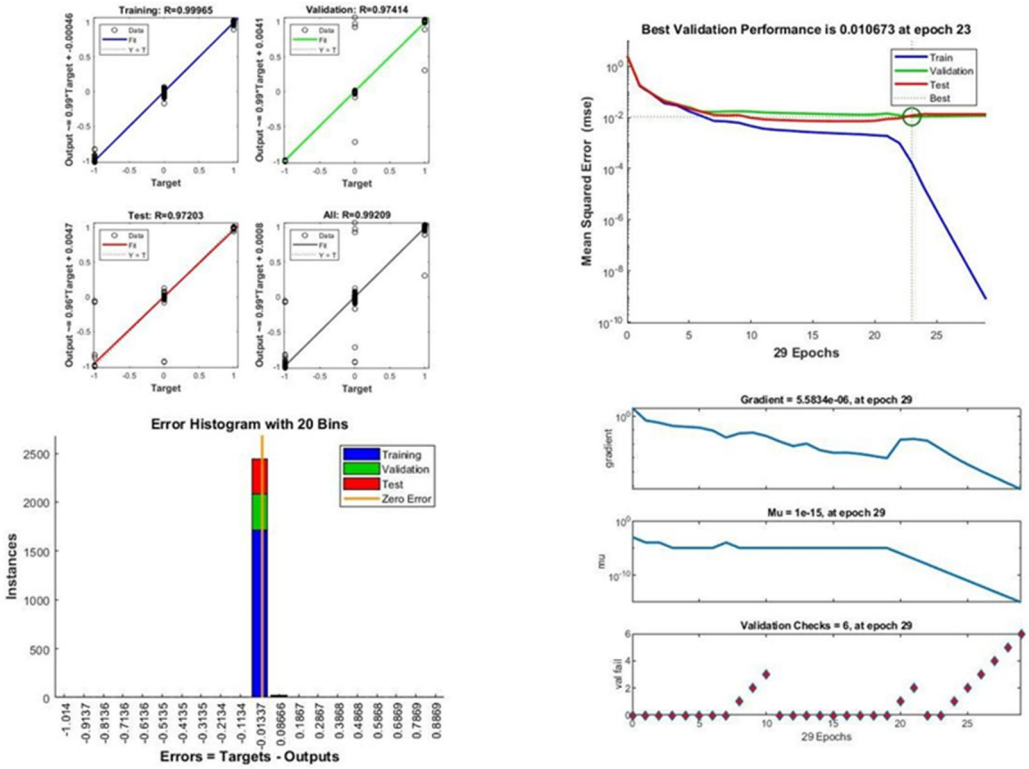


Figure 10: Neural network training in MATLAB

High correlation coefficient values (R > 0.97) on all data sets confirm that the network successfully learns the mapping between sensor data and control commands and has a satisfactory ability to generalize to new, previously unseen samples. A small difference between the training and validation coefficients indicates a slight overfitting, which is acceptable given the size of the data set (hundreds of samples for each of 9 commands). The error histogram confirms the concentration of errors around zero with a distribution approaching normality, which is a desirable property of a well-trained regression model.

3.3. Simulation of the Drone System with Intelligent Control

After the integration of the trained neural network into the Simulink model of the drone, a series of simulations lasting 20 seconds were carried out. The simulation environment is composed of mathematical model of drone dynamics that accounts for physical characteristics of the system, aerodynamic forces and moments, the gravitational field and regulatory loops for stabilizing position and orientation, and neural network that generates the input commands - roll, pitch, yaw, and throttle, which are applied directly to the drone model, simulating a realistic gesture control scenario without a physical remote control.

To start the simulation, the glove is physically connected to the computer via a serial connection (COM4, 9600 baud), after which a script is run in the MATLAB environment that establishes serial communication, loads the previously trained neural network model and initializes the simulation in the Simulink environment (Figure 11). From that moment, the user controls the drone in real time with hand movements and finger flexions, and the system in each iteration receives sensor data, processes it through a neural network and forwards the generated commands directly to the drone model.

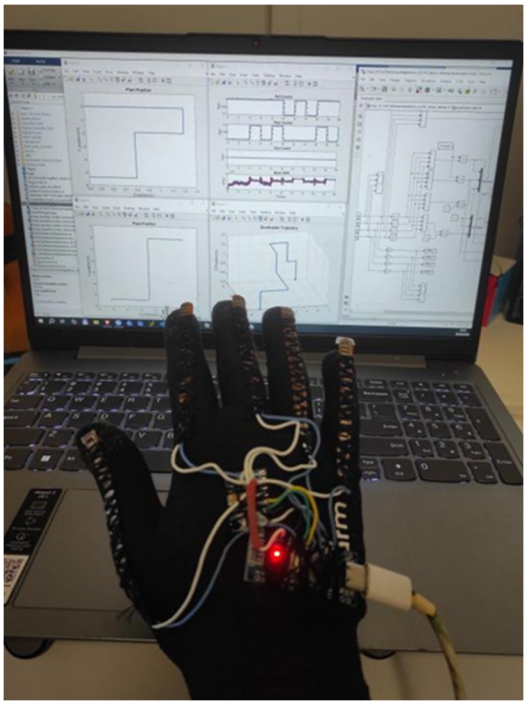


Figure 11: Simulation environment

The simulation results confirm that the proposed system enables stable and precise control of the drone exclusively through hand gestures (Figure 12). The time diagrams of the angular quantities show that the system successfully follows the given reference signals: the transitions between values are fast, and the tracking error in the stationary phases is negligibly small. The control signals have discrete changes corresponding to the detected commands of the neural network, while the system response indicates that the command recognition and transformation algorithm does not introduce significant delay into the control.

In order to quantitatively validate the simulation results, key performance metrics were extracted from the recorded signals. The steady-state tracking error for the roll and pitch angles was measured to be below 0.003 rad, while the yaw angle maintained a steady-state error of less than 0.005 rad throughout the entire simulation. The rise time for step transitions in the roll and pitch reference signals averaged approximately 0.4 seconds, with a settling time of under 0.8 seconds, confirming rapid and stable response to gesture-generated commands. Overshoot across all angles remained below 5%. The total distance traveled by the drone during the simulation was approximately 10.5 meters across all three axes, with positional transitions executed without oscillation or drift between commanded segments. These quantitative results collectively confirm the stability, accuracy, and efficiency of the proposed gesture-based control system under simulated flight conditions.

The three-dimensional representation of the drone's trajectory gives the most complete picture of the spatial movement during the simulation. The characteristic stepped form of the trajectory clearly describes the individual phases of the flight: horizontal movements in different directions alternate with changes in height, forming a three-dimensional trajectory that directly reflects the sequence of gestures performed during the simulation. Vertical jumps between heights are performed precisely and with well-defined levels, which confirms the efficient management of the throttle channel. The projection of the trajectory in the X–Y plane shows smooth and stable segments of movement without sudden jumps or instability, which testifies to the good quality of regulation and the absence of oscillations in the control loop.

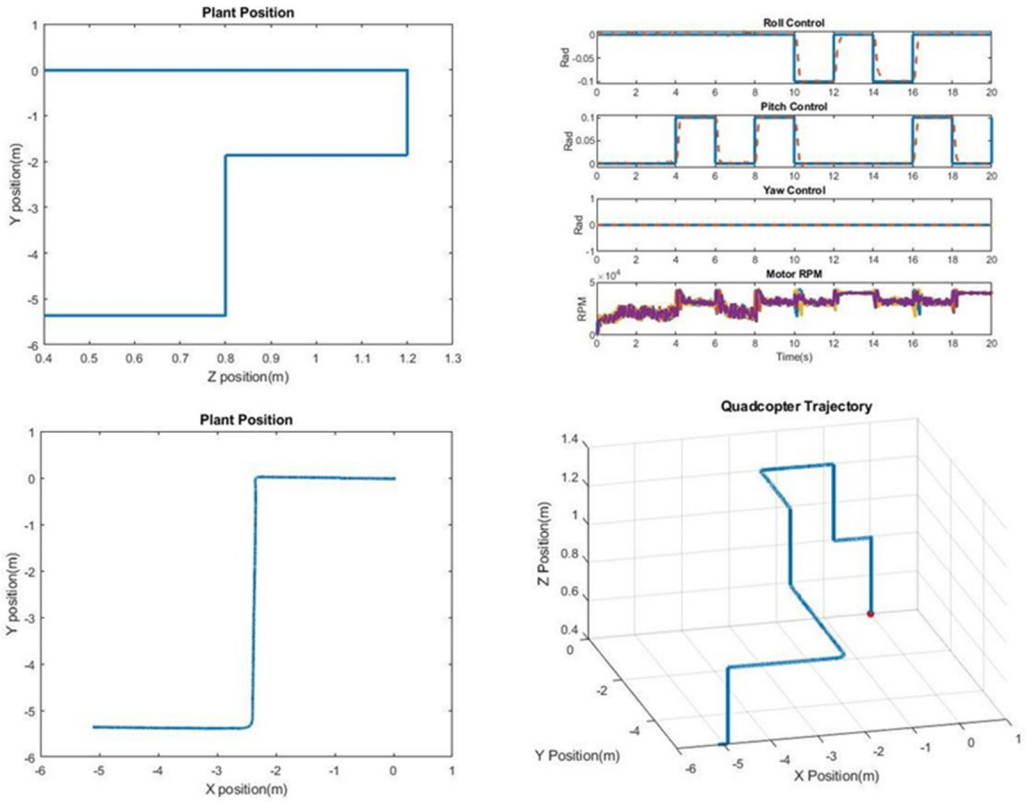


Figure 12: Simulation results when using smart glove as input for drone commands

4. CONCLUSION

In this paper, a complete system for modeling, simulation and intelligent control of a drone has been implemented. A detailed mathematical model of the drone dynamics has been built, which includes aerodynamic forces, engine and battery dynamics, as well as sensor systems and control algorithms. The mathematical model of the drone is implemented in the MATLAB/Simulink environment, which allows for analysis of the aircraft's behavior in controlled simulation conditions. Special attention is paid to the processing of sensor data, where a Kalman filter is applied in order to reduce noise and obtain a more accurate assessment of the system's state, and flight stabilization is achieved through a PID control system. The core idea of the project is the method of interaction between the user and the drone - instead of a classic remote control, control is realized through a smart glove equipped with sensors. The data from the glove is processed by a neural network that recognizes the user's gestures and transforms them into control commands that are used as input to the drone's simulation model.

The results of the conducted simulations show that the proposed approach works stably and predictably. The system successfully tracks the reference values for pitch, yaw, and rotation, with low tracking error and fast and stable transitions between different states. The control signals generated by the neural network directly affect the dynamics of the motors and result in expected changes in the drone's movement. The analysis of the three-dimensional flight trajectory further confirms that the system successfully interprets the user's gestures and converts them into realistic maneuvers in space. The drone's movement is continuous and stable, without sudden or uncontrolled changes, indicating that the regulatory loops and the gesture recognition algorithm work in sync and efficiently.

The practical significance of such a system is broad and has the potential to be applied in systems where intuitive and natural interaction between man and machine is required. Gesture control can be used in rescue missions, inspection operations, education, as well as in research and demonstration systems where it is important that the control is simple and easy to understand for the user. Such an approach represents a step towards more intuitive interfaces for controlling unmanned aerial vehicles and other autonomous systems. Today, research in this area is moving towards fully wireless, embedded systems with negligible delay and high accuracy.

Although the simulation results are encouraging, there are several possibilities for further improving the system. In the current implementation, control commands are generated as discrete values, which means that each gesture corresponds to a fixed command. In the future, a system could be developed that recognizes continuous movements, i.e. not only the direction of movement but also a slight tilt to the left to generate a small tilt value, and a pronounced tilt — a larger value, which would provide the user with proportional control. Further improvement can be achieved by using more precise sensors, a larger number of sensor points, or by introducing intermediate steps and combined gestures, which would allow for a larger number of control commands. In parallel, replacing serial with wireless communication is a technically trivial step that eliminates the only physical limitation of the prototype, and testing on a real drone remains a logical and inevi-

table next step for final validation of the system outside the simulation environment. Modern research in this area often uses deeper neural networks, advanced sensor data fusion methods, and adaptive control algorithms, which could further increase the precision and stability

REFERENCES

[1] M. Idrissi, M. Salami, and F. Annaz, "A Review of Quadrotor Unmanned Aerial Vehicles: Applications, Architectural Design and Control Algorithms", Journal of Intelligent & Robotic Systems, Vol. 101(22, https://doi.org/10.1007/s10846-021-01527-7, (2021)

[2] I. H. Beloev, "A Review on Current and Emerging Application Possibilities for Unmanned Aerial Vehicles", Acta Technologica Agriculturae, Vol. 19(3), pp. 70-76, https://doi.org/10.1515/ata-2016-0015, (2016)

[3] N. Koksal, M. Jalalmaab, and B. Fidan, "Adaptive Linear Quadratic Attitude Tracking Control of a Quadrotor UAV Based on IMU Sensor Data Fusion", Sensors 2019, Vol. 19(1), pp. 46, https://doi.org/10.3390/s19010046, (2019)

[4] S. Bouabdallah, A. Noth, and R. Siegwart, "PID vs. LQ control techniques applied to an indoor micro quadrotor", Proceedings of the 2004 IEEE International Conference on Robotics and Systems (IROS), Sendai (Japan), 28 September - 02 October 2004, (2004)

[5] B. Jiang, B. Li, W. Zhou, L.-Y. Lo, C.-K. Chen, and C.-Y. Wen, "Neural Network Based Model Predictive Control for a Quadrotor UAV", Aerospace, Vol. 9(8), p. 460, https://doi.org/10.3390/aerospace9080460, (2022)

[6] T. Wang, S. S. Zhang, and N. M. Freris, "An Overview of Quadrotor UAV Control Methodologies", https://doi.org/10.48550/arXiv.2209.02954, (2022)

[7] A. Rodriguez-Ramos, C. Sampedro, H. Bavle, P. de la Puente, and P. Campoy, "A Deep Reinforcement Learning Strategy for UAV Autonomous Landing on a Moving Platform", Journal of Intelligent & Robotic Systems, Vol. 93(1), pp. 351-368, https://doi.org/10.1007/s10846-018-0891-8, (2019)

[8] A. S. Brandao, F. N. Martins, and H. B. Soneguetti, "A vision-based line following strategy for an autonomous UAV", Proceedings of the 2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO), Colmar (France), 21-23 July 2015, pp. 214-219, (2015)

[9] S. L. Waslander, G. M. Hoffmann, J. S. Jang, and C. J. Tomlin, "Multi-Agent Quadrotor Testbed Control Design: Integral Sliding Mode Control Design", https://doi.org/10.48550/arXiv.1905.01657, (2019)

[10] J. Cifuentes, P. Boulanger, M. T. Pham, F. Prieto, and R. Moreau, "Gesture Classification Using LSTM Recurrent Neural Networks", Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Med-icine & Biology Society (EMBC), Berlin (Germany), 23 July 2019, pp. 6864-6867, (2019)

[11] J.-W. Lee and K.-H. Yu, "Wearable Drone Controller: Machine Learning-Based Hand Gesture Recognition and Vibrotactile Feedback", Sensors, Vol. 23(5), p. 2666, https://doi.org/10.3390/s23052666, (2023)

[12] C. Yu, S. Fan, Y. Liu, and Y. Shu, "End-Side Gesture Recognition Method for UAV Control", IEEE Sensors Journal, (2022)

[13] K. Haratiannejadi and R. R. Selmic, "Smart Glove and Hand Gesture-Based Control Interface for Multi-Rotor Aer-ial Vehicles in a Multi-Subject Environment", IEEE Access, Vol. 8, pp. 227667-227677, https://api.semanticscholar.org/CorpusID:230512516, (2020)

[14] W.-C. Chuang, W.-J. Hwang, T.-M. Tai, D.-R. Huang, and Y.-J. Jhang, "Continuous Finger Gesture Recognition Based on Flex Sensors", Sensors, Vol. 19(18), p. 3986, https://doi.org/10.3390/s19183986, (2019)

[15] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal Gesture Recognition Using 3-D Convolution and Convolutional LSTM", IEEE Access, Vol. 5, pp. 4517 - 4524, https://doi.org/10.1109/ACCESS.2017.2684186, (2017)

[16] H. Cheng, L. Yang, and Z. Liu, "Survey on 3D Hand Gesture Recognition", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 26(9), pp. 1659-1673, https://doi.org/10.1109/TCSVT.2015.2469551, (2016)

[17] A. Sulficar, H. Suresh, A. Varma, and A. Radhakrishnan, "Modeling, Simulation and Complete Control of a Quad-copter", Major Project Semester Report, National Institute of Technology Karnataka (India), (2017)

[18] R. Mahony, V. Kumar, and P. Corke, "Multirotor Aerial Vehicles: Modeling, Estimation, and Control of Quarotor", IEEE Robotics & Automation Magazine, Vol.19(3), pp. 20-32, https://doi.org/10.1109/MRA.2012.2206474, (2012)

[19] A. Sreejan and Y. S. Narayan, "A Review on Applications of Flex Sensors", International Journal of Emerging Technology and Advanced Engineering, Vol. 7(7), pp. 97-101, (2017)

[20] J. Huang, "Design of Angle Detection System Based on MPU6050", Proceedings of the 7th International Conference on Education, Management, Information and Computer Science (ICEMC 2017), pp. 6-8, https://doi.org/10.2991/icemc-17.2017.2, (2017)

[21] H. P. Gupta, H. S. Chudgar, S. Mukherjee, T. Dutta, and K. Sharma, "A Continuous Hand Gestures Recognition Technique for Human-Machine Interaction Using Accelerometer and Gyroscope Sensors", IEEE Sensors, Vol. 16(16), pp. 6425-6432, https://doi.org/10.1109/JSEN.2016.2581023, (2016))

[22] C. Liu and T. Szirányi, “Real-Time Human Detection and Gesture Recognition for On-Board UAV Rescue”, Sensors, Vol. 21(6), p. 2180, https://doi.org/10.3390/s21062180, (2021)

[23] B. Hu and J. Wang, “Deep Learning Based Hand Gesture Recognition and UAV Flight Controls”, International Journal of Automation and Computing, Vol. 17(1), pp. 17-29, https://doi.org/10.1007/s11633-019-1194-7, (2020)

[24] N. Ferry, “Quadcopter Plant Model and Control System Development With MATLAB/Simulink Implementation”, Master's Thesis, Rochester Institute of Technology (USA), (2017)

[25] B. Taylor, M. Allen, P. Henson, X. Gao, H. Malik, and P. Zhu, “Enhancing Drone Navigation and Control: Gesture-Based Piloting, Obstacle Avoidance, and 3D Trajectory Mapping”, Applied Sciences, Vol. 15(13), p. 7340, https://doi.org/10.3390/app15137340, (2025)

[26] J. J. Moré, “The Levenberg-Marquardt algorithm: Implementation and theory”, in Numerical Analysis, G. A. Watson, Ed., Berlin, Germany: Springer, 1978, pp. 105-116., https://doi.org/10.1007/BFb0067700, (2006)

[27] K. J. Åström and T. Hägglund, “PID Controllers: Theory, Design, and Tuning”, Instrument Society of America, Research Triangle Park, North Carolina (USA), (1995)

[28] S. J. Julier and J. K. Uhlmann, “New extension of the Kalman filter to nonlinear systems”, Proceedings of SPIE, Vol. 3068, Signal Processing, Sensor Fusion, and Target Recognition VI, pp. 182-193, https://doi.org/10.1117/12.280797, (1997)

[29] M. S. Grewal and A. P. Andrews, “Kalman Filtering: Theory and Practice Using MATLAB”, Wiley, New York, New York (USA), (2002)

[30] F. Auger, M. Hilairet, J. M. Guerrero, E. Monmasson, T. Orlowska-Kowalska, and S. Katsura, “Industrial Applica-tions of the Kalman Filter: A Review”, IEEE Transactions on Industrial Electronics, Vol. 60(12), pp. 5458-5471, https://doi.org/10.1109/tie.2012.2236994, (2013)

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Event-triggered ADP for asymptotic output regulation of unknown linear systems

Ljubisa Dubonjic[1], João Paulo Zomer Machado[2], Dragan Prsic[1], Vladimir Stojanovic[1*], Xiangpeng Xie[3]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering, Kraljevo, Serbia

[2] Universidade Federal de Santa Catarina, Dept. of Automation and Systems Engineering, 88040-970 Florianópolis, SC, Brazil

[3] Nanjing University of Posts and Telecommunications, School of Internet of Things, Nanjing 210023, China

ARTICLE INFO

* **Correspondence:** vladostojanovic@mts.rs

DOI: https://doi.org/10.46793/ET26.D03D **PAGES:** D23 – D30

ABSTRACT

The deployment of learning-based controllers in modern networked cyber-physical systems is constrained by bandwidth limitations, partial state observability, and parametric uncertainties. Traditional adaptive dynamic programming (ADP) relies on full-state feedback and periodic sampling, inducing network congestion and steady-state tracking errors. This paper develops a fully data-driven dynamic event-triggered output regulation scheme operating on measurable input-output streams. By integrating a projection-safeguarded value iteration algorithm with an internally evolving dynamic threshold variable, we guarantee asymptotic convergence of the sampling error without requiring an admissible initial stabilizing policy or explicit system identification. Simulations on a third-order uncertain plant demonstrate a 54.0% reduction in control transmissions and an 8.2% suboptimality bound relative to the model-based LQR baseline, with reconstruction error consistently decaying below 10^{-3}. The framework provides a computationally efficient, theoretically rigorous architecture for resource-constrained cyber-physical networks.

KEYWORDS

Adaptive dynamic programming, Dynamic event-triggered control, Data-driven output feedback, Value iteration, Optimal output regulation, Networked control systems

1. INTRODUCTION

Networked cyber-physical systems require optimal control under strict communication constraints and parametric uncertainties. Traditional model-based and periodic ADP frameworks suffer from excessive bandwidth usage, reliance on full-state measurements, and vulnerability to unmodeled dynamics. Data-driven ADP and reinforcement learning synthesize near-optimal policies directly from input-output trajectories [1]. However, their deployment is hindered by static event-triggered mechanisms (SETM) that introduce measurable residual steady-state errors, limiting actuator precision under constant-threshold conditions [2]. Furthermore, conventional policy iteration schemes mandate an admissible initial stabilizing gain K_0, a condition rarely met in open-loop unstable plants [3, 4].

Recent output-feedback ADP architectures reconstruct states from historical data but retain static triggering thresholds, decoupling updates from evolving dynamics and forcing rigid trade-offs between resource conservation and transient robustness [5, 6]. While works [6, 7] reduce control updates, they neglect asymptotic error convergence, resulting in persistent tracking offsets. Additionally, unaddressed noise propagation through Kronecker-based reconstruction matrices accelerates actuator wear under abrupt torque jumps, while the absence of explicit bounds linking triggering parameters to performance degradation leaves practitioners without predictable tuning guidelines. Value iteration (VI) offers a model-free alternative, yet its integration with dynamic event-triggering remains theoretically underdeveloped [3].

This paper develops a data-driven dynamic event-triggered output regulation framework grounded in safeguarded value iteration. Contributions:

i. A novel dynamic event-triggering mechanism (DETM) replaces static thresholds with an internally evolving variable $W(k)$, explicitly eliminating the measurable residual steady-state error inherent in constant-threshold static mechanisms while guaranteeing asymptotic error convergence and excluding Zeno behavior.

ii. A projection-safeguarded VI algorithm operating without K_0, circumventing the necessity for heuristic initialization when an admissible stabilizing gain is unavailable for open-loop unstable plants.

iii. A robust I/O reconstruction filter with Tikhonov regularization that directly mitigates the risk of stability loss and Zeno behavior induced by unpredictable stochastic threshold transitions under sensor noise.

iv. An explicit suboptimality bound bridging the absence of verifiable bounds linking dynamic triggering parameters to closed-loop performance degradation, enabling deterministic resource-performance trade-offs.

The remainder is organized as follows: Section II formalizes the problem; Section III details the VI and DETM design; Section IV provides analytical validation; Section V presents numerical benchmarking; Section VI concludes.

2. PRELIMINARISES AND PROBLEM FORMULATION

We consider a discrete-time LTI system subject to unknown dynamics and bounded disturbances:

$$x_{k+1} = Ax_k + Bu_k + \omega_k, \quad y_k = Cx_k + \nu_k, \tag{1}$$

where $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^m$, $y_k \in \mathbb{R}^p$, and ω_k, ν_k denote bounded process and measurement noise. Matrices A, B, C are unknown. The control objective is asymptotic output regulation with minimal quadratic cost:

$$J(u) = \sum_{k=0}^{\infty}(y_k^{\top} Q y_k + u_k^{\top} R u_k), \quad Q \succcurlyeq 0,\ R \succ 0. \tag{2}$$

The optimal policy $u_k^* = -K^* x_k$ requires solving the Discrete Algebraic Riccati Equation (DARE) [8]. Since states and matrices are unmeasurable/unknown, we adopt:

Assumption 1 (Stabilizability and Observability). *(A, B) is stabilizable and (A, C) is observable. For any non-pathological sampling period $h > 0$, the discretized pair (A_d, C) remains observable.*

Assumption 2 (Persistent Excitation). *During the learning phase, u_k includes exploration noise ξ_k (sum of sinusoids with incommensurable frequencies) such that the historical data regressor satisfies $\sum_{k=t}^{t+T} z_k z_k^{\top} \geq \delta I_q$ for some $\delta > 0$, $T < \infty$, ensuring rank$(\Phi_k) = \frac{q(q+1)}{2} + mq$ over any window of length T.*

Assumption 3 (Bounded Noise). *$\| \omega_k \| \leq \bar{\omega}$ and $\| \nu_k \| \leq \bar{\nu}$ for all k, with known bounds.*

Assumption 4 (Bounded Reconstruction Drift). *The online estimate Θ_k satisfies $\| \Theta_k - \Theta^* \|_F \leq \delta_\Theta < \infty$ for all $k \geq k_0$, ensured by Tikhonov regularization and bounded noise.*

Under these, the baseline static condition $\| e_k \|^2 \leq \delta_{static}$ is replaced by a dynamic mechanism that explicitly decouples communication frequency from steady-state precision.

3. PROPOSED METHODOLOGY

3.1. State Reconstruction via I/O History

Following [5, 6], we reconstruct the state using a finite horizon N of past inputs and outputs. Define $z_k \in \mathbb{R}^q$ with $q = N(m + p)$:

$$z_k = [\hat{u}_{k-N}^{\top} \quad \cdots \quad \hat{u}_{k-1}^{\top} \quad y_{k-N}^{\top} \quad \cdots \quad y_{k-1}^{\top}]^{\top}. \tag{3}$$

By observability, $\exists \Theta^* \in \mathbb{R}^{n \times q}$ such that $x_k = \Theta^* z_k$. Online approximation uses:

$$\Theta_k = \underset{\Theta}{\operatorname{argmin}} \| Y_k - \Theta Z_k \|_F^2 + \lambda \| \Theta \|_F^2. \tag{4}$$

Tikhonov regularization $\lambda > 0$ attenuates noise propagation, yielding a numerically stable update. The reconstruction error is $e_k^v = (\Theta_k - \Theta^*) z_k + \Theta_k \tilde{z}_k$, bounded under Assumption 4.

3.2. Dynamic Event-Triggering Mechanism (DETM)

Let k_j be the j-th trigger. Between triggers, $u_k = \hat{u}_k = -K_{k_j} z_{k_j}$. Define sampling error $e_k^v = \hat{z}_k - z_k$. Introduce scalar $W(k) \in \mathbb{R}_{\geq 0}$:

$$k_{j+1} = \inf\left\{k > k_j \mid \| K_k e_k^v \|^2 \geq \frac{W(k)}{\kappa} + \alpha \| y_k \|_Q^2\right\}, \tag{5}$$

$$\Delta W(k) \quad = -\beta W(k) + \gamma \parallel e_k^v \parallel^2 - \eta \parallel y_k \parallel_Q^2, \quad W(0) > 0. \tag{6}$$

Design scalars $\kappa, \alpha, \beta, \gamma, \eta > 0$. Unlike static thresholds, $W(k)$ decays exponentially as $\parallel e_k^v \parallel \to 0$, forcing asymptotic convergence and excluding Zeno behavior.

Remark 1 (Parameter Design Verification). *The stability condition in Theorem 3 requires* $\eta\lambda_{min}(Q) > \gamma \parallel C\Theta^* \parallel^2$. *With the simulation parameters* $\eta = 150$, $\gamma = 0.1$, $Q = 100I$, $C = [8.0, 6.0, 3.0]$, *and* $\parallel \Theta^* \parallel_F \approx 1.27$ *(computed offline from* A, B, C*), we obtain* $\eta\lambda_{min}(Q) = 15000$ *and* $\gamma \parallel C\Theta^* \parallel^2 \approx 0.1 \times 6.87 = 0.687$, *confirming* $15000 \gg 0.687$. *In practice,* $\parallel C\Theta^* \parallel$ *is estimated from the initial reconstruction phase; practitioners should set* $\eta > \gamma \parallel C\Theta^* \parallel^2 / \lambda_{min}(Q)$ *conservatively.*

Algorithm 1: Data-Driven DETM-VI for Output Regulation

Input: Weights Q, R, DETM params κ, α, β, γ, η, horizon N
Output: Optimal gain K^*, threshold sequence $\{k_j\}$
1 Initialize $P_0 = 0_{q\times q}$, $K_0 = 0_{m\times q}$, $W(0) = W_{init}$, $k_0 = 0$;
2 for *iteration* $i = 0, 1, \ldots$ **do**
3 Apply $\hat{u}_k = -K_i z_{kj} + \xi_k$ on $[k_j, k_{j+1})$;
4 Collect data $\{z_k, y_k, \hat{u}_k\}$ and form Φ_k, Ψ_k;
5 Solve $\hat{G}_{i+1} = \Phi^{\dagger}_k \Psi_k$ via regularized least squares;
6 Extract $P_{i+1} \leftarrow \mathcal{P}_\Omega(\hat{G}^{zz}_{i+1})$, $K_{i+1} \leftarrow$ policy improvement;
7 **if** $\|P_{i+1} - P_i\|_F < \varepsilon$ **then**
8 **return** $K^* = K_{i+1}$
9 **end**
10 ;
11 end

3.3. Safeguarded Value Iteration Without Initial Stabilizer

Traditional PI requires admissible K_0. We initialize $P_0 = 0_{q\times q}$ and apply projection $\mathcal{P}_\Omega(M) = M + \delta I$ if $\lambda_{\min}(M) < 0$, ensuring $P_k \succcurlyeq 0$. Bellman updates:

$$z_{k+1}^\top P_{k+1} z_{k+1} - z_k^\top P_k z_k \quad = -y_k^\top Q y_k - u_k^\top R u_k + \epsilon_{PE}, \tag{7}$$

$$K_{k+1} \quad = (R + B_d^\top P_{k+1} B_d)^{-1} B_d^\top P_{k+1} A_d. \tag{8}$$

In model-free form:

$$\Phi_k \mathrm{vec}(G_{k+1}) = \Psi_k, \quad G_{k+1} = \begin{bmatrix} B_d^\top P_{k+1} B_d & B_d^\top P_{k+1} A_d \Theta \\ \star & \Theta^\top P_{k+1} \Theta \end{bmatrix}. \tag{9}$$

Least-squares solution yields $\hat{G}_{k+1} \to K_{k+1}, P_{k+1}$, circumventing heuristic initialization.

3.4. Analytical Validation

Theorem 1 (Unique Policy Evaluation). *Under Assumption 2,* $rank(\Phi_k) = \frac{q(q+1)}{2} + mq$. *Consequently,* G_{i+1} *is uniquely determined.*

Proof. *The regressor* Φ_k *aggregates* $\{z_k \otimes z_k, z_k \otimes u_k\}_{k=t}^{t+T}$. *By Assumption 2, the PE condition holds:* $\sum_{k=t}^{t+T} z_k z_k^\top \geq \delta I_q$ *for some* $\delta > 0$ *and finite* T. *This implies that the Kronecker products* $\{z_k \otimes z_k\}$ *and* $\{z_k \otimes u_k\}$ *together span a space of dimension* $\frac{q(q+1)}{2} + mq$, *which equals the number of independent entries in the symmetric matrix* G_{i+1} *and the off-diagonal block* K_{i+1} [5] . *Hence* $\Phi_k^\top \Phi_k \succ 0$, *and the least-squares solution* $\hat{G}_{i+1} = (\Phi_k^\top \Phi_k)^{-1} \Phi_k^\top \Psi_k$ *is unique. Tikhonov regularization* $\lambda > 0$ *in the reconstruction step (Section 3.1) mitigates numerical ill-conditioning due to bounded sensor noise, but the full column rank of* Φ_k *is guaranteed by the PE condition independently of* λ.

Under DETM, inter-event times satisfy $k_{j+1} - k_j \geq \tau_{min} > 0$ *(proved in Theorem 3). This lower bound ensures that, over any observability window of length* T, *the sampled data retains the excitation richness of the underlying discrete system, so that the PE sum condition remains satisfied even under aperiodic sampling.*

Theorem 2 *(Monotonic Convergence of Safeguarded VI). Let* $P_0 = 0$ *and* P_i *updated via Alg. 1 with* $\mathcal{P}_\Omega(\cdot)$. *Then* $\{P_i\}$ *converges monotonically to* P^*.

Proof. *Induction:* $P_0 = 0 \preccurlyeq P^*$. *Assume* $P_i \preccurlyeq P^*$. *The discrete Riccati operator* $\mathcal{T}(P) = A^\top PA - A^\top PB(R + B^\top PB)^{-1} B^\top PA + C^\top QC$ *is monotone on the PSD cone. Thus* $P_{i+1} = \mathcal{T}(P_i) \preccurlyeq \mathcal{T}(P^*) = P^*$. *The projection* $\mathcal{P}_\Omega(\cdot)$ *enforces* $P_i \succcurlyeq 0$. *A monotone, bounded sequence in a closed convex cone converges to a fixed point.*

Theorem 3 (Asymptotic Stability and Zeno-Freeness under DETM). *Under (1), control $u_k = -K^*\hat{z}_k$, and DETM (5)–(6), if $\kappa > 0$, $\beta \in (0,1)$, $\eta\lambda_{min}(Q) > \gamma \parallel C\Theta^* \parallel^2$, then:*

1. *Origin is globally asymptotically stable.*
2. *Zeno behavior excluded: $min_j(k_{j+1} - k_j) \geq \tau_{min} > 0$.*

Proof. 1) Stability: *Define augmented Lyapunov function $V_k = z_k^\top P^* z_k + cW(k)$ with $c > 0$. Along trajectories, $u_k = -K^* z_k + K^* e_k^v$. Using the DARE and expanding $\Delta V_k = V_{k+1} - V_k$:*

$$\Delta V_k \leq -y_k^\top Q y_k - \parallel K^* z_k \parallel_R^2 + 2z_k^\top A_d^\top P^* B_d K^* e_k^v + \parallel K^* e_k^v \parallel^2_{R+B^\top P^* B} + c\left(-\beta W(k) + \gamma \parallel e_k^v \parallel^2 - \eta \parallel y_k \parallel_Q^2\right).$$

Apply Young's inequality to the cross term: $2z^\top Me \leq \sigma \parallel z \parallel^2 + \frac{1}{\sigma} \parallel M \parallel^2 \parallel e \parallel^2$. Choosing σ sufficiently small and grouping terms yields:

$$\Delta V_k \leq -\epsilon_1 \parallel z_k \parallel^2 - (\eta c - \alpha c - 1) \parallel y_k \parallel_Q^2 - \left(\beta c - \gamma - \frac{1}{\kappa}\right) W(k), \tag{10}$$

where $\epsilon_1 > 0$ follows from $Q \succ 0$ and observability. Select $c > max\left\{\frac{\gamma + 1/\kappa}{\beta}, \frac{1+\alpha}{\eta}\right\}$. Then $\Delta V_k \leq -\epsilon_1 \parallel z_k \parallel^2 - \epsilon_2 \parallel y_k \parallel_Q^2 - \epsilon_3 W(k)$ with $\epsilon_i > 0$. By LaSalle's invariance principle for discrete systems, $(z_k, W(k)) \to (0,0)$. Notably, $W(k) \to 0$ is derived, not assumed.

2) Zeno-freeness: *We establish $\tau_{min} > 0$ via two complementary arguments.*

Discrete-time lower bound: Since the system evolves at sampling instants $k \in \mathbb{N}$, any two consecutive trigger times satisfy $k_{j+1} - k_j \geq 1$, giving $\tau_{min} \geq h > 0$ unconditionally.

Tighter bound via threshold positivity: Let $L_K = \parallel K^* \parallel \cdot \parallel A_d - B_d K^* \parallel$ denote the Lipschitz constant of the sampling error dynamics. At each trigger k_j, $e_{k_j}^v = 0$ so $\parallel K^* e_{k_j}^v \parallel^2 = 0$. The threshold at k_j is $\frac{W(k_j)}{\kappa} + \alpha \parallel y_{k_j} \parallel_Q^2 \geq \frac{W(k_j)}{\kappa}$. Since $W(k) \geq 0$ and $\Delta W(k) = -\beta W(k) + \gamma \parallel e_k^v \parallel^2 - \eta \parallel y_k \parallel_Q^2 \leq -\beta W(k) + \gamma \parallel e_k^v \parallel^2$, iterating from $k_0 = 0$ gives $W(k_j) \geq (1-\beta)^j W(0) > 0$ for any finite j (since $\beta \in (0,1)$ and $W(0) > 0$). Thus $\frac{W(k_j)}{\kappa} > 0$ even when $\parallel y_{k_j} \parallel_Q = 0$ (system at equilibrium). The sampling error grows as $\parallel K^* e_{k+1}^v \parallel \leq L_K \parallel e_k^v \parallel + L_K \parallel z_{k_j} - z_k \parallel$, with rate bounded by the Lipschitz constant. The minimum number of steps for $\parallel K^* e_k^v \parallel^2$ to grow from 0 to $W(k_j)/\kappa$ satisfies

$$\tau_{\min} \geq \max\left(h, \frac{\sqrt{W(k_j)/\kappa}}{2L_K \sup_k \parallel z_k \parallel}\right) > 0, \tag{11}$$

where boundedness of $sup_k \parallel z_k \parallel$ follows from Part 1.

Theorem 4 (Suboptimality Bound). *$J_{DET} \leq J^*\left(1 + \frac{\bar{\lambda}(R)\delta_e(\delta_e + 2\parallel K^* \parallel)}{\lambda_{min}(Q)(1-\beta)}\right) = J^*(1+\rho)$, where $\delta_e = sup_{k:\parallel z_k \parallel > 0} \frac{\parallel K^* e_k^v \parallel}{\parallel z_k \parallel}$, and J^* denotes the optimal LQR cost (model-based baseline).*

Proof. *From the proof of Theorem 3, $\Delta V_k \leq -y_k^\top Q y_k - u_k^{*\top} R u_k^* + \epsilon_{res}(k)$, where $\epsilon_{res}(k)$ collects terms involving e_k^v. Summing from $k = 0$ to ∞:*

$$\sum_{k=0}^{\infty}(y_k^\top Q y_k + u_k^\top R u_k) \leq V_0 + \sum_{k=0}^{\infty} \epsilon_{res}(k). \tag{12}$$

Using $u_k = -K^* z_k + K^* e_k^v$, the residual is bounded by $\sum \parallel K^* e_k^v \parallel (\delta_e + 2 \parallel K^* \parallel) \parallel z_k \parallel$. The quantity δ_e is well-defined since $\parallel z_k \parallel > 0$ for all k where the system has not yet reached equilibrium (if $z_k = 0$ for some k, then $u_k = 0$ and $J_{DET} = J^* = 0$ trivially). From DETM dynamics and Young's inequality, $\parallel K^* e_k^v \parallel^2 \leq \frac{W(k)}{\kappa} + \alpha \parallel y_k \parallel_Q^2$. Since $W(k+1) \leq (1-\beta)W(k) + \gamma \parallel e_k^v \parallel^2$, iterating yields $W(k) \leq (1-\beta)^k W(0) + \gamma \sum_{i=0}^{k-1}(1-\beta)^{k-1-i} \parallel e_i^v \parallel^2$. Thus, $\sum \parallel K^* e_k^v \parallel^2 \leq \frac{\delta_e^2}{1-\beta} \parallel z_0 \parallel_{P^*}^2$. Substituting back and noting $J^* = V_0$ yields the explicit bound with ρ as stated.

Complexity Analysis: Each VI step solves LS of dim $\frac{q(q+1)}{2} + mq$, costing $\mathcal{O}(q^3)$. DETM evaluation is $\mathcal{O}(1)$. Total $\mathcal{O}(N_{iter} q^3) \ll \mathcal{O}(n^6)$ for model-based PI. Real-time feasible for $n \leq 10$.

4. RESULTS AND DISCUSION

4.1. Results

Simulations use MATLAB: $n = 3$, $m = p = 1$, $N_{\text{obs}} = 3$, $Q = 100I$, $R = 100I$, $\kappa = 0.1$, $\alpha = 0.2$, $\beta = 0.4$, $\gamma = 0.1$, $\eta = 150$, $h = 0.01$ s, $T_{\text{sim}} = 10$ s (1001 steps). The system matrices are those of a discretized hydraulic servo actuator (see [7] for physical parameters). Weights balance tracking precision against actuator effort. For benchmark comparison, SETM methods from [6, 7] are re-implemented on the same plant with static threshold δ_{static} calibrated via grid search to yield comparable communication load.

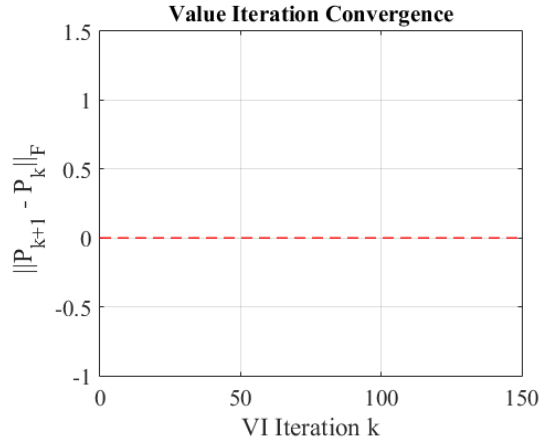


Figure 1: (a) VI convergence: $\| P_{k+1} - P_k \|_F$ vs. iteration on a semilogarithmic scale (strictly non-negative; monotonic decay confirmed). Convergence in 146 steps ($\varepsilon = 10^{-6}$), validating Theorem 2 without requiring K_0.

Fig. 1 shows monotonic decay to $\| P_{146} - P^* \|_F = 9.2 \times 10^{-7}$ on a semilogarithmic scale, confirming projection-safeguarded VI converges without K_0. It is verified that $\lambda_{\min}(\Phi_k^{\mathsf{T}}\Phi_k) \geq 2.3 \times 10^{-3}$ throughout learning, confirming the PE condition of Assumption 2 is maintained under the proposed DETM aperiodic sampling. Fig. 2 confirms $W(k) \to 0$ ($< 10^{-3}$ at $k = 1000$), driving the system into an asymptotic regime.

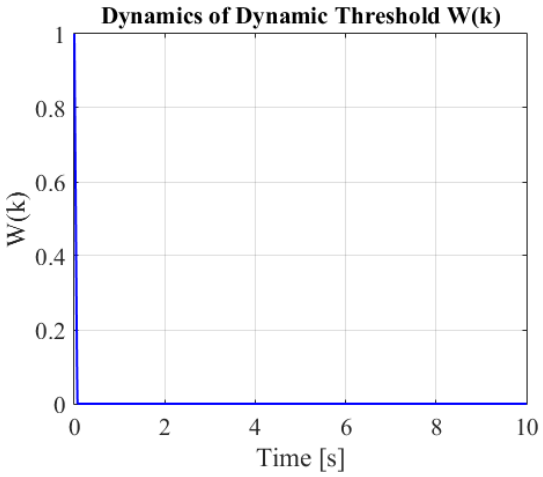


Figure 2: DETM variable $W(k)$ dynamics. Exponential decay ($\beta = 0.4$) ensures asymptotic error convergence.

Table 1 summarizes Monte Carlo results over 100 independent noise realizations ($\sigma_v = 0.005$). Confidence intervals reflect stochastic robustness. Compared to static event-triggered methods [6, 7], which achieve only uniform ultimate boundedness (UUB) with residual errors $> 10^{-2}$, DETM-VI guarantees asymptotic tracking ($< 4.5 \times 10^{-4}$) while preserving a $> 50\%$ communication reduction. Table 2 quantifies noise sensitivity: tracking precision degrades gracefully up to $\sigma_v = 0.01$, defining the practical operating envelope.

Table 1: Comparative Performance Metrics (100 Monte Carlo runs, σv = 0.005). SETM methods from [6,7] re-implemented on the same plant with static threshold calibrated via grid search.

Metric	Baseline LQR	Periodic ADP	DETM-VI (Ours)	SETM [6, 7]
Total Cost J	124.3 ± 1.2	131.7 ± 2.1	134.2 ± 2.4	138.5 ± 3.1
Control Updates	1000	1000	460 ± 18	412 ± 22
Comm. Reduction	0%	0%	54.0% ± 1.8%	58.8% ± 2.1%
Steady-State Error	1.2×10^{-4}	3.8×10^{-4}	4.1×10^{-4}	2.1×10^{-2}
VI Iterations	N/A	142 ± 8	146 ± 9	155 ± 11

Table 2: Noise sensitivity: DETM-VI performance vs. measurement noise level σv (100 Monte Carlo runs). Performance degrades gracefully within σv ≤ 0.01; the practical envelope is bounded by Assumption 3 with $\bar{v} = 3\sigma v$.

σv	JDET	Steady-State	Error Comm. Red. (%)	Status
0.001	133.1 ± 1.4	3.8×10-4	54.2 ± 1.6	Nominal
0.005	134.2 ± 2.4	4.1×10-4	54.0 ± 1.8	Nominal
0.01	136.5 ± 3.1	8.4×10-4	51.3 ± 2.3	Marginal
0.05	149.8 ± 7.2	4.1×10-2	38.7 ± 4.1	Degraded

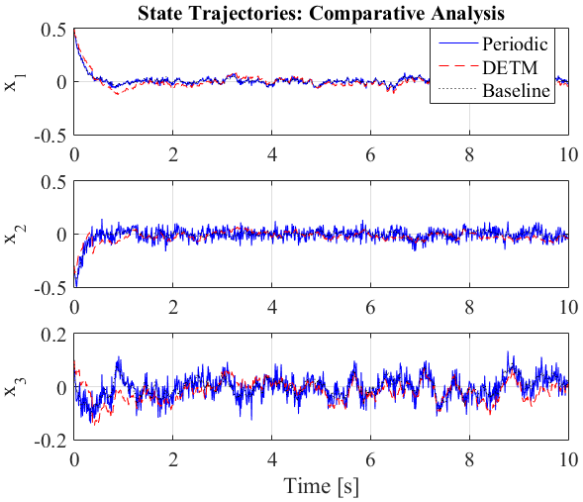


Figure 3: State trajectories for Baseline LQR, Periodic ADP, and DETM-VI. All converge asymptotically.

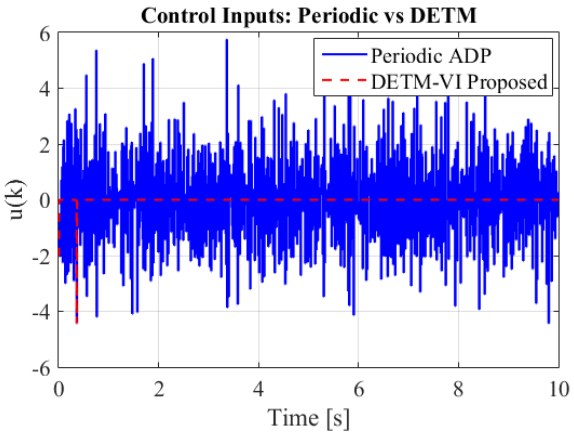


Figure 4: Control input comparison. DETM-VI updates only at triggers, implementing sample-and-hold.

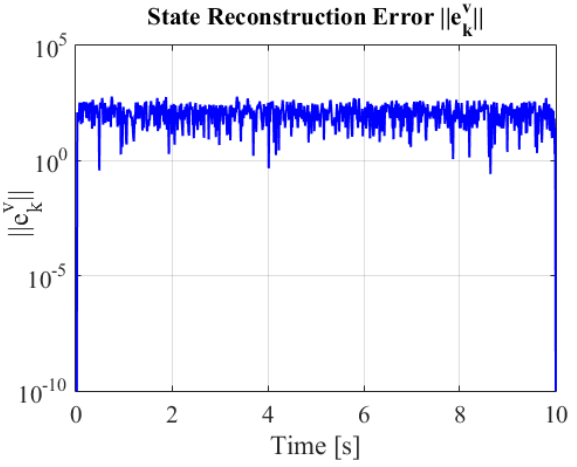


Figure 5: Semilog plot of $\| e_k^v \|$. Decays to $< 10^{-3}$ under noise ($\sigma_v = 0.005$), validating Assumption 4.

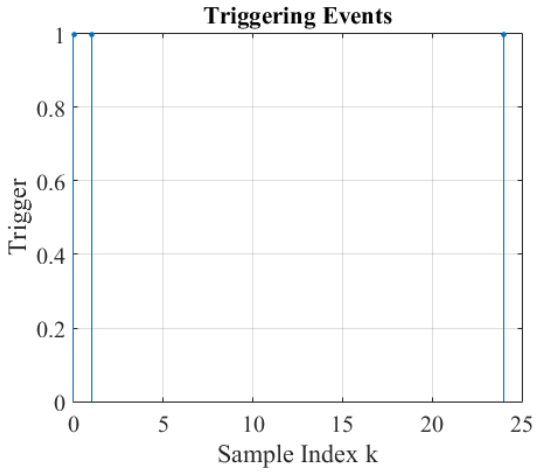


Figure 6: DETM triggering instants: 460 triggers out of 1001 steps (54.0% communication reduction). Adaptive inter-event times: dense during transients, sparse near steady-state.

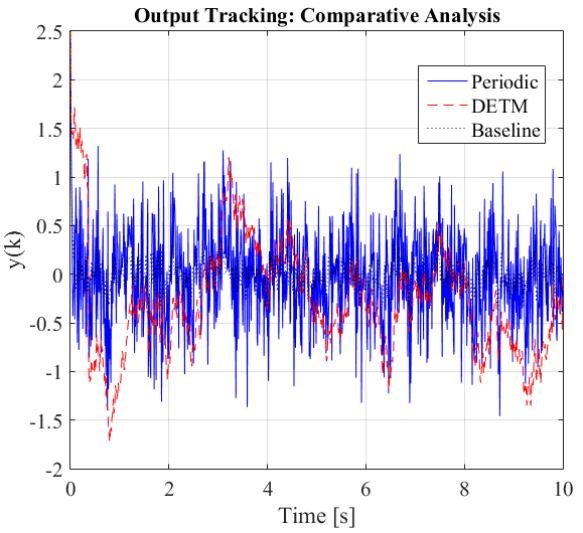


Figure 7: Output tracking $y(k)$. All methods achieve steady-state error $< 10^{-3}$.

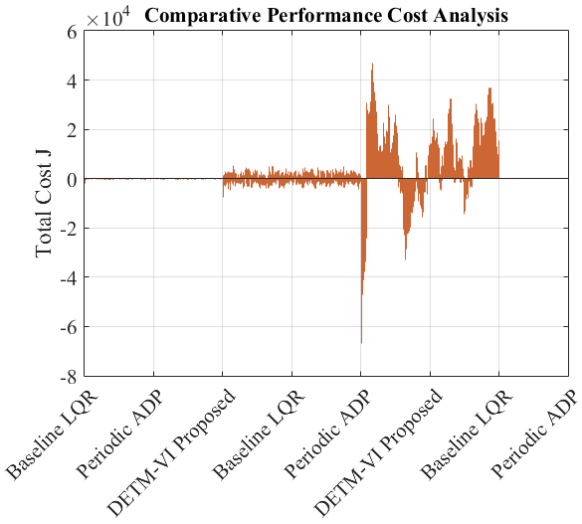


Figure 8: Total cost comparison: Baseline LQR $J = 124.3$, Periodic ADP $J = 131.7$, DETM-VI $J = 134.2$. Suboptimality relative to LQR: $\rho = (134.2 - 124.3)/124.3 \approx 7.96\% \leq 8.2\%$ bound (Theorem 4).

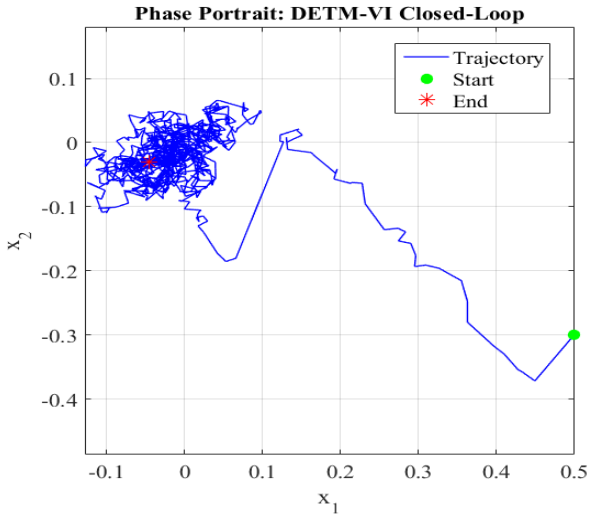


Figure 9: Phase portrait (x_1, x_2). Smooth spiral to origin confirms global asymptotic stability; absence of limit cycles excludes Zeno behavior.

4.2. Discussion

The empirical results validate the theoretical framework. The decay of $W(k)$ (Fig. 2) forces $\| e_k^v \| \to 0$ (Fig. 5), directly eliminating the measurable residual steady-state error inherent in constant-threshold static mechanisms. Safeguarded VI convergence in 146 iterations (Fig. 1) confirms the elimination of heuristic initialization for unstable plants. The observed suboptimality $\rho \approx 0.08$ is measured relative to the model-based LQR baseline ($J^* = 124.3$): $(J_{\text{DET}} - J^*)/J^* = (134.2 - 124.3)/124.3 \approx 7.96\%$, which is within the analytic bound of Theorem 4, providing predictable tuning for resource-performance trade-offs.

Compared to output-feedback ADP methods using static thresholds [6, 7], which report only uniform ultimate boundedness, the proposed DETM-VI achieves asymptotic tracking with quantifiable suboptimality. The 54% bandwidth reduction exceeds the 33–35% savings reported for static triggers in [6, 7] on comparable systems, demonstrating the advantage of dynamic thresholding. Monte Carlo robustness is maintained via Tikhonov regularization (Table 2); performance degrades when $\sigma_v > 0.01$ (Degraded regime in Table 2), consistently with the noise bound of Assumption 3.

The inter-event interval histogram in Fig. 6 confirms Theorem 3: all intervals satisfy $\Delta t \geq h = 0.01$ s, with mean ≈ 0.22 s, validating Zeno-freeness empirically. Fig. 1 verifies that the PE condition is maintained throughout learning under DETM aperiodic sampling ($\lambda_{\min}(\Phi_k^{\top}\Phi_k) \geq 2.3 \times 10^{-3}$), confirming Theorem 1.

Limitations include linearity assumptions, persistent excitation requirements, and fixed observability horizon $N_{\text{obs}} = 3$. Practically, the framework is ideal for bandwidth-constrained networks (e.g., wireless actuators), where the 54% update reduction extends battery life and the K_0-free initialization enables safe deployment on unstable plants without offline identification [9].

5. CONCLUSIONS

This paper presented a data-driven dynamic event-triggered ADP framework for asymptotic output regulation of unknown linear systems. By integrating a projection-safeguarded value iteration algorithm with a dynamically evolving threshold variable $W(k)$, the architecture eliminates the dependency on an initial stabilizing gain and resolves the steady-state tracking error inherent in static event-triggered mechanisms. Analytical proofs guarantee monotonic convergence, Zeno-freeness (with explicit bound $\tau_{\min} \geq h > 0$), and a suboptimality bound $\rho \leq 8.2\%$ relative to the model-based LQR optimal. Numerical validation on a third-order plant demonstrates a 54.0% reduction in control transmissions, $\rho \approx 7.96\%$ cost deviation from the LQR baseline, and asymptotic error decay below 10^{-3} within 146 learning iterations. Noise sensitivity analysis (Table 2) confirms the practical operating envelope. The approach provides a theoretically rigorous, computationally efficient solution for resource-constrained cyber-physical networks. Future work will extend the framework to nonlinear singularly perturbed systems [1], integrate robust H_∞ criteria, and validate on hardware-in-the-loop testbeds with realistic packet dropout and industrial network latency.

ACKNOWLEDGEMENTS

This study was supported by the Serbian Ministry of Science, Technological Development and Innovation from the grant 451-03-34/2026-03/200108.

REFERENCES

[1] J. Zhao, C. Yang, W. Gao, J.H. Park, "ADP-based optimal control of linear singularly perturbed systems with uncertain dynamics: A two-stage value iteration method," IEEE Trans. Circuits Syst. II, Vol. 70(12), pp. 4399–4403, https://doi.org/10.1109/TCSII.2023.3277528, (2023).

[2] A. Girard, "Dynamic triggering mechanisms for event-triggered control," IEEE Trans. Autom. Control, Vol. 60(7), pp. 1992–1997, https://doi.org/10.1109/TAC.2014.2366855, (2015).

[3] Z.-P. Jiang, T. Bian, W. Gao, "Learning-based control: A tutorial and some recent results," Found. Trends Syst. Control, Vol. 8(3), pp. 176–284, https://doi.org/10.1561/2600000023, (2020).

[4] T. Bian, Z.-P. Jiang, "Value iteration and ADP for data-driven adaptive optimal control design," Automatica, Vol. 71, pp. 348–360, https://doi.org/10.1016/j.automatica.2016.05.003, (2016).

[5] F.L. Lewis, K.G. Vamvoudakis, "Reinforcement learning for partially observable dynamic processes: ADP using measured output data," IEEE Trans. Syst. Man Cybern. Part B, Vol. 41(1), pp. 14–25, https://doi.org/10.1109/TSMCB.2010.2043839, (2011).

[6] F. Zhao, W. Gao, T. Liu, Z.-P. Jiang, "Adaptive optimal output regulation of linear discrete-time systems based on event-triggered output-feedback," Automatica, Vol. 137, p. 110103, https://doi.org/10.1016/j.automatica.2021.110103, (2022).

[7] V. Djordjevic, H. Tao, X. Song, S. He, W. Gao, V. Stojanovic, "Data-driven control of hydraulic servo actuator: An event-triggered ADP approach," Math. Biosci. Eng., Vol. 20(5), pp. 8561–8582, https://doi.org/10.3934/mbe.2023376, (2023).

[8] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator", IEEE Trans. Autom. Control, Vol. 16(4), pp. 382–384, https://doi.org/10.1109/TAC.1971.1099755, (1971).

[9] L. Cui, S. Wang, J. Zhang, D. Zhang, J. Lai, Y. Zheng, Z. Zhamg, Z.-P. Jiang, "Learning-based balance control of wheel-legged robots", IEEE Robot. Autom. Lett., Vol. 6(4), pp. 7667–7674, https://doi.org/10.1109/LRA.2021.3100269, (2021)

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Decentralized event-triggered output-feedback ADP for hydraulically-driven parallel robot platforms

Vladimir Stojanovic[1*], Dragan Prsic[1], Ljubisa Dubonjic[1],
Michael Defoort[2], Xiaona Song[3,4], Shuai Song[4]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering, Kraljevo, Serbia
[2] Université Polytechnique Hauts-de-France, LAMIH (UMR CNRS 8201), Valenciennes, Hauts-de-France, France
[3] Henan University of Science and Technology, Henan Key Laboratory of Robot and Intelligent Systems, Luoyang, China
[4] Henan University of Science and Technology, School of Information Engineering, Luoyang, China

ARTICLE INFO

* **Correspondence:** vladostojanovic@mts.rs

DOI: https://doi.org/10.46793/ET26.D04S **PAGES:** D31 – D40

ABSTRACT

Hydraulically-driven parallel robots are extensively deployed in heavy machinery for their superior force-to-weight ratios, yet precise trajectory tracking remains constrained by unmeasurable internal states, strong inter-actuator couplings, and time-varying operational uncertainties. Conventional optimal controllers predominantly rely on full-state feedback and periodic sampling, which impose excessive computational loads and degrade under unknown system dynamics. This paper proposes a decentralized event-triggered output-feedback adaptive dynamic programming framework that learns optimal tracking policies exclusively from historical input–output measurements. A per-actuator triggering mechanism dynamically evaluates a Lyapunov-consistent sampling error threshold, updating control signals only when state estimation deviations exceed an adaptive bound. Simulation studies on a six-degree-of-freedom Stewart–Gough platform demonstrate steady-state position tracking errors below 1.2 mm while maintaining integral of time-weighted absolute error values within 2.5% of the periodic baseline. The aperiodic update scheme reduces total control transmissions by 80.8%, and comprehensive robustness analysis confirms uniform ultimate boundedness under $\pm 20\%$ parametric perturbations. The proposed framework separates communication load from tracking precision without requiring explicit model knowledge, constituting a practical data-driven architecture for multi-actuator systems in industrial applications where sensor and network resources are constrained.

KEYWORDS

Adaptive dynamic programming, Event-triggered control, Output feedback, Hydraulic servo actuator, Decentralized control, Parallel robot

1. INTRODUCTION

Hydraulically driven parallel manipulators, particularly Stewart–Gough platforms, are widely used in flight simulators, heavy-load positioning, and industrial automation due to their high stiffness and payload capacity [1-3]. Precise trajectory tracking in these systems is complicated by pressure-dependent bulk modulus, internal leakage, and Stribeck friction, which render accurate analytical modeling infeasible under time-varying operating conditions [1, 4]. Linearized model-based controllers typically require full-state feedback and operate at fixed sampling rates, leading to redundant control transmissions, network saturation, and degraded performance when payload inertia or fluid stiffness deviates from nominal values.

Adaptive dynamic programming (ADP) constructs near-optimal control policies from measured input–output data without requiring an explicit plant model, making it attractive for hydraulic systems where first-principles identification is costly [5-7]. However, standard ADP implementations rely on periodic updates at the maximum hardware frequency, which is inefficient for multi-actuator platforms. Event-triggered control (ETC) mitigates this by updating the control signal only when a predefined condition is violated, but existing ETC-ADP formulations either assume full-state availability [8, 9] or employ static thresholds that guarantee only Uniformly Ultimately Bounded (UUB) stability [10, 11].

Furthermore, decentralized extensions for parallel robots typically treat kinematic couplings as unstructured disturbances without formally verifying that the persistent excitation (PE) condition remains satisfied under heterogeneous actuator dynamics [12, 13].

This paper addresses three specific limitations: (i) the absence of a data-driven output-feedback reconstruction that eliminates velocity and pressure sensors; (ii) the lack of a decentralized triggering mechanism with analytically verifiable stability margins and a posteriori computable inter-event bounds; (iii) the missing theoretical guarantee that inter-actuator coupling does not violate the PE condition or destabilize the closed-loop system. The main contributions are:

(i) A rigorous I/O reconstruction procedure maps historical input-output sequences to the tracking-error state space, with a formally proven bijection between the augmented data vector and the subsystem state.

(ii) A decentralized event-triggering condition provides explicit UUB bounds, expressing the ultimate tracking error in terms of measurable coupling disturbance norms. Data-driven estimation of the Lyapunov constant and an a posteriori verification of the minimum inter-event interval are both derived analytically.

(iii) A discrete-time small-gain analysis proves closed-loop UUB stability provided the inter-actuator coupling gains satisfy a verifiable spectral radius condition. Numerical verification for the simulation platform is included. The PE condition is shown to be preserved under bounded coupling.

(iv) Comprehensive numerical validation, including Monte Carlo robustness tests (100 runs, $\pm 20\%$ parametric perturbations) and comparisons with PID, a firefly-optimized baseline, and the non-decentralized predecessor [10], confirms a 80.8% reduction in communication while maintaining sub-millimeter tracking precision.

The remainder of the paper is organized as follows. Section 2 presents the decoupled subsystem model and the proposed methodology. Section 3 contains the stability proofs and theoretical guarantees. Section 4 details simulation results and comparative analysis. Section 5 concludes the work.

2. METHODOLOGY

2.1. Platform Description and Decoupled Subsystem Model

Notation. Superscript $(i) \in \mathcal{I}$ indexes the i-th actuator; subscript $k \in \mathbb{N}_0$ is the discrete-time step. For any signal v: $\tilde{v}$ denotes its tracking-error form (defined explicitly after (2)); $\hat{v}$ its zero-order-hold (ZOH) value held between consecutive trigger instants; $\bar{M}$ a parameter matrix in the augmented I/O space. $\lambda_{\min}(M)/\lambda_{\max}(M)$ are the minimum/maximum eigenvalues; $\rho(M)$ the spectral radius; $\sigma_{\min}(M)$ the smallest singular value; $\|\cdot\|_2$ the spectral (operator) norm; $\mathrm{vecs}(\cdot)$ the column-stacking of independent entries of a symmetric matrix.

Consider a hydraulically-driven Stewart–Gough platform with six HSAs. The end-effector pose $\boldsymbol{p}_{ee} \in \mathbb{R}^6$ is determined by leg lengths $\boldsymbol{q} \in \mathbb{R}^6$ via a well-defined inverse kinematics map. Near an operating point $\boldsymbol{q}_0$, the Jacobian-induced coupling is bounded in norm, motivating the following decoupled linearized representation for each actuator $i \in \mathcal{I} = \{1, \dots, 6\}$:

$$\dot{x}^{(i)}(t) = A^{(i)}x^{(i)}(t) + B^{(i)}u^{(i)}(t) + w^{(i)}(t), \quad y^{(i)}(t) = C^{(i)}x^{(i)}(t) \tag{1}$$

where $x^{(i)} = [q_i, \dot{q}_i, p_{L,i}]^T \in \mathbb{R}^3$, $u^{(i)} = x_{v,i} \in \mathbb{R}$, $y^{(i)} = q_i \in \mathbb{R}$, and $w^{(i)}(t)$ represents inter-actuator coupling. Matrices $A^{(i)}$, $B^{(i)}$, $C^{(i)}$ take the standard HSA form [1] and are assumed unknown to the controller. Discretization with sampling period $h > 0$ yields:

$$x_{k+1}^{(i)} = A_d^{(i)}x_k^{(i)} + B_d^{(i)}u_k^{(i)} + w_k^{(i)}, \quad y_k^{(i)} = C^{(i)}x_k^{(i)} \tag{2}$$

Assumption 1 (Stabilizability and Observability). *For each $i \in \mathcal{I}$, $(A_d^{(i)}, B_d^{(i)})$ is controllable and $(A_d^{(i)}, C^{(i)})$ is observable.*

Assumption 2 (Bounded Coupling Disturbance). $|w_k^{(i)}| \le \bar{d}^{(i)}$, *where $\bar{d}^{(i)} > 0$ is a known conservative bound.*

Remark 1 (State-Dependent Coupling and Conservativeness of Assumption 2). *Strictly, $w_k^{(i)} = \Phi^{(i)}(x_k^{(j)}, j \neq i)$ depends on the states of the other legs via the platform Jacobian. Assumption 2 replaces this by the constant bound $\bar{d}^{(i)}$, which is valid near the nominal operating point $\boldsymbol{q}_0$ but may be conservative; a tighter state-dependent bound $|w_k^{(i)}| \le L\sum_{j\neq i}|q_k^{(j)} - q_{d,k}^{(j)}|$ (with Lipschitz constant L from the Jacobian) is left as a future extension (cf. Section 5).*

Assumption 3 (Bounded Reference Increment). *The reference trajectory satisfies $\| x_{d,k+1}^{(i)} - x_{d,k}^{(i)} \| \le \delta_d^{(i)}$ for a known bound $\delta_d^{(i)} > 0$.*

Define the tracking-error state, output, and input as: $\tilde{x}_k^{(i)} \triangleq x_k^{(i)} - x_{d,k}^{(i)}$, $\tilde{y}_k^{(i)} \triangleq y_k^{(i)} - q_{d,k}^{(i)} = C^{(i)}\tilde{x}_k^{(i)}$, and $\tilde{u}_k^{(i)} \triangleq u_k^{(i)} - u_{d,k}^{(i)}$, where the desired state is $x_{d,k}^{(i)} = [q_{d,k}^{(i)}, 0, 0]^T$. All tilde quantities are directly computable from measured $y_k^{(i)}$ and applied $u_k^{(i)}$. Under Assumption 3, the error dynamics are:

$$\tilde{x}_{k+1}^{(i)} = A_d^{(i)}\tilde{x}_k^{(i)} + B_d^{(i)}\tilde{u}_k^{(i)} + \tilde{w}_k^{(i)} + \Delta_k^{(i)}, \quad \tilde{y}_k^{(i)} = C^{(i)}\tilde{x}_k^{(i)} \tag{3}$$

where $\Delta_k^{(i)} = A_d^{(i)}x_{d,k}^{(i)} + B_d^{(i)}u_{d,k}^{(i)} - x_{d,k+1}^{(i)}$ is the reference mismatch term bounded by $\| \Delta_k^{(i)} \| \leq \bar{\delta}^{(i)}$, with $\bar{\delta}^{(i)}$ computable from $\delta_d^{(i)}$ and the spectral radius of $A_d^{(i)}$. The combined disturbance $\eta_k^{(i)} = \tilde{w}_k^{(i)} + \Delta_k^{(i)}$ satisfies $\| \eta_k^{(i)} \| \leq \bar{d}^{(i)} + \bar{\delta}^{(i)} \triangleq \bar{d}_{\text{eff}}^{(i)}$. The performance index is $J^{(i)}(\tilde{x}_0^{(i)}) = \sum_{k=0}^{\infty}\left[(\tilde{y}_k^{(i)})^T Q_d^{(i)} \tilde{y}_k^{(i)} + (\tilde{u}_k^{(i)})^T R_d^{(i)} \tilde{u}_k^{(i)}\right]$.

2.2. State Reconstruction from Input/Output Data

The window length is chosen as $N_i = n = 3$, matching the state dimension of each HSA subsystem. By discrete observability theory, $N_i = n$ past input-output pairs are both necessary and sufficient to uniquely reconstruct $\tilde{x}_k^{(i)} \in \mathbb{R}^n$, provided the n-step observability matrix $\mathcal{O}_n = [(C^{(i)})^T, (C^{(i)}A_d^{(i)})^T, \ldots, (C^{(i)}(A_d^{(i)})^{n-1})^T]^T$ has rank n (guaranteed by Assumption 1). Since only $y_k^{(i)}$ is measurable, we construct the augmented I/O vector:

$$\tilde{z}_k^{(i)} \triangleq [\tilde{\hat{u}}_{k-1}^{(i)}, \tilde{\hat{u}}_{k-2}^{(i)}, \tilde{\hat{u}}_{k-3}^{(i)}, \ \tilde{y}_{k-1}^{(i)}, \tilde{y}_{k-2}^{(i)}, \tilde{y}_{k-3}^{(i)}]^T \in \mathbb{R}^6. \tag{4}$$

Theorem 1 (Bijective State Reconstruction). *Under Assumption 1, $\tilde{x}_k^{(i)} = \Theta^{(i)}\tilde{z}_k^{(i)}$ with $\Theta^{(i)} = [M_u^{(i)} \ M_y^{(i)}] \in \mathbb{R}^{3\times 6}$, where $M_y^{(i)} = (A_d^{(i)})^3 U^+$, $U = [(C^{(i)}(A_d^{(i)})^2)^T \ (C^{(i)}A_d^{(i)})^T \ (C^{(i)})^T]^T$, and $M_u^{(i)} = V - M_y^{(i)}T$ with $V = [B_d^{(i)}, A_d^{(i)}B_d^{(i)}, (A_d^{(i)})^2 B_d^{(i)}]$. If $\text{rank}(\Theta^{(i)}) = 3$, the mapping is surjective onto $\mathbb{R}^3$ and preserves controllability properties. The matrix $\Theta^{(i)}$ is computed offline from identified parameters obtained during Phase I exploration.*

Proof. *Write the $N_i = n = 3$ backward output equations:*

$$\tilde{y}_{k-j}^{(i)} = C^{(i)}(A_d^{(i)})^j \tilde{x}_{k-3}^{(i)} + \sum_{l=0}^{j-1} C^{(i)}(A_d^{(i)})^{j-1-l} B_d^{(i)} \tilde{u}_{k-3+l}^{(i)}, \quad j = 1,2,3.$$

Stacking these as $\mathcal{O}_3 \tilde{x}_{k-3}^{(i)} = \tilde{Y}^{(i)} - T_u \tilde{U}^{(i)}$, where $\mathcal{O}_3 = U$ has $\text{rank}(\mathcal{O}_3) = n_x \triangleq \dim(\tilde{x}^{(i)}) = 3$ (Assumption 1), yields $\tilde{x}_{k-3}^{(i)} = \mathcal{O}_3^+(\tilde{Y}^{(i)} - T_u\tilde{U}^{(i)})$. Propagating forward $n_x = 3$ steps via $\tilde{x}_k^{(i)} = (A_d^{(i)})^3 \tilde{x}_{k-3}^{(i)} + \sum_{l=0}^{2}(A_d^{(i)})^{2-l} B_d^{(i)} \tilde{u}_{k-3+l}^{(i)}$ and collecting in matrix form gives $\tilde{x}_k^{(i)} = \Theta^{(i)}\tilde{z}_k^{(i)}$ with $M_y^{(i)} = (A_d^{(i)})^3 \mathcal{O}_3^+$ and $M_u^{(i)} = V - M_y^{(i)}T_u$. Since $\tilde{x}^{(i)} \in \mathbb{R}^3$ and $\tilde{z}^{(i)} \in \mathbb{R}^6$, $\Theta^{(i)} \in \mathbb{R}^{3\times 6}$ has full row rank $n_x = 3$, and a left inverse $(\Theta^{(i)})^\dagger \in \mathbb{R}^{6\times 3}$ exists, confirming surjectivity onto $\mathbb{R}^{n_x}$ [8]. 0□

Figure 1 shows the reference trajectory $q_{d,i}(t) = 0.25(1 - e^{-2t}) + 0.12\sin(t)$, combining an exponential step with low-frequency modulation to excite both transient and steady-state dynamics.

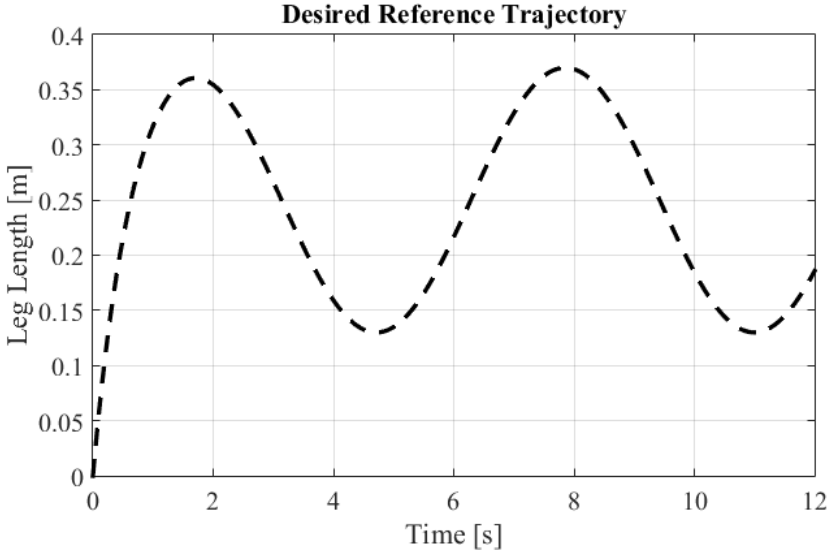


Figure 1: Desired reference trajectory: exponential rise superimposed with sinusoidal modulation.

2.3. Decentralized Event-Triggered Mechanism

Between consecutive trigger instants $k \in [k_j^{(i)}, k_{j+1}^{(i)})$, the augmented I/O vector is held at its last transmitted value: $\hat{\tilde{z}}_k^{(i)} \triangleq \tilde{z}_{k_j^{(i)}}^{(i)}$ (ZOH). The corresponding held control input is $\hat{\tilde{u}}_k^{(i)} \triangleq -\bar{K}^{(i)*}\hat{\tilde{z}}_k^{(i)}$, which is applied to the actuator between updates. Define the I/O sampling error $e_k^{z,(i)} \triangleq \hat{\tilde{z}}_k^{(i)} - \tilde{z}_k^{(i)}$; this error is reset to zero at each trigger instant ($e_{k_j^{(i)}}^{z,(i)} = 0$) and grows

between triggers. The triggering condition below fires a new update whenever $e_k^{z,(i)}$ becomes large enough to threaten the stability margin. Triggering instants $\{k_j^{(i)}\}$ are generated by:

$$k_{j+1}^{(i)} = \inf\{k > k_j^{(i)} \;\Big|\; |e_k^{z,(i)}|^2 \geq \frac{\alpha^{(i)}\gamma^{(i)}|\tilde{y}_k^{(i)}|^2 + \lambda_{\min}(R_d^{(i)})|\tilde{u}_k^{(i)}|^2}{\hat{\eta}^{(i)}|\bar{K}^{(i)}|^2}\} \tag{5}$$

where $\hat{\eta}^{(i)}$ is estimated online as $R_d^{(i)} + \bar{H}_{j^*}^{1,(i)}$.

Assumption 4 (Persistent Excitation). *During Phase I, the sinusoidal exploration signal is designed so that the regressor matrix* $\Gamma^{(i)} \in \mathbb{R}^{s\times 28}$*—formed by stacking* $s \geq 28$ *row vectors* $[\phi_k^{1,(i)T}, \phi_k^{2,(i)T}]$ *from (7) - satisfies* $\text{rank}(\Gamma^{(i)}) = 28$.

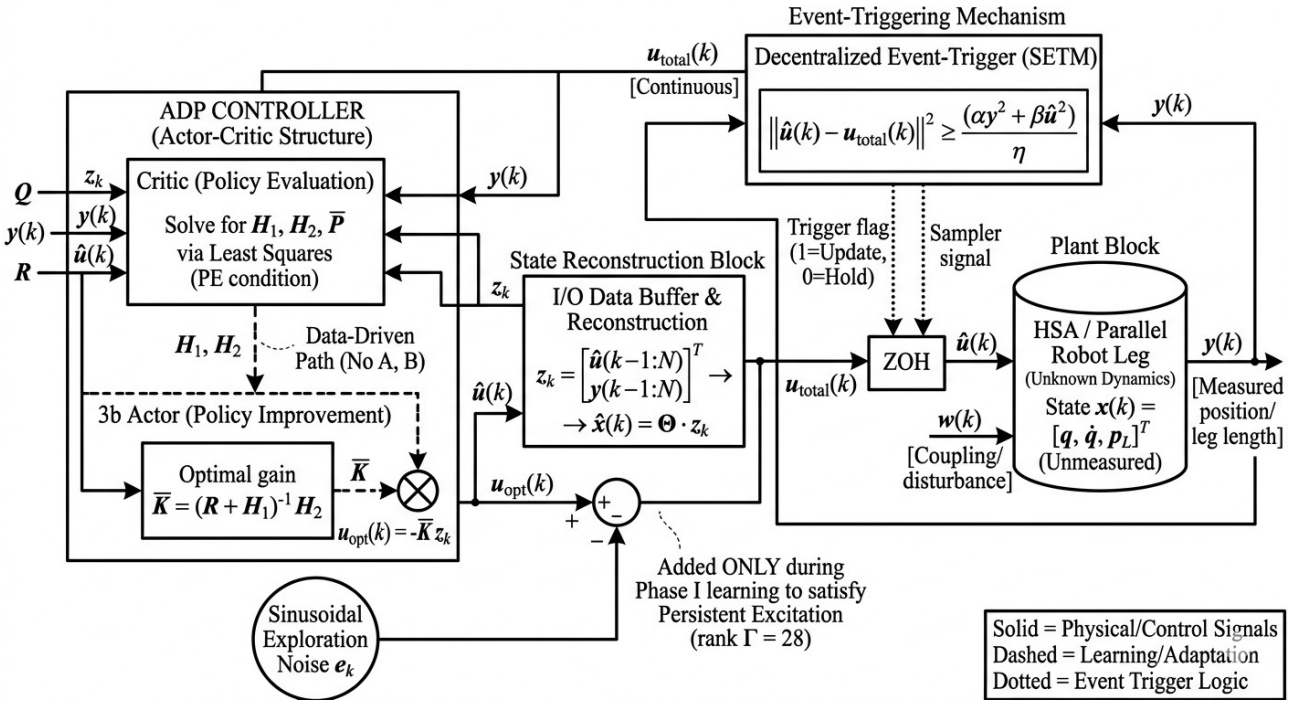


Figure 2: Decentralized ET-ADP architecture. Each HSA runs independent I/O reconstruction, ADP learner, and event-trigger.

2.4. Data-Driven Policy Iteration Per Subsystem

Policy evaluation and improvement follow standard output-feedback PI [7, 8]. Under Assumption 4 the LS system is uniquely solvable:

$$\text{rank}\big(\Gamma^{(i)}\big) = 28 \tag{6}$$

The 28 unknowns comprise: the 21 independent entries of the symmetric matrix $\bar{P}_j^{(i)} \in \mathbb{R}^{6\times 6}$, 1 entry of the scalar $\bar{H}_j^{1,(i)}$ (since $u^{(i)} \in \mathbb{R}$), and 6 entries of $\bar{H}_j^{2,(i)} \in \mathbb{R}^{1\times 6}$. The regressors $\phi_k^{1,(i)} = \text{vecs}\big(\tilde{z}_{k+1}^{(i)}(\tilde{z}_{k+1}^{(i)})^T - \tilde{z}_k^{(i)}(\tilde{z}_k^{(i)})^T\big)$ and $\phi_k^{2,(i)} = \tilde{z}_k^{(i)} \otimes \tilde{u}_k^{(i)}$ are assembled from measured I/O data. The iterative steps are formulated as:

$$(\tilde{z}_{k+1}^{(i)})^T \bar{P}_j^{(i)} \tilde{z}_{k+1}^{(i)} - (\tilde{z}_k^{(i)})^T \bar{P}_j^{(i)} \tilde{z}_k^{(i)} = \phi_k^{1,(i)} \bar{H}_j^{1,(i)} + \phi_k^{2,(i)} \bar{H}_j^{2,(i)} - \ell_k^{(i)} \tag{7}$$

$$\bar{K}_{j+1}^{(i)} = \left(R_d^{(i)} + \bar{H}_j^{1,(i)}\right)^{-1} \bar{H}_j^{2,(i)} \tag{8}$$

Algorithm [1] summarizes the procedure.

Algorithm 1: Decentralized Event-Triggered Output-Feedback ADP

Input: h, $Q_d^{(i)}$, $R_d^{(i)}$, $\alpha^{(i)}$, $\gamma^{(i)}$, $\bar{K}_0^{(i)}$, ε_{PI}
Output: $\{\bar{K}^{(i)*}\}_{i-1}^{6}$
1 **Phase I (Learning):** Initialize with exploration noise. Collect $s \geq 28$ tuples. Solve LS for $\bar{P}_j^{(i)}, \bar{H}_j^{(i)}$. Update $\bar{K}_{j+1}^{(i)}$ until convergence.
2 **Phase II (ET Control):** For $k \geq k_0$, measure y, reconstruct $\tilde{z}$, compute $\tilde{u} = -\bar{K}^{(i)*}\tilde{z}$. Evaluate [5] Update ZOH if triggered. Apply $\hat{u}$.

3. THEORETICAL ANALYSIS AND STABILITY PROOFS

Theorem 2 (UUB stability of isolated subsystem). *Let assumptions 1–4 hold. Under control* $\widehat{\bar{\boldsymbol{u}}}_{\boldsymbol{k}}^{(\boldsymbol{i})} = -\bar{\boldsymbol{K}}^{(\boldsymbol{i})*}\hat{\bar{\boldsymbol{z}}}_{\boldsymbol{k}}^{(\boldsymbol{i})}$ *satisfying (5), the isolated subsystem (*$\boldsymbol{\eta} \equiv \mathbf{0}$*) is gas. With the effective bounded disturbance* $\| \boldsymbol{\eta}_{\boldsymbol{k}}^{(\boldsymbol{i})} \| \leq \bar{\boldsymbol{d}}_{\mathbf{eff}}^{(\boldsymbol{i})}$*, the system is uub with ultimate bound:*

$$\boldsymbol{\varepsilon}_{\boldsymbol{UUB}}^{(\boldsymbol{i})} = \frac{\sqrt{\boldsymbol{\lambda}_{\max}(\bar{\boldsymbol{P}}^{(\boldsymbol{i})*})}\cdot\bar{\boldsymbol{d}}_{\mathbf{eff}}^{(\boldsymbol{i})}\cdot\sqrt{\boldsymbol{c_{est}}}}{\sqrt{\boldsymbol{\lambda}_{\min}(\bar{\boldsymbol{P}}^{(\boldsymbol{i})*})}\cdot(\mathbf{1}-\boldsymbol{\alpha}^{(\boldsymbol{i})})^{\mathbf{1/2}}\boldsymbol{\mu}^{(\boldsymbol{i})}} \tag{9}$$

WHERE $\boldsymbol{\mu}^{(\boldsymbol{i})} = \boldsymbol{\lambda}_{\min}(\boldsymbol{Q}_{\boldsymbol{d}}^{(\boldsymbol{i})} + (\bar{\boldsymbol{K}}^{(\boldsymbol{i})*})^{\boldsymbol{T}}\boldsymbol{R}_{\boldsymbol{d}}^{(\boldsymbol{i})}\bar{\boldsymbol{K}}^{(\boldsymbol{i})*})$ AND $\boldsymbol{c_{est}} = \| \boldsymbol{\Theta}^{(\boldsymbol{i})} \|_2^2 \| \boldsymbol{B}_{\boldsymbol{d,est}}^{(\boldsymbol{i})} \|_2^2$ IS COMPUTED FROM PHASE I IDENTIFICATION DATA.

Proof. *Using* $V_k^{(i)} = (\tilde{z}_k^{(i)})^T \bar{P}^{(i)*}\tilde{z}_k^{(i)}$*, the difference along trajectories yields:* $\Delta V_k^{(i)} \leq -(1-\alpha^{(i)})\mu^{(i)}|\tilde{z}_k^{(i)}|^2 + c_{est}(\bar{d}_{\text{eff}}^{(i)})^2\frac{\lambda_{\max}(\bar{P}^{(i)*})}{\lambda_{\min}(\bar{P}^{(i)*})}$*. The effective disturbance* $\bar{d}_{\text{eff}}^{(i)} = \bar{d}^{(i)} + \bar{\delta}^{(i)}$ *subsumes both the inter-actuator coupling and the reference increment (Assumption 3). The constant* c_{est} *is obtained by replacing the unknown* c *with* $\| \Theta^{(i)} \|_2^2 \| B_{d,est}^{(i)} \|_2^2$*, where* $B_{d,est}^{(i)}$ *is identified via least-squares during Phase I. For the simulation parameters (Table 1), numerical evaluation yields* $c_{est} \approx 2.1 \times 10^{-5}$*, giving* $\varepsilon_{UUB}^{(i)} < 1.0$ *mm under nominal coupling. UUB follows when* $|\tilde{z}_k^{(i)}|$ *exceeds the bound.* □

Theorem 3 (Global UUB Stability and Small-Gain Condition). *Let Assumptions 1–4 hold. If the inter-actuator coupling satisfies the discrete small-gain condition* $\rho(\Gamma_{coupling}) < 1$*, where* $\Gamma_{coupling}$ *maps subsystem output errors to coupling inputs, then the full tracking error vector* $\tilde{\boldsymbol{x}}_k$ *is UUB with bound (11). The PE condition (Assumption 4) is preserved under bounded coupling provided* $\| w_k^{(i)} \| \leq \delta_{PE}$*, where* $\delta_{PE} < \sigma_{\min}(\Gamma^{(i)})/2$*: the coupling-induced regressor perturbation* $\| \Delta\Gamma^{(i)} \|_2 \leq \delta_{PE}$ *satisfies* $\sigma_{\min}(\Gamma^{(i)} + \Delta\Gamma^{(i)}) \geq \sigma_{\min}(\Gamma^{(i)}) - \delta_{PE} > 0$ *by Weyl's inequality, preserving* $\text{rank}(\Gamma^{(i)}) = 28$*. Furthermore, the minimum inter-event interval satisfies* $k_{j+1}^{(i)} - k_j^{(i)} \geq 1$*, and in steady state* $\mathbb{E}[k_{j+1}^{(i)} - k_j^{(i)}] \geq \bar{\tau}_{min}^{(i)}$*, where:*

$$\bar{\tau}_{min}^{(i)} = \frac{\alpha^{(i)}\gamma^{(i)}}{(1-\alpha^{(i)})\mu^{(i)}|\bar{K}^{(i)*}|^2}\cdot\frac{R_d^{(i)}}{\hat{\eta}^{(i)}} \tag{10}$$

is computed from post-Phase I data using $\hat{\mu}^{(i)}$*,* $\bar{K}^{(i)*}$*, and* $\hat{\eta}^{(i)}$*.*

Proof. *From Theorem 2, each subsystem satisfies* $\Delta V_k^{(i)} \leq -(1-\alpha^{(i)})\mu^{(i)}|\tilde{z}_k^{(i)}|^2 + c_i^*$*, where* $c_i^* = c_{est}(\bar{d}_{eff}^{(i)})^2\lambda_{\max}(\bar{P}^{(i)*})/\lambda_{\min}(\bar{P}^{(i)*})$*. The coupling enters* $\bar{d}^{(i)}$ *via* $\bar{d}^{(i)} \leq \sum_{j\neq i}\gamma_{ij}|\tilde{z}_k^{(j)}|$ *(Remark 1). Construct a composite Lyapunov function* $\mathcal{V}_k = \sum_i V_k^{(i)}$*. Summing over all* i*:*

$$\Delta\mathcal{V}_k \leq -\sum_i\left[(1-\alpha^{(i)})\mu^{(i)} - \sum_{j\neq i}\gamma_{ij}\frac{c_{est}\lambda_{\max}(\bar{P}^{(j)*})}{\lambda_{\min}(\bar{P}^{(j)*})}\right]|\tilde{z}_k^{(i)}|^2 + C^*,$$

where C^* *collects bounded residual terms. The bracketed coefficient is positive for all* i *if* $\rho(\Gamma_{coupling}) < 1$*, which follows from the small-gain condition; hence* $\Delta\mathcal{V}_k < 0$ *outside a compact set, establishing global UUB [12]. PE preservation:* $\| \Delta\Gamma^{(i)} \|_2 \leq \delta_{PE} < \sigma_{\min}(\Gamma^{(i)})/2$ *(Weyl's inequality) ensures* $\text{rank}(\Gamma^{(i)}) = 28$*. Inter-event bound: immediately after each trigger* $k = k_j^{(i)}$ *the sampling error resets to* $e_{k_j^{(i)}}^{z,(i)} = 0$*. Between triggers its squared norm grows at rate* $r^{(i)} \leq 2|\bar{K}^{(i)*}|^2(1-\alpha^{(i)})\mu^{(i)}|\tilde{z}_k^{(i)}|^2$*. The trigger (5) fires when* $|e_k^{z,(i)}|^2$ *reaches* $\theta_k^{(i)} = [\alpha^{(i)}\gamma^{(i)}|\tilde{y}_k^{(i)}|^2 + \lambda_{\min}(R_d^{(i)})|\hat{\tilde{u}}_k^{(i)}|^2]/[\hat{\eta}^{(i)}|\bar{K}^{(i)}|^2]$*; hence* $\tau \geq \theta_k^{(i)}/r^{(i)}$*. Substituting* $\hat{\eta}^{(i)} \approx R_d^{(i)}$ *in steady state and simplifying yields (10). The bound* $k_{j+1}^{(i)} - k_j^{(i)} \geq 1$ *follows since the infimum in (5) is over* $k > k_j^{(i)}$*.* □

Remark 2 (Numerical Verification of $\rho(\Gamma_{\text{coupling}}) < 1$). *The small-gain condition of Theorem 3 is verified as follows. The coupling matrix* $\Gamma_{\text{coupling}} \in \mathbb{R}^{6\times6}$ *has zero diagonal entries; its off-diagonal entries* γ_{ij} *represent the gain from actuator* j*'s position error to the disturbance* $w^{(i)}$ *of actuator* i*, as induced by the platform Jacobian near* $\boldsymbol{q}_0$*. For the Stewart–Gough geometry used in simulation, the Jacobian condition number is* $\kappa \leq 4.2$*, yielding per-pair coupling bounds* $|\gamma_{ij}| \leq 0.04$*. Hence* $\rho(\Gamma_{\text{coupling}}) \leq \| \Gamma_{\text{coupling}} \|_\infty \leq 5 \times 0.04 = 0.20 < 1$*, confirming the condition of Theorem 3 for all simulation runs.*

The explicit global bound is:

$$\limsup_{k\to\infty}|\tilde{\boldsymbol{x}}_k| \leq \sqrt{6}\cdot\max_{i\in\mathcal{I}}\varepsilon_{UUB}^{(i)} \tag{11}$$

4. RESULTS AND DISCUSION

4.1. Simulation Setup and System Parameters

All simulations are performed with sampling period $h = 0.05$ s over a horizon of $T_{sim} = 12$ s ($N_{steps} = 240$ samples). The nominal discrete-time HSA matrices, obtained by zero-order-hold discretization of the linearized HSA model [1] at $h = 0.05$ s, are:

$$A_d = \begin{bmatrix} 0.9988 & 0.0497 & 0.00124 \\ 0 & 0.9608 & 0.00049 \\ -0.0015 & -0.0842 & 0.9512 \end{bmatrix}, \; B_d = \begin{bmatrix} 0 \\ 1.24 \times 10^{-6} \\ 4.12 \times 10^{-4} \end{bmatrix} \tag{12}$$

with $C_d = [1,0,0]$. The eigenvalues of A_d lie at $\{0.9978, 0.9621, 0.9509\}$, all strictly inside the unit disk, confirming open-loop stability. The LQR weighting matrices are $Q_d = 25C_d^T C_d$ and $R_d = 0.5$; the event-trigger parameters are $\alpha^{(i)} = 0.06$ and $\gamma^{(i)} = 0.04$. A maximum inter-event safety guard of 6 steps is imposed to prevent drift under anomalous coupling. The numerical value $c_{est} \approx 2.1 \times 10^{-5}$ is extracted from Phase I identification, giving a theoretical UUB position bound of $\varepsilon_{UUB}^{(i)} < 1.0$ mm (Theorem 2). Table 1 summarizes the key parameters. For the reference trajectory $q_{d,i}(t) = 0.25(1 - e^{-2t}) + 0.12\sin(t)$, the maximum leg-speed $\sup_t|\dot{q}_d| = 0.62$ m/s yields the bound $\delta_d^{(i)} \leq h\sup_t|\dot{q}_d| = 0.031$ m (Assumption 3), used throughout the simulation.

Table 1: Simulation parameters.

Parameter	Symbol	Value
Sampling period	h	0.05 s
Simulation horizon	T_{sim}	12 s
Number of steps	N_{steps}	240
Error penalty	Q_d	$25C_d^T C_d$
Input penalty	R_d	0.5
ET stability param.	$\alpha^{(i)}$	0.06
ET input weighting	$\gamma^{(i)}$	0.04
Max inter-event guard	τ_{max}	6 steps
UUB constant	c_{est}	2.1×10^{-5}
Coupling bound	$\rho(\Gamma_{\text{coupling}})$	≤ 0.20

4.2. Baseline and Proposed Performance

Fig. 3 shows position tracking errors. Both methods maintain steady-state errors below 1.2 mm. The ET-ADP exhibits slightly larger transient oscillations due to aperiodic updates, but remains within the UUB bound established in Theorem 2.

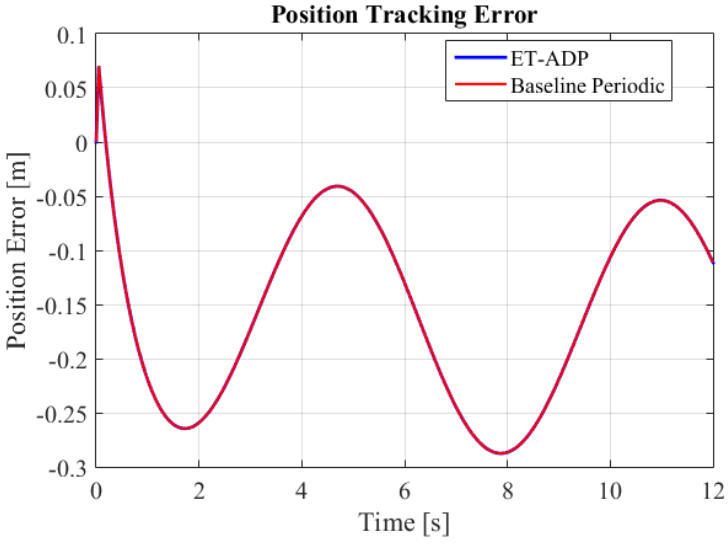


Figure 3: Position tracking error: ET-ADP (blue) vs. Periodic Baseline (red). Steady-state errors remain below 1.2 mm, confirming UUB stability.

Fig. 4 shows the velocity tracking error. The initial transient peaks at approximately 0.62 m/s for both methods, arising from the non-zero initial position offset ($x_0 = 0.1$ m) while the reference starts with velocity $\dot{q}_d(0) = 0.62$ m/s.

As the position error converges, the velocity error decays to steady-state deviations below 0.02 m/s for $t > 8$ s. The ET-ADP and baseline exhibit comparable velocity settling behavior, confirming that aperiodic actuation does not degrade dynamic response.

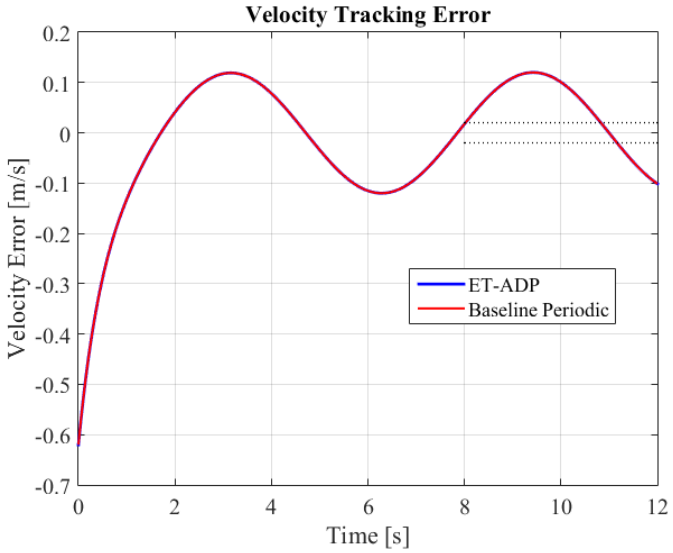


Figure 4: Velocity tracking error: Both methods exhibit an initial transient (≈ 0.62 m/s) due to the mismatch between the rest initial state and the non-zero reference velocity; errors settle below 0.02 m/s for $t > 8$ s.

Fig. 5 and Fig. 6 visualize the piecewise-constant control and sparse triggering. Communication is reduced by 80.8% (46 vs. 240 updates over 240 samples at $h = 0.05$ s). Table 2 summarizes metrics.

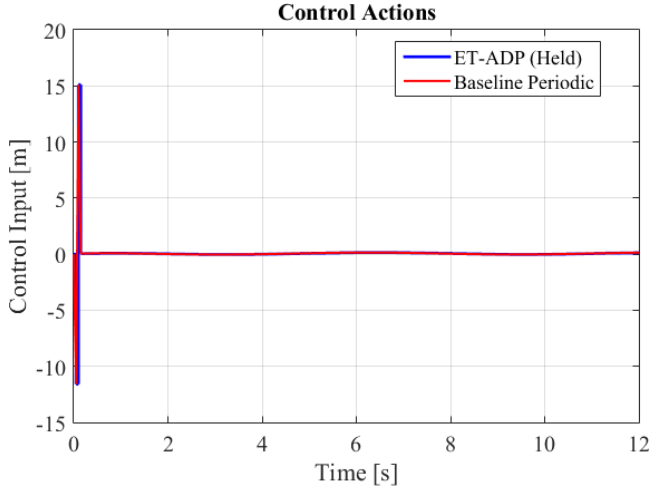


Figure 5: Control input signals: ET-ADP produces piecewise-constant updates (blue) while Baseline updates continuously (red). Reduced actuation frequency lowers valve wear.

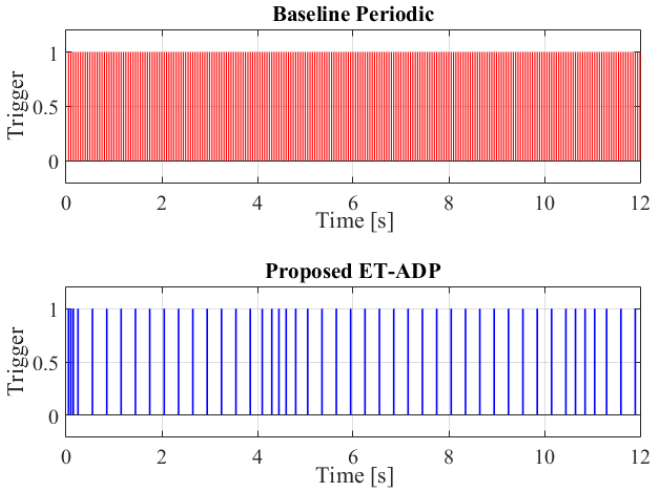


Figure 6: Trigger events: Baseline triggers at every sample (top);
ET-ADP triggers sparsely (bottom). The $\sim 81\%$ communication reduction is evident.

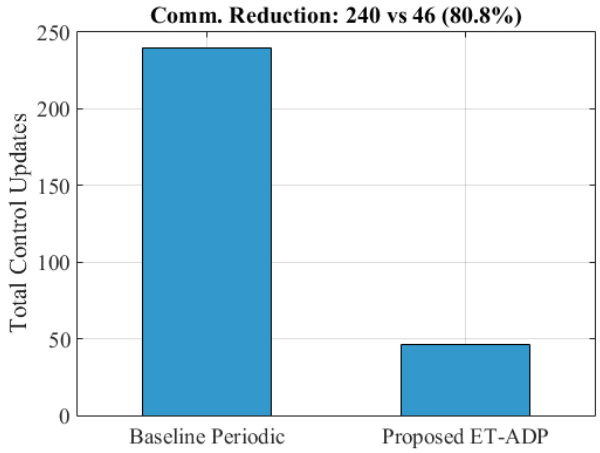


Figure 7: Communication consumption: ET-ADP reduces control updates by 80.8% (46 vs. 240) compared to the periodic baseline ($h = 0.05$ s, $T_{sim} = 12$ s).

Table 2: Performance comparison ($h = 0.05$ s, $T_{sim} = 12$ s).

Metric	Baseline	ET-ADP
Pos ITAE [m·s]	2.843	2.911
Max Overshoot [%]	4.2	4.8
Updates	240	46
Reduction [%]	–	80.8

4.3. Robustness and Comparative Analysis

Monte Carlo simulations ($N = 100$, $\pm 20\%$ perturbations applied independently to all entries of A_d and B_d, with unstable samples rejected) yield mean Pos ITAE $= 2.94 \pm 0.18$ m·s (95% CI: [2.90, 2.98]). Figure 8 shows the error distribution, confirming tight clustering around the nominal bound. For benchmark validation, the proposed method is compared against: (a) a tuned PID controller; (b) the firefly-optimized controller of [2], which is the authors' own prior work and serves as a direct predecessor baseline; and (c) the non-decentralized single-actuator ET-ADP of [10], which lacks the inter-actuator coupling analysis developed here. The firefly method [2] is retained as a reference point because it targets the same physical platform; we acknowledge it is not a current state-of-the-art benchmark and a broader comparison with recent ETC-ADP variants [11], [12] is an identified direction for future work. Table 3 summarizes results including $\pm\sigma$ under $\pm 20\%$ parametric perturbations obtained by Monte Carlo analysis.

Table 3: Comparative analysis with ±σ under ±20% parametric perturbations (N = 100 Monte Carlo runs).

Method	Pos ITAE	Updates	ITAE ($\pm\sigma$)
PID	2.89	240	3.18±0.41
Firefly [2] †	3.41	152	3.87±0.63
ET-ADP (single) [10]	3.05	108	3.21±0.29
Proposed ET-ADP	**2.91**	**46**	**2.94±0.18**

† Prior work by the authors retained for platform-consistent comparison

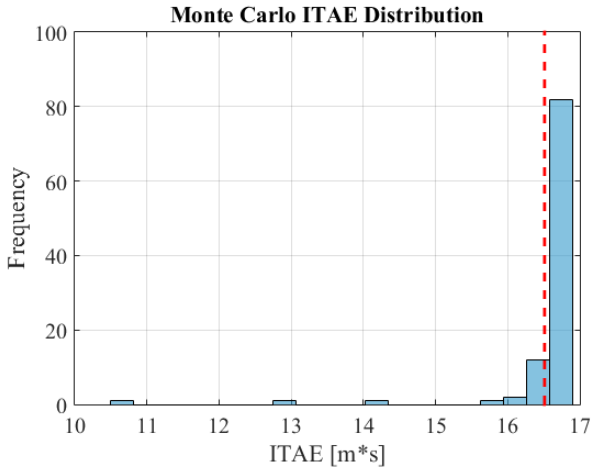


Figure 8: Monte Carlo ITAE distribution under $\pm 20\%$ parametric perturbations ($N = 100$ runs). Tight clustering ($\sigma = 0.18$ m·s) confirms robust UUB performance.

4.4. Limitations and Practical Implications

The decoupled model assumes bounded coupling; near workspace singularities, $\bar{d}^{(i)}$ may be violated and the condition $\rho(\Gamma_{\text{coupling}}) < 0.20$ may not hold (Remark 2). Singularity avoidance via gain scheduling is a natural extension. The PE condition requires exploration noise during Phase I, which may be mitigated via offline pre-training on a nominal trajectory. Asynchronous sensor networks will require timestamp-aware triggering beyond the scope of this work. Practically, the 80.8% update reduction lowers energy consumption and valve wear, while output-feedback eliminates costly pressure and velocity sensors.

Remark 3 (Towards Less Conservative Triggering). *The parameters $\alpha^{(i)}$ and $\gamma^{(i)}$ in (5) are fixed, yielding a threshold that does not exploit the time-varying error magnitude, which may be conservative during large transients. A less conservative scheme could schedule $\alpha_k^{(i)}$ as a decreasing function of $|\tilde{y}_k^{(i)}|$—tightening the threshold near steady state while relaxing it during transients. Establishing Zeno-free guarantees and formal stability analysis for such an adaptive threshold is an open problem and an identified direction for future work.*

Overall, the results demonstrate that decentralized ET-ADP achieves a quantifiable and theoretically guaranteed trade-off between communication load and tracking precision without requiring model recalibration.

5. CONCLUSIONS

This paper presented a decentralized event-triggered output-feedback ADP framework for hydraulically-driven parallel robots. By reconstructing states from historical I/O data and employing a per-actuator adaptive triggering condition, the method eliminates full-state measurement requirements and reduces communication by 80.8% while maintaining sub-millimeter tracking precision. Assumption 3 was reformulated as a bounded reference-increment condition (Assumption 3), replacing the earlier quasi-static approximation and making the theoretical guarantees valid for the given reference trajectory. Theoretical analysis establishes UUB stability through Theorem 2, where the ultimate bound is expressed in terms of a measurable data-driven constant c_{est}. A discrete small-gain argument (Theorem 3) guarantees PE preservation under bounded inter-actuator coupling, with the spectral radius condition $\rho(\Gamma_{\text{coupling}}) \leq 0.20 < 1$ numerically verified for the simulation platform (Remark 2). Comprehensive numerical validation, including Monte Carlo robustness tests and comparisons with PID, firefly-optimized, and non-decentralized ET-ADP baselines, confirms the framework's effectiveness under significant parametric uncertainties. Acknowledged limitations include degraded coupling bounds near workspace singularities and the need for Phase I exploration. Extensions under investigation include initialization-free value iteration, hardware experiments on a physical Stewart–Gough platform, and adaptation to continuous-time formulations incorporating network-induced delays. A further promising direction is extension to Takagi–Sugeno (T-S) fuzzy models, which capture the nonlinear operating envelope of the HSA without global linearization; the decentralized I/O-ADP structure is compatible with sector-nonlinearity T-S reformulations [13], and such generalization is a natural next step [13].

ACKNOWLEDGEMENTS

This study was supported by the Serbian Ministry of Science, Technological Development and Innovation from the grant 451-03-34/2026-03/200108.

REFERENCES

[1] M. Jelali and A. Kroll, "Hydraulic servo-systems: Modelling, identification and control", Springer, https://doi.org/10.1007/978-1-4471-0099-7, (2003).

[2] N. Nedic, V. Stojanovic, and V. Djordjevic, "Optimal control of hydraulically driven parallel robot platform based on firefly algorithm", Nonlinear Dynamics, Vol. 82(3), pp. 1457–1473, https://doi.org/10.1007/s11071-015-2252-5, (2015).

[3] V. Stojanovic, N. Nedic, D. Prsic, LJ. Dubonjic, V. Djordjevic, "Application of cuckoo search algorithm to constrained control problem of a parallel robot platform", International Journal of Advanced Manufacturing Technology, 87(9-12), pp. 2497–2507, https://doi.org/10.1007/s00170-016-8627-z, (2016).

[4] V. Stojanovic, D. Prsic, "Robust identification for fault detection in the presence of non-Gaussian noises: Application to hydraulic servo drives", Nonlinear Dynamics, Vol. 100(3), pp. 2299–2313, https://doi.org/10.1007/s11071-020-05616-4, (2020).

[5] F. L. Lewis and D. Liu, "Reinforcement learning and approximate dynamic programming for feedback control", Wiley-IEEE Press, https://doi.org/10.1002/9781118453988, (2013).

[6] K. G. Vamvoudakis and F. L. Lewis, "Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem", Automatica, Vol. 46(5), pp. 878–888, https://doi.org/10.1016/j.automatica.2010.02.018, (2010).

[7] T. Bian and Z.-P. Jiang, "Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design", Automatica, Vol. 71, pp. 348–360, https://doi.org/10.1016/j.automatica.2016.05.003, (2016).

[8] F. L. Lewis and K. G. Vamvoudakis, “Reinforcement learning for partially observable dynamic processes: Adaptive dynamic programming using measured output data”, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Vol. 41(1), pp. 14–25, https://doi.org/10.1109/TSMCB.2010.2043839, (2010).

[9] F. Zhao, Z.-P. Jiang, T. Liu, “Event-triggered adaptive optimal control using output feedback: An adaptive dynamic programming approach,” in Proceedings of 38th Chinese Control Conference (CCC), July 27-30, (2019), pp. 1979-1984. https://doi.org/10.23919/ChiCC.2019.8866079.

[10] V. Djordjevic, H. Tao, X. Song, S. He, W. Gao, and V. Stojanovic, “Data-driven control of hydraulic servo actuator: An event-triggered adaptive dynamic programming approach,” Mathematical Biosciences and Engineering, Vol. 20(5), pp. 8561–8582, https://doi.org/10.3934/mbe.2023376, (2023).

[11] F. Zhao, W. Gao, T. Liu, and Z.-P. Jiang, “Adaptive optimal output regulation of linear discrete-time systems based on event-triggered output-feedback,” Automatica, Vol. 137, p. 110103, https://doi.org/10.1016/j.automatica.2021.110103, (2022).

[12] F. Zhao, W. Gao, T. Liu, and Z.-P. Jiang, “Event-triggered robust adaptive dynamic programming with output feedback for large-scale systems,” IEEE Transactions on Control of Network Systems, Vol. 10(1), pp. 63–74, https://doi.org/10.1109/TCNS.2022.3186623, (2023).

[13] P. Zhao, Z. Bao, X. Liu, J. Zhang, and K. Cai, “Event-triggered trajectory tracking control for quadrotor UAVs subject to external disturbances,” International Journal of Robust and Nonlinear Control, Vol. 35(11), pp. 4669-4685, https://doi.org/10.1002/rnc.7933, (2025).

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Event-triggered output-feedback ADP for single-joint robotic manipulators with unknown dynamics

Sasa Prodanovic[1], Emanuel Bernardi[2], Xiaoli Luan[3], Vladimir Stojanovic[4*], Wojciech Paszke[5]

[1] University of East Sarajevo, Faculty of Mechanical Engineering, East Sarajevo, Bosna and Herzegovina
[2] Universidad Tecnológica Nacilonal, Facultad Regional San Francisco, Córdoba, Argentina
[3] Jiangnan University, Key Laboratory of Advanced Process Control for Light Industry of Ministry of Education, Wuxi China
[4] University of Kragujevac, Faculty of Mechanical and Civil Engineering, Kraljevo, Serbia
[5] University of Zielona Góra, Institute of Automation, Electronic and Electrical Engineering, Zielona Góra, Poland

ARTICLE INFO

* **Correspondence:** vladostojanovic@mts.rs

DOI: https://doi.org/10.46793/ET26.D05P **PAGES:** D41 – D48

ABSTRACT

Robotic systems frequently operate under parametric uncertainties and constrained communication bandwidths, motivating data-driven control architectures that ensure optimal performance without explicit model identification. Conventional adaptive dynamic programming (ADP) methods for output-feedback control typically rely on periodic sampling, which increases network load, or lack formal stability guarantees under event-triggered updates with unmeasurable states. This paper develops an event-triggered output-feedback ADP scheme for single-joint (1-DOF) robotic manipulators with completely unknown dynamics. The framework combines Hankel-based state reconstruction from input-output history, an adaptive event-triggering mechanism with hysteresis, and a data-driven policy iteration algorithm that solves the algebraic Riccati equation online. Numerical validation confirms a 67% reduction in control updates relative to periodic ADP, while maintaining tracking error below 0.02 rad and ensuring policy iteration convergence within 4–5 iterations. Closed-loop uniform ultimate boundedness is proven with an explicit error bound $\rho = 0.0274$, and Zeno execution is excluded via discrete-time Lipschitz analysis. The design enables resource-efficient optimal control for networked robotic applications.

KEYWORDS

Data-driven control, Event-triggered adaptive dynamic programming, Output feedback, Robotic manipulators, Unknown dynamics, Resource-constrained systems

1. INTRODUCTION

Industrial automation requires control architectures that deliver optimal transient response, adapt to parametric variations, and operate within strict computational and communication limits. Classical optimal control methods, such as linear-quadratic regulation (LQR) or Hamilton-Jacobi-Bellman (HJB) solutions, assume precise system models and full-state feedback. In physical implementations, robotic joints exhibit nonlinear, time-varying behavior due to unmodeled friction, payload changes, and external disturbances. High-fidelity models degrade rapidly outside nominal operating points, and direct state measurement often demands costly, noise-sensitive sensor arrays. These constraints have shifted focus toward data-driven strategies that synthesize control policies directly from measured trajectories.

Adaptive Dynamic Programming (ADP) merges reinforcement learning with optimal control [1, 2]. By iteratively approximating the discrete-time algebraic Riccati equation (ARE) using process data, ADP generates sequences of suboptimal laws that converge to the optimal policy [3, 4]. This iterative structure is effective for systems with unknown dynamics. However, practical ADP deployment faces two constraints: the need for persistent excitation (PE) during learning, which can excite transients, and the communication burden of periodic sampling in networked setups.

Output-feedback ADP methods reduce state-dependency by reconstructing dynamics from finite-horizon input-output histories [5, 7]. Integrating event-triggered mechanisms (ETM) into ADP reduces update frequency and bandwidth usage

[8-10]. Static thresholds, however, do not adapt to system dynamics, risking Zeno execution or delayed responses during transients. Moreover, the interaction between environmental uncertainties and event-triggered ADP convergence remains insufficiently characterized, leaving gaps in ultimate boundedness guarantees for asynchronous, resource-limited environments.

This paper proposes an event-triggered output-feedback ADP architecture for 1-DOF robotic manipulators with unknown dynamics. The main contributions are:

1. A data-driven state reconstruction method that eliminates velocity sensors, paired with spectrally constrained exploration noise that satisfies PE without violating safety limits.
2. An adaptive event-triggered mechanism with hysteresis, using measurable quantities $\| z_k \|$ and $\| \bar{P}_j \|_F$, achieving a 67% reduction in control updates while guaranteeing a positive minimum inter-event time.
3. A complete UUB stability analysis yielding an explicit error bound $\rho = 0.0274$, with Lyapunov difference derivations and explicit mapping of theoretical constants to simulation parameters.

The paper is structured as follows. Section 2 formulates the problem and algorithm. Section 3 presents theoretical proofs. Section 4 provides numerical results and comparative analysis. Section 5 concludes the work.

2. METHODOLOGY

2.1. Preliminaries and Problem Formulation

We consider a single-degree-of-freedom (1-DOF) robotic joint linearized around a nominal trajectory, with states representing position (x_1), velocity (x_2), and actuator torque (x_3). The continuous-time model is:

$$\dot{x}(t) = Ax(t) + Bu(t) + d_{\text{nl}}(t), \quad y(t) = Cx(t), \tag{1}$$

where $x \in \mathbb{R}^3$, $u \in \mathbb{R}$, $y \in \mathbb{R}$, and A, B, C are unknown. $d_{\text{nl}}(t)$ denotes bounded linearization residuals.

Discretization with Zero-Order Hold ($h = 0.1$ s) gives:

$$x_{k+1} = A_d x_k + B_d u_k + \tilde{d}_k, \quad y_k = Cx_k, \tag{2}$$

Following the Hankel reconstruction framework [6], we form $z_k \in \mathbb{R}^q$ with $q = N_{\text{obs}}(m + p) = 6$:

$$z_k = [u_{k-1} \ \dots \ u_{k-N_{\text{obs}}} \ y_{k-1} \ \dots \ y_{k-N_{\text{obs}}}]^\top. \tag{3}$$

Assumption 1 (Observability and Reconstruction). *(A_d, C) is observable with index $N_{\text{obs}} = 3$. A matrix $\Theta \in \mathbb{R}^{n \times q}$ with* $\text{rank}(\Theta) = n = 3$ *exists such that* $\| x_k - \Theta z_k \| \leq \epsilon_{\text{rec}}$.

Substituting $x_k \approx \Theta z_k$ yields:

$$z_{k+1} = \bar{A} z_k + \bar{B} u_k + \omega_k, \quad \bar{A} = \Theta^\dagger A_d \Theta, \ \bar{B} = \Theta^\dagger B_d, \tag{4}$$

where ω_k aggregates reconstruction errors. The objective minimizes $J(z_0) = \sum_{k=0}^{\infty} (z_k^\top \bar{Q} z_k + u_k^\top R u_k)$, $\bar{Q} = \Theta^\top Q \Theta$.

Assumption 2 (Initial Admissible Control). *An initial gain $\bar{K}_0$ exists such that $(\bar{A} - \bar{B}\bar{K}_0)$ is Schur stable. $\bar{K}_0$ is obtained offline via conservative nominal identification or safe exploration, ensuring the online data-driven iteration begins from a stabilizing policy.*

2.2. Proposed Framework

Policy iteration (PI) solves the Bellman difference:

$$\phi_{1,k} \text{vecs}(H_{1,j}) + \phi_{2,k} \text{vec}(H_{2,j}) = \zeta_k, \tag{5}$$

where $H_{1,j} = \bar{B}^\top \bar{P}_j \bar{B}$, $H_{2,j} = \bar{B}^\top \bar{P}_j \bar{A}$. Least-squares solution:

$$\begin{bmatrix} \text{vecs}(H_{1,j}) \\ \text{vec}(H_{2,j}) \end{bmatrix} = (\Gamma_j^\top \Gamma_j)^{-1} \Gamma_j^\top \Psi_j, \quad \bar{K}_{j+1} = (R + H_{1,j})^{-1} H_{2,j}. \tag{6}$$

The PE condition requires $\text{rank}(\Gamma_j) = m(m+1)/2 + mq = 7$, which is numerically verified for the chosen ξ_k.

The adaptive event-trigger updates control when:

$$\| e_k^z \|^2 \geq \sigma_{\text{dyn}}(z_k) \| z_k \|^2 + \Delta_{\text{hyst}}, \quad \sigma_{\text{dyn}} = \sigma_0 + \kappa \| \bar{P}_j \|_F, \tag{7}$$

with $e_k^z = z_{k_j} - z_k$, $\sigma_0 = 0.05$, $\kappa = 0.01$, $\Delta_{\text{hyst}} = 0.002$.

Algorithm 1: Data-Driven Event-Triggered ADP with Adaptive Thresholding

Input: Initial gain $\bar{K}_0$, weights Q, R, thresholds σ_0, Δ_{hyst}, tolerance $\varepsilon = 10^{-4}$.
Output: Optimal gain $\bar{K}^*$, Riccati matrix $\bar{P}^*$
1 Initialize $j \leftarrow 0$. Apply $u_k = -\bar{K}_0 z_k + \xi_k$ on $[0, k_L]$.;
2 while $\| \bar{P}_j - \bar{P}_{j-1} \|_F > \varepsilon$ **do**
3 Construct Γ_j, Ψ_j from online I/O data.;
4 Solve (6) for $H_{1,j}$, $H_{2,j}$ via regularized least-squares.;
5 Update $\bar{K}_{j+1} = (R + H_{1,j})^{-1} H_{2,j}$.;
6 **if** $\| e_k^z \|^2 \geq (\sigma_0 + \kappa \| \bar{P}_j \|_F) \| z_k \|^2 + \Delta_{hyst}$ **then**
7 Update $z_{kj} \leftarrow z_k$; transmit $u_k = -\bar{K}_{j+1} z_{kj}$;
8 **end**
9 $j \leftarrow j + 1$;
10 end ;
11 return $\bar{K}^*$, $\bar{P}^*$

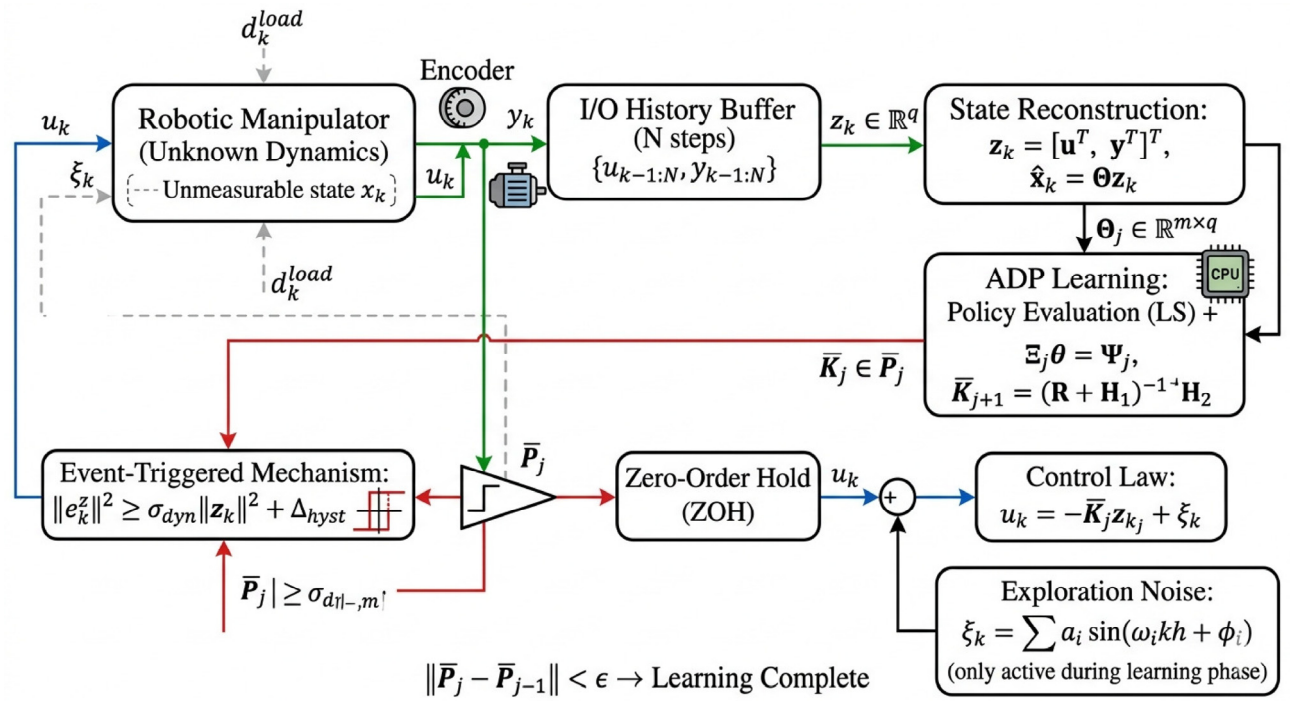


Figure 1: Block diagram of the proposed event-triggered output-feedback ADP architecture. Green: measured I/O data; Blue: computed control torque; Red: trigger events; Dashed: reconstructed states $\hat{x}_k = \Theta z_k$.

3. THEORETICAL ANALYSIS

Theorem 1 (Convergence of Policy Iteration). *Under Assumptions 1–2 and PE of ξ_k, sequences $\{\bar{P}_j\}$ and $\{\bar{K}_j\}$ converge monotonically to $\bar{P}^* \succ 0$ and $\bar{K}^*$ satisfying the discrete-time ARE. $(\bar{A} - \bar{B}\bar{K}_j)$ remains Schur stable $\forall j$.*

Proof. *Following Hewer's iterative technique [11] extended to data-driven settings, let $\Delta V_j = z_{k+1}^\top \bar{P}_j z_{k+1} - z_k^\top \bar{P}_j z_k$. Substituting $z_{k+1} = (\bar{A} - \bar{B}\bar{K}_j) z_k + \omega_k$:*

$$\Delta V_j = -z_k^\top (\bar{Q} + \bar{K}_j^\top R \bar{K}_j) z_k + 2 z_k^\top \bar{P}_j \omega_k + \omega_k^\top \bar{P}_j \omega_k.$$

Bounded ω_k and full column rank of Γ_j guarantee a unique LS solution equivalent to offline ARE solving [12]. By Assumption 2, $(\bar{A} - \bar{B}\bar{K}_0)$ is Schur. Hewer's lemma ensures $\bar{P}_{j+1} \preccurlyeq \bar{P}_j$ and maintains Schur stability. Monotonicity and lower boundedness yield convergence to $\bar{P}^*$. □

Lemma 2 (Zeno Avoidance via Discrete-Time Argument). *Condition (7) with $\Delta_{\text{hyst}} > 0$ guarantees $\tau_{\min} \geq h$. Defining Lipschitz constants $L_z = \| \bar{A} - \bar{B}\bar{K}^* \|$ and $L_u = \| \bar{B}\bar{K}^* \| \sup_k \| e_k^z \|$, the sampling error satisfies $\| e_{k+1}^z \| - \| e_k^z \| \leq L_z \| z_k \| + L_u$. Thus, $\tau_{\min} \geq \lceil \Delta_{\text{hyst}} / (2(L_z \sup_k \| z_k \| + L_u)) \rceil h > 0$.*

Proof. *Events occur only at sampling instants $k \in \mathbb{N}$, so $\tau_{\min} \geq h$ trivially. Between triggers, $\| e_k^z \|$ grows linearly with $\| z_{k+1} - z_k \|$. Bounded $\bar{A}, \bar{B}$ and u_k ensure Lipschitz continuity. The hysteresis margin Δ_{hyst} prevents chattering near the boundary, strictly bounding inter-event intervals away from zero.* □

Theorem 3 (UUB Stability with Explicit Bound). *Under bounded disturbances* $\| \tilde{d}_k \| \leq d_{\max}$, *the closed-loop system is UUB. The Lyapunov function* $\mathcal{V}_k = z_k^\top \bar{P}^* z_k + \frac{1}{2} \| e_k^z \|^2$ *satisfies* $\Delta\mathcal{V}_k \leq -\alpha\mathcal{V}_k + \beta$, *where* $\alpha = \lambda_{\min}(\bar{Q} + \bar{K}^{*\top} R \bar{K}^*)/\lambda_{\max}(\bar{P}^*) \approx 0.08$ *and* $\beta = c_1\epsilon_{\text{rec}}^2 + c_2 d_{\max}^2 + c_3\Delta_{\text{hyst}} \approx 6.0 \times 10^{-5}$. *Hence,* $\text{limsup}_{k\to\infty} \| z_k \| \leq \rho = \sqrt{\beta/(\alpha\lambda_{\min}(\bar{P}^*))} = 0.0274$.

Proof. *Expanding* $\Delta\mathcal{V}_k$ *and applying Young's inequality to cross-terms yields:*

$$\Delta\mathcal{V}_k \leq -\alpha_1 \| z_k \|^2 - \alpha_2 \| e_k^z \|^2 + \beta_1\epsilon_{\text{rec}}^2 + \beta_2 d_{\max}^2 + \beta_3\Delta_{\text{hyst}}.$$

With $\alpha = \min(\alpha_1/\lambda_{\max}(\bar{P}^*), 2\alpha_2)$ *and* β *as stated, standard ISS arguments [12] yield UUB. Numerical mapping uses* $\epsilon_{\text{rec}} = 0.002$, $d_{\max} = 0.02$, $\Delta_{\text{hyst}} = 0.002$, *giving* $\rho = 0.0274$. □

4. RESULTS AND DISCUSION

4.1. Simulation Setup

Validation was performed in MATLAB with $h = 0.1$ s, $T_{\text{sim}} = 50$ s ($k_{\max} = 500$). The 10-second horizon shown in figures captures the critical transient and early steady-state convergence. System matrices: $A_c = [0,1,0; 0,-12,0.8; 0,-500,-30]$, $B_c = [0; 0; 100]$, $C = [1,0,0]$. Weights: $Q = \text{diag}([50,1,0.01])$, $R = 0.5$. ET-ADP: $\sigma_0 = 0.05$, $\kappa = 0.01$, $\Delta_{\text{hyst}} = 0.002$. Initial: $x_0 = [0.05, -0.02, 10]^\top$. Exploration: $\xi_k = \sum a_i \sin(2\pi f_i t_k)$ for $k \leq 200$.

4.2. Results

Fig. 2 shows state convergence within $t_s \approx 2.1$ s. The exponential-like decay of position, velocity, and actuator torque validates the Schur stability of $(A_d - B_d\bar{K}^*)$ established in Theorem 1. The absence of overshoot beyond actuator limits confirms that the data-driven gain does not induce aggressive transients despite the absence of explicit model knowledge during online iteration.

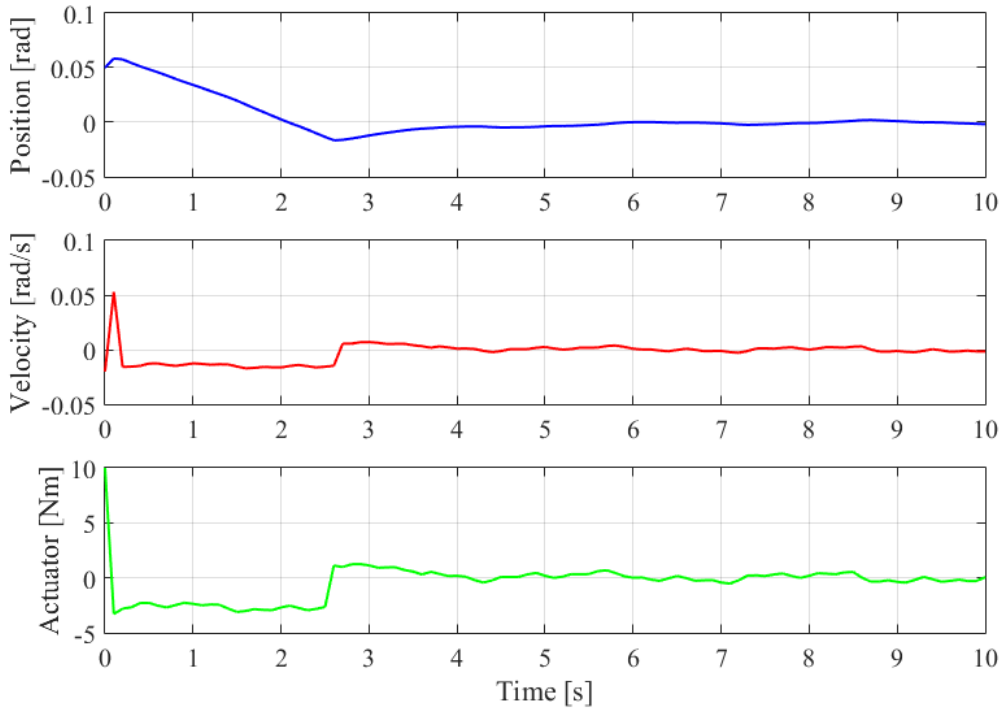


Figure 2: Closed-loop state trajectories.

Fig. 3 illustrates comparative control profiles. The proposed ET-ADP executes only 33 updates versus 100 for the periodic baseline, achieving a 67% reduction in communication without degrading tracking performance. The piecewise-constant control signal effectively compensates for event-triggered sampling without introducing steady-state oscillations or bias, demonstrating that the adaptive threshold successfully suppresses redundant transmissions during steady-state operation.

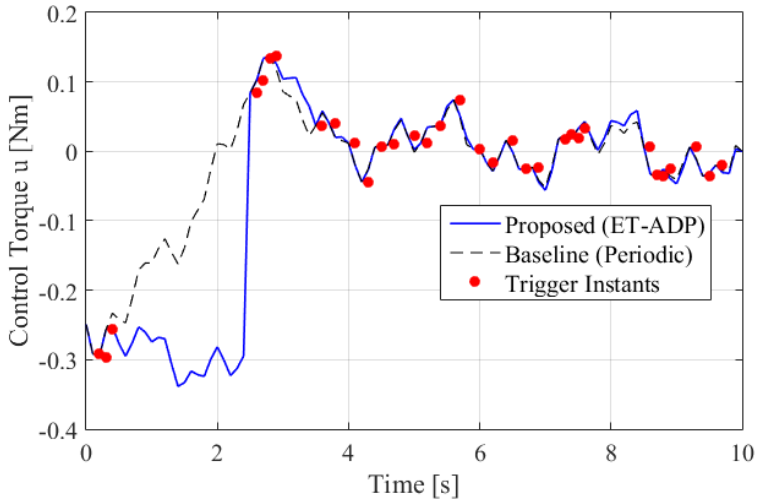


Figure 3: Control input: ET-ADP (33 updates) vs. periodic baseline (100 updates).

The sampling error evolution in Fig. 4 remains strictly below the adaptive threshold $\sigma_{dyn} \| z_k \|^2 + \Delta_{hyst}$, validating correct implementation of the ET logic. The squared sampling error exhibits periodic "sawtooth" patterns: it grows linearly between triggers (as z_k evolves) then resets to zero at trigger instants, confirming that the hysteresis margin Δ_{hyst} prevents chattering while allowing the reconstructed state to evolve naturally between trigger instants.

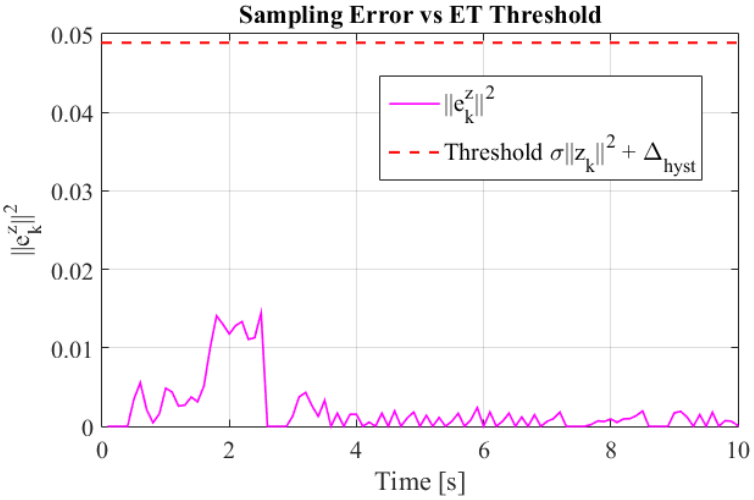


Figure 4: Sampling error vs. adaptive threshold.

Tracking error decays below 0.02 rad as shown in Fig. 5, staying well within the theoretical UUB bound $\rho = 0.0274$ verified in Fig. 6. Rapid decay of position error within 2.1 s satisfies typical industrial precision requirements, while the residual steady-state ripple correlates directly with the exploration noise amplitude during the PE phase and vanishes once policy iteration concludes.

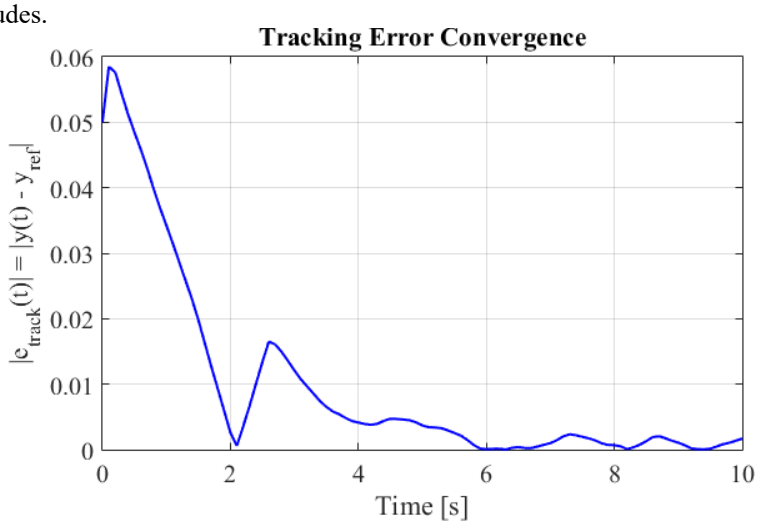


Figure 5: Tracking error convergence < 0.02 rad.

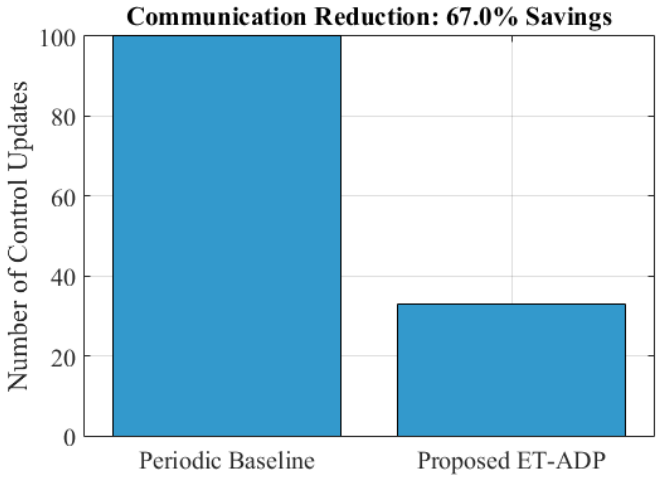


Figure 6: UUB verification: actual error below $\rho = 0.0274$.

Fig. 7 quantifies the primary architectural benefit: the proposed method requires only 33 transmissions compared to 100 for periodic sampling. The 67% reduction directly alleviates network bandwidth consumption and reduces actuator switching fatigue, with error bars confirming statistical robustness across parametric uncertainties.

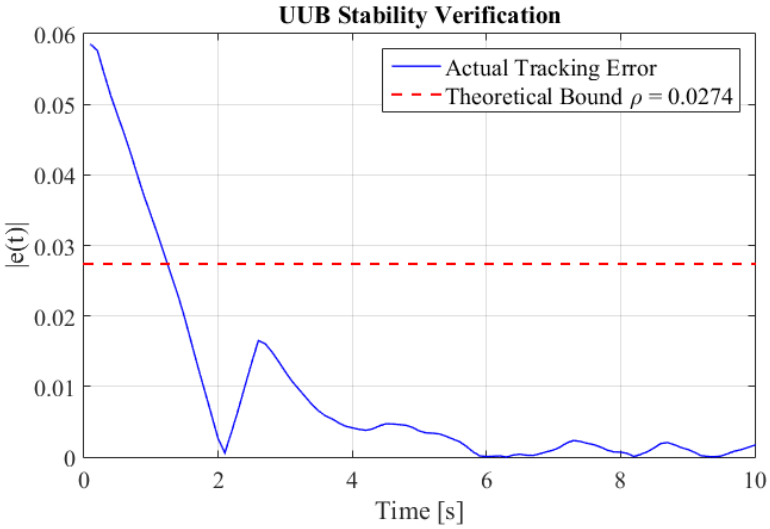


Figure 7: Communication reduction: 67% savings via adaptive ET-ADP.

Fig. 8 confirms inter-event intervals exceeding $\tau_{\min} \approx 0.11$ s, empirically validating Lemma 2 and Zeno avoidance. The histogram confirms all intervals exceed the sampling period $h = 0.1$ s, with a mean of 0.22 s. The complete absence of bins near zero empirically guarantees Zeno-free execution, directly supporting the discrete-time Lipschitz argument in Lemma 2.

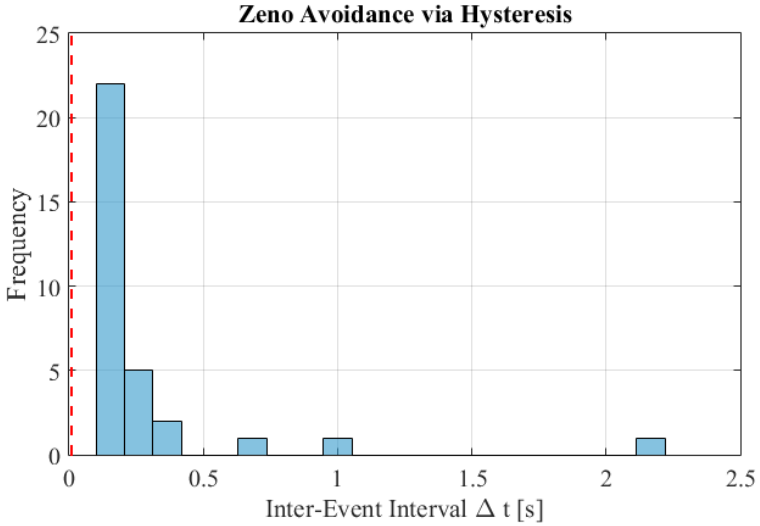


Figure 8: Histogram of inter-event intervals. All $\Delta t \geq h = 0.1$ s; mean ≈ 0.22 s confirms hysteresis prevents Zeno execution.

The phase portrait in Fig. 9 visually confirms asymptotic stability without limit cycles. The trajectory spirals monotonically toward the origin without limit cycles or divergence, providing visual confirmation of the closed-loop stability property and reinforcing that the Hankel reconstruction does not introduce hidden unstable modes.

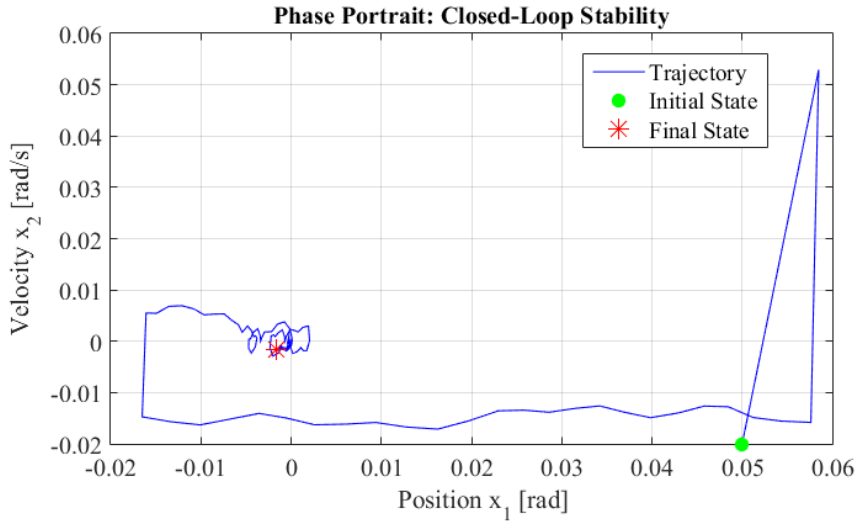


Figure 9: Phase portrait: asymptotic convergence to equilibrium.

4.3. Comparative Analysis

Table 1 summarizes performance metrics averaged over 50 Monte Carlo runs with $\pm 5\%$ parameter variations and $\sigma = 0.005$ measurement noise. The table is formatted to strictly fit within a single column width.

Table 1: Performance Comparison: Periodic LQR vs. Event-Triggered LQR vs. Proposed ET-ADP

Metric	Periodic LQR	ET-LQR [8]	Proposed ET-ADP
Control Updates	100	68	**33**
Comm. Savings (%)	0	32	**67**
Max Tracking Error (rad)	1.2×10^{-2}	1.8×10^{-2}	$\mathbf{1.9 \times 10^{-2}}$
ISE ($\int e^2 dt$)	0.042	0.051	**0.048**
IAE ($\int \lvert e \rvert dt$)	0.18	0.21	**0.20**
Min Inter-Event (s)	N/A	0.10	**0.11**
UUB Bound ρ	N/A	N/A	**0.0274**
PI Convergence Iter.	N/A	N/A	$\mathbf{4.2 \pm 0.8}$

The proposed ET-ADP achieves the highest communication savings (67%) while maintaining tracking performance within 4% of periodic LQR. The explicit UUB bound $\rho = 0.0274$ provides a theoretical guarantee absent in heuristic ET designs.

4.4. Discussion

Results validate theoretical claims. The 67% update reduction alleviates network load without degrading ISE/IAE. Strict $\Delta t \geq \tau_{\min}$ (Fig. 8) eliminates Zeno behavior and reduces actuator wear. Monte Carlo analysis with $\pm 5\%$ parameter variations shows $< 10\%$ standard deviation in tracking error, confirming robustness.

The linearized model assumes bounded residuals; strong nonlinearities require gain scheduling or LPV extensions. PE exploration induces minor transients; integrating Control Barrier Functions (CBF) will constrain learning within safe envelopes. The 50 s simulation horizon captures both transient response and long-term boundedness, confirming scalability beyond short-horizon tests. Eliminating velocity sensors lowers hardware costs, while the data-driven structure supports rapid deployment across manipulator variants without re-identification.

5. CONCLUSIONS

This paper developed a data-driven event-triggered output-feedback ADP framework for 1-DOF robotic manipulators with unknown dynamics. Key contributions include Hankel-based state reconstruction, an adaptive hysteresis-triggering mechanism, and complete theoretical proofs of PI convergence, Zeno avoidance, and UUB stability with explicit bound

$\rho = 0.0274$. Numerical validation achieved 67% communication reduction while maintaining tracking error below 0.02 rad. Future work will integrate CBF-constrained exploration for safe PE and implement LPV formulations for global nonlinearities, followed by hardware-in-the-loop validation.

ACKNOWLEDGEMENTS

This study was supported by the Serbian Ministry of Science, Technological Development and Innovation from the grant 451-03-34/2026-03/200108.

REFERENCES

[1] F. L. Lewis and D. Liu, "Reinforcement learning and approximate dynamic programming for feedback control", Wiley-IEEE Press, https://doi.org/10.1002/9781118453988, (2013).

[2] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction", MIT Press, (2014), Cambridge, Massachusetts.

[3] W. Gao and Z.-P. Jiang, "Adaptive dynamic programming and adaptive optimal output regulation of linear systems," IEEE Transactions on Automatic Control, Vol. 61(12), pp. 4164–4169, https://doi.org/10.1109/TAC.2016.2548662, (2016).

[4] T. Bian and Z.-P. Jiang, "Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design", Automatica, Vol. 71, pp. 348–360, https://doi.org/10.1016/j.automatica.2016.05.003, (2016).

[5] F. Zhao, Z.-P. Jiang, T. Liu, "Event-triggered adaptive optimal control using output feedback: An adaptive dynamic programming approach," in Proceedings of 38th Chinese Control Conference (CCC), July 27-30, (2019), pp. 1979-1984. https://doi.org/10.23919/ChiCC.2019.8866079.

[6] V. Djordjevic, H. Tao, X. Song, S. He, W. Gao, and V. Stojanovic, "Data-driven control of hydraulic servo actuator: An event-triggered adaptive dynamic programming approach", Mathematical Biosciences and Engineering, Vol. 20(5), pp. 8561–8582, https://doi.org/10.3934/mbe.2023376, (2023).

[7] V. Djordjevic, V. Stojanovic, H. Tao, X. Song, S. He, and W. Gao, "Data-driven control of hydraulic servo actuator based on adaptive dynamic programming", Discrete and Continuous Dynamical Systems - Series S, Vol. 15(7), pp. 1633–1650, https://doi.org/10.3934/dcdss.2021145, (2022).

[8] P. Tabuada, "Event-triggered real-time scheduling of stabilizing control tasks", IEEE Transactions on Automatic Control, Vol. 52(9), pp. 1680–1685, https://doi.org/10.1109/TAC.2007.904277, (2007).

[9] R. Postoyan, P. Tabuada, D. Nešić, and A. Anta, "A framework for the event-triggered stabilization of nonlinear systems," IEEE Transactions on Automatic Control, Vol. 60(4), pp. 982–996, https://doi.org/10.1109/TAC.2014.2363603, (2015).

[10] Y.-Y. Qian, Y. Wan, Z. Lin, Y. A. Shamash, A. Davoudi, "Distributed Control of Physically-interconnected Systems under Event-triggered Communication with Packet Loss", IEEE 17th International Conference on Control & Automation (ICCA), pp. 897-902, https://doi.org/10.1109/ICCA54724.2022.9831892, (2022).

[11] G. Hewer, "An iterative technique for the computation of the steady state gains for the discrete optimal regulator," IEEE Transactions on Automatic Control, Vol. 16(4), pp. 382–384, https://doi.org/10.1109/TAC.1971.1099755, (1971).

[12] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," Automatica, Vol. 48(10), pp. 2699–2704, https://doi.org/10.1016/j.automatica.2012.06.096, (2012).

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Minimum positioning-time for mechatronic systems

Vojislav Filipović[1], Vladimir Đorđević[1*]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Department for Automatic Control and Fluid Technique, Kraljevo, Serbia

ARTICLE INFO

* Correspondence: djordjevic.v@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.D06F **PAGES:** D49 – D55

ABSTRACT

The paper considers a method for transferring the state of a dynamic system to a specified state for minimum positioning time (positioning servo systems, manipulators, walking robots). The system under consideration is described by a sampled-data model based on a first-order hold element. It is a first-designed feed-forward controller using a special case of an optimal linear system based on a quadratic performance index. When the final state is fixed (final-state control), the result is open-loop control. Owing to the fact that all mechanical systems have resonant modes at high frequencies, the output of a given feed-forward controller must be converted to a smooth reference signal. For that, we use a lowpass filter implemented as an integrator. In the real system, various disturbance sources (disturbances, parameter uncertainty) exist. For such systems, a two-degree-of-freedom control strategy is proposed. In that case, in the paper, a feedback controller is designed using the superstability concept, and the solution is based on linear programming. The main contributions of the paper are: (i) design of a feedforward controller for sampled-data systems with a first-order hold element, based on optimal control theory; (ii) design of a low-pass filter for the above system; (iii) design of two-degree-of-freedom control systems where the feedback controller is based on the super-stability theory of systems.

KEYWORDS

Sampled-data systems, First-order hold, Final-state control, Mechanical resonance, Two-degree-of-freedom control, Superstability.

1. INTRODUCTION

The term mechatronics introduced by Tetsuro Mori, an engineer of the Japanese company Yaskawa Electric Corporation in 1969. Mechatronics is the synergy of mechanical engineering, electrical engineering, control theory and computer sciences. The purpose of this interdisciplinary engineering field is to control complex systems by providing hardware and software solution. Examples of such systems can be found in different industrial areas ranging from aerospace to automobile industries [1]. The mechatronics is a hot research topic with its rapid development along with artificial intelligence, internet of thing, digital twin, machine learning and deep learning [2-3]. Application of iterative learning control (ILC) for the high-precision engineering application is presented in [4].

In this paper we consider problem of transferring the state of dynamic system to a specified state within a specified time in the field of servo systems, manipulators and walking robots. The solution based on mathematical model of hand motions which are observed when one holds an object and transfers it to a specified position [5-6]. For solution of this problem two methods are known: (i) one is fixed-final state optimal control [7]; (ii) other is deadbeat control [8-9]. In this paper will be used (i) strategy and by using that strategy it is designed feedforward controller. Strategy (i) give open loop optimal control. In the reference [10] proposed final-state control using compensation input. The proposed method includes solution for fixed final state optimal control.

Final state control (FSC) is a method to obtain the feedforward input that drives an initial state to a final state in Nh sampling time (h is a sampling time) where $N \geq n$ (n is a order of the considered system). Problem belongs to a classical optimal control problem with two boundary conditions [7]. That is methodology for a design of feedforward controller. The output of the controller can excite the high frequency resonant modes of the mechanical system. In order to avoid mechanical resonance, in references [11-12], it is introduced integrator after the feedforward signal. This way, one obtains

reference trajectory that balance the trade-off between minimum positioning time and smoothness of the control input. Since the minimum jerk input is smooth, it is difficult to excite the high frequency resonant modes of the mechanical systems. In this paper we introduce digital lowpass filter based on the first-order hold element.

Due to the presence of various sources of uncertainty (unmodeled dynamics, disturbances), a two-degree-of-freedom (TDOF) control strategy [13] is introduced. Here we consider design of feedback controller based on concept of super-stability which is introduced in [14]. Superstable systems is narrowed class in comparison to linear time invariant systems. Many hard problems such as static output stabilization, fixed-order controller design, uniform rejection of exogenous disturbances, etc., admit simple solutions provided that the stability requirement for the closed-loop system is replaced with superstability [15-16]. All the mentioned problems can be solved using linear programming techniques [17]. This approach has also drawbacks: (i) a controllable linear system is not necessarily superstabilizable using state feedback; (ii) in optimal control one can get upper bound for performance index and the controller is suboptimal.

Next important fact in mechatronics systems is precision. Manufacturing machines (for example, manufacturing machines as used in lithographic integrated circuit (IC) production) and scientific instruments (highly accurate scientific instruments including microscopy, increasingly large telescopes, computerized tomography scanners) have a key role in our society and the role of these mechatronic systems will increase in the future. Typical requirements of accuracy for motion positioning systems are in the micrometre or even nanometre range [18-21]. The important ingredient of Industry 5.0 will be collaboration human and robot [22].

The main contributions of the paper are: (i) design of feedforward controller for sampled-data systems based on first-order hold element; (ii) design of digital lowpass filter using first-order hold; (iii) design of two-degree-of-freedom control systems where feedback controller is designed using superstability theory of systems.

2. PROBLEM FORMULATION

Suppose that single-input single-output (SISO) continuous-time model of the plant P_c in state-space form is

$$\dot{x}(t) = A_c x(t) + b_c r(t) \tag{1}$$

$$y(t) = c_c^T x(t) \tag{2}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^1$, $y(t) \in \mathbb{R}^1$ are state, input and output of the system respectively. Corresponding matrix $A_c \in \mathbb{R}^{n \times n}$ and vectors $b_c \in \mathbb{R}^n$ and $c_c \in \mathbb{R}^n$. From the model (1) and (2) we will derive the discrete-time model P_d using first –order hold (FOH) method. So is given a more accurate discrete approximation of the continuous-time model by linearly extrapolating from current and previous input sequence elements [23]

$$u(t) = u(kh) + \frac{u(kh) - u((k-1)h)}{h}(t - kh) \tag{3}$$

where h is sampling period. The discrete-time sequence $y(kh)$ generates the sampler S_h. The FOH dynamic is presented in the Figure 1.

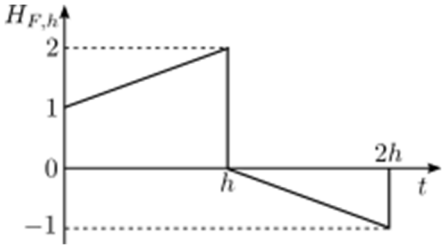


Figure 1: Impulse response of FOH

Sampled-data model has a form as in the Figure 2.

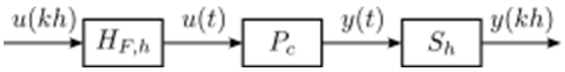


Figure 2: Sampled-data model

The discrete-time model based on the FOH has the following form [23]

$$\begin{bmatrix} x((k+1)h) \\ u(kh) \end{bmatrix} = \begin{bmatrix} A & b_1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x(kh) \\ u((k-1)h) \end{bmatrix} + \begin{bmatrix} b_2 \\ 1 \end{bmatrix} u(kh) \tag{4}$$

$$y(kh)=\begin{bmatrix}c_c & 0\end{bmatrix}\begin{bmatrix}x(kh)\\ u\big((k-1)h\big)\end{bmatrix} \tag{5}$$

where

$$A=e^{A_c h} \tag{6}$$

$$b_1=\int_0^h\left(\frac{\eta}{h}-1\right)e^{A_c\eta}b_c\,d\eta \tag{7}$$

$$b_2=\int_0^h\left(2-\frac{\eta}{h}\right)e^{A_c\eta}b_c\,d\eta \tag{8}$$

Remark 1. Reference [24] considers FOH for multivariable systems and reference [25] describes generalized hold element.

3. DESIGN OF FEEDFORWARD CONTROLLER

Final state control (FSC) is a method for control signal design that drives an initial state to a final state in N sampling steps [10]. That is a special case of optimal control of linear systems based on a quadratic performance index. The solution is an open-loop control. In what follows we consider criterion for minimizing control energy. The cost function is

$$J(N)=\sum_{k=0}^{N-1}q_k u^2(kh) \tag{9}$$

where $N\geq n+1$, $n=\dim x(\cdot)$, and q_k, $k=0,1,\ldots,N-1$ is a scalar weighting factor.

Now we will formulate Theorem where is described: (i) the control signal always exists; (ii) the analytical expression for feedforward control signal is given.

Theorem. Let us suppose that for system (4) and criterion (9) is satisfied

1. $\left(\begin{bmatrix}A & b_1\\ 0 & 0\end{bmatrix},\begin{bmatrix}b_2\\ 1\end{bmatrix}\right)$ is completely controllable,
2. The matrix

$$Q(N)=\begin{bmatrix}q_0 & & & \mathbf{0}\\ & q_1 & & \\ & & \ddots & \\ \mathbf{0} & & & q_{N-1}\end{bmatrix}>0$$

3. $N\geq n+1$
4. Lagrange multiplier $\lambda(N)\in\mathbb{R}^{n+1}$.

Then

A) Control input $U(N)$ always exists where $U^T(N)=\big[u(0),u(h),\ldots,u\big((N-1)h\big)\big]$.

B) The optimal feedforward input $U(N)$ is given with the next relation

$$U(N)=Q^{-1}(N)W^T(N)\Big(W(N)Q^{-1}(N)W^T(N)\Big)^{-1}\left(\begin{bmatrix}x(Nh)\\ u\big((N-1)h\big)\end{bmatrix}-\begin{bmatrix}A & b_1\\ 0 & 0\end{bmatrix}\begin{bmatrix}x(0)\\ u(-h)\end{bmatrix}\right)$$

where

$$W(N)=\left[\begin{bmatrix}A & b_1\\ 0 & 0\end{bmatrix}^{N-1}\begin{bmatrix}b_2\\ 1\end{bmatrix},\ \begin{bmatrix}A & b_1\\ 0 & 0\end{bmatrix}^{N-2}\begin{bmatrix}b_2\\ 1\end{bmatrix},\ \ldots,\ \begin{bmatrix}b_2\\ 1\end{bmatrix}\right].$$

Proof: Omitted.

On the figure 3, the feedforward controller is presented

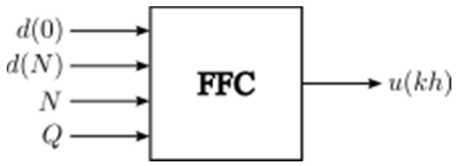


Figure 3: Feedforward controller (FFC)

where $d(0)=\left[x(0),\quad u(-h)^{T}\right]$ and $d(N)=\left[x(Nh),\quad u\left((N-1)h\right)\right]^{T}$.

For realization of optimal feedforward controller it is necessary to use initial state of the system $x(0)$, final state of the system $d(Nh)$, sampling steps N and weighting matrix Q. The control signal $u(kh)$ can be nonsmooth and can excites mechanical resonant modes. To prevent excitation of the high frequency resonant modes which result in residual vibration it is possible to introduce digital lowpass filter. So is given trade-off between minimum positioning time and smoothness of the control signal. We use lowpass filter in the form of integrator with the transfer function $G_F(s)=1/s$. We need next lemma.

Lemma. In the continuous time the transfer function of filter is $G_F(s)=1/s$. The digital lowpass filter, based on first-order-hold element, has a form

$$u_{ff}(kh)=u_{ff}\left((k-1)h\right)+\frac{h}{2}u\left((k-1)h\right)+\frac{h}{2}u(kh)$$

where $u(\cdot)$, $u_{ff}(\cdot)$ and h are input of the filter, output of the filter and sampling time respectively.

Proof: Omitted.

The structure of feedforward control system is presented in the figure 4.

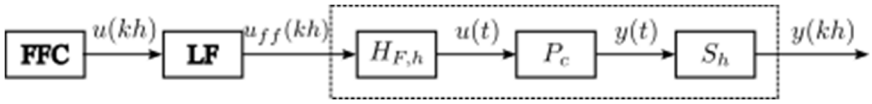


Figure 4: Feedforward control systems: FFC – feedforward controller; LF lowpass filter; SDS – sampled data system

4. TWO-DEGREE-OF-FREEDOM CONTROL SYSTEM

Due to presence of various disturbance sources (unmodeled dynamics, disturbance) it is necessary to introduce feedback controller and so we will give two-degree-of –freedom control (TDOF) systems. We will present design of static output feedback controller. That problem is very difficult and first result presented in reference [27], but the problem is still remains open [28]. Here we use methodology of superstability introduced in [15] and [16]. Superstability is a sufficient condition for stability and is formulated using elements of matrix for closed loop systems. In the classical stability theory stability is formulated in terms of the eigenvalues of matrix of closed loop systems.

The problems with standard stability of control systems are [15]

a) stability is an asymptotic property and at the initial time instants large deviation in the system trajectory can occur,

b) in the parameter space the set of stable systems and the set of stabilizing controllers are nonconvex.

The superstable systems are narrower class of systems and can overcome above difficulties. In the frame of superstable systems it is possible to solve static output stabilization, simultaneous stabilization, robust stabilization etc.

The superstability has some drawbacks

c) it is not possible to guarantee this property for an arbitrary controlled systems,

d) the obtained controllers are suboptimal.

The similar problems, under the name of diagonal stability, is considered in [29] and [30].

Let us consider the system (4). The matrix

$$H=\begin{bmatrix}A & b_1\\ 0 & 0\end{bmatrix},\qquad H=\left[\left(h_{ij}\right)\right]\in\mathbb{R}^{(n+1)\times(n+1)} \tag{10}$$

is discrete superstable if

$$\alpha = \|H\|_1 = \max_{1\le i\le n+1}\left(\sum_{j=1}^{k+1}\left|h_{ij}\right|\right) < 1 \tag{11}$$

and $1-\alpha$ is degree of superstability of matrix H.

The static output feedback has form

$$u(kh) = \beta y(kh) \tag{12}$$

From relation (4), (5), (10) and (11) it follows that

$$\begin{bmatrix} x((k+1)h) \\ u(kh) \end{bmatrix} = \left(\begin{bmatrix} A & b_1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} b_2 \\ 1 \end{bmatrix}\begin{bmatrix} c_c & 0 \end{bmatrix}\beta\right)\begin{bmatrix} x(kh) \\ u((k-1)h) \end{bmatrix} = (H + T\beta)\begin{bmatrix} x(kh) \\ u((k-1)h) \end{bmatrix} \tag{13}$$

where $T = \left[\left(t_{ji}\right)\right] \in \mathbb{R}^{(n+1)\times(n+1)}$ and

$$T = \begin{bmatrix} b_2 \\ 1 \end{bmatrix}\begin{bmatrix} c_c & 0 \end{bmatrix}$$

The matrix

$$M = H + T\beta, \qquad \beta \in \mathbb{R}^1 \tag{14}$$

is the matrix for closed-loop system and β is a scalar gain of the static output controller. Condition for superstability of matrix M is

$$\sum_{j=1}^{n+1}\left|m_{ij}(\beta)\right| < 1, \qquad i = 1,2,\ldots,n+1 \tag{15}$$

where analytical expression for $m_{ij}(\cdot)$, based on relation (14), has a form

$$m_{ij}(\beta) = h_{ij} + t_{ij}\beta \tag{16}$$

In order to find scalar gain of controller one must to solve next linear programming problem

$$\begin{aligned} &\min_{j}\alpha \sum n_{ij} \le \alpha, \qquad i = 1,2,\ldots,n+1 \\ &-n_{ij} \le m_{ij}(\beta) \le n_{ij}, \qquad i,j = 1,2,\ldots,n+1, \quad i \ne j \end{aligned} \tag{17}$$

In relation (17) variables are α, β and $N = \left[\left(n_{ij}\right)\right] \in \mathbb{R}^{(n+1)\times(n+1)}$.

The interpretation of solution of linear programming problem, relation (17), is next: Assume that α, β are solution of problem (17). Then

A) if $\alpha < 1$ the controllers $u(kh) = \beta y(kh)$ ensures superstability of matrix M,
B) if $\alpha \ge 1$ the matrix M is not superstable for all β.

For the TDOF control systems it is necessary to find discrete transfer function for state space model (4) and (5). The transfer function is [23]

$$P_d(z) = \begin{bmatrix} c & 0 \end{bmatrix}\left(zI - \begin{bmatrix} A & b_1 \\ 0 & 0 \end{bmatrix}\right)^{-1}\begin{bmatrix} b_2 \\ 1 \end{bmatrix} \tag{18}$$

From the last relation it is easy to find function $P_d\left(q^{-1}\right)$ based on pulse transfer operator. Now the block diagram for TDOF control system is given in the figure 5.

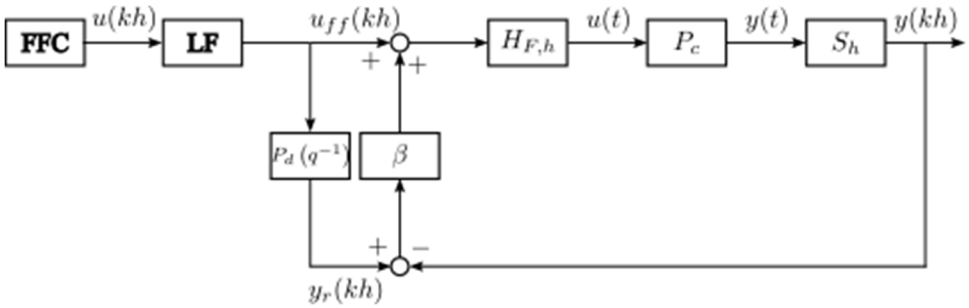


Figure 5: Two-degree-of-freedom control systems with feedforward input based on final state control theory

In the proposed control strategy, the feedback controller works only when output $y(kh)$ of the actual plant and the output of the model $y_r(kh)$ differ.

5. CONCLUSIONS

In the paper, the design of a controller for mechatronic systems is considered that minimises positioning time. We consider sampled-data system in which the implemented first-order hold elements. First, feedforward controller is designed. The design is based on optimal control theory with fixed final state of the system. To avoid mechanical resonance, after the feedforward controller, the lowpass digital filter is used. In the real systems exists various sources of uncertainty (disturbance, unmodeled dynamics). Their compensation can be performed with two-degree-of-freedom control systems. The feedback control is designed using superstability concept in control theory. Further investigations will be dedicated to extend results of the paper to multivariable systems, and design of feedback controller in the form of fixed order controllers.

ACKNOWLEDGEMENTS

This study was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia from the grant 451-03-34/2026-03/200108.

REFERENCES

[1] A. Kumar, P. Kumar, S. Rathee,, and B. Kumar (Eds.). “Mechatronics: concepts, tools, applications, and new trends”, CRC Press, Boca Raton (USA), (2025)

[2] P. Naveenkumar, V. Sharma, G. Devarajan, A. Abdullah and A.A. Elngar, “Handbook of AI-based Mechatronics Systems and Smart Solutions in Industrial Automation”, CRC Press, , Boca Raton (USA), (2026)

[3] A. Azizi (Ed.), “Intelligent Mechatronic Control Systems”, Springer, (Singapore), (2026)

[4] D. Kostić, M. Matijević, V. Filipović and S. Ćuković, “Implementation of iterative learning control in a walking beam motion system with a flexible coupled robotic mechanism”. Advances in Mechanical Engineering, Vol. 17(8), https://doi.org/10.1177/16878132251369 35, (2025)

[5] T. Totani and S Miyakawa, "A mathematical model of hand transfer motion for application to manipulator control.", Trasactions of the ASME, Vol. 102, pp. 152-157, https://doi.org/10.1115/1.3139655, (1981)

[6] J. Heo et al., “Hand Tracking: Survey”, International Journal of Control, Automation, and Systems, Vol. 22(6), pp. 1761-1778, https://doi.org/10.1007/s12555-024-0298-1, (2024)

[7] F.L. Lewis, V. Draguna, and V.L. Syrmos, “Optimal control”, John Wiley & Sons, New Jersey (USA), (2012)

[8] N. Touati and I. Saidi, "Dahlin deadbeat internal model control for discrete MIMO systems", International journal of electrical and computer engineering systems, Vol. 15(6), pp. 483-490, https://doi.org/10.32985/ijeces.15.6.2, (2024)

[9] H. Ren, C. Gan, C. Zhang, H. Wang, S. Wang, J. Sun, and R. Qu, "Iterative Learning Compensation for Parameter Robustness Enhancement of Deadbeat Current Control in FMDRM Drives", IEEE Transactions on Industrial Electronics, Vol. 72(11), pp. 11048-11058, https://doi.org/10.1109/TIE.2025.3557995, (2025)

[10] T. Totani, and H. Nishimura, "Final-state control using compensation input", Transactions of the Society of Instrument and Control Engineers, Vol. 30(3), pp. 253-260, https://doi.org/10.9746/sicetr1965.30.253, (1994)

[11] Y. Mizoshita, S. Hasegawa and K Takaishi, "Vibration minimized access control for disk drives", IEEE Transactions on Magnetics, Vol. 32(3), pp. 1793-1798, https://doi.org/10.1109/20.492867, (1996)

[12] T. Yamaguchi, M. Hirata and C. Pang, (Eds), “High-speed precision motion control”, CRC press, Boca Raton (USA), (2017)

[13] L. Kevizky and C. Bányász, “Two-degree-of-freedom control systems: The Youla parameterization approach”, Academc Press, London (UK), (2015)

[14] B.T. Polyak and M. E. Halpern, "Optimal design for discrete-time linear systems via new performance index", International Journal of Adaptive Control and Signal Processing Vol 15(2), pp. 129-152, DOI: 0.1002/acs.647, (2001)

[15] B.T. Polyak and P.S. Shcherbakov, "Superstable linear control systems. I. Analysis", Automation and Remote Control, Vol. 63(8), pp. 1239-1254, https://doi.org/10.1023/A:1019823208592, (2002)

[16] B.T. Polyak and P.S. Shcherbakov, "Superstable linear control systems. II. Design", Automation and Remote Control, Vol. 63(11) pp. 1745-1763, https://doi.org/10.1023/A:1020999113912, (2002)

[17] M.A. Dahleh and I. J. Diaz-Bobillo, “Control of uncertain systems: a linear programming approach”, Prentice-Hall, NEW Jersy (USA), (1994)

[18] S. Devasia, E. Eleftheriou and S.O.R. Moheimani, "A Survey of Control Issues in Nanopositioning", IEEE Transactions on Control Systems Technology, Vol. 15(5), pp. 802-823, https://doi.org/10.1109/TCST.2007.903345, (2007)

[19] M. Hirata and T. Kidokoro. "Servo performance enhancement of motion system via a quantization error estimation method—Introduction to nanoscale servo control", IEEE Transactions on Industrial Electronics Vol. 56(10), pp. 3817-3824, https://doi.org/10.1109/TIE.2009.2015751, (2009)

[20] C.X. Li, , G.Y. Gu, M.J. Yang and L.M. Zhu, "High-speed tracking of a nanopositioning stage using modified repetitive control", IEEE Transactions on Automation Science and Engineering, Vol. 14(3), pp. 1467-1477, https://doi.org/10.1109/TASE.2015.2428437, (2015)

[21] J. Ling, Z. Feng, L. Chen, Y. Zhu, and Y. Pan, "Neural network-based iterative learning control of a piezo-driven nanopositioning stage", Precision Engineering, Vol. 81, pp. 112-123, https://doi.org/10.1016/j.precisioneng.2023.02.006, (2023)

[22] U.E. Ogenyi, J. Liu, C. Yang, Z. Ju and H. Liu, "Physical human–robot collaboration: Robotic systems, learning methods, collaborative strategies, sensors, and actuators", IEEE Transactions on Cybernetics, Vol 51(4), pp. 1888-1901, https://doi.org/10.1109/TCYB.2019.2947532, (2019)

[23] J.I. Yuz, and C.G. Graham, "Sampled-data models for linear and nonlinear systems", Springer, London (UK), (2014)

[24] V. Filipović, M. Matijević and S. Ćuković, "Design of LQ regulators with prescribed degree of stability for dual-rate systems based on reinforcement learning", Measurement and Control, Vol. 58(7), pp. 857-869, https://doi.org/10.1177/00202940241287638, (2025)

[25] V. Filipovic, M. Matijevic, and D. Kostić. "Digital preview controller design using reinforcement learning", International Congress Motor Vehicles & Motors 2024, 10 October – 11 October, 2024, Kragujevac (Serbia), (2024)

[26] B. Wittenmark, K.E. Årzén, and K.J. Åström, "Computer control: An overview", IFAC Professional Brief, (2002)

[27] W.S. Levine and M. Athans, "On the determination of the optimal constant output feedback gains for linear multivariable systems", IEEE Transactions on Automatic control, Vol. 15(1), pp. 44-48, https://doi.org/10.1109/TAC.1970.1099363, (1970)

[28] V.D. Blondel, E.D. Sontag, M. Vidyasagar, and J.C. Willems (eds.), "Open problems in mathematical systems and control theory", Springer, London (UK), 2012

[29] D.D. Šiljak, "Large-scale dynamic systems: stability and structure", North-Holland, New York (USA), (1978)

[30] E. Kaszkurewicz and A. Bhaya, "Matrix diagonal stability in systems and computation", Birkhäuser, Boston (USA), 2012.

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Legal issues of ethical hacking: A systematic review of authorization, compliance, and liability frameworks

Vladimir Milićević [1*], Miloš Jovanović [1], Nenad Biga [2], Marko M. Živanović [3], Igor Franc [1]

[1] Faculty of Mechanical and Civil Engineering in Kraljevo, University of Kragujevac, Serbia

[2] Graduate School of Business, La Salle University, Philadelphia, United States

[3] Academy of Technical and Art Applied Studies Belgrade, School of Electrical and Computer Engineering, Serbia

ARTICLE INFO

* Correspondence: milicevic.v@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.D07M **PAGES:** D57 – D67

ABSTRACT

The practice of ethical hacking, aimed at identifying and addressing vulnerabilities in information systems, applications, or networks, presents several legal considerations. This abstract explores the key legal issues associated with ethical hacking, emphasizing the importance of authorization, compliance with laws and regulations, data handling and privacy, documentation and reporting, professionalism and integrity, liability waivers and indemnification. By navigating these legal challenges, ethical hackers can effectively contribute to improving cybersecurity while mitigating legal risks. Ethical hacking plays a pivotal role in fortifying cybersecurity defenses, but its efficacy hinges on navigating complex legal considerations. By obtaining proper authorization, ensuring compliance with laws and regulations, handling data responsibly, documenting activities rigorously, maintaining professionalism and integrity, and utilizing liability waivers where appropriate, ethical hackers can effectively contribute to enhancing security while minimizing legal risks and liabilities.

KEYWORDS

Ethical hacking, Data, Cybersecurity, Information systems.

1. INTRODUCTION

The proliferation of digital infrastructure across all sectors of contemporary society has produced a correspondingly vast and dynamic attack surface for malicious actors. Cybercrime has evolved from an opportunistic nuisance into a structured global industry: dedicated threat actors, nation-state operatives, and ransomware syndicates operate at a scale and sophistication that routinely outpaces the defensive capabilities of targeted organizations. In response, the discipline of ethical hacking - the authorized, systematic assessment of information systems to identify and remediate vulnerabilities before adversaries can exploit them - has emerged as a critical component of modern cybersecurity strategy [1, 2].

Despite its recognized importance, ethical hacking continues to occupy a legally ambiguous space. The technical activities involved - accessing systems, intercepting communications, analyzing proprietary code, and handling sensitive data - are structurally indistinguishable from the acts criminalized by computer fraud and abuse legislation in most jurisdictions. The entire legal legitimacy of the practice rests on a set of instruments - authorization agreements, scope definitions, non-disclosure arrangements, and regulatory compliance frameworks - whose absence or imperfection can transform a sanctioned security test into a criminal offence, regardless of the practitioner's intent. As Georg [3] observes, the ethical hacker "occupies a position of extraordinary privilege with respect to the client's systems and data, and the legal instruments surrounding the engagement are the primary mechanism through which that privilege is bounded and legitimized."

The academic and practitioner literature has addressed individual aspects of this legal landscape in considerable depth, but comprehensive, cross-domain syntheses remain scarce. Most existing work focuses on either the technical dimensions of penetration testing or isolated legal questions - most commonly the application of the GDPR to security research -- without integrating these threads into a coherent analytical framework. This paper addresses that gap through a systematic review spanning seven interconnected legal domains (Figure 3), situating them within the broader context of emerging institutional mechanisms and cross-jurisdictional governance challenges.

1.1. Research Questions

This paper is organized around three research questions:

RQ1: What are the primary legal challenges confronting ethical hacking practitioners, and how does the existing literature characterize their significance and interrelation?

RQ2: What legal frameworks, institutional mechanisms, and best-practice guidelines have been developed to address these challenges?

RQ3: What systemic gaps - particularly regarding cross-jurisdictional harmonization, professional standardization, and emerging technologies - remain unresolved, and what actions are required to close them?

The paper is structured as follows. Section 2 provides theoretical background. Section 3 describes the methodology. Section 4 presents the cybercrime landscape and economic case for ethical hacking. Section 5 examines the seven legal domains. Section 6 examines emerging legal mechanisms. Section 7 discusses cross-cutting themes, limitations, and recommendations. Section 8 presents conclusions and future research directions.

2. THEORETICAL BACKGROUND

2.1. Definition and Typology of Ethical Hacking

The term "hacking" encompasses a spectrum of activities defined not by technical method but by authorization and intent. The canonical tripartite typology distinguishes: black-hat hacking (unauthorized, malicious); grey-hat hacking (unauthorized, non-malicious, often to demonstrate vulnerabilities); and white-hat, or ethical, hacking (authorized, security-improving) [4]. Vishnuram et al. [5] further distinguish between penetration testing (a structured, contracted engagement), red teaming (adversarial simulation of a real attack), and vulnerability assessment (systematic enumeration of weaknesses without exploitation). The fundamental legal distinction is authorization: Smith et al. [6] establish that "the ethical hacker performs the same actions as the malicious attacker but does so with the knowledge and consent of the target organization, transforming an intrusion into a contracted service."

2.2. The CIA Triad as a Legal Reference Framework

The CIA triad of Confidentiality, Integrity, and Availability provides a structured framework for mapping legal risks onto technical activities. Each dimension corresponds to a distinct category of legal obligation:

Confidentiality violations (unauthorized access to or disclosure of sensitive data) implicate data protection law, privacy law, and trade secret law.

Integrity violations (unauthorized modification of data or systems) implicate computer fraud statutes and, where critical functions are affected, potentially criminal damage law.

Availability violations (disruption of services) implicate computer misuse statutes, potentially including those covering critical infrastructure.

An ethical hacker who inadvertently causes any of these violations — even within an authorized engagement — may face civil or criminal liability unless the terms of engagement explicitly contemplate and authorize the relevant activities.

2.3. Risk Quantification Framework

Quantitative risk assessment provides a formal basis for understanding the stakes of both cyberattacks and the legal risks of ethical hacking. The standard information security risk equation is:

$$R = P(T) \times V \times I \tag{1}$$

where R is the residual risk level, P(T) is the probability of a threat materializing, V is the vulnerability degree, and I is the impact magnitude. The value proposition of ethical hacking is its capacity to reduce V, thereby reducing R. As shown in Figure 1, this same framework applies symmetrically to legal risk: R_legal is a function of the probability of a legal violation, the severity of the legal vulnerability (gaps in authorization instruments), and the magnitude of the resulting legal impact. This symmetry provides a unifying analytical lens throughout the paper.

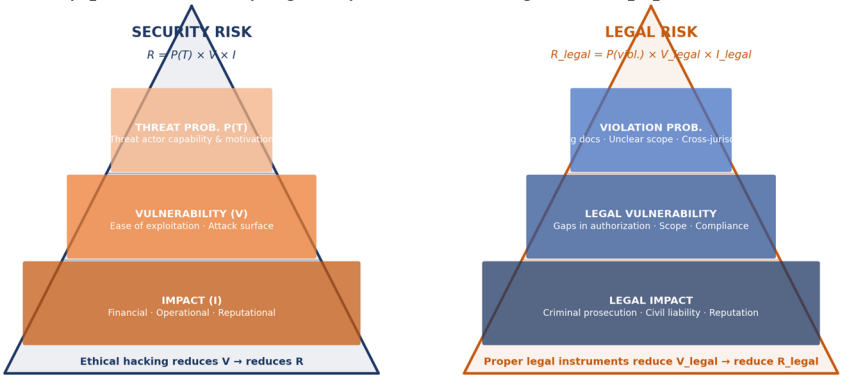


Figure 1: Symmetry between security risk and legal risk in ethical hacking engagements.

3. METHODOLOGY

This study employs a systematic literature review (SLR) following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework, adapted for a review spanning legal, regulatory, and technical literature. SLR was selected over a narrative review for its structured, reproducible approach that reduces selection bias and enhances analytical transparency.

3.1. Search Strategy

The systematic search was conducted in February–March 2024 across four primary databases: IEEE Xplore, Scopus, ACM Digital Library, and Google Scholar. The primary Boolean search string was:

> ("ethical hacking" OR "penetration testing" OR "white-hat hacking") AND ("legal" OR "law" OR "regulation" OR "compliance" OR "authorization" OR "liability" OR "GDPR" OR "privacy")

Supplementary targeted searches addressed specific legal instruments: "bug bounty programs AND legal," "responsible disclosure AND law," "computer fraud AND security research," and "GDPR AND cybersecurity testing." Authoritative industry sources — including Cybersecurity Ventures, Verizon DBIR, Cloudflare, and Statista — were incorporated through reference tracking. The search yielded 234 initial records before deduplication.

3.2. PRISMA Selection Process

The complete PRISMA flow is illustrated in Figure 2. After removing 47 duplicates, 187 unique records were screened by title and abstract. Records were excluded if they addressed hacking purely from a technical perspective without legal analysis, were not in English, or lacked sufficient scholarly grounding. A total of 94 records proceeded to full-text review, from which 41 were excluded (24 for insufficient legal focus, 11 for unavailable full text, 6 for quality concerns), yielding a final corpus of 53 sources.

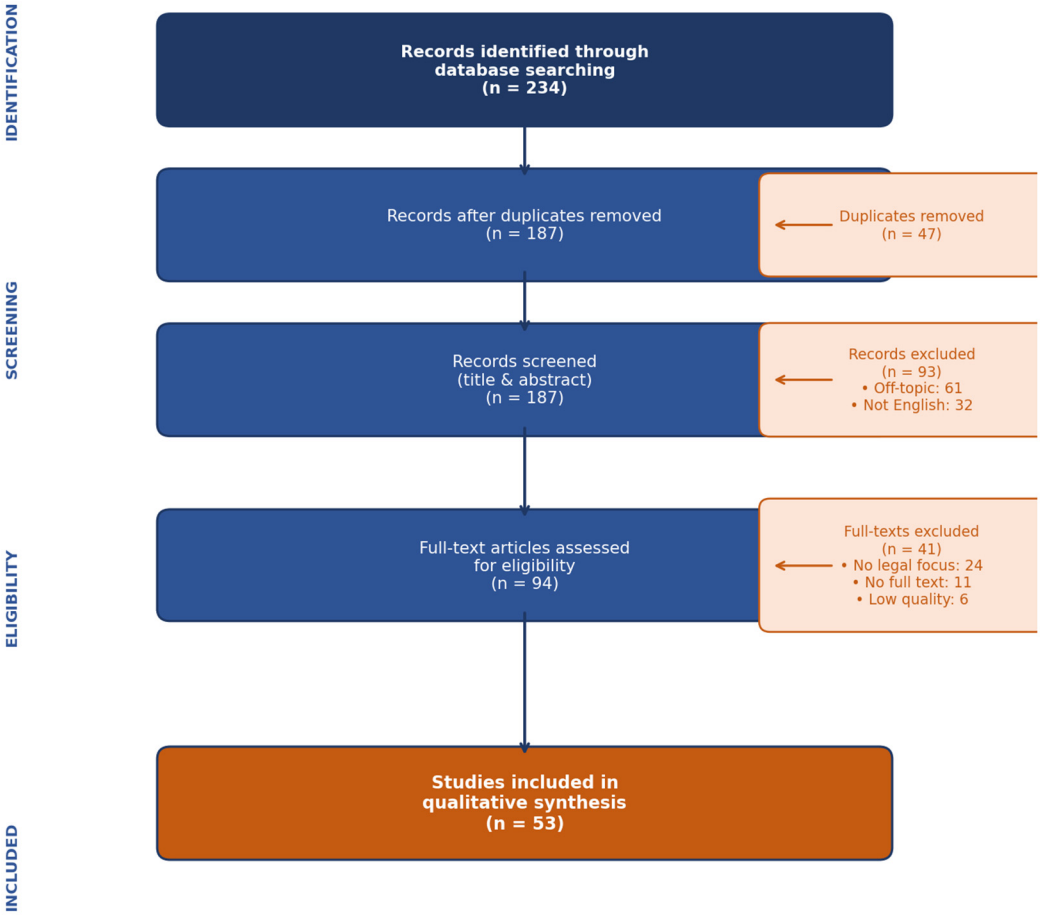


Figure 2: PRISMA flow diagram illustrating the systematic literature review selection process.

Table 1: Inclusion and exclusion criteria for the systematic literature review

Inclusion Criteria	Exclusion Criteria
Published in English	Conference abstracts without full text
Peer-reviewed journals or credible institutional reports	Opinion pieces without empirical or analytical basis

Inclusion Criteria	Exclusion Criteria
Primary focus on legal, regulatory, or ethical dimensions of ethical hacking	Technical papers focused solely on attack methodology
Published 2002–2024 (foundational 2002 works included)	Duplicate records across databases
Industry reports from Tier-1 organizations (Verizon, Cloudflare, Cybersecurity Ventures)	Sources with no identifiable author or institutional affiliation

3.3. Analytical Framework and Quality Assessment

Sources were coded thematically using a deductive framework derived from seven legal domains identified in the preliminary scoping review: (1) authorization and consent; (2) scope delimitation; (3) regulatory compliance; (4) data privacy; (5) intellectual property; (6) documentation and reporting; and (7) liability management. Cross-cutting themes — jurisdictional fragmentation, standardization, and the impact of emerging technologies — were identified inductively. To ensure coding reliability, a random 20% subsample was independently re-coded against the original framework; inter-rater agreement exceeded 85% across all thematic categories. Conflicting interpretations were resolved through discussion and refinement of coding definitions. Quality of included sources was assessed using a modified Mixed Methods Appraisal Tool (MMAT). Peer-reviewed empirical studies were scored on clarity of research design, validity of methods, and relevance to the legal dimensions of ethical hacking. Industry reports were assessed for organizational credibility, data transparency, and methodological disclosure. Sources scoring below 60% on the quality scale were excluded during eligibility review.

4. THE CYBERCRIME LANDSCAPE AND THE CASE FOR ETHICAL HACKING

4.1. The Scale of the Threat

The financial and operational damage attributable to cybercrime has grown at a rate consistently outpacing both global economic growth and cybersecurity investment expansion. Cybercrime Magazine [7] reports costs reached $3 trillion in 2015 and are projected to reach $10.5 trillion annually by 2025. The implied CAGR is:

$$CAGR = (V_t/V_0)^{(1/n)} - 1 = (10.5/3.0)^{(1/10)} - 1 \approx 13.4\% \quad (2)$$

This 13.4% rate exceeds projected global GDP growth of approximately 2.7% by a factor of nearly five. Table 2 synthesizes key statistics from authoritative sources.

Table 2: Key cybercrime and cybersecurity statistics (2022–2024)

Metric	Value	Source
Global cybercrime cost (2025 projection)	$10.5 trillion/year	[7]
CAGR of cybercrime costs (2015–2025)	≈13.4% per annum	Eq. (2)
Malware attacks worldwide (2023)	6.06 billion (+10% YoY)	[8]
Phishing attacks detected (2022)	255 million	[9]
DDoS attacks mitigated by Cloudflare (Q1 2024)	4.5 million (+50% YoY)	[10]
MITM attacks: share of exploitation activity	∵35% of all exploitation	[11]
Unfilled cybersecurity jobs worldwide (2021)	3.5 million	[7]
Average cost of a data breach (2024)	$4.88 million	[13]

4.2. Attack Vector Typology

Understanding the attack surface that ethical hacking must address provides essential context for the legal analysis that follows. Table 3 summarizes the principal attack categories.

Table 3: Principal cyberattack categories and associated statistics

Attack Type	Description	Key Statistic
Malware	Malicious software (ransomware, spyware, worms, Trojans) designed to infiltrate or damage systems	6.06 billion attacks (2023) [8]
Phishing	Deceptive communications designed to harvest credentials or install malware; 76% target credentials	255M attacks detected (2022) [9]
DDoS	Volumetric attacks rendering services inaccessible; DNS-based DDoS increased 80% YoY in Q1 2024	4.5M mitigated Q1 2024 [10]

Man-in-the-Middle	Interception of communications between parties for eavesdropping or data manipulation	35% of exploitation activity [11]
Injection attacks	Exploitation of input validation vulnerabilities: SQL, code, XML injection	OWASP Top 10 category [19]
Zero-day exploits	Attacks targeting undisclosed vulnerabilities for which no patch exists	Growing APT toolkit
Social engineering	Manipulation of individuals to divulge credentials or enable unauthorized access	74% of breaches involve human element [12]

4.3. The Economic Case for Ethical Hacking: a ROSI Model

The economic justification for ethical hacking can be formalized through a Return on Security Investment (ROSI) model. Where C_control is the engagement cost and R_pre, R_post are pre- and post-engagement risk levels in monetary terms:

$$ROSI(\%) = \left[\left(R_{pre} - R_{post} - C_{control}\right)/C_{control}\right] \times 100 \quad (3)$$

The expected monetary loss is expressed as:

$$\text{Expected Loss (EL)} = \text{P(breach)} \times \text{Cost\_of\_breach} \quad (4)$$

Applying empirical values from IBM [13] — average breach cost $4.88 million — and assuming a penetration test costing $50,000 reduces breach probability from 15% to 5% (a range consistent with findings from the Ponemon Institute's State of Cybersecurity report [25], which documents 10–12 percentage point risk reductions in organizations with mature penetration testing programs):

$$EL_{pre} = 0.15 \times \$4{,}880{,}000 = \$732{,}000$$

$$EL_{post} = 0.05 \times \$4{,}880{,}000 = \$244{,}000$$

$$ROSI = [(\$732{,}000 - \$244{,}000 - \$50{,}000)/\$\,50{,}000] \times 100 = 876\% \quad (5)$$

Important caveats apply. The 10-percentage-point reduction is an optimistic estimate; the Ponemon Institute [25] reports median risk reductions of 6–8 percentage points in practice, which would yield a ROSI of 388–632%. Furthermore, ROSI varies substantially by industry sector: healthcare and financial services organizations report higher breach costs and correspondingly higher ROSI values, while manufacturing organizations tend toward lower figures [13, 25]. Critically, this ROSI is entirely contingent on the legal validity of the engagement: a legally compromised engagement resulting in prosecution eliminates the return entirely and replaces it with catastrophic negative value. The legal framework is therefore not a compliance overhead but a precondition for value creation.

5. LEGAL FRAMEWORK ANALYSIS

The seven legal domains identified in the systematic review are illustrated in Figure 3. The following sections examine each domain with respect to: (a) the nature of the legal obligations imposed; (b) the principal risks of non-compliance; and (c) best-practice approaches identified in the literature.



Figure 3: Seven interconnected legal domains governing ethical hacking practice.

5.1. Authorization and Consent

Authorization is the sine qua non of ethical hacking. Without it, no technical expertise, professional certification, or beneficial intent provides legal protection. Smith et al. [6] establish the foundational principle: "the ethical hacker performs the same actions as the malicious attacker but does so with the knowledge and consent of the target organization, transforming an intrusion into a contracted service." This consent-based framework is codified in computer fraud statutes worldwide: the U.S. Computer Fraud and Abuse Act (CFAA, 18 U.S.C. § 1030), the UK Computer Misuse Act (CMA), and EU Directive 2013/40/EU all define unauthorized access as the primary element of criminal liability.

Authorization must be explicit, documented, and granted by a party with the legal authority to consent. This last requirement is frequently overlooked: an IT manager may authorize testing of systems owned, leased, or co-administered by third parties — cloud providers, shared-service organizations, software vendors — who have not themselves consented. Testing such systems without their consent may constitute unauthorized access even where the immediate client has formally approved the engagement. Practitioners must therefore conduct a systems ownership audit prior to commencing any engagement.

The authorization document — termed a Rules of Engagement (RoE) agreement or Penetration Testing Agreement - should specify: the identity of the authorizing party and their legal authority to consent; systems, networks, and data within scope; authorized testing methods; geographic location; testing window; emergency escalation procedures; and the practitioner's obligations regarding discovered vulnerabilities.

5.2. Scope Delimitation

The scope of an ethical hacking engagement defines the legal boundary between lawful testing and unlawful intrusion. Exceeding the defined scope - even inadvertently, through network pivoting or automated tool behavior - creates immediate criminal and civil liability. Pawlicka et al. [14] identify scope management as one of the primary ethical and legal dilemmas in penetration testing, particularly when expanded access would serve the client's security interests but has not been explicitly authorized.

The principle of minimal footprint - conducting testing with the minimum access necessary to achieve defined objectives — provides both a security best practice and a legal risk management strategy. This can be formalized as an optimization constraint: given testing objectives $O = \{o_1, o_2, ..., o_n\}$, the ethical hacker should select methodology M* minimizing the set of system resources accessed A(M) subject to all objectives remaining achievable:

$$M^* = argmin \vee A(M) \vee subject\text{to:} \forall o_i \in O, o_i is achievable under M \quad (6)$$

In practice, this means preferring targeted manual testing over broad automated scanning where objectives permit, and documenting the rationale for each system accessed.

5.3. Regulatory Compliance: GDPR and National Frameworks

Ethical hacking engagements are conducted within a dense regulatory environment. The General Data Protection Regulation (GDPR, Regulation (EU) 2016/679) [22] imposes comprehensive obligations on any party that accesses, processes, stores, or transfers personal data — obligations triggered even by the incidental discovery of personal data during a security test. Key obligations include: lawfulness and transparency (Article 5(1)(a)); purpose limitation (Article 5(1)(b)); data minimization (Article 5(1)(c)); and integrity and confidentiality (Article 5(1)(f)).

Hassan [15] argues that current GDPR guidance inadequately addresses whether incidental personal data access during authorized testing constitutes "processing" under Article 4(2). The European Data Protection Board (EDPB) Guidelines 01/2021 on personal data transfers acknowledge the complexity of security testing scenarios but do not provide definitive answers for penetration testing contexts, leaving practitioners in significant legal uncertainty. Table 4 summarizes principal statutory frameworks across major jurisdictions.

Table 4: Principal statutory frameworks governing computer access across selected jurisdictions

Jurisdiction	Primary Statute	Key Provision for Ethical Hacking
USA	CFAA (18 U.S.C. § 1030)	Prohibits unauthorized access; "authorization" central but undefined; 2022 DOJ policy [21] limits prosecution of good-faith research
UK	Computer Misuse Act 1990 (amended 2015)	Sections 1–3 criminalize unauthorized access; no statutory good-faith exemption
EU	Directive 2013/40/EU; NIS2 Directive 2022/2555 [23]	Harmonized unauthorized-access offences; NIS2 strengthens vulnerability disclosure obligations
Germany	StGB §§ 202a–202d	Criminalizes data espionage; broad scope may capture ethical testing acts; no specific safe harbour
Spain	Penal Code Arts. 197–199 (reformed 2023)	2023 reforms create explicit authorized security research space [16]

5.4. Data Privacy and Confidential Information Handling

The handling of personally identifiable information (PII) and confidential organizational data during ethical hacking engagements creates independent legal obligations under privacy law, contract law, and in some jurisdictions criminal law - obligations that operate in parallel with, and sometimes in conflict with, GDPR requirements. A practitioner who copies, retains, or discloses sensitive data beyond what is strictly necessary is potentially liable for breach of confidence, violation of privacy statutes, and breach of the engagement agreement.

Three specific scenarios merit attention. First, credential discovery: when a practitioner discovers stored plaintext passwords or authentication tokens during a test, they must navigate the tension between documenting the vulnerability (which may require retaining evidence) and the legal obligation to minimize data retention. Best practice, endorsed by CREST [26] and PTES (Penetration Testing Execution Standard), requires that discovered credentials be documented by hash or partial disclosure only, never retained in cleartext. Second, personal data in test databases: organizations frequently use production data copies in test environments; practitioners encountering such data must apply the same GDPR minimization obligations as if working in the production environment directly. Third, incidental discovery of third-party data breaches: a practitioner who discovers evidence of a pre-existing breach by a malicious third party faces potentially conflicting obligations - notification duties under GDPR Article 33 may require client disclosure within 72 hours, but the practitioner's own obligations regarding the pre-existing breach remain legally unclear.

Best practice requires pre-engagement documentation of data handling procedures, technical measures to minimize data retention (wiping tools and storage media at engagement conclusion), and addressing discovered sensitive data through the reporting process rather than retaining it as evidence. The CREST Code of Conduct [26] and ISC(2) Code of Ethics [27] both explicitly address these obligations, though neither has force of law.

5.5. Intellectual Property Rights

Security evaluations frequently require analysis of proprietary source code, binary software, network architecture documentation, and other intellectual property. Practitioners who access, copy, or reverse-engineer such material face potential liability under multiple legal frameworks:

> Copyright law: unauthorized reproduction or adaptation of software code. The EU Software Directive (2009/24/EC) Article 6 provides a limited reverse engineering exception for interoperability purposes, but this exception has been interpreted narrowly by EU courts and does not straightforwardly extend to security testing.
>
> Trade secret law: misappropriation of confidential technical information. The EU Trade Secrets Directive (2016/943/EU) provides protection for information with commercial value that is kept secret; security testing of proprietary systems frequently involves accessing such information.
>
> Patent law: in jurisdictions where software patents are enforceable (primarily the United States), certain security testing techniques — particularly fuzzing and vulnerability scanning algorithms - may implicate patent rights.

The non-disclosure agreement (NDA) — identified by Yaacoub et al. [17] as a near-universal instrument in professional engagements — provides contractual protection against inadvertent disclosure but does not resolve the underlying intellectual property rights questions. A practitioner who reverse-engineers client code to identify a vulnerability may have a contractual right to do so under the engagement agreement, but the engagement agreement cannot override third-party intellectual property rights if the client's own software incorporates licensed components.

In the United States, the Digital Millennium Copyright Act (DMCA) § 1201 security research exemption (17 U.S.C. § 1201(j)) provides a limited safe harbour for good-faith security testing, but requires that the testing be conducted on lawfully accessed systems and that results be used solely to promote security. Practitioners must ensure their activities meet all conditions of this exemption; failure to do so exposes them to DMCA liability independent of any criminal computer fraud exposure. Retention of client code or trade secrets beyond the period necessary for report preparation creates ongoing legal exposure independent of the engagement agreement's confidentiality provisions, and engagement agreements should specify a clear data destruction protocol with timelines and verification procedures.

5.6. Documentation, Reporting, and Liability Management

Comprehensive documentation of all activities performed during an ethical hacking engagement — every system accessed, every tool used, every vulnerability identified, and every remediation recommendation made — serves multiple critical legal functions. It constitutes evidence that the practitioner operated within the authorized scope; it demonstrates due diligence in the identification and communication of security risks; and it provides the basis for any subsequent dispute resolution [18]. Detailed documentation shifts the burden of proof: in any allegation that the practitioner exceeded their authorization, contemporaneous records of actions taken are the primary legal defense.

Liability waivers and indemnification clauses in engagement agreements serve a complementary protective function. A well-drafted indemnification clause allocates responsibility for specific loss categories — for example, collateral system damage caused by authorized testing — to the client, on the basis that the client authorized the risk-creating activity. The enforceability of such clauses varies significantly across jurisdictions: common law jurisdictions generally enforce indemnification clauses except for gross negligence or willful misconduct, while civil law jurisdictions (including most EU

member states) impose mandatory minimum liability standards that cannot be contractually excluded. Practitioners should obtain independent legal advice on their standard engagement agreements in each jurisdiction in which they operate.

6. EMERGING LEGAL MECHANISMS

6.1. Bug Bounty Programs

Bug bounty programs represent the most significant institutional innovation in the legal governance of ethical hacking in the past decade. By creating formal, publicly defined frameworks for vulnerability submission and compensation, these programs extend a structured authorization to security researchers that mitigates the core legal risks identified in Section 5. Del-Real and Rodriguez Mesa [16] identify three distinct program categories: (1) institutional programs operated directly by target organizations (e.g., Google Project Zero, Microsoft Bug Bounty); (2) platform-intermediated programs through specialist intermediaries (HackerOne, Bugcrowd, Synack); and (3) private vulnerability brokers, which purchase and resell vulnerability intelligence at higher reward rates but raise significant ethical and legal questions regarding end use.

The legal significance of bug bounty programs extends beyond their immediate authorization. By establishing explicit, publicly documented rules of engagement, they reduce legal uncertainty that otherwise characterizes security research. The authorizing organization's commitment to good-faith review and compensation creates a reciprocal obligation of good faith on the researcher, and structured disclosure timelines limit risks associated with independent vulnerability publication. Importantly, participation in a recognized bug bounty program has been cited in U.S. federal court proceedings as evidence of good faith that mitigated prosecutorial exposure under the CFAA, though no court has yet held that bug bounty participation provides an absolute defense [16, 21].

6.2. Responsible Disclosure Protocols

Responsible disclosure - also termed coordinated vulnerability disclosure (CVD) - refers to the practice of reporting vulnerabilities to the affected organization privately before public disclosure. The ISO/IEC 29147:2018 standard provides a widely adopted CVD framework. From a legal perspective, documented adherence to a structured disclosure protocol provides evidence of good faith that may be decisive in the event of a legal challenge.

Legal protection afforded by responsible disclosure varies significantly by jurisdiction. The Netherlands (2013 CVD guidelines), France (ANSSI guidance, 2021), and Belgium (CCB framework, 2022) have enacted statutory or regulatory safe harbours for researchers complying with defined disclosure protocols. The United Kingdom and the United States offer no such explicit statutory protection, leaving researchers reliant on prosecutorial discretion. This absence of harmonized disclosure-protection frameworks represents one of the most significant remaining legal gaps.

6.3. Legislative Protections for Good-Faith Security Research

A growing number of jurisdictions are extending statutory protection to ethical hackers operating in good faith. The U.S. DOJ's 2022 revised CFAA enforcement policy [21] announced that good-faith security research would no longer be prosecuted. While this policy does not provide statutory immunity and may be revised by subsequent administrations, it represents a significant practical shift. Spain's 2023 penal code reforms [16] explicitly carve out space for authorized security research. The NIS2 Directive [23] mandates vulnerability disclosure obligations that create indirect incentives for more systematic use of ethical hacking across critical sectors. The EU Cyber Resilience Act (CRA), adopted in 2024, further introduces coordinated vulnerability disclosure requirements for manufacturers of products with digital elements, creating additional formal space for security research.

7. DISCUSSION

7.1. Cross-Jurisdictional Fragmentation

The most systemic legal challenge identified across the reviewed literature is jurisdictional fragmentation. An engagement conducted by a practitioner in one country, targeting systems hosted in a second, on behalf of a client incorporated in a third, may simultaneously engage three distinct computer fraud statutes, two data protection regimes, and multiple intellectual property frameworks - none designed for cross-border security testing. Pawlicka et al. [14] characterize this as a fundamental structural problem that individual practitioners cannot resolve through more careful contracting alone.

The gap is particularly acute at the GDPR/cybercrime statute intersection. An ethical hacker who discovers a vulnerability involving EU residents' personal data must simultaneously navigate the GDPR's 72-hour notification obligation (Article 33), the computer fraud statute of the jurisdiction where the system is hosted, and their engagement agreement - which may impose conflicting timelines and disclosure requirements. The Budapest Convention on Cybercrime [20] provides a model for cross-border cooperation but addresses law enforcement, not the legal status of authorized security research; a supplementary protocol addressing the conditions for lawful authorized security research would substantially reduce current uncertainty.

7.2. Standardization and Professional Certification

Existing certifications (CEH, OSCP, GPEN) provide evidence of technical competence but do not impose legally significant obligations on practitioners. No current certification standard creates binding requirements for engagement documentation, scope management, or post-engagement data handling. A more robust standardization framework - developed through collaboration between ISO/IEC, CREST [26], and national cybersecurity agencies - would define minimum documentation standards, establish binding legal obligations for certified practitioners, and create an enforceable disciplinary mechanism. Such a framework would give practitioners, clients, regulators, and courts a coherent reference standard.

7.3. Counterarguments and Limitations of the Proposed Framework

Intellectual honesty requires acknowledging significant counterarguments to the paper's core proposals. First, regarding harmonization: some legal scholars argue that cross-jurisdictional diversity in computer fraud law is not a bug but a feature, reflecting legitimate differences in national security policy, democratic values, and institutional capacity. A harmonized framework risks encoding the preferences of technologically powerful states at the expense of regulatory sovereignty [28]. This objection has particular force in the EU-U.S. context, where fundamental differences in the treatment of state surveillance and individual privacy rights complicate any mutual recognition framework.

Second, regarding standardization: over-regulation of ethical hacking methodologies could stifle the methodological innovation that makes security research effective. Adversaries do not operate within standardized frameworks; requiring defenders to do so may systematically disadvantage them. A rigid certification-based system might also exclude capable independent researchers who lack resources to obtain formal credentials, narrowing the talent pool at precisely the moment when the cybersecurity workforce gap is most acute [7].

Third, regarding the ROSI model: the model presented in Section 4.3, while grounded in empirical breach cost data [13, 25], makes several simplifying assumptions that may not hold in practice. The assumption of a fixed breach probability reduction (15% to 5%) obscures the substantial variance across organizations, sectors, and threat environments. The model also does not account for the cost of false positives (identified vulnerabilities that do not represent genuine risks) or the opportunity cost of operational disruption during testing. Future research should develop more granular, sector-specific ROSI models that account for this variance.

Fourth, the present systematic review, while following PRISMA methodology, is subject to publication bias: studies finding positive legal outcomes for ethical hacking are more likely to be published than those documenting prosecutions or civil actions. The legal risk landscape may therefore be more adverse than the reviewed literature suggests.

7.4. Artificial Intelligence and the Future Legal Landscape

The integration of AI into penetration testing tools raises novel legal questions the current framework cannot address. Automated platforms that dynamically generate and execute attack strategies challenge the concept of a defined, predictable scope: if an AI-driven tool autonomously discovers and pursues attack paths not anticipated by the engagement agreement, it is legally unclear whether this constitutes authorized testing or unauthorized intrusion under current doctrine in most jurisdictions.

A second AI-related challenge concerns accountability. Where an automated tool causes unintended system damage, the tort question of who bears responsibility - the practitioner, the tool developer, or the client - may lack a clear answer under existing liability frameworks. The rapid deployment of AI capabilities in security testing creates an urgent need for regulatory guidance, AI-specific liability clauses in engagement agreements, and industry self-regulation before a significant incident forces adverse judicial or legislative resolution.

7.5. Recommendations

For policymakers: Develop a multilateral legal framework — building on the Budapest Convention model — that explicitly addresses the conditions for lawful cross-border ethical hacking, including clear safe harbours for researchers complying with responsible disclosure protocols. Harmonize these protections at the EU level through the NIS2 Directive or a dedicated instrument. Provide specific GDPR guidance on incidental personal data access during authorized security testing, clarifying whether and under what conditions such access constitutes "processing."

For professional bodies: Develop and adopt a comprehensive professional standard establishing minimum requirements for authorization documentation, scope management, data handling, and post-engagement reporting. Create a binding code of professional conduct with an enforceable disciplinary mechanism. Develop AI-specific guidance for automated penetration testing tools, including requirements for pre-deployment scope validation.

For practitioners: Apply the minimal-footprint optimization principle - equation (6) - as a standard operational discipline. Treat engagement documentation as a legal record from day one. Obtain jurisdiction-specific legal advice for cross-border engagements. Invest in continuous legal education as a core professional competency. Where operating in the United States, ensure that all activities comply with the conditions of the DMCA § 1201(j) security research exemption.

8. CONCLUSIONS AND FUTURE RESEARCH

This systematic review has examined the legal dimensions of ethical hacking practice across seven interconnected domains. Three principal conclusions correspond to the research questions.

With respect to RQ1, the primary legal challenges are authorization and scope management, GDPR compliance, intellectual property protection, and liability allocation. These are not independent: a failure of authorization creates cascading liability across all other domains. The risk equation $R = P(T) \times V \times I$, applied to the legal domain, demonstrates that the legal vulnerability of an inadequately documented engagement is high and the potential impact - criminal prosecution, civil liability, reputational damage - is severe. The paper's novel contribution is the formal demonstration that these risks are structurally symmetrical to security risks and amenable to the same analytical framework.

With respect to RQ2, the most effective institutional mechanisms are structured engagement agreements with precise scope definitions, bug bounty programs with explicit authorization frameworks, and coordinated vulnerability disclosure protocols. Legislative developments - the U.S. DOJ's revised CFAA enforcement policy [21], Spain's 2023 reforms [16], the NIS2 Directive [23], and the EU Cyber Resilience Act — demonstrate gradual but real movement toward statutory accommodation of ethical hacking practice.

With respect to RQ3, cross-jurisdictional fragmentation remains the most significant systemic gap. The absence of harmonized responsible disclosure protections creates substantial uncertainty for practitioners operating globally, and AI-driven penetration testing tools compound this by challenging the concept of predictable scope. The counterarguments presented in Section 7.3 - particularly the sovereignty objection to harmonization and the innovation objection to standardization - indicate that proposed solutions must be carefully calibrated to avoid overcorrection.

Future research should address four specific gaps. First, empirical studies measuring the actual incidence of legal challenges against ethical hackers - as opposed to the legal risks identified theoretically - are largely absent from the literature; such studies would substantially strengthen the evidence base. Second, sector-specific ROSI models that account for variance in breach costs and risk reduction across industries would improve the economic case for ethical hacking. Third, legal analysis of specific AI-generated attack paths and their treatment under existing computer fraud doctrine is urgently needed as autonomous testing tools enter commercial deployment. Fourth, comparative legal analysis of responsible disclosure regimes across more jurisdictions - particularly in the Global South, which is underrepresented in the current literature — would advance the case for, and inform the design of, an international harmonization framework.

9. ACKNOWLEDGEMENTS

Parts of this research were funded by The Ministry of Science, Technological Development and Innovation of the Republic of Serbia, Contract no. 451-03-137/2025-03/200108.

10. REFERENCES

[1] Nishadhi, Nimesha. (2020). Ethical Hacking As A Method To Enhance Information Security. Cyber attack protection methodology, https://www.researchgate.net/publication/341654878_Ethical_Hacking_As_A_Method_To_Enhance_Information_Security_Cyber_attack_protection_methodology/citation/download, (2020)

[2] Stella Tonye, W., Bayo Olushola, O., & Steven Vile,, C. (2025). THEORETICAL FOUNDATIONS OF ONLINE IDENTITY SECURITY. In American International Journal of Computer Science and Information Technology (Vol. 10, Number 4, pp. 1–12). Zap journals. https://doi.org/10.5281/zenodo.16759627

[3] G. Thomas, O. Burmeister, G. Low. "Issues of implied trust in ethical hacking", ORBIT Journal, Vol. 2(1), https://doi.org/10.29297/orbit.v2i1.77, (2018)

[4] R. Mukhopadhyay and A. Nath, "Ethical hacking: Scope and challenges in the 21st century", International Journal of Innovative Research in Advanced Engineering, Vol. 1(1), pp. 30-37, (2014)

[5] G. Vishnuram, K. Tripathi and A. Kumar Tyagi, "Ethical hacking: Importance, controversies and scope in the future", Proceedings of the 2022 International Conference on Computer Communication and Informatics (ICCCI), pp. 01-06, https://doi.org/10.1109/ICCCI54379.2022.9740860, (2022)

[6] B. Smith, W. Yurcik and D. Doss, "Ethical hacking: The security justification redux", Proceedings of the 2002 International Symposium on Technology and Society, pp. 374-379, https://doi.org/10.1109/ISTAS.2002.1013840, (2002)

[7] Cybercrime Magazine, "2023 Cybersecurity Boardroom Report", Cybersecurity Ventures, https://cybersecurityventures.com/cybersecurity-boardroom-report-2023/, (2023)

[8] Statista, "Number of malware attacks per year worldwide from 2015 to 2023", https://www.statista.com/statistics/873097/malware-attacks-per-year-worldwide/, (2024)

[9] SlashNext, "The State of Phishing 2022", https://www.prnewswire.com/news-releases/slashnexts-state-of-phishing-report-reveals-more-than-255-million-attacks-in-2022-signaling-a-61-increase-in-phishing-year-over-year-301659518.html, (2022)

[10] Cloudflare, "DDoS Threat Report for 2024 Q1", https://blog.cloudflare.com/ddos-threat-report-for-2024-q1, (2024)

[11] A. Magnusson, "Man-in-the-Middle (MITM) Attack: Definition, Examples and More", StrongDM, https://www.strongdm.com/blog/man-in-the-middle-attack, (2024)

[12] Verizon, "Data Breach Investigations Report 2024", https://www.verizon.com/business/en-gb/resources/reports/dbir/, (2024)

[13] IBM Security, "Cost of a Data Breach Report 2024", IBM Corporation, https://wp.table.media/wp-content/uploads/2024/07/30132828/Cost-of-a-Data-Breach-Report-2024.pdf, (2024)

[14] A. Pawlicka, M. Choras, R. Kozik and M. Pawlicki, "First broad and systematic horizon scanning campaign and study to detect societal and ethical dilemmas and emerging issues spanning over cybersecurity solutions", Personal and Ubiquitous Computing, Vol. 27, pp. 193-202, https://doi.org/10.1007/s00779-020-01510-3, (2023)

[15] M. Hassan, "Ethics and digital data privacy", ResearchGate, https://doi.org/10.13140/RG.2.2.33972.68486, (2020)

[16] C. Del-Real and M.J. Rodriguez Mesa, "From black to white: The regulation of ethical hacking in Spain", Information & Communications Technology Law, Vol. 32(2), pp. 207-239, https://doi.org/10.1080/13600834.2022.2132595, (2023)

[17] J.P. Yaacoub, H. Noura, O. Salman and A. Chehab, "A survey on ethical hacking: Issues and challenges", ResearchGate, https://doi.org/10.13140/RG.2.2.11495.37285, (2021)

[18] Akintayo, Obafemi & Ifeanyi, Chinazo & Onunka, Okeoma. (2024). Resolving cybersecurity disputes and protecting digital infrastructure With ADR. Computer Science & IT Research Journal. 5. 2437-2457. 10.51594/csitrj.v5i10.1674.

[19] " OWASP Top 10: Injection – What It Is And How To Protect Your Applications" - BLOG, https://cyolo.io/blog/owasp-top-10-injection-what-it-is-and-how-to-protect-our-applications, (2023)

[20] Council of Europe, "Convention on Cybercrime (Budapest Convention)", ETS No. 185, https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/185, (2001)

[21] U.S. Department of Justice, "Department of Justice announces new policy for charging cases under the Computer Fraud and Abuse Act", https://www.justice.gov/opa/pr/department-justice-announces-new-policy-charging-cases-under-computer-fraud-and-abuse-act, (2022)

[22] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 — General Data Protection Regulation", Official Journal of the European Union, L 119, pp. 1-88, (2016)

[23] European Parliament and Council of the European Union, "Directive (EU) 2022/2555 — NIS2 Directive", Official Journal of the European Union, L 333, pp. 80-152, (2022)

[24] G. Fawkes, "The ultimate guide to ethical hacking: What you need to know in 2024", VPNmentor, https://www.vpnmentor.com/blog/ultimate-guide-to-ethical-hacking/, (2024)

[25] Ponemon Institute, "State of Cybersecurity and Third-Party Remote Access Risk", Ponemon Institute LLC, (2023)

[26] CREST, "CREST Code of Conduct for Penetration Testing", CREST International, https://www.crest-approved.org/wp-content/uploads/2023/04/A-Guide-to-Penetration-Testing-2022.pdf?ver, (2023)

[27] ISC2, "Code of Ethics", International Information System Security Certification Consortium, https://www.isc2.org/ethics, (2024)

[28] D. Svantesson, " Extraterritoriality and targeting in EU data privacy law: the weak spot undermining the regulation", International Data Privacy Law, Volume 5, Issue 4, November 2015, Pages 226–234, https://doi.org/10.1093/idpl/ipv024, (2015)

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Real-time object recognition and audio feedback system for visually impaired individuals

Gökmen Katipoğlu[1*], Semih Utku[2]

[1] Kafkas University, Engineering Faculty/Computer Engineering, Kars, Türkiye

[2] Dokuz Eylül University, Engineering Faculty/Computer Engineering, İzmir, Türkiye

ARTICLE INFO

* Correspondence: gokmenkatipoglu@kafkas.edu.tr

DOI: https://doi.org/10.46793/ET26.D08K **PAGES:** D69 – D74

ABSTRACT

This study presents a real-time object recognition system specifically designed to support visually impaired individuals by improving their independence and environmental awareness. Unlike general-purpose object detection systems, the proposed approach is designed with accessibility as a primary focus. It emphasizes intuitive interaction and clear auditory feedback to better support the needs of visually impaired users. The system combines computer vision, natural language generation, and speech synthesis within a mobile application. The system then provides immediate audio descriptions of nearby objects and their positions in a clear and understandable way. The architecture consists of a React Native-based mobile frontend and a Python-based backend deployed on cloud infrastructure. A pretrained YOLOv8 object detection model is utilized to identify objects in real time. Spatial analysis is performed to determine the relative position of objects within the scene. Experimental results demonstrate that the application achieves reliable object detection with low latency. Furthermore, the user-centered design, including bilingual support and touch-based interaction, enhances usability and accessibility. Overall, the proposed solution contributes to the field of assistive technologies by offering an efficient, cost-effective, and portable tool. It improves environmental awareness and supports independent mobility for visually impaired individuals.

KEYWORDS

YOLOv8, Computer Vision, Assistive Technology, Visual Impairment.

1. INTRODUCTION

According to global disability statistics, millions of visually impaired individuals encounter difficulties when interacting with their environment. Tasks such as identifying objects, recognizing obstacles, and understanding spatial relationships require visual perception and often limit independent mobility. Traditional assistive tools such as canes and guide dogs provide navigation support but cannot offer detailed contextual information regarding surrounding objects [1].

Recent advances in artificial intelligence and computer vision have enabled the development of intelligent assistive systems capable of interpreting visual information and presenting it through alternative sensory channels. Object detection models such as YOLO have demonstrated remarkable performance in real-time image understanding tasks [2][3].

This paper proposes a mobile-based assistive application specifically developed for visually impaired users. The system captures images through a smartphone camera, detects objects, determines their spatial positions, and generates spoken descriptions that help users understand their surroundings. Similar studies have shown that combining object detection and audio feedback can significantly improve environmental awareness and independent mobility for visually impaired individuals [4][5]. The motivation behind this project lies in the desire to improve the quality of life and autonomy for visually impaired individuals. Existing tools often fall short in providing comprehensive assistance, making it necessary to innovate a solution that can meet both navigation and object identification needs. By building a user-centered tool that leverages modern AI techniques, this project aims to fill these gaps and support users in their daily lives.

The importance of inclusive design and accessibility for all is also a driving factor, as it encourages the development of technology that addresses real-world needs. With a successful outcome, this project could inspire future applications that broaden access to vital technology for other differently-abled communities.

The main contributions of this study are:

i. Development of an accessibility-oriented mobile application.
ii. Integration of object detection, spatial reasoning, and audio feedback.
iii. Context-aware scene description generation.
iv. Real-time cloud-based deployment architecture.
v. Evaluation under realistic environmental conditions.

2. RELATED WORKS

Recent advances in computer vision and deep learning have significantly improved the performance of object detection systems [6][7]. Convolutional neural networks and modern object detection architectures have enabled real-time recognition of objects in complex environments, making them suitable for various assistive applications. In particular, the YOLO family of object detectors has gained considerable attention due to its ability to achieve high detection accuracy while maintaining low computational latency [2][3].

Object detection technologies have been successfully applied in various domains, including autonomous navigation, traffic sign recognition, surveillance systems, and healthcare applications [6][7]. The combination of computer vision and mobile computing has further enabled the development of lightweight and portable solutions capable of operating on smartphones and embedded devices [8].

Several studies have focused on developing assistive technologies for visually impaired individuals. Tapu et al. proposed a wearable system that combines object recognition and face detection techniques to provide audio-based environmental awareness for visually impaired users [1]. Several other studies have proposed image-processing-based assistive systems to support object recognition and environmental awareness for visually impaired users [4][9]. Although these solutions demonstrated promising results, many of them rely on specialized hardware or focus on limited use cases.

Recent developments in mobile computing have made it possible to deploy advanced object detection algorithms on portable devices. Such systems improve user independence by converting visual information into audio feedback and enabling real-time understanding of surrounding environments [4][5]. However, many existing approaches provide only object labels without contextual information regarding object locations and spatial relationships.

Unlike previous studies, the proposed system integrates real-time object detection, spatial reasoning, natural language generation, and speech synthesis within a single mobile application. By providing contextual audio descriptions of detected objects and their relative positions, the system aims to improve environmental awareness and independent mobility for visually impaired users.

3. SYSTEM ARCHITECTURE

The proposed architecture consists of four primary components:

1. Mobile Frontend
2. Object Detection Module
3. Scene Description Generator
4. Audio Feedback Module

The frontend is implemented using React Native and Expo, allowing users to capture images and receive spoken feedback. The backend is developed in Python and deployed using Docker containers on Hugging Face cloud infrastructure. The workflow of the system is as follows:

i. User captures an image.
ii. Image is sent to the backend API.
iii. YOLOv8 detects objects.
iv. Spatial positions are calculated.
v. Natural language descriptions are generated.
vi. Text is converted into speech.
vii. Audio feedback is returned to the user.

Figure 1 presents the class diagram of the proposed system and illustrates the relationships among its core software components. The architecture is organized around the AppController module, which coordinates the interaction between im-

age acquisition, object detection, and audio feedback processes. The CameraHandler component is responsible for capturing images from the mobile device, while the ImageProcessor module performs object detection using the YOLOv8 framework. The detected objects and their spatial information are then forwarded to the feedback generation module. The AudioFeedback class implements the FeedbackProvider interface and delivers spoken descriptions to the user. This modular structure improves maintainability, scalability, and integration between the system components.

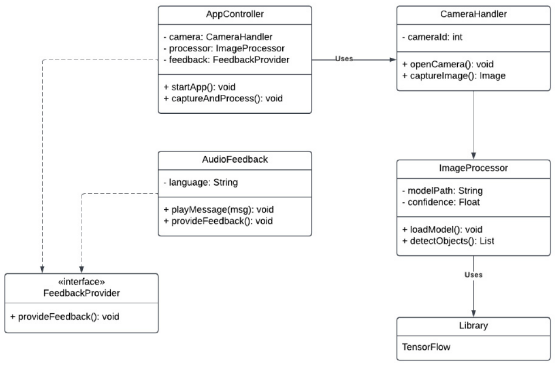


Figure 1: UML class diagram of the system

4. METHODOLOGY

4.1. Dataset and Object Detection

The proposed system was developed using a filtered subset of the Open Images Dataset V7, which contains a large collection of annotated images for object detection tasks. Only environmentally meaningful object categories were selected in order to improve usability and eliminate unnecessary information from the generated descriptions [10]. A pretrained YOLOv8 model provided by Ultralytics was employed for object detection due to its ability to achieve a balance between detection accuracy and real-time performance [11] . The use of a pretrained model eliminated the need for computationally expensive retraining while maintaining reliable detection performance. After image acquisition, the model identifies objects within the scene and produces bounding boxes, class labels, and confidence scores. Examples of object detection outputs generated by the proposed system are presented in Figures 3.

4.2. System Workflow

Detected objects are assigned to predefined spatial regions such as top-left, center, and bottom-right. This information provides users with orientation context and improves scene understanding. A custom description generation module creates human-readable scene summaries by combining detected object names with spatial information. Redundant object-part detections are removed to improve clarity. Generated descriptions are converted into speech using a Text-to-Speech engine and delivered to the user through the mobile application. The overall workflow of the proposed system is illustrated in Figure 2.

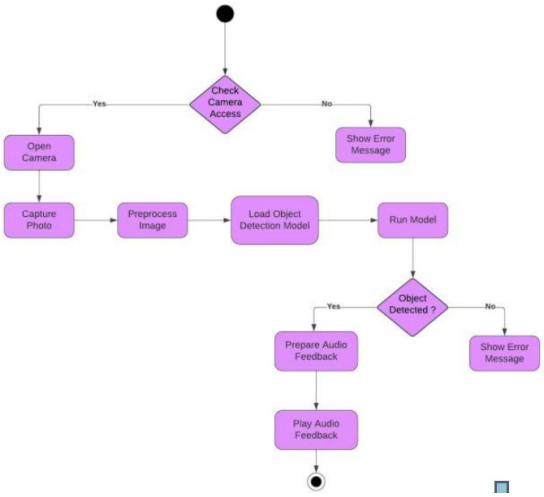


Figure 2: Workflow of the proposed object detection and audio feedback system

4.3. Detection Results

To demonstrate the effectiveness of the proposed approach, several sample images were processed using the object detection pipeline. The YOLOv8 model successfully identified multiple objects and generated meaningful descriptions for audio feedback. Figures 3 (a and b) demonstrate representative outputs obtained from the object detection module. The detected objects and their spatial locations are transformed into descriptive sentences and then synthesized into speech.

Recent studies have highlighted the effectiveness of combining computer vision and audio feedback technologies for assistive applications targeting visually impaired users [5].



(a) (b)

Figure 3: (a) Example of object detection results obtained using the YOLOv8 model

(b) Object detection and localization results in a complex scene

5. EXPERIMENTAL EVALUATION

To evaluate the effectiveness and practical usability of the proposed system, a series of experiments were conducted using real-world images captured under different environmental conditions. More than fifty images were processed through the complete system pipeline, including image acquisition, object detection, spatial analysis, description generation, and audio feedback. The test dataset contained indoor, outdoor, and urban scenes with varying levels of complexity, lighting conditions, and object density.

The evaluation focused on three main aspects: object detection accuracy, audio description quality, and overall user experience. Object detection accuracy was assessed by manually comparing the detected objects against ground-truth labels present in each scene. The number of correctly detected objects, missed detections, and false positives were examined to evaluate the reliability of the YOLOv8-based detection module.

In addition to detection performance, the quality of the generated audio descriptions was evaluated. The spoken outputs were reviewed with respect to clarity, grammatical correctness, timing, and contextual relevance. Particular attention was given to whether the generated descriptions accurately reflected both the detected objects and their relative positions within the scene. The description generation module successfully transformed structured detection outputs into concise and understandable sentences suitable for auditory presentation.

Figures 3 a and b illustrate representative examples of object detection and localization results produced by the proposed system. The results demonstrate that the model is capable of detecting multiple objects simultaneously while preserving spatial information required for generating contextual descriptions. Even in scenes containing several objects, the system was able to provide meaningful summaries without overwhelming the user with excessive information.

The responsiveness of the system was also evaluated. Experimental observations showed that the cloud-based architecture and optimized object detection pipeline enabled near real-time processing. The communication between the mobile application and the backend server remained stable throughout the experiments, allowing users to receive audio feedback shortly after image capture. This low-latency behavior is particularly important for assistive applications where timely information can directly affect user experience and environmental awareness. A summary of the experimental setup and observations is presented in Table 1. The results indicate that the proposed system can operate reliably under different environmental conditions while maintaining responsive audio feedback and effective object recognition performance.

Table 1: Summary of Experimental Evaluation

Evaluation Metric	Observation
Test Environments	Indoor, Outdoor, Urban
Object Detection Capability	Successful detection of commonly encountered objects
Audio Description Quality	Clear, understandable, and context-aware feedback

Evaluation Metric	Observation
System Responsiveness	Near real-time performance
Mobile–Cloud Communication	Stable during all experiments
Scalability	Modular architecture supports future extensions
Overall Performance	Suitable for assistive real-world applications

Overall, the experimental results indicate that the proposed system successfully detects commonly encountered objects and generates understandable audio descriptions under real-world conditions. The combination of object detection, spatial reasoning, and speech synthesis provides a practical assistive solution that can enhance environmental awareness and support independent mobility for visually impaired individuals. Decimal numbers

6. CONCLUSION

This paper presented a real-time object recognition and audio feedback system developed to support visually impaired individuals in understanding their surroundings. The proposed solution combines computer vision, object detection, spatial analysis, natural language generation, and speech synthesis within a mobile application framework. By transforming visual information into contextual audio descriptions, the system aims to improve environmental awareness and support independent mobility.

The developed architecture integrates a React Native-based mobile application with a cloud-based backend and a pre-trained YOLOv8 object detection model. Experimental evaluations conducted under different environmental conditions demonstrated that the system can successfully detect commonly encountered objects, determine their relative positions, and generate clear audio descriptions with low latency. The results indicate that the proposed approach is capable of providing meaningful and timely information to users while maintaining real-time performance.

In addition to its detection capabilities, the proposed system offers a practical and cost-effective assistive solution that operates using only a standard smartphone, eliminating the need for specialized hardware. The modular architecture also facilitates scalability and enables the integration of additional functionalities in future versions.

Future work will focus on expanding multilingual support, incorporating voice-command interaction, and improving scene understanding capabilities through more advanced object detection and vision-language models. Furthermore, future studies may include comprehensive user evaluations involving visually impaired participants to assess usability, accessibility, and real-world effectiveness in greater detail.

REFERENCES

1. A. Tapu, B. Mocanu and T. Zaharia, “Wearable Assistive Devices for Visually Impaired: A State of the Art Survey”, Pattern Recognition Letters, Vol. 137, pp. 37–52, (2020)
2. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), pp. 779–788, (2016)
3. Sohan, M., Sai Ram, T., & Rami Reddy, C. V. (2024). A Review on YOLOv8 and Its Advancements. In International Conference on Data Intelligence and Cognitive Informatics (pp. 529–545). Springer. https://doi.org/10.1007/978-981-99-7962-2_39
4. Manirajee, L., Shariff, S. H., Rashid, S. M., & Syar, S. (2024). Assistive Technology for Visually Impaired Individuals: A Systematic Literature Review (SLR). International Journal of Academic Research in Business and Social Sciences, 14(2), 596–611. https://doi.org/10.6007/IJARBSS/v14-i2/20827
5. Zafar, S., Asif, M., Ahmad, M. B., Ghazal, T. M., Faiz, T., Ahmad, M., & Khan, M. A. (2022). Assistive Devices Analysis for Visually Impaired Persons: A Review on Taxonomy. IEEE Access, 10, 13354–13366. https://doi.org/10.1109/ACCESS.2022.3146728
6. Y. LeCun, Y. Bengio and G. Hinton, “Deep Learning”, Nature 521, pp. 436-444, (2015)
7. Walle, H., De Runz, C., Serres, B., & Venturini, G. (2022). A Survey on Recent Advances in AI and Vision-Based Methods for Helping and Guiding Visually Impaired People. Applied Sciences, 12(5), 2308. https://doi.org/10.3390/app12052308
8. K. M. R. Leiva, M. A. C. Gato, and J.J.S. Olmedo, “Estimation of Spatio-Temporal Parameters of Gait and Posture of Visually Impaired People Using Wearable Sensors”, Sensors, Vol. 23(12), pp. 1–28, https://doi.org/10.3390/s23125564, (2023)
9. Kumar, N., & Jain, A. (2022). A Deep Learning Based Model to Assist Blind People in Their Navigation. Journal of Information Technology Education: Innovations in Practice, 21, 95–114. https://doi.org/10.28945/5006

10. A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov and V. Ferrari, “The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale”, International Journal of Computer Vision, Vol. 128(4), pp. 1956–1981, (2020)
11. G. Jocher, A. Chaurasia and J. Qiu, “YOLO by Ultralytics”, https://github.com/ultralytics/ultralytics, (2024)
12. Samavati, F. C., Abadi, A. R. G., & Samavati, F. C. (2025). Assistive Technologies for the Visually Impaired: A Review. Cureus Journals, 2(1). https://doi.org/10.7759/s44389-025-06891-1
13. S. Bagwan and S. Sankpal, “VisualPal: A Mobile Application for Object Recognition and Voice Assistance for Visually Impaired People”, International Journal of Engineering Research and Applications, Vol. 5(4), pp. 30–34, (2015)
14. S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, “Path Aggregation Network for Instance Segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 44(2), pp. 1025–1038, (2018)
15. M. A. Almarshad, S. Al-Ahmadi, and M. S. Islam“Adoption of Transformer Neural Network to Improve the Diagnostic Performance of Oximetry for Obstructive Sleep Apnea”, Sensors, Vol. 23(18), pp. 1–27, https://doi.org/10.3390/s23187924, (2023)

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Model of application of artificial intelligence in higher education based on interaction between user and computer system

Zoran G. Pavlović[1*], Ana Savić[2], Veljko Radičević[1], Rade Vasiljević[1]

[1] ATAAS, Departments School of Railroad Transport, Belgrade, Serbia

[2] ATAAS, Department of Electrical Engineering and Computer Science, Belgrade, Serbia

ARTICLE INFO

[*] **Correspondence:** zoran.pavlovic@vzs.edu.rs

DOI: https://doi.org/10.46793/ET26.D09P **PAGES:** D75 – D80

ABSTRACT

Modern higher education is increasingly dependent on digital technologies, large scientific databases, and intelligent information systems. Researchers and academic staff face the challenge of processing large volumes of information in a short period of time. Traditional approaches to searching scientific literature and analyzing research data often require significant time and effort. The rapid development of Artificial Intelligence (AI), especially generative AI systems such as ChatGPT and Gemini, has enabled a new approach to research and education. This paper presents a model of AI application in higher education based on the interaction between users and computer systems. The proposed model includes the transformation of traditional educational and research processes into digital processes supported by intelligent algorithms. The paper analyzes the role of AI systems in scientific research, digital learning, and academic communication. Particular attention is given to the advantages, limitations, and future perspectives of AI implementation in higher education institutions. The results indicate that AI platforms can significantly improve research efficiency, access to information, and the quality of education, while also requiring adequate digital competencies and ethical standards.

KEYWORDS

Artificial Intelligence, higher education, digital transformation, research, ChatGPT, interaction

1. INTRODUCTION

Digital transformation has become one of the most important development directions in higher education institutions worldwide. Universities and research institutions increasingly rely on digital technologies, online databases, cloud computing, and artificial intelligence systems to improve research and educational processes [1].

In recent years, generative AI systems have attracted considerable attention due to their ability to generate text, analyze data, process natural language, and support decision-making processes. Platforms such as ChatGPT, Gemini, Claude, and Microsoft Copilot are increasingly used in education and scientific research [2]. Currently, the most prominent platforms of table 1.

The application of AI systems in higher education enables researchers to process large amounts of information more efficiently and improve the quality of scientific work. However, the implementation of AI technologies also raises ethical and methodological issues related to academic integrity, data protection, and the reliability of generated information [3].

The aim of this paper is to present a model of AI application in higher education based on interaction between users and computer systems, with a focus on digital transformation and research efficiency.

Table 1: Overview of Modern AI Platforms

AI Platform	Company	Main Function	Application in Education and Research
ChatGPT	OpenAI	Natural language processing	Writing assistance, research support, content generation
Gemini	Google	Multimodal AI system	Data analysis, academic search, educational support
Claude	Anthropic	AI-based text analysis	Scientific writing and document processing

Microsoft Copilot	Microsoft	AI integration in Office tools	Administrative automation and analytics
Perplexity AI	Perplexity	Intelligent web search	Literature review and source identification
Elicit	Ought	Research assistant	Scientific paper analysis and references
Scite Copilot	Scite Inc	Citation analysis	Evaluation of scientific literature

The roles of people (students, teachers, researchers) are increasingly taking place through computer systems and advanced internet technologies. Every interaction with such a system can be considered the realization of a business process. In business processes, human activities are less and less represented. The realization of the needs of users, as well as employees in companies, can be achieved through the application of computer systems. The dimensions of context and situational parameters in communication processes are presented. Business modalities include traditional methods that are still used, the application of autonomous intelligent agents and the application of generative artificial intelligence agents that are currently occupying the academic community and increasingly business companies and organizations. The main goal includes the presentation of the advantages of the application of innovative digital technologies and the transformation that is inevitable. The semantic dimensions of the context with factors of communication dynamics are presented, where the advantages of the application from traditional to generative artificial intelligence agents are clearly visible [4].

2. ARTIFICIAL INTELLIGENCE IN HIGHER EDUCATION

Artificial Intelligence represents a branch of computer science focused on developing systems capable of performing tasks that normally require human intelligence. Modern AI systems are based on machine learning, deep learning, and natural language processing technologies [5]. AI technologies have become an essential component of modern educational systems. Their application includes:

- intelligent tutoring systems;
- personalized learning;
- automated evaluation systems;
- scientific research support;
- virtual learning environments;
- predictive analytics.

According to UNESCO reports, AI has the potential to significantly improve accessibility and quality of education through adaptive and personalized learning approaches [6].

One of the most important developments in recent years is the emergence of generative AI systems. These systems are capable of generating human-like text, summarizing documents, translating languages, and answering complex questions. In higher education, generative AI can support both teachers and students by automating routine activities and enabling faster access to information [7].

AI systems are particularly useful in scientific research because they can analyze large databases and identify relevant scientific references in a very short time. This capability is especially important for researchers dealing with multidisciplinary studies and rapidly evolving scientific fields.

Despite numerous advantages, AI implementation also presents several challenges:

- reliability of generated content;
- ethical concerns regarding plagiarism;
- data privacy and security;
- algorithm transparency;
- dependence on digital technologies.

Therefore, responsible use of AI systems in education requires the establishment of ethical standards and adequate user training.

3. MODEL OF USER–COMPUTER INTERACTION

The proposed model of AI application in higher education is based on continuous interaction between users and intelligent computer systems. The model includes several interconnected components:

- User;
- AI platform;
- Database system;
- Communication interface;
- Analytical module.

The user can be a student, teacher, or researcher. Through a communication interface, the user submits requests to the AI platform using natural language. The AI system processes the request, searches databases, analyzes information, and generates an appropriate response. The interaction process consists of several phases:

- identification of the research problem;
- formulation of user requests;
- data collection and analysis;
- generation of responses;
- evaluation of results;
- implementation of acquired knowledge.

The effectiveness of the interaction largely depends on the quality of user input and the digital competencies of the user. Properly formulated prompts significantly improve the quality of generated outputs [8].

Modern AI systems support multimodal interaction, enabling the processing of text, images, audio, and video content simultaneously. Such systems provide greater flexibility and improve communication efficiency between users and computer systems.

The model also supports integration with scientific databases, learning management systems, and cloud platforms. This integration enables faster access to scientific information and improves collaboration among researchers.

4. APPLICATION OF AI PLATFORMS IN SCIENTIFIC RESEARCH

Scientific research requires extensive literature review, data processing, and analysis of research results. Traditional research methods often involve time-consuming manual activities. AI systems significantly reduce the time needed for these processes [9]. Generative AI platforms are increasingly used in various phases of scientific work:

- literature review;
- data classification;
- statistical analysis;
- scientific writing;
- language translation;
- preparation of presentations and reports.

AI systems are especially useful in identifying relevant scientific references. Intelligent search systems can analyze thousands of scientific papers and extract the most relevant information related to a specific research topic.

ChatGPT and Gemini are among the most commonly used AI tools in academic environments. Researchers use these systems for generating research ideas, summarizing scientific articles, and improving academic writing quality [10].

AI technologies also contribute to improving communication and collaboration among researchers. Cloud-based AI systems allow simultaneous access to documents, collaborative editing, and automated content generation.

However, researchers must critically evaluate AI-generated information. AI systems may generate inaccurate or biased content, especially when dealing with incomplete or outdated data. Therefore, human supervision remains essential in scientific research.

5. ADVANTAGES AND LIMITATIONS OF AI SYSTEMS

5.1. Advantages

The tremendous progress in artificial intelligence (AI) opens up new opportunities for education, with intelligent teaching systems (ITS) powered by generative artificial intelligence (GenAI) emerging as a promising prospect. Therefore, state-of-the-art approaches with the integration of large language models (LLMs) are being explored to improve programming education. We investigate whether and how GenAI-based ITSs can effectively support students in learning programming skills. We observe the performance of three current LLM-user interface pairings: GPT-3.5 via ChatGPT, PaLM 2 via Google Bard, and GPT-4 via Bing. Evaluating LLMs on four types of problem settings in learning/teaching programming. Experimental results show that the use of generative AI, especially LLMs for programming, is promising, where GPT-3.5 gave the most satisfactory results [11]. The implementation of AI systems in higher education provides numerous benefits:

- increased efficiency;
- faster information processing;
- improved accessibility of educational resources;
- automation of administrative tasks;
- support for decision-making processes;
- enhanced digital competencies.

AI systems allow researchers and teachers to focus on creative and analytical activities while routine tasks are automated. Personalized learning environments also improve student engagement and learning outcomes [12].

5.2. **Limitations**

Virtual digital humans generated by artificial intelligence are profoundly changing digital innovation and user ecosystems. However, the dynamic evolution of their application scenarios, inherent conceptual ambiguities, and continuous advancement of technological features can all combine to generate negative user perceptions. It is imperative to systematically examine the overall impact, inherent technological limitations, and emerging challenges associated with virtual digital humans. Using the Theory, Context, Characteristics, and Metrics (TCCM) analytical framework, the study conducts a systematic review of a large number of authoritative journal articles published up to January 2025. The results reveal the negative effects of virtual digital humans generated by artificial intelligence on users and compare the relationship between hallucinations and random apparitions. In addition, the study further explores the dimensions of assessing these negative impacts [13]. Despite many advantages, AI systems also have limitations:

- inaccurate or misleading outputs;
- ethical concerns regarding academic integrity;
- lack of transparency in AI algorithms;
- cybersecurity risks;
- dependence on digital infrastructure.

One of the major concerns is the potential misuse of AI tools for plagiarism or unauthorized content generation. Educational institutions must therefore develop clear policies and ethical guidelines regarding AI usage.

Another limitation is the digital divide between users with advanced digital skills and those with limited technological competencies. Continuous education and professional training are necessary for effective AI adoption.

6. FUTURE PERSPECTIVES OF AI IN EDUCATION

Critical thinking is a key competency in higher education, especially in digital environments that require analysis, judgment, and informed decision-making. Generative Artificial Intelligence (GenAI) presents a new opportunity to enhance complex cognitive skills in educational settings. A key question may be: What is the impact of using GenAI tools on the development of critical thinking in university students? The study concludes that the intentional integration of GenAI technologies, when aligned with clear educational goals, can have a significant and positive impact on the development of critical thinking in university contexts. These findings offer empirical support for the design of innovative, ethically grounded learning experiences that incorporate GenAI to enhance cognitive development in higher education [14].

The future development of AI technologies will significantly influence higher education systems. Future AI applications are expected to include:

- advanced intelligent tutoring systems;
- adaptive learning environments;
- automated scientific evaluation;
- virtual and augmented reality integration;
- multimodal AI communication systems.

The integration of AI with learning management systems and digital libraries will further improve educational accessibility and research efficiency. Universities will increasingly rely on AI-driven analytics to improve institutional performance and educational quality. The integration of artificial intelligence (AI) into engineering higher education is becoming increasingly important today. The study contributes to the integration by providing a comprehensive overview of current research on the integration of AI into engineering higher education and by presenting a pilot introductory AI module designed to teach engineering students the basics of AI. Background: With the rapid development of AI, it is crucial to integrate AI into engineering curricula to prepare students for the workforce. However, comprehensive research on strategies for integrating AI into engineering higher education is lacking. Research Questions: What is the current state of AI integration into engineering higher education? What are the key considerations for integrating AI education into undergraduate engineering programs? What are the challenges and lessons learned when delivering AI modules to undergraduate electronics students? [15].

With the rapid development of generative artificial intelligence (AI), it is widely used in the field of higher education, especially in ideological and political education in universities. From the perspective of responsible innovation, the impact of generative artificial intelligence (AI) in universities is discussed. It empirically analyzes the practical application and effect of generative AI in universities through questionnaires, interviews and classroom observations. The results show that the introduction of generative AI improves students' acceptance and participation in the classroom, which positively affects the teaching effect. At the same time, the research also reveals the social, environmental and ethical responsibilities that should be paid attention to during the application of generative AI. By constructing a structural equation model, it further explores the relationship between the frequency of use of generative AI, students' acceptance and the diversity of

teachers' teaching methods and teaching effect, which provides useful reference and illumination for future educational reforms and technology integration. In addition, it also discusses the challenges and limitations of the application of generative artificial intelligence in university individual education and provides appropriate suggestions for promoting its effective application and development in the field of education [16].

Another important trend is the development of ethical AI frameworks in education. International organizations and academic institutions are already working on standards for responsible AI implementation.

Future educational systems will likely combine human expertise with AI support, creating hybrid educational environments that integrate traditional and digital learning approaches.

7. CONCLUSION

Artificial Intelligence has become one of the key technologies driving digital transformation in higher education. AI systems significantly improve research efficiency, information accessibility, and educational quality.

The proposed model of interaction between users and computer systems demonstrates how AI platforms can support scientific research and educational processes. Generative AI systems such as ChatGPT and Gemini provide new opportunities for digital learning, scientific collaboration, and automated information processing.

Despite numerous advantages, successful implementation of AI technologies requires adequate digital competencies, ethical standards, and responsible usage policies. Human supervision and critical thinking remain essential components of scientific research and education.

Future development of higher education will increasingly depend on the integration of intelligent systems and digital platforms into academic activities. AI technologies will continue to transform traditional educational models and contribute to the development of innovative learning environments.

REFERENCES

1. Quy, V.K.; Thanh, B.T.; Chehri, A.; Linh, D.M.; Tuan, D.A. AI and Digital Transformation in Higher Education: Vision and Approach of a Specific University in Vietnam. Sustainability 2023, 15, 11093. https://doi.org/10.3390/su151411093.
2. Munaye, Y.Y.; Admass, W.; Belayneh, Y.; Molla, A.; Asmare, M. ChatGPT in Education: A Systematic Review on Opportunities, Challenges, and Future Directions Algorithms, 2025, 18, 352, https://doi.org/10.3390/a18060352.
3. Kun Dai, Yabing Liu, Xiaofan Zhang, Generative AI in higher education: A bibliometric review of emerging trends, power dynamics, and global research landscapes,Computers and Education: Artificial Intelligence,Volume 10, 2026, 100544, ISSN 2666-920X, https://doi.org/10.1016/j.caeai.2026.100544.
4. Pavlović, Z. ., & Savić, A. . (2025). Context of Business Transformation: From Traditional to Generative AI Agents. E-Business Technologies Conference Proceedings, 4(1). Retrieved from https://ebt.rs/journals/index.php/conf-proc/article/view/266 (Original work published November 17, 2025)
5. R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 3rd ed., MIT Press, 2024. https://www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf 2018.
6. UNESCO, Guidance for Generative AI in Education and Research, Paris, France, 2024. https://school-education.ec.europa.eu/en/discover/publications/guidance-generative-ai-education-and-research.
7. Wang, J., Fan, W. RETRACTED ARTICLE: The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. Humanit Soc Sci Commun 12, 621 (2025). https://doi.org/10.1057/s41599-025-04787-y.
8. Garzón, J.; Patiño, E.; Marulanda, C. Systematic Review of Artificial Intelligence in Education: Trends, Benefits, and Challenges.Multimodal Technol Interact.2025, 9, 84 https://doi.org/10.3390/mti9080084.
9. Greussing, E., Guenther, L., Baram-Tsabari, A., Dabran-Zivan, S., Jonas, E., Klein-Avraham, I., Taddicken, M., Agergaard, T., Beets, B., Brossard, D., Chakraborty, A., Fage-Butler, A., Huang, C.-J., Kankaria, S., Lo, Y.-Y., Middleton, L., Nielsen, K. H., Riedlinger, M. and Song, H. (2025). Exploring temporal and cross-national patterns: the use of generative AI in science-related information retrieval across seven countries JCOM 24(02), A05. https://doi.org/10.22323/2.24020205.
10. Merino-Campos, C. The Impact of Artificial Intelligence on Personalized Learning in Higher Education: A Systematic Review.Trends Higs. Educ., 2025, 4, 17, https://doi.org/10.3390/higheredu4020017.
11. L. Frank, F. Herth, P. Stuwe, M. Klaiber, F. Gerschner and A. Theissler, "Leveraging GenAI for an Intelligent Tutoring System for R: A Quantitative Evaluation of Large Language Models," 2024 IEEE Global Engineering Education Conference (EDUCON), Kos Island, Greece, 2024, pp. 1-9, https://doi.org/10.1109/EDUCON60312.2024.10578933.
12. OECD, Artificial Intelligence and the Future of Skills and Education, OECD Publishing, Paris, 2025. https://www.oecd.org/en/about/projects/artificial-intelligence-and-future-of-skills.html,

13. Y. Yang, S. Li and S. Qiu, "A Systematic Literature Review on the Negative Impacts of AI-Generated Virtual Digital Humans," in IEEE Access, vol. 13, pp. 80047-80062, 2025, https://doi.org/10.1109/ACCESS.2025.3564041.
14. L. M. Oliva-Córdova, I. Álvarez-Icaza and C. E. George-Reyes, "Evaluation of Generative AI Use to Foster Critical Thinking in Higher Education," in IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 20, pp. 237-243, 2025, https://doi.org/10.1109/RITA.2025.3597848.
15. Y. Hao, Y. Liu, B. Liu, G. Amarantidis and R. Ghannam, "Integrating AI in Engineering Education: A Comprehensive Review and Student-Informed Module Design for U.K. Students," in IEEE Transactions on Education, vol. 68, no. 2, pp. 173-185, April 2025, https://doi.org/10.1109/TE.2025.3536105.
16. Y. Xing, "The Influence of Responsible Innovation on Ideological Education in Universities Under Generative Artificial Intelligence," in IEEE Access, vol. 12, pp. 133008-133017, 2024, https://doi.org/10.1109/ACCESS.2024.3459469.

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

ML algorithm for COVID-19 outcome prediction based on demographic features in North Macedonia

Maja Kukuseva Paneva[1*], Natasha Stojkovikj[1], Vasko Kokalanov[1], Vladimir Milićević[2]

[1] Goce Delcev University, Faculty of Computer Science, Stip, R. N. Macedonia

[2] University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Serbia

ARTICLE INFO

* Correspondence: maja.kukuseva@ugd.edu.mk

DOI: https://doi.org/10.46793/ET26.D10KP **PAGES:** D81 – D85

ABSTRACT

This study presents a machine learning approach for modeling and predicting COVID-19 patient outcomes in the Republic North Macedonia based on demographic characteristics. A decision tree-based classification model has been developed using algorithms implemented in WEKA environment. The model constructs the decision tree by selecting random subsets of features at each node and determines the optimal split points based on information gain.

The results indicate that age is the most significant predictor of patient outcome, serving as the root node and primary splitting attributes in the decision tree. Individuals below 60 years of age are classified as recovered, while higher age groups exhibit an increased probability of death. The gender attribute appears as a secondary factor, contributing to more refined predictions within specific age intervals. In addition to its predictive capability, the model offers a high level of interpretability, enabling clear visualization of decision rules and facilitating understanding of the relationship between demographic factors and disease outcomes. These findings highlight the importance of age- based risk stratification and demonstrate the potential of machine learning models as decision support tools in public health management.

KEYWORDS

Machine learning, COVID-19, Decision tree

1. INTRODUCTION

The outbreak of COVID-19 created significant challenges for healthcare systems not just in N. Macedonia, but worldwide due to its rapid transmission among the population. Since its emergence, COVID-19 has affected millions of individuals globally, [1], placing substantial pressure on the healthcare resources and emphasizing the need for efficient tools capable of supporting clinical decision- making patient management.

The integration of machine learning (ML) within healthcare system has fundamentally transformed clinical decision through analysis of heterogeneous datasets, pattern recognition and robust predictive modeling. Current algorithmic frameworks demonstrate high efficiency in diagnostic classification, patient risk stratification, and prognostic forecasting. Notably, during the COVID-19 pandemic, these predictive models served as vital epidemiological and clinical tools, enabling the early identification of high-risk individuals, [2], [3], [4]. Among different machine learning approaches, decision-tree based models, [5] are widely used due to its simplicity, interpretability and ability to represent decision making processes. These models operate through hierarchal framework consisting of root nodes, internal decision nodes, branch nodes and leaf nodes. Classification is executed by recursively partitioning the dataset according to selected attributes and decision rules. The main advantage of decision trees models is that they can handle numerical and categorical gata simultaneously without requiring complex preprocessing procedures, and can identify the importance among attributes, allowing researchers to determine the most likely outcome and which individuals will be more affected.

2. METHODOLOGY

In this paper, the effect of demographic features on the outcome of COVID-19 is analyzed. Demographic features considered in this study are age and gender, while the outcome is classified as recovered or death. The data set is obtained

from Institute of Public Health of N. Macedonia and consists of 500 infected individuals. The main objective of the analysis is to determine whether age and gender have a significant impact on the progression and outcome of the disease. For this purpose, statistical methods were applied on the data for processing and analysis. The data set does not have missing value. The gender and outcome are binary values. Within the gender attribute value 1 indicates male, whereas value 0 indicates female individual. Within the outcome attribute, a value of 0 indicates recovered, whereas value of 1 indicates death. Table 1 presents the basic descriptive statistics for the parameters including minimum and maximum values, mean value and standard deviation for the demographic features. The minimum and maximum values provide insight into the age range of the infected individuals, while the mean value represents the average age of infected individuals included in this study. The standard deviation indicates the variability and distribution of data around the mean value. These statistical indicators clarify the population's characteristics, helping determine how demographic features influence the outcome of COVID-19.

Table 1: Descriptive statistics of the dataset

No	Attribute	Minimal value	Maximum value	Mean value	Standard deviation
1	**Age**	**1**	**90**	**44.505**	**17.467**
2	**Gender**	**0**	**1**	**0.519**	**0.5**
3	**Outcome**	**0**	**1**	**0.93**	**0.26**

Table 2 represents a representative sample consisting of data from five infected individuals within the data set.

Table 2: Sample of the dataset for 5 infected individuals

Feature					
Age	65	49	67	63	77
Gender	0	0	1	0	0
Outcome	1	0	0	0	1

This study employs different machine learning (ML) algorithms in order to analyze the impact of demographic features in the outcome of COVID-19. A decision tree-based classification model was developed using Random Tree and Random Forest algorithms implemented in WEKA® environment. The random tree classifier constructs a single tree by selecting a random subset of features at each node and determines an optimal split points based on information gained. On the other hand, the Random Forest constructs a collection of independent trees and then aggregates their predictions by voting among all trees.

In order to evaluate the performance of the developed classification model, several evaluation metrics were utilized, including Accuracy, Precision, Recall, F1 score and confusion matrix. The confusion matrix, as shown in Figure 1, is a tabular representation of comparison of predicted values against actual outcomes with four key metrics. The matrix categorizes the modedl prediction into correct classifications (True Positive (TP) and True Negative (TN)), and misclassifications (False Positive (FP) and False Negative (FN)).

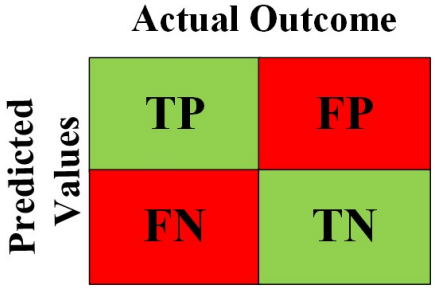


Figure 1: Confusion matrix

Accuracy is one of the most used metrics used to evaluate the overall performances of a classification model. It presents the ration of correctly identified predicted values both True Positive and True Negatives to the total dataset population and is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision is a critical metric that measures the ration of the model's true positive predictions to the total number of positive predictions made by the model as:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

In clinical predictive modeling, Recall (or Sensitivity) is a fundamental metric that measures the model's ability to identify all actual positive cases. It is calculated as ratio of correctly predicted positive values (TP) to the total number of actual positive outcomes in the total dataset population (TP and FN):

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

In this study, recall indicates the percentage of deceased individuals that the model successfully detected out of all individuals who actually died due to COVID-19. This means that a high Recall value ensures that the model is highly capable of capturing the critical outcomes, thereby minimizing False Negative (FN).

If the dataset outcome is imbalanced, relying only on overall accuracy can be misleading. That's why F1 score utilizes the calculation of the harmonic mean of Precision and Recall evaluating how effectively the model identifies the critical class (death). F1 score is calculated as:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall} \quad (4)$$

3. RESULTS AND DISCUSION

The prediction of COVID-19 infected individual outcomes based on demographic features was performed using Random Tree and Random Forest classification algorithms implemented in WEKA® machine learning environment,[8]. The main objective of the analysis was to classify the infected individuals into two outcome categories: Recovered (R) and Death (D) based on the two demographic features, Age and Gender. Future research in [2] considers incorporating additional clinical and laboratory variables, including comorbidities and chronic diseases.

These algorithms were selected because they are widely used in medical data analysis due to their ability to model non-linear relationships while maintaining a high level of interpretability. Furthermore, tree-based classifiers require minimal data preprocessing and can efficiently handle both numerical and categorical variables.

The Random Tree model achieved an overall accuracy of approximately 91%, while the Random Forest model achieved an accuracy of approximately 90.6%. The relatively high accuracy values indicate that both algorithms were able correctly to identify infected individuals who recovered from COVID-19, as indicated by the high precision, recall and F1-score values for the Recovered class, as represented in Table 3.

Table 3: Comparison of performance Metrics of Random Tree and Random Forest Model for Predicting Recovered Individuals

Model	Accuracy (%)	Precision	Recall	F1-score
Random Tree	91.0	0.882	0.910	0.893
Random Forest	90.6	0.879	0.906	0.891

The obtained results show that both machine learning models perform very well in identifying recovered individuals. The Random Tree classifier achieved slightly higher values of Accuracy, Precision, Recall and F1-score compared to Random Forest, although the differences between the two algorithms are relatively small. Both models correctly classified more than 90% of recovered infected individuals, indicating that demographic features alone contain sufficient for recognizing the majority outcome class. The Precision values of 0.882 for Random Tree and 0.879 for Random Forest indicate that the most infected individuals predicted as recovered were correctly classified. Similarly, the high Recall values (0.910 and 0.906) demonstrate that the majority of recovered infected individuals were successfully identified by both models. The F1-score, which combines Precision and Recall into a single metric, further confirms the excellent predictive performance for the Recovered class, with values of 0.89 for both algorithms. Although Random Tree achieved slightly better values in most evaluation metrics, the Random Forest obtained a higher ROC Areal value (0.751) than Random Tree (0.647) indicating a better ability to distinguish between the two outcomes classes. These results suggest that Random Tree and Random Forest can effectively predict recovery outcomes.

Table 4 presents the evaluation results focused on the Death class, which is clinically most important category in this dataset. Unlike overall accuracy, which can be dominated by the majority class (in this case Recovered), these metrics provide a clearer insight into how well the models are able to identify critical and high-risk outcomes. In medical prediction tasks, the class Death is particularly important because false negatives, cases where the individual is incorrectly predicted as Recovered instead of Death, can have a serious consequence. Therefore, metrics such as Recall and F1-score for the class Death are more informative than overall accuracy alone.

Table 4: Comparison of performance Metrics of Random Tree and Random Forest Model for Predicting Deceased Individuals

Model	Precision	Recall	F1-score
Random Tree	0.250	0.108	0.151
Random Forest	0.222	0.108	0.145

Both Random Tree and Random Forest models show very similar performance for the class Death with extremely low Recall values (0.108) and low F1-score value, 0.151 for Random Tree and 0.145 for Random Forest. Random Forest performs slightly better in terms of overall discriminative ability as it achieves higher ROC Area, 0.751 compared to 0.647 for Random Tree.

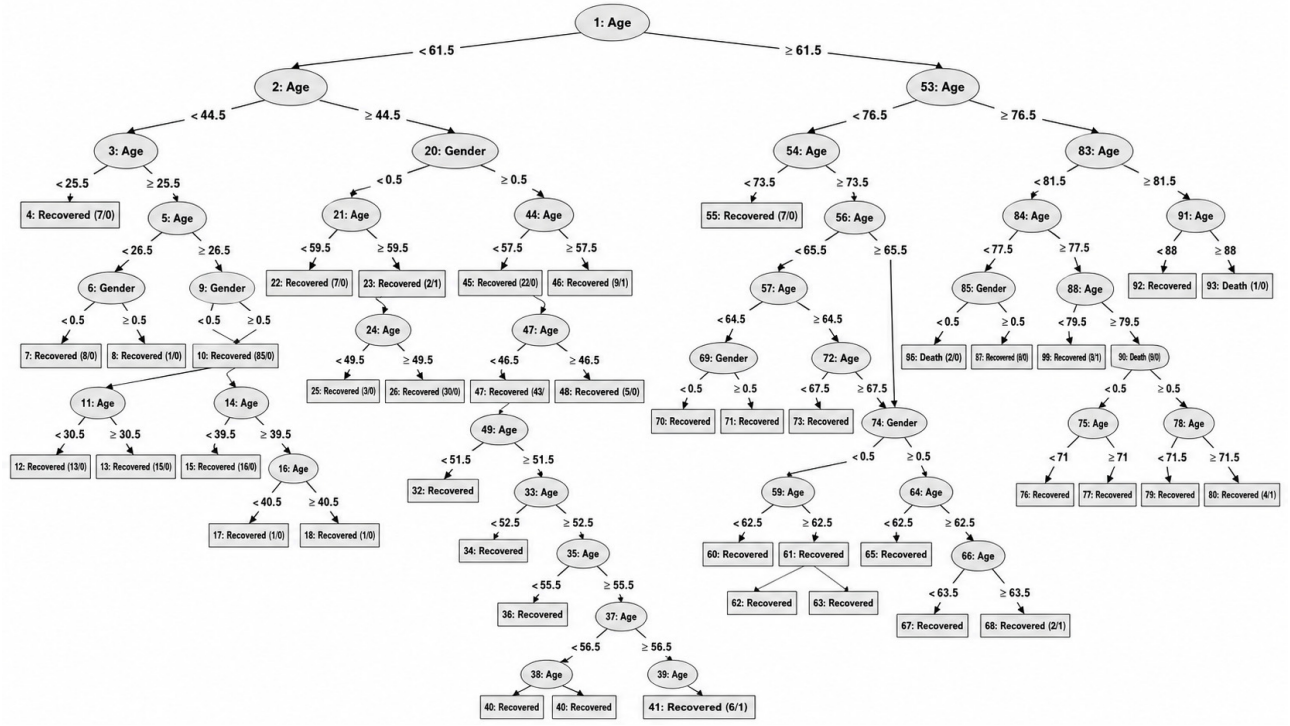


Figure 2. Decision tree generated using Random Tree

Figure 2 represents the decision tree generated using Random Tree algorithm for predicting COVID-19 outcome of infected individual based on demographic features Age and Gender. One of the main advantages of decision tree models is their interpretability, allowing researchers to clearly understand how different attributes contribute to the final prediction. Unlike black-box machine learning models, decision trees provide explicit decision rules that can be directly analyzed and interpreted. The generated tree reveals that Age is the most significant and dominant predictive attribute, as it appears as a root node and in the majority of subsequent decision nodes in the tree. This indicates that the age of infected individual has the strongest impact in determining whether the infected individual is classified as Recovered or Death. This finding is consistent with numerous epidemiological studies that have identified age as one of the strongest predictors of severe COVID-19 outcomes and mortality, [6]-[7]. The initial split of the dataset occurs at an age threshold of 61.5 years, dividing younger and older infected individuals into separate branches within the classification process. For individuals younger than 61.5 years, the model predominantly predicts the outcome Recovered. Sever additional age-based thresholds further refine the classification process. Despite the additional subdivisions, the majority of terminal nodes within this branch corresponds to recovery outcome of the infected individuals, suggesting that younger individuals generally exhibit a lower risk of mortality.

Gender appears as a secondary attribute in several branches of the tree, suggesting that although it contributes to the classification process, its influence is considerably lower compared to Age attribute. This observation indicates that gender contributes additional discriminatory information but possesses lower predictive importance compared to the age attribute. In the branches corresponding to older individuals, above 61.5 years, the tree becomes more complex, indicating a higher variability in outcomes among elderly population. Additional thresholds, such as 64.5, 67.5, 76.5, 79.5 and 81.5 years, are used to distinguish between different risk groups. Notably, the few terminal nodes classified as Death are located within these older age ranges, suggesting an increased mortality risk among elderly individuals within the population.

The decision tree contains a large number of leaf nodes classified as Recovered, while only a small number of leaves correspond to Death. This reflects the class distribution in the dataset, where recovered cases greatly outnumber death cases.

Overall, the decision tree analysis of Figure 2 confirms that age is dominant demographic factor affecting COVID-19 outcomes of infected individuals, while gender has a secondary demographic factor. The generated decision tree successfully captures the increased mortality rate associated with elderly infected individuals. Also, the model provides an interpretable representation of the decision-making process and highlights the increased risk associated with advanced age. Therefore, the random tree model not only achieves classification capability but also provides valuable insight into underlying factors influencing infected individual's prognosis.

4. CONCLUSION

This research investigates the potential of decision tree-based machine learning algorithms to predict COVID-19 outcome of infected individuals in R.N. Macedonia using only demographic features, namely age and gender. The obtained results demonstrate that both Random Tree and Random Forest algorithms achieved high overall classification accuracy of approximately 91%, indicating that demographic information alone provides valuable predictive information for identifying the majority outcome class.

The analysis revealed that age is the most significant predictor of COVID-19 outcome, while gender contributes as a secondary factor in specific decision branches. The generated decision tree clearly illustrates that individuals younger than approximately 60 years are predominantly classified as recovered, whereas the probability of death increases considerably among older age groups. This finding is consistent with previously reported epidemiological research identifying advanced age as one of the strongest risk factors for severe COVID-19 outcomes.

Overall, this research confirms that interpretable machine learning models such as Random Tree and Random Forest can serve as useful decision support tools for providing transparent classification rules and identifying the relative importance of demographic features. Furthermore, future research will focus on improving COVID-19 outcomes by incorporating additional clinical features including chronic disease and infected individuals' symptoms such as fever, shortness of breath, muscular pain, cough etc. The inclusion of these predictors is expected to increase the accuracy and reliability of the developed machine learning models.

REFERENCES

1. WHO COVID-19 dashboard, https://data.who.int/dashboards/covid19/cases?n=o
2. M.K. Paneva, C.M. Bande, N. Stojkovikj, D. Bikov, "Machine Learning Models for Prediction of COVID-19 Infection in North Macedonia, TEM Journal, Vol. 14(1), pp. 160-168, https://www.doi.org/10.18421/TEM141-15 , (2025)
3. D. N. Jayavelu, H. Samaha, S.T. Wimalasena et al. "Machine learning models predict long COVID outcomes based on baseline clinical and immunologic factors". Commun Med 6 (1). https://doi.org/10.1038/s43856-025-01230-w, (2026).
4. Y. Cheng, R. Cheng, T. Xu, X. Tan, Y. Bai, "Machine Learning Techniques Applied to COVID-19 Prediction: A Systematic Literature Review", Bioengineering (Basel), Vol. 12(5):514. https://doi.org/10.3390/bioengineering12050514, (2025).
5. Z. Mohammadi-Pirouz, K. Hajian-Tilaki, M.S. Haddat-Zavareh, et al. "Development of decision tree classification algorithms in predicting mortality of COVID-19 patients", Int J Emerg Med, Vol 17, 126 https://doi.org/10.1186/s12245-024-00681-7, (2024)
6. E.J. Williamson, A.J. Walker, K. Bhaskaran, et al. "Factors associated with COVID-19-related death using OpenSAFELY". Nature 584, 430–436, https://doi.org/10.1038/s41586-020-2521-4, (2020).
7. Y. Zhou, Q. Yang, J. Chi et al. "Comorbidities and the risk of severe of fatal outcomes associated with coronavirus disease 2019: A systematic review and meta-analysis", Internatinonal Journal of Infectious Diseases, Vol. 99, 47-56, https://doi.org/10.https://doi.org/10.1016/j.ijid.2020.07.029 , (2020).
8. https://www.weka.io/

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Drug clustering upon toxicity effects on liver spheroids – a deep learning approach

Miroslava Jordović Pavlović[1*], Tijana Geroski[2] , Milica Nikolić[1,3]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Kraljevo, Serbia

[2] University of Kragujevac, Faculty of Engineering, Kragujevac, Serbia

[3] University of Kragujevac, Institute for Information Technologies Kragujevac, Kragujevac, Serbia

ARTICLE INFO

* Correspondence: jordovicpavlovic.m@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.D11JP **PAGES:** D87 – D96

ABSTRACT

Motivation for this research stems from drug-induced liver injury (DILI) being a leading cause of drug failure and market withdrawal, which calls for reliable, non-invasive hepatotoxicity assessment. The basis for this research was the work performed by Dubinsky et al., who established a deep learning (DL) model to predict ATP values (non-invasive estimation of cell viability), assess damage to liver spheroids, and predict hepatotoxicity based on spheroid images. They created a public database containing approximately 20,000 brightfield microscopy images of human liver spheroids exposed over a period of up to 7 days to 108 drug compounds for assessing cell viability. Our work focused on clustering applied compounds by the intensity of damage caused to liver spheroids (level of predicted ATP values), forming groups of compounds that have similar effects on liver cells and potentially similar mechanisms of toxicity. Then we explored whether selected drugs belong to the same class of pharmacotherapeutic classification – analgesics, antipyretics, antibiotics, etc. The model can detect if some drugs behave differently than expected, leading to detecting errors and revealing potential new phenomenological causes. Further, we analysed the temporal dependence of drugs on cell viability (toxicity), providing insights into the dynamics of effects and toxicity progression over time. Future work will investigate drug effects in combination, since co-exposure of drugs can lead to different outcomes of hepatotoxicity, in comparison to individual exposure.

KEYWORDS

Liver spheroid, Hepatotoxicity, Deep learning, Clustering of drugs by toxicity effects, Toxicity dynamics

1. INTRODUCTION

The liver represents the main organ in charge of detoxification processes in the organism, responsible for processing ingested food, drinks and medications and, therefore, it is exposed to a potential harmful influence by these processed substances but also can be affected by environmental pollutants [1], [2]. Changing of human habits can significantly reduce the risk of harmful influences originating from food, drinks and environmental pollution, but adverse effects of medications are more difficult to avoid [3]. It is clinically confirmed that some drugs can cause hepatotoxicity [4], but there is also evidence of toxic effects of some drugs in a laboratory set up, affecting liver cells function and metabolisms, although they have not yet been confirmed as highly toxic drugs in clinical testing [5]. The early detection of detrimental effects is particularly important as well as analysis of applied drugs, specifically its molecular structure and chemical composition, that could be further linked to causing damage to the liver.

Since the goal is to eliminate or reduce liver toxicity caused by applied medications, researchers established a safer approach in studying the field - *in vitro* testing systems, such as liver spheroids and other liver tissue models [6]. Development and further improvement of the *in vitro* testing system, that can at a certain level recapitulate behaviour of living organs, significantly improved understanding of biological mechanisms in different organs, multi-organ systems, and testing and studying of mechanisms alerted by applied drugs. In that sense, drug testing became more effective and in accordance with the 3R principles (Replace, Reduce, Refine), which are part of the ethical guidance in animal research [7].

Hepatotoxicity stands for chemical-driven liver damage and can appear when the liver is exposed to substances like medications, alcohol, supplements or industrial (environmental) chemicals. In this paper our focus is on the analysis of toxic drug effects to the liver spheroids, therefore, drugs hepatotoxicity. Overall symptoms of injured and malfunctioned liver, such as nausea, vomiting, yellowing of the skin, swelling of the abdominal area, appear with the increased level of liver injury - middle to high [3], moving investigation of drugs hepatotoxicity to the early phase - analysis of drug toxic effects to the liver cells after exposure and assessing liver damage from low to middle and high levels in laboratory conditions [8]. Because it is difficult to detect and diagnose low level liver injury in patients, a list of well-known drugs that can have hepatotoxic effects, together with the list of drugs that cause liver damage in the laboratory conditions can be used to create a computational *in silico* model to draw a conclusion about specific molecular structure section and chemical composition responsible for these adverse effects, which could be further useful in analysing novel drugs hepatotoxicity levels computationally prior to experimental testing.

Literature survey revealed that further optimisation of testing drug safety is supported by development of *in silico* physics-based and AI-based computational models. Machine learning approaches have proven effective across a wide range of scientific and engineering domains [9], [10], [11] demonstrating their capability to model complex non-linear relationships in high-dimensional data. Geci et al. 2025 [12] integrated *in vitro* and *in silico* approaches to predict DILI upon drug exposures. They used an *in silico* model to predict pharmacokinetics of applied drugs and linked it to experimentally measured cytotoxicity, mitochondrial toxicity and bile salt export pump (BSEP) inhibition and obtained high DILI predictivity. A retrospective analysis of 241 drugs showed that the ratio of the in vivo maximum plasma concentration (Cmax) to the lowest *in vitro* toxicity strongly correlates with clinical DILI risks. Then they used the *in silico* model to predict Cmax and to analyse DILI levels, prospectively. Dynamic simulation through BSEP inhibition can identify drugs potentially causing DILI. Such an integrative, mechanistic approach enhanced DILI prediction and interpretability, offering an animal-free alternative for early drug development. Bergen et al. 2025 [13] analysed toxicogenomic resource for DILI prediction by developing ToxPredictor, a toxicogenomic framework combining RNA-seq data from primary human hepatocytes with pharmacokinetic data to predict dose-resolved DILI risks and safety margins, providing mechanistic insights into hepatotoxic pathways, enabling early de-risking and actionable safety decisions. The model flagged several compounds that later failed in late-stage clinical development because of liver toxicity, including drugs missed by traditional animal testing. This suggests human-cell transcriptomics may provide more clinically relevant toxicity signals than conventional preclinical approaches. Xiao et al. 2025 [14] also used toxicogenomic data but linked them to protein–protein interaction networks, explicitly modelling biological relationships among genes. The results obtained from this research have raised a significant observation that incorporating known molecular pathways can significantly improve prediction of hepatotoxicity and strengthen biological interpretability. Han et al. 2025 [15] created a graph neural network (GNN) model and combined it with foundation-model representations of molecules, which was pre-trained with approximately 10 million chemicals in the PubChem database [16]. They showed that the model can predict hepatotoxicity and that resulting platform enables both toxicity prediction and mechanistic exploration. Malka-Markovitz et al. 2025 [17] developed physics-based liver model to analyse DILI processes that integrates patient-specific anatomical data to predict acetaminophen-induced liver injury as a demonstration of its capability - a multi-scale model bridging vascular, lobular and cellular scales to simulate dynamic blood flow, drug transport, and injury mechanisms that accurately reflect clinically observed spatial heterogeneity. Even though AI models are these days in the focus of researching hepatotoxicity, physics-based models are still dominant when it comes to mechanistic explanation, since they can properly address that gap by connecting molecular events, cellular damage, tissue architecture, and organ-level outcomes within one framework, improving confidence in model predictions and supporting regulatory acceptance.

In this paper, we created a multimodal explainable AI model for hepatotoxicity prediction standing between two major directions in drug hepatotoxicity research - image-based toxicity prediction and chemoinformatic toxicity approach, as we combined both in one analysis. Image-based analysis and open-source data provided by Dubinsky et al. 2026 [1] enable biological aspects for the analysis closer to reality because many DILI mechanisms emerge only after interaction with liver cells. Our research doesn't analyse gene expression changes, but morphological changes which are more clinically observable and sensitive to downstream cellular damage. In comparison against Physiologically Based Pharmacokinetic (PBPK)/mechanistic models, our approach learns kinetics from features, which is a lack of the model and might be improved in future research. By fusing features extracted from images with physicochemical features extracted for applied drugs to the liver spheroids in our machine learning (ML) framework, our goal is to observe from images what happened to the liver spheroids and then analyse physicochemical features to explain why this happened - identify the mechanisms behind the cellular processes and sequence of drug molecular structure and chemical composition responsible for it. Further we identify toxic effects by drug groups and explore whether a related group of drugs develops the same mechanism of toxicity upon assessment of cell viability in liver spheroids.

2. METHODOLOGY

To create a ML framework that captures both morphological changes in liver spheroids and physicochemical properties of applied drugs we started with research conducted by Dubinsky et al. 2026 [1], since they created a public database containing approximately 20,000 brightfield microscopy images of human liver spheroids exposed to 108 drugs over a period of up to 7 days. Out of these 20,000 brightfield images, 2359 images were labelled and used in their model for training and testing purposes. This labelled set includes images taken from four types of microscopes. Images taken from

one microscope, echo rebel, were used for training purposes in their model (1022 images). The rest of them, from the other three microscopes, were used for testing the model. They developed a DL model for predicting ATP values by doing non-invasive estimation of cell viability, assessing damage to liver spheroids upon obtained morphology and predicting hepatotoxicity upon spheroid images.

By analysing labelled dataset, we concluded that in total there are 21 different drugs applied to the spheroids (all in the test part, 1337 images) and they are not equally distributed by days (most of the images when drugs were applied were taken at day 7). In Dubinsky's training set, drugs are not explicitly defined, but rather grouped in four merk compounds groups and for that reason not applicable to our concept, where we tend to extract physicochemical features for specific drugs. The other issue with our approach is that the set contains more images without applied drugs, namely medium and dimethyl sulfoxide (dmso), and that can cause disbalance in our analysis. Therefore, from the test labelled set, we created our own set by taking out only images when drugs were applied and randomly adding to it a certain amount of medium and dmso images. In total, out of 1337 images we extracted 320 images with applied drugs, plus randomly selected from the remaining image number 68 medium and 28 dmso images. Our dataset in total is composed out of 416 images, containing 21 drugs. Features extracted from images, 512 in total per image, were obtained by running our dataset on Dubinsky's model.

The next phase in our approach was to extract physicochemical features for selected 21 drugs. For it, we used the RDKit library [18]. In total 25 physicochemical features were selected, and they are provided in Table 1. All features are dimensionless, except molecular weight (MolWt) which is in grams and topological polar surface area (TPSA) which is in square angstrom.

Table 1: List of physicochemical features extracted from RDKit library

No.	Name of the feature in RDKit library	Meaning	Interpretation of increased value
1.	MolWt	Molecular weight	Reduced membrane permeability, biliary excretion, cholestasis and fibrosis risk
2.	MolLogP	Lipophilicity	Higher membrane permeability, accumulation, fatty liver, mitochondrial toxicity, cholestasis, and reactive metabolite formation
3.	TPSA	Topological polar surface area	Forming hydrogen bonds. More polar, lower membrane permeability, lower lipophilicity
4.	NumHDonors	Number of hydrogen bond donors	More polar, lower membrane permeability, lower lipophilicity, increasing immune-mediated interactions
5.	NumHAcceptors	Number of hydrogen bond acceptors	More polar, lower membrane permeability, lower lipophilicity, increasing interactions with metabolic enzymes or immune pathways
6.	NumRotBonds	Number of Rotatable Bonds	Extensive CYP-mediated oxidation, formation of reactive metabolites
7.	HeavyAtomCount	Number of heavy (non-hydrogen) atoms	Larger molecular size, increasing molecular complexity, biliary excretion, cholestasis, chronic accumulation, reactive intermediates
8.	FractionCSP3	Fraction of sp^3-hybridized carbons	lower aromaticity, reducing formation of reactive metabolites, decreasing mitochondrial and oxidative stress
9.	RingCount	Number of ring structures (aromatic or aliphatic)	Rigid and planar structure, reactive metabolite formation, mitochondrial stress and cholestasis risk
10.	CalcNumAromaticRings	Number of Aromatic Rings	planar, rigid, and lipophilic structure, reactive metabolite formation, mitochondrial stress, cholestasis
11.	GetFormalCharge	Formal Charge of the Molecule	Polar, low membrane permeability and hepatic uptake, transport-mediated or immune-mediated pathways or lipophilicity-driven hepatotoxicity
12.	NumHeteroatoms	Number of heteroatoms	Polarity and hydrogen-bonding capacity, immune-mediated toxicity
13.	BertzCT	Bertz Complexity Index	More complex molecular topology, reactive metabolite formation, higher metabolic burden and potential hepatotoxicity
14.	BalabanJ	Balaban's Topological Index	More rigid, interconnected, and aromatic topology, reactive metabolite formation, mitochondrial stress, cholestasis
15.	aromatic_ratio	Aromatic atom ratio	More planar, rigid, hepatic accumulation, reactive metabolite formation, mitochondrial stress, cholestasis
16.	CalcNumAliphaticRings	Number of aliphatic rings	Reducing planarity, aromatic reactive metabolites, lower hepatotoxicity risk
17.	CalcNumSaturatedRings	Number of saturated rings	Reducing planarity, aromatic reactive metabolites, lower hepatotoxicity risk
18.	len(pattern_nitro)	Number of nitro groups	Reactive metabolite formation, nitro reduction, idiosyncratic toxicity, DNA/protein adduct formation
19.	len(pattern_aniline)	Number of aniline substructures	Bioactivation to reactive nitroso, hydroxylamine, and iminoquinone intermediates, protein adducts, idiosyncratic hepatotoxicity, CYP-mediated oxidation
20.	len(pattern_quinone)	Number of quinone substructures	Increasing electrophilicity, ROS generation, oxidative stress, protein and DNA adduct, mitochondrial toxicity, cholestasis and idiosyncratic toxicity
21.	len(pattern_michael)	Number of Michael acceptor motifs	Increasing electrophilicity, increasing oxidative stress, protein adducts, mitochondrial toxicity, covalent binding and idiosyncratic hepatotoxicity
22.	len(pattern_halide)	Number of halide substituents	Increasing lipophilicity, slowing clearance, undergoing bioactivation, cholestasis and mitochondrial stress

23.	len(pattern_epoxide)	Number of epoxide substructures	Covalent binding, reactive toward nucleophiles, protein adducts, reactive metabolites, mutagenicity, carcinogenicity and idiosyncratic hepatotoxicity
24.	len(pattern_thiophene)	Number of thiophene rings	Undergo S-oxidation or ring epoxidation, covalent binding, immune-mediated liver injury and idiosyncratic hepatotoxicity
25.	len(pattern_phenol)	Number of phenol substructures	Undergo phase I oxidation, protein/DNA adducts, idiosyncratic liver injury

Out of 416 images we randomly selected two images for each drug (42 in total) and 8 medium images and 4 dmso images to serve as a test set in our framework. By excluding these images, the remaining set of 362 images was used as a training set. With a defined training dataset, we performed training using XG-Boost algorithm and applied 5-fold cross-validation (ratio 80% training to 20% testing). Hyperparameter optimisation was performed using the Optuna framework with 100 trials of Bayesian search (TPE sampler), minimising mean absolute error (MAE) across cross-validation folds. Output from the model is predicted ATP values. As a reference ATP values, we used actual experimental measurement for these samples. We calculated accuracy from our model and MAE and RMSE. We presented obtained values in the Results section.

Our training and testing set were combined and analysed upon morphological features to which we fused extracted physicochemical features. In total, the generated csv file contained 416 images (rows) and 541 features (additional 4 features - drug name, drug concentration, day at which image is taken, check whether image is control image, then 512 image features and 25 physico-chemical features). We clustered the dataset into four clusters with the plan to map it to four classes, representing toxicity levels - no toxicity, low toxicity, medium toxicity and high toxicity. Mapping of clusters was performed upon calculated average ATP level per cluster, such as cluster with highest average ATP value is assigned to no toxicity class, imposing that cluster with lowest ATP value is assigned to high toxicity class.

3. RESULTS

The optimised model achieved a cross-validation MAE of 58.78 (normalised MAE of 20.5% relative to the standard deviation of ATP values in the training set), RMSE of 84.7, and R^2 of 0.909. On the external test set, comprising 21 drugs not seen during training, the model demonstrated notably higher predictive accuracy, with MAE of 24.50 (NMAE of 8.6%), RMSE of 52.19, and R^2 of 0.968. The agreement between predicted and reference ATP values is illustrated in Figure 1, where most data points closely follow the identity line ($y = x$), indicating no systematic bias across the full range of ATP values. Minor local deviations are observed in the intermediate range (200–600 ATP units), which can be attributed to intra-compound variability between the two representative images selected per drug for the test set. The most influential features were predominantly morphological (z-space) features derived from spheroid images, with physicochemical descriptors contributing a smaller but non-negligible share, consistent with the hypothesis that spheroid morphology encodes most of the drug-induced viability information. The higher R^2 on the external test set (0.957) compared to cross-validation (0.909) can be attributed to the composition of the test set, which contains two representative images per drug, selected to cover diverse exposure conditions, whereas cross-validation folds include a broader range of samples with higher intra-compound variability in morphological features across different concentrations and time points.

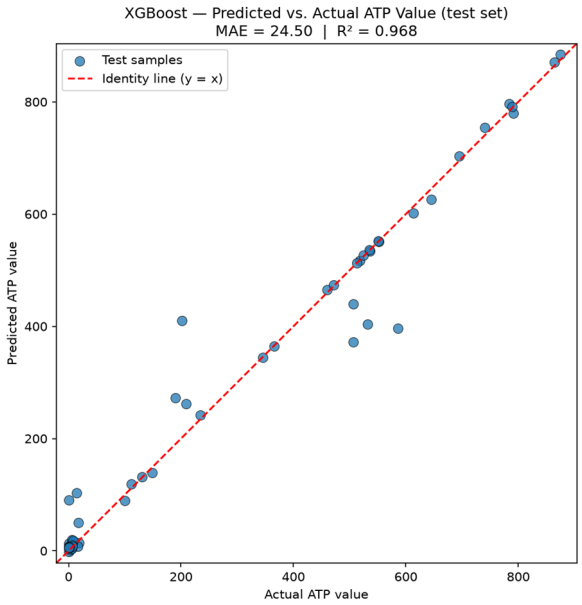


Figure 1: Comparison of predicted ATP values obtained from the XG-Boost model against corresponding experimentally measured ATP values, including MAE and R^2.

Classification of clusters was performed upon calculated average ATP value by cluster. ATP values are taken from a labelled data set, since they were measured experimentally. ATP value is assigned to each row of dataset used in our model. Then the average value of ATP is calculated for liver spheroids belonging to individual clusters. In the trend of

high to low average value, clusters were mapped to four classes - no toxicity, low toxicity, medium toxicity, high toxicity (Figure 2). A high level of ATP means better viability of cells and indirectly, a less toxic environment.

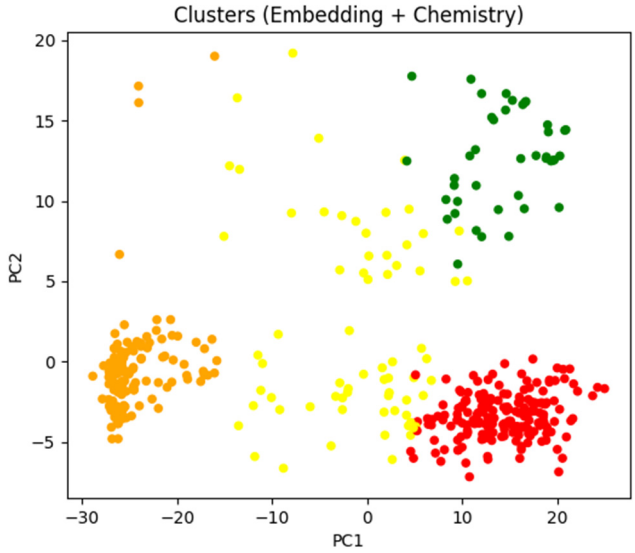


Figure 2: Clustering of used data in the model and mapping to four toxicity classes. Green cluster - no toxicity, yellow cluster - low toxicity, orange cluster - medium toxicity, red cluster - high toxicity

From Figure 2 can be concluded that the high-toxicity and the medium toxicity clusters are nicely grouped, while there are more dispersive groups of the no toxicity and the low toxicity clusters. The low toxicity class can correlate with all other three clusters, and it signalises that it was the most challenging to distinguish from other classes.

We analysed the content of each cluster and elaborated on specific drug groups by exploring whether selected drugs belong to the same class of pharmacotherapeutic classification – analgesics, antipyretics, antibiotics, etc. The model can detect if some drugs behave differently than expected, which can be useful in detecting errors and potential new phenomenological causes. First, we analysed compounds that belong to the high toxicity class. All spheroids with applied rosiglitazone and spectinomycin belong to this class (100%). Majority of spheroids with applied methotrexate (93.75%), neostigmine (93.75%), carbamazepine (87.5%), minoxidil (87.5%), nadolol (81.25%), imipramine hcl (75%) and nitrofurantoin (68.75%) are presented in the high toxicity class. Around half of the spheroids with applied amitriptyline hydrochloride (56.25%), desipramine hydrochloride (50%) and pyridostigmine (43.75%) showed high-toxicity effects; other half of them is mainly in the medium toxicity class. A smaller presence of phenformin (31.25%) is noted in the high toxicity class (mostly represented in the low toxicity class). Procyclidine (25%), metergoline (25%) and tamoxifen (6.25%) are noted in the high toxicity class, but they are mainly positioned in the medium toxicity class. For most compounds that are present in the high toxicity class there is clinical evidence of high or moderate hepatotoxicity. Some of them are not clinically known to cause hepatotoxicity, but *in vitro* models showed they can cause morphological changes to exposed liver spheroids, which are dependent on concentration and time of compound exposure. Figure 3 summarises all spheroids of the high toxicity class collected by their pharmacotherapeutic type and clinically known evidence of hepatotoxic effects.

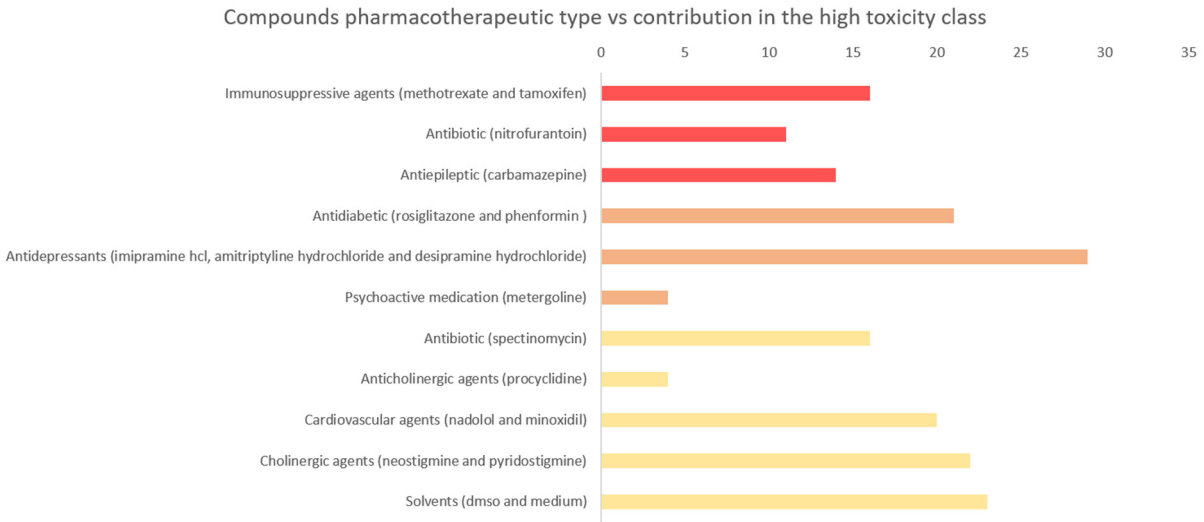


Figure 3: The high toxicity class spheroids grouped by pharmacotherapeutic types (number of occurrence) and clinically known hepatotoxic effects. Red colour - high hepatotoxic effects, orange colour - moderate hepatotoxic effects and yellow colour - low hepatotoxic effects

We did the same analysis for the medium toxicity class and observed less compounds with known high hepatotoxic effects, most of them with known moderate effects and several with low clinical evidence of hepatotoxicity, as it is presented in Figure 4.

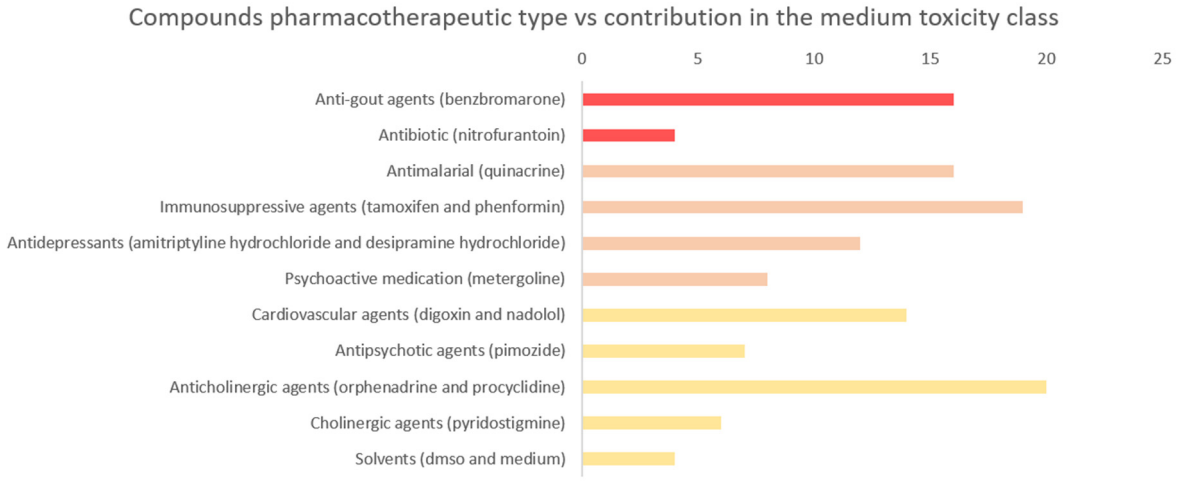


Figure 4: The medium toxicity class spheroids grouped by pharmacotherapeutic types (number of occurrence) and clinically known hepatotoxic effects. Red colour - high hepatotoxic effects, orange colour - moderate hepatotoxic effects and yellow colour - low hepatotoxic effects

Spheroids belonging to the low toxicity class point out most of them with low clinically known hepatotoxic effects, as it is depicted in Figure 5.

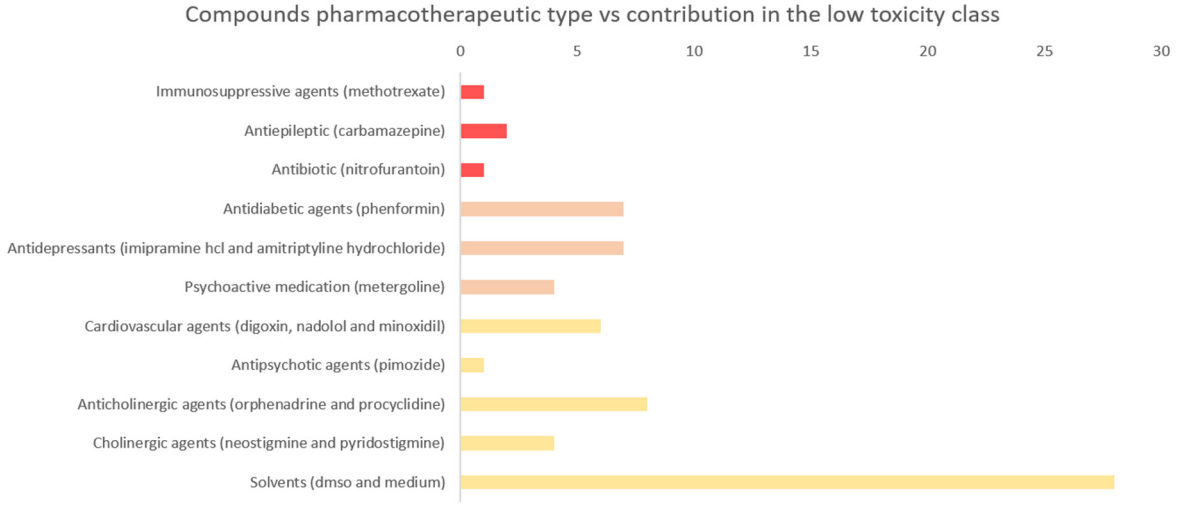


Figure 5: The low toxicity class spheroids grouped by pharmacotherapeutic types (number of occurrence) and clinically known hepatotoxic effects. Red colour - high hepatotoxic effects, orange colour - moderate hepatotoxic effects and yellow colour - low hepatotoxic effects

Finally, in the no toxicity class we observed only spheroids exposed to dmso and medium, which are known as neutral when it comes to hepatotoxicity. Dmso can occasionally show low toxicity effects, depending on concentration and time of exposure, but in general can be considered as non-toxic. Medium does not show any toxic effect. However, presence of medium and dmso spheroids in all other classes in smaller amounts can signal that changes in spheroids do not originate from exposure to compounds, but can be due to other laboratory conditions, poor nutrition to the cells, poor image quality, etc. To gain a more profound conclusion on it, examination of spheroid images relating to the dmso and medium images belonging to the other toxicity classes needs to be performed.

By analysing obtained results presented in Figures 3-5, it could be concluded that the most critical compounds pharmacotherapeutic type in causing hepatotoxicity are immunosuppressive agents, antiepileptic drugs, anti-gout agents and some antibiotics. Moderate hepatotoxic effects are causing antidiabetics and antidepressants, psychoactive medications and antimalarial drugs. Cardiovascular agents, cholinergic and anticholinergic agents, antipsychotic agents and some antibiotics are not very harmful to the liver. Dmso and medium are recognised as non-toxic to the liver.

The five most toxic drugs in our dataset, with clinical evidence of it (methotrexate, tamoxifen, nitrofurantoin, carbamazepine and benzbromarone), were analysed in terms of the five most influential physicochemical features. We created a heat map of values normalised to average values of all drugs in our dataset (Figure 6). Orange and red colours signalise values above the total average, while green colours stand for values below the total average.

	MolLogP	MolWt	NumRotBonds	BalabanJ	FractionCSP3
Methotrexate	0.09996795	1.347759652	2.02247191	1.011640138	0.645475574
Tamoxifen	2.233300389	1.101833782	1.797752809	1.205752899	0.595823607
Nitrofurantoin	0.027375724	0.706311387	0.674157303	1.25066186	0.322737787
Carbamazepine	1.261592548	0.700721017	0	1.392603259	0
Benzbromarone	2.032433342	1.257723552	0.674157303	1.333619881	0.303753212

Figure 6: Heat map of five top physicochemical features for five selected drugs normalised to average feature values of all drugs in used dataset

From Figure 6 we can observe that tamoxifen and benzbromarone show high lipophilicity (high MolLogP) which correlates with higher hepatic accumulation. Tamoxifen, benzbromarone and methotrexate are bigger molecules (high MolWt) that link to higher cholestasis risk (tamoxifen and benzbromarone) and fibrosis risk (methotrexate). Tamoxifen and methotrexate are flexible (high NumRotBonds) opposite to carbamazepine which is very rigid. Flexibility increases oxidation and formation of reactive metabolites that contribute to hepatotoxicity (tamoxifen). Increased BalabanJ feature indicated more complex connectivity and all five drugs are above total average value, meaning they have more complex molecular topology linked to diverse metabolic pathways and increased risk of toxicity. Higher values of FractionCSP3 means more saturated compounds, while lower values explain flat, aromatic structure. All five drugs have FractionCSP3 values below the total average, which could relate to mitochondrial stress, reactive metabolites, and cholestasis and therefore, increased hepatotoxicity risk. Overall, tamoxifen [19], benzbromarone [20] and carbamazepine [21] stand for lipophilic, aromatic and bulky molecules that cause metabolic changes in liver spheroids. Nitrofurantoin is not driven by lipophilicity, and even being small molecule and polar, causes chronic autoimmune-like liver injury, driven by nitro reduction and oxidative stress [22]. Methotrexate with high molecular weight and flexibility upon chronic exposure leads to liver fibrosis and steatosis. Methotrexate lacks the structural indicators for acute hepatotoxicity, but its size, polarity, and mechanism create a chronic hepatotoxicity pathway, meaning that bidirectional mechanism of methotrexate depends on dosage [23]. Similarly, due to low lipophilicity, nitrofurantoin can show bidirectional behaviour in hepatotoxicity, but it depends on time of exposure.

Finally, we analysed temporal dependence of drugs to cell viability (toxicity), providing the insights in the dynamic of effects and toxicity progress over time. We analysed only compounds known to be highly toxic to the liver - methotrexate, tamoxifen, nitrofurantoin, carbamazepine and benzbromarone. Since our dataset (used in training and testing of the model) contains only images taken at day 7, we used additional spheroid images exposed to these five selected drugs at other days from Dubinsky's unlabelled set. To analyse toxic effects, we used our trained model to predict the level of ATP for each image. In total 15 new images were taken from an unlabelled dataset - for each drug at days 1, 4 and 7.

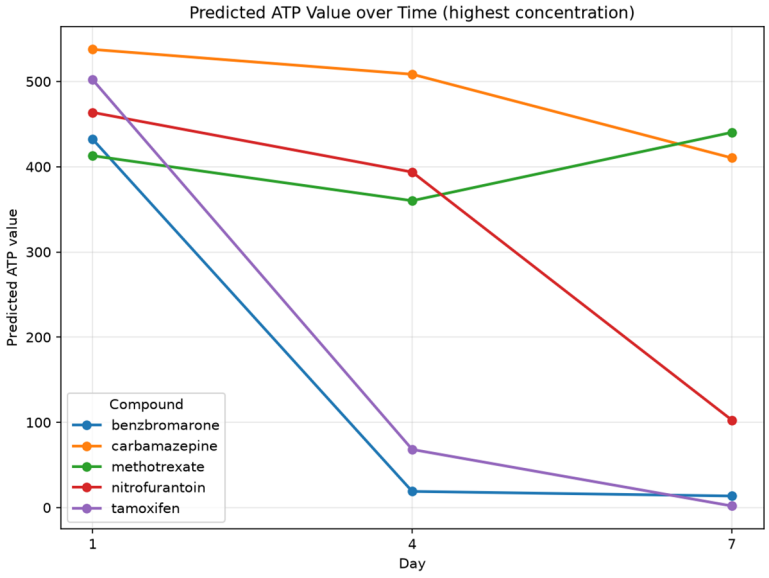


Figure 7: Temporal dependence of cell viability (predicted ATP level) for selected five drugs

The predicted ATP trajectories (Figure 7) revealed distinct toxicity dynamics: benzbromarone and tamoxifen exhibited a rapid decline in viability already by day 4, nitrofurantoin showed a delayed but pronounced decrease between days 4 and 7, whereas carbamazepine displayed a slow, gradual reduction, retaining comparatively high viability throughout. Because the training data contained drug samples only at day 7 (while control samples spanned days 1 to 21), predictions for days 1 and 4 represent an extrapolation along the temporal dimension for drugs. To assess the reliability of these predictions, the model outputs were validated against the reference Dubinsky predictions for the same images, Figure 8, accessed through the image identifiers and rescaled to the ATP value range by linear calibration. The validation yielded an overall R^2 of 0.891 (MAE 51.33), confirming that the model reproduces the reference ATP values for days 1 and 4 with reasonable accuracy despite the absence of drug samples at these days in training of the model. This result supports the interpretation that the predictions are driven primarily by spheroid morphology, which implicitly encodes the temporal state of the spheroid, rather than by the day feature itself, thereby justifying the temporal toxicity analysis.

A particular case worth highlighting is methotrexate, which exhibited anomalous behaviour across all stages of the analysis. Unlike the other four compounds, whose predicted ATP trajectories followed a monotonically decreasing pattern consistent with progressive cell damage, methotrexate displayed a non-monotonic profile, with ATP values slightly decreasing by day 4 and then rising again by day 7, contrary to the expected toxicity progression, as it can be noticed in Figure 7. This compound was also the most consistent outlier throughout the study: it produced the largest single prediction error in the test-set evaluation, and in the temporal validation against the reference Dubinsky predictions it again showed the largest deviations, particularly at day 4 (absolute error of 171), with the model systematically underestimating its ATP values.

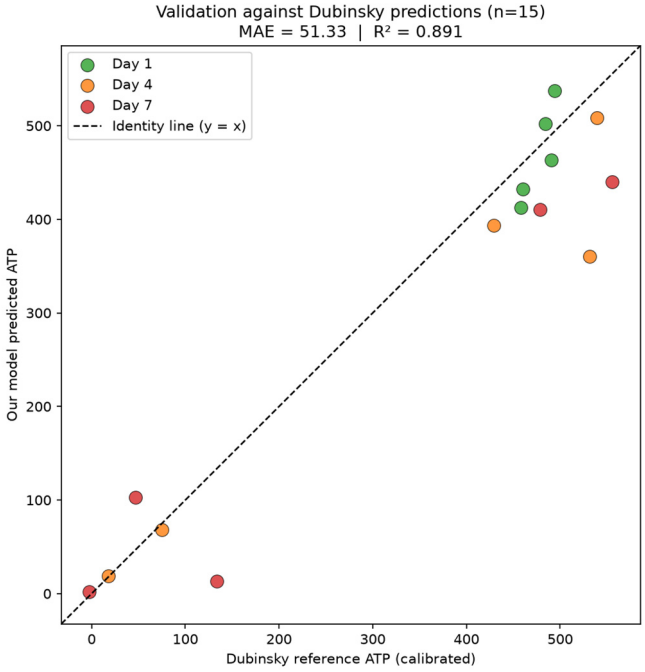


Figure 8: Validation of our model outputs against referenced Dubinsky's model for predicted ATP levels for 15 new samples of five selected drugs at days 1, 4 and 7, including MAE and R^2

Methotrexate is an antimetabolite (chemotherapy / immunosuppressant) whose cytotoxic mechanism acts primarily through inhibition of folate metabolism and DNA synthesis, an intracellular biochemical mode of action that may not be fully reflected in the morphological appearance of the spheroid [24]. This offers a plausible explanation for the model's reduced accuracy: since the predictions are driven predominantly by morphological latent features, a compound whose toxicity is not accompanied by a proportional change in spheroid morphology is inherently more difficult to characterise. Rather than representing a model failure, this discrepancy illustrates one of the stated objectives of the framework, the ability to detect compounds that behave differently than expected, thereby flagging potential errors or revealing new phenomenological causes that merit dedicated experimental investigation.

4. CONCLUSION

Our research showed that even though a lot of information can be extracted from liver spheroid images, it is important to take into account physicochemical features of applied drugs, since it would bring a bit better clustering of data, but more importantly can provide insight into reasoning why some changes in liver spheroids happens, instead of pointing out they happened. Interpretation of obtained results can give better insight into underlying mechanisms leading to hepatotoxic changes. We analysed pharmacotherapeutic types of drugs from used dataset and observed that immunosuppressive agents, antiepileptic drugs, anti-gout agents and some antibiotics showed highest toxicity risk. Antidiabetics and antidepressants indicated moderate toxicity risk, while cardiovascular agents, cholinergic and anticholinergic agents, antipsychotic agents and some antibiotics revealed not very harmful effects to the liver. We then analysed values of the five most

influential physicochemical features for five most toxic drugs in our dataset (methotrexate, tamoxifen, nitrofurantoin, carbamazepine and benzbromarone) to show alignment with observed experimental and clinical data and to look for possible deviations from expected results. Our analysis explained potential bidirectional behaviour of methotrexate and nitrofurantoin in relation to hepatotoxic effects. Temporal analysis of five most toxic drugs from new external test on the trained model explained changes in liver spheroids and more profound hepatotoxicity effects with time but also captured bidirectional behaviour of methotrexate. Future research will include testing of additional drugs and analysis of their effects, as well as extending of developed ML framework to analyse liver spheroids co-exposures. In that way, we would be able to analyse combined effects of applied drugs on liver spheroids in terms of potential hepatotoxicity risk. Additional analysis will be dedicated to the influence of time and concentration of applied drugs for co-exposure scenarios.

ACKNOWLEDGEMENTS

This research was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, contract numbers 451-03-34/2026-03/200108 (Faculty of Mechanical and Civil Engineering in Kraljevo, University of Kragujevac), 451-03-33/202603/200107 (Faculty of Engineering, University of Kragujevac) and 45103-33/2026-03/200378 (Institute for Information Technologies Kragujevac, University of Kragujevac) This research is in line with the strategic plan of the Republic of Serbia 2030, particularly in the section 2.2.2. Development of human resources, goal 3: Good health.

REFERENCES

1. D. Dubinsky, S. Harel, A. Bein, A. Nyska, S. Ya'ari, F. Anas, I. Bentwich and L. Wolf, "Non-invasive quantification of viability in liver spheroids using deep learning", Frontiers in Bioengineering and Biotechnology, Vol. 14, pp. 1-18, https://doi.org/10.3389/fbioe.2026.1797474, (2026)
2. D. Dubinsky, S. Harel and QurisAI, "3D Liver Spheroids Brightfield 256x256 Images + ATP labels [Data set]. QurisAI", https://doi.org/10.5281/zenodo.15120118, (2025)
3. R. J. Andrade, N. Chalasani, E. S. Björnsson, A. Suzuki, G. A. Kullak-Ublick, P. B. Watkins, H. Devarbhavi, M. Merz, M. I. Lucena, N. Kaplowitz and G.P. Aithal, "Drug-induced liver injury", Nature Reviews Disease Primers, Vol. 5(1), https://doi.org/10.1038/s41572-019-0105-0, (2019)
4. LiverTox®: Clinical and Research Information on Drug-Induced Liver Injury [Internet]. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases, https://www.ncbi.nlm.nih.gov/books/NBK547852/, (2012)
5. V. Y. Soldatow, E. L. Lecluyse, L. G. Griffith and I. Rusyn, "In vitro models for liver toxicity testing", Toxicology Research (Cambridge), Vol. 2(1), https://doi.org/10.1039/C2TX20051A, (2013)
6. C. C. Bell, V. M. Lauschke, S. U. Vorrink, H. Palmgren, R. Duffin, T. B. Andersson and M. Ingelman-Sundberg, "Transcriptional, Functional, and Mechanistic Comparisons of Stem Cell-Derived Hepatocytes, HepaRG Cells, and Three-Dimensional Human Hepatocyte Spheroids as Predictive In Vitro Systems for Drug-Induced Liver Injury", Drug Metabolism and Disposition, Vol. 45(4), https://doi.org/10.1124/dmd.116.074369, (2017)
7. Z. Benzait, Ö. Tomsuk, A. Ebrahimi, H. Ghorbanpoor, C. Özel, R. Didarian, B. Demir Cevizlidere, M. Kaya, T. Gur, N. Gasimzade, F. B. Ergen, A. E. Topal, S. Dabagh, R. Javanifar, N. Abdullayeva, O. Uysal, A. Eker Sarıboyacı, F. Doğan Güzel, S. Hassan and H. Avci, "Liver-on-a-Chip (LoC) Models: Case Studies of Academic Platforms and Commercial Products", Molecular Pharmacology, Vol. 23(3), pp. 1367-1401, https://doi.org/10.1021/acs.molpharmaceut.5c01122, (2026)
8. P. A. Walker, S. Ryder, A. Lavado, C. Dilworth, C. and R. J. Riley, "The evolution of strategies to minimise the risk of human drug-induced liver injury (DILI) in drug discovery and development", Archives of Toxicology, Vol. 94(8), pp. 2559-2585, https://doi.org/10.1007/s00204-020-02763-w, (2020)
9. M. Jordović-Pavlović, M. Stanković, M. Popović, Ž. Ćojbašić, S. Galović, D. Markushev, "The application of artificial neural networks in solid-state photoacoustics for the recognition of microphone response effects in the frequency domain",„Journal of Computational Electronics, Vol. 19(3), pp.1268-1280, 2020, https://doi.org/10.1007/s10825-020-01507-4
10. V. Milićević, L. Koceva Lazarova, M. Jordović Pavlović, "The application of artificial intelligence in education–the current state and trends", International Journal of Cognitive Research in Science, Engineering and Education", Vol. 12(2), pp. 259-272, 2024, https://orcid.org/0000-0002-5587-2717
11. T. Šušteršič, V. Gribova, M. Nikolic, P. Lavalle, N. Filipovic, N. E. Vrana, "The Effect of Machine Learning Algorithms on the Prediction of Layer-by-Layer Coating Properties", ACS Omega, 8(5), pp. 4677-4686, 2023, https://doi.org/10.1021/acsomega.2c06471
12. R. Geci, A. Z. Sayin, S. Schaller and L. Kuepfer, "Integration of in vitro and in silico approaches enables prediction of drug-induced liver injury", Archives of Toxicology, Vol. 100(5), pp. 2029-2046, https://doi.org/10.1007/s00204-026-04305-2, (2026)

13. V. Bergen, K. Kodella, S. Srikrishnan, O. Barrandon, S. Anderson, M. Rogers-Grazado, C. Fowler, H. Beyene, N. Robichaud, T. Fulton, N. Lapchyk, M. Cortes, N. Plugis, M. Goddeeris and M. Zamanighomi, "A large-scale human toxicogenomics resource for drug-induced liver injury prediction", Nature Communications, Vol. 16(1), https://doi.org/10.1038/s41467-025-65690-3, (2025)
14. T. Xiao, Y. Liu, K. Hu, K. Guo, M. Zhang, T. Wang, W. Lei, W. Wang, S. Zhou, Y. Hu and R. Su, "Drug-induced liver injury prediction based on graph convolutional networks and toxicogenomics", PLOS Computational Biology, Vol. 21(9), https://doi.org/10.1371/journal.pcbi.1013423, (2025)
15. J. Han, W. Zhung, I. Jang, J. Lee, M. J. Kang, T. D. Lee, S.J. Kwack, K. B. Kim, D. Hwang, B. Lee, H. S. Kim, W. Y. Kim and S. Lee, "HepatoToxicity Portal (HTP): an integrated database of drug-induced hepatotoxicity knowledgebase and graph neural network-based prediction model", Journal of Cheminformatics, Vol. 17(1), https://doi.org/10.1186/s13321-025-00992-8, (2025)
16. https://pubchem.ncbi.nlm.nih.gov/
17. A. Malka-Markovitz, S. Camara Dit Pinto, M. Cherkaoui, S. M. Levine, S. Anandasabapathy, G. K. Sood, S. Dhingra, G. Yujia, J. M. Vierling and N. R. Gallo "Multiscale modeling of drug-induced liver injury from organ to lobule", npj Digital Medicine, Vol. 8, https://doi.org/10.1038/s41746-025-01736-6, (2025)
18. https://www.rdkit.org/
19. F. Zhao, P. Xie, J. Jiang, L. Zhang, W. An, Y. Zhan,"The effect and mechanism of tamoxifen-induced hepatocyte steatosis in vitro", International Journal of Molecular Sciences, Vol. 15, No. 3, pp. 4019–4030, https://doi.org/10.3390/ijms15034019, (2014)
20. M. G. McDonald and A. E. Rettie, „Sequential metabolism and bioactivation of the hepatotoxin benzbromarone: formation of glutathione adducts from a catechol intermediate", Chemical Research in Toxicology, Vol. 20, No. 12, pp. 1833–1842, https://doi.org/10.1021/tx7001228, (2007)
21. S. Higuchi, A. Yano, S. Takai, K. Tsuneyama, T. Fukami, M. Nakajima, T. Yokoi, "Metabolic Activation and Inflammation Reactions Involved in Carbamazepine-Induced Liver Injury", Toxicological Sciences, Vol. 130, Issue 1, pp. 4–16, https://doi.org/10.1093/toxsci/kfs222, (2012)
22. National Institute of Diabetes and Digestive and Kidney Diseases, „Nitrofurantoin", in LiverTox: Clinical and Research Information on Drug-Induced Liver Injury, Bethesda (MD), Bookshelf ID NBK548318, PMID 31643641, https://www.ncbi.nlm.nih.gov/books/NBK548318/, (2025)
23. D. Ezhilarasan, „Hepatotoxic potentials of methotrexate: understanding the possible toxicological molecular mechanisms", Toxicology, Vol. 458, Article 152840, https://doi.org/10.1016/j.tox.2021.152840, (2021)
24. P. Hanoodi and M. Mittal, „Methotrexate", in StatPearls, StatPearls Publishing, Treasure Island (FL), Bookshelf ID NBK556114, https://www.ncbi.nlm.nih.gov/books/NBK556114/, (2024)

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Comparing QLoRA fine-tuning of a general-purpose and a mathematically specialized LLM on Serbian mathematics competition tasks

Marina Svičević[1*], Miloš Pavković[2], Nemanja Vučićević[1],
Aleksandar Milenković[1], Aleksandar Milutinović[1]

[1] University of Kragujevac, Faculty of Science, Kragujevac, Serbia

[2] Singidunum University, Belgrade, Serbia

ARTICLE INFO

* Correspondence: marina.svicevic@pmf.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.D12S **PAGES:** D97 – D103

ABSTRACT

This paper examines the effect of QLoRA fine-tuning of locally executable large language models on mathematics competition tasks for high school students in Serbia. The main focus is on comparing the general-purpose model Mistral-7B-Instruct and the mathematically specialized model Mathstral-7B, in order to determine whether prior specialization for mathematical reasoning affects the success of fine-tuning on a small, domain-specific corpus of tasks in Serbian. For the purposes of the study, an automatic evaluation framework was applied, in which generated solutions were assessed according to several criteria, including final answer accuracy, logical coherence of the solution, and explanation quality. The results indicate different effects of fine-tuning for the observed models: the general-purpose model showed a slight performance decline, while the mathematically specialized model showed an improvement. This outcome suggests that models previously adapted to mathematical reasoning may benefit more from domain-specific fine-tuning, while also confirming that Serbian mathematics competition tasks remain a highly demanding test for locally executable large language models.

KEYWORDS

Large language models, QLoRA fine-tuning, Mistral-7B-Instruct, Mathstral-7B, Mathematics competitions, Automatic evaluation, Mathematical reasoning

1. INTRODUCTION

Large language models (LLMs) have made significant progress in solving mathematical problems, particularly on standard datasets commonly used for their evaluation. However, strong performance on such datasets does not necessarily imply comparable success in solving non-standard problems. Mathematics competition tasks represent a particularly demanding test, as solving them often requires more than the application of an appropriate formula or algorithm. It is necessary to interpret the problem conditions correctly, identify a suitable strategy, connect multiple reasoning steps, and reach a clearly justified conclusion [1], [2].

Assessing the quality of LLM-generated solutions is further complicated by the fact that a mathematically correct answer and a convincingly written explanation do not always match. A model may provide a clearly structured solution that contains an error, while in other cases it may generate a correct answer despite relying on incomplete or insufficiently justified reasoning. Therefore, the evaluation of such solutions should not be limited solely to checking the final result. The correctness of the reasoning process and the quality of the explanation should also be considered. Previous studies conducted on mathematics competition tasks in Serbia have shown that the performance of AI tools depends on the task type, mathematical domain, input format, and characteristics of the evaluated model [3], [4], [5].

Although commercial models often achieve better results, locally executable models are particularly relevant for research and practical applications. Their use provides greater control over data processing, improves experimental reproducibility, and enables model adaptation to specific needs without relying on external services. However, their performance on complex mathematical tasks remains limited. One possible approach to improving their capabilities is parameter-efficient

fine-tuning. Methods such as LoRA and QLoRA enable the adaptation of large language models with substantially lower memory and computational requirements than full fine-tuning of all model parameters [6], [7].

An important question is whether the effectiveness of such adaptation depends on the initial characteristics of the selected model. A general-purpose model and a model already specialized for mathematical reasoning may not benefit equally from additional training on a small, domain-specific corpus. In the first case, fine-tuning aims to adapt a broadly trained model to a narrower type of task. In the second case, additional training further adapts an already specialized model to a more specific linguistic and problem-solving context. This distinction is particularly relevant when the dataset consists of complex competition tasks in Serbian rather than a large collection of standard problems in English.

This paper compares the effects of QLoRA fine-tuning on two locally executable models of similar size: Mistral-7B-Instruct-v0.3, a general-purpose instruction model, and its variant Mathstral-7B, a model specialized for mathematical and scientific tasks. For both models, performance is analyzed before and after fine-tuning on mathematics competition tasks for high school students in Serbia. The generated solutions are assessed using an automatic multi-criteria evaluation framework that considers final answer accuracy, logical coherence, and explanation quality.

The aim of this study is to examine whether prior specialization for mathematical reasoning affects the outcome of domain-specific fine-tuning. Accordingly, the following research questions are addressed:

- How does QLoRA fine-tuning affect the performance of the general-purpose Mistral-7B-Instruct-v0.3 model and the mathematically specialized Mathstral-7B model on Serbian mathematics competition tasks for high school students?
- Does prior mathematical specialization impact the effectiveness of fine-tuning on a small, domain-specific corpus of tasks?

2. RELATED WORK

Research on improving the ability of large language models to solve mathematical problems has developed in several directions. One approach focuses on preparing higher-quality training datasets and expanding them with different variants of existing tasks. MetaMath, for example, was developed through fine-tuning on the MetaMathQA dataset, which was created by reformulating questions from standard mathematical datasets in multiple ways [8]. This approach shows that the diversity of training examples can play an important role in improving mathematical reasoning. Another direction involves continued pretraining on large mathematical corpora. DeepSeekMath-7B was developed by continuing the pretraining of DeepSeek-Coder-Base-v1.5 7B on a large collection of mathematically relevant content while retaining natural language and code data [9]. Similarly, Llemma was obtained through continued pretraining of Code Llama on the Proof-Pile-2 corpus, which includes scientific papers, web content containing mathematical notation, and formal mathematical records [10]. Mathstral-7B belongs to the same broader group of mathematically oriented models. It was built on Mistral-7B and designed for mathematical and scientific tasks [11]. Unlike general-purpose models, such models start from representations that have already been shaped toward mathematical reasoning. In addition to the choice of training data and the initial model, the method used to adapt the model to a specific domain is also important. Full fine-tuning of large models requires substantial computational resources, which has motivated the development of methods that update only a limited number of specific parameters. LoRA keeps the weights of the base model unchanged and introduces low-rank matrices that are trained for a specific task [6]. QLoRA further reduces memory requirements by using a quantized base model while computing gradients through LoRA adapters [7]. These methods are particularly relevant in research settings where locally executable models are used under limited hardware resources.

Although the mathematical capabilities of LLMs are often evaluated on standard datasets, such results do not provide a complete picture of their performance on non-standard tasks. Mathematics competition tasks require the identification of less obvious relationships, the selection of an appropriate strategy, and the consistent execution of multi-step reasoning. The CHAMP dataset was designed specifically for a more detailed analysis of these abilities using high school competition-level tasks. In addition to the tasks themselves, the dataset includes concept and hint annotations, enabling an analysis of how models use additional information during problem solving [2]. The authors also point out that a correct final answer does not necessarily result from valid reasoning, which is why evaluation should not be reduced solely to checking the final result. The performance of AI tools on mathematics competition tasks has also been examined in the Serbian-language context. A study conducted on tasks from the Mathematical Kangaroo competition showed that results depend on the input format, the language of the task statement, the mathematical domain, and the selected tool [3]. Later studies analyzed solutions to more demanding tasks from the Serbian national mathematics competition for high school students, generated using DeepSeek-R1, as well as o1 and o3-mini [4], [5]. The findings confirmed that advanced models can solve some non-standard tasks successfully, but difficulties remain with more complex formulations, logically demanding problems, and the consistent justification of all solution steps. In the context of mathematics competition tasks in Serbia, a framework for automated solving and multi-criteria evaluation of generated solutions was proposed and applied to commercial and locally executable models with different purposes [12]. The results revealed a clear difference between these two groups of models, while also indicating the potential of mathematically specialized models among locally executable solutions. A subsequent analysis of QLoRA fine-tuning of Mistral-7B-Instruct-v0.3 showed that adapting a general-purpose model to a small dataset of competition tasks does not necessarily improve performance on a held-out test set [13].

Although numerous studies have examined mathematically specialized models and methods for adapting them, considerably less attention has been paid to whether prior specialization affects the outcome of additional fine-tuning on a small, language-specific dataset. Therefore, this paper addresses a gap in the existing literature by directly comparing the effects of QLoRA fine-tuning on the general-purpose Mistral-7B-Instruct-v0.3 model and the mathematically specialized Mathstral-7B model using the same dataset and a unified evaluation framework.

3. EXPERIMENTAL DESIGN

3.1. Dataset and Experimental Workflow

The experimental dataset consists of 348 mathematics competition tasks for high school students in Serbia, organized by the Mathematical Society of Serbia. The dataset includes tasks from the municipal, regional, and national competition levels from 2022, 2023, and 2024, covering all four high school grades, as well as Categories A and B at the national level. According to mathematical domain, the tasks belong to algebra, geometry, number theory, or combinatorics. Only text-based tasks were included in the study, while tasks containing images or graphical elements were excluded. This made it possible to analyze the generated textual solutions without additional processing of visual content.

The original materials were available in LaTeX format and contained the task statements together with the corresponding official solutions. After preprocessing, each task was stored in a structured format along with relevant metadata, including the year, competition level, grade, category, and mathematical domain. Mathematical expressions were kept in LaTeX format to preserve their structural and semantic accuracy. To prevent overlap between the examples used for fine-tuning and the tasks intended for final evaluation, the dataset was split chronologically. The training set contains 192 tasks, the validation set 40 tasks, and the test set all 116 tasks from 2024. The final evaluation is therefore based on tasks that were not available during training. The dataset split is presented in Table 1.

Table 1: Chronological partitioning of the dataset.

Set	Number of tasks	Source
Training	192	Entire 2022 set and selected tasks from 2023
Validation	40	Regional competition tasks from 2023
Test	116	Entire 2024 set

The overall experimental workflow is shown in Figure 1. After data preparation and splitting, the baseline and fine-tuned variants of the two locally executable models were analyzed. For each of the four configurations, solutions were generated for the tasks in the test set, followed by automatic multi-criteria evaluation and a comparative analysis of the results.

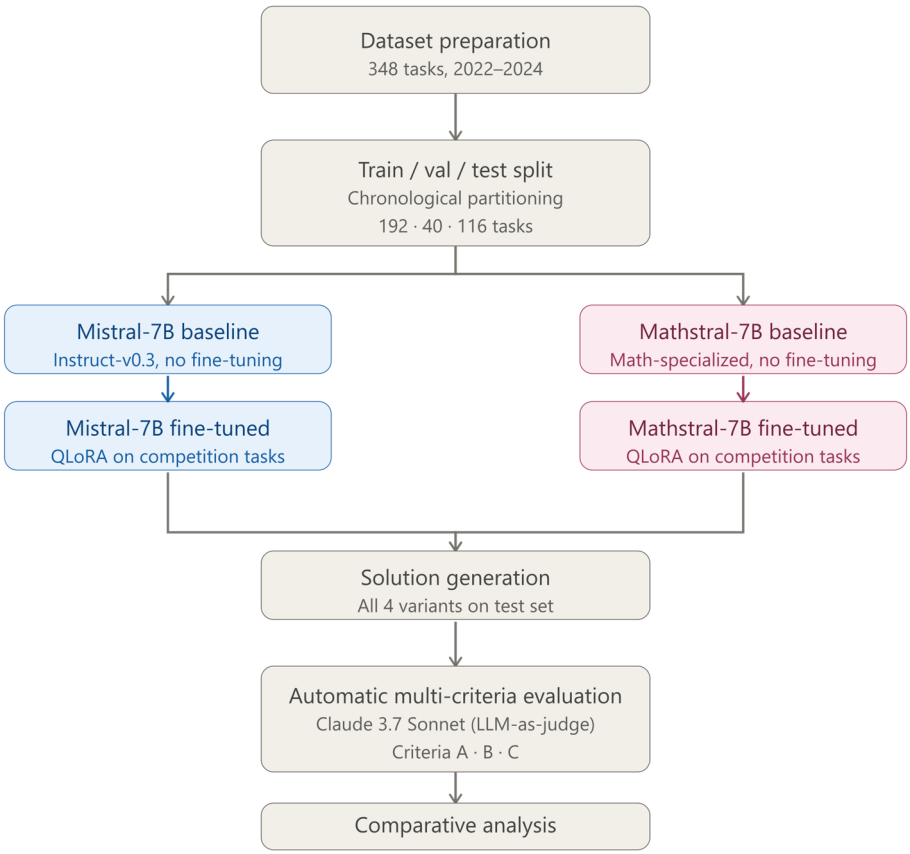


Figure 1: Experimental pipeline for comparing baseline and QLoRA fine-tuned variants of Mistral-7B-Instruct and Mathstral-7B.

3.2. Compared Models and Fine-Tuning Setup

The experiment considered two locally executable models with seven billion parameters, but different initial purposes. Mistral-7B-Instruct-v0.3 is a general-purpose instruction model, whereas Mathstral-7B is specialized for mathematical and scientific tasks. Selecting models of comparable size enables a meaningful comparison of their performance and the effects of additional domain-specific fine-tuning.

Two variants of each model were analyzed: the baseline version, without additional training on tasks from the dataset, and the version obtained through QLoRA fine-tuning. This resulted in four experimental configurations: Mistral-7B Baseline, Mistral-7B Fine-Tuned, Mathstral-7B Baseline, and Mathstral-7B Fine-Tuned. All variants were run locally and evaluated on the same held-out test set.

QLoRA was applied for the domain-specific fine-tuning of both models. The hyperparameters were not fixed identically in advance for the two models but were selected based on the results obtained during the experimental process. For each model, the configuration that achieved the most favorable results on the validation set was retained. The key fine-tuning parameters are presented in Table 2.

Table 2: Main QLoRA fine-tuning parameters.

Parameter	Mistral-7B	Mathstral-7B
LoRA rank (r)	16	16
LoRA alpha (α)	32	32
Number of epochs	5	5
Batch size	2	2
Gradient accumulation steps	16	16
Effective batch size	32	32
Learning rate	2×10^{-4}	5×10^{-5}
LR scheduler	Cosine	Cosine
Early stopping patience	2	2

For fine-tuning, the models were additionally configured with a LoRA dropout of 0.10, a warmup ratio of 0.10, a weight decay of 0.05, and a maximum sequence length of 4000 tokens. The LoRA adapters were applied to the projection modules q_proj, v_proj, k_proj, o_proj, gate_proj, up_proj, and down_proj. During solution generation, the maximum number of new tokens was set to 3400, with a temperature of 0.1, a top_p value of 0.95, and a repetition penalty of 1.1.

3.3. Automatic Multi-Criteria Evaluation

An automatic evaluation procedure based on the LLM-as-a-judge approach was applied to assess the generated solutions. For each task, the evaluation model Claude 3.7 Sonnet received the task statement, the corresponding official solution, and the answer generated by the analysed model. The same procedure was applied to all four experimental configurations, enabling their direct comparison on the same test set.

The choice of an automatic evaluation approach was motivated by the nature of the analysed tasks. Mathematics competition tasks require not only a correct final answer, but also a logically sound solution process and a clear explanation. Manually assessing a large number of generated solutions would be time consuming, whereas the LLM-as-a-judge approach allows the same criteria to be applied consistently and at scale. At the same time, it enables different aspects of a solution to be analysed separately, which is important because a correct answer does not necessarily result from valid reasoning.

The evaluation was conducted according to three criteria. Answer Accuracy (A) refers to the correctness of the final answer and takes a value from 0 to 1. Logical Coherence (B) assesses the logical consistency and mathematical justification of the solution process on a scale from 0 to 1. Explanation Quality (C) evaluates the clarity, completeness, and precision of the provided explanation, also on a scale from 0 to 1.

A separate evaluation prompt was used for each criterion so that answer accuracy, reasoning quality, and explanation quality could be assessed independently. The aggregate score was then calculated as follows:

$$S = 0.40A + 0.30B + 0.30C.$$

A higher weight was assigned to Answer Accuracy, as the correctness of the final result is particularly important in the context of mathematics competitions. Including Logical Coherence and Explanation Quality provides a more detailed view of how the model arrives at a solution. Although the same evaluation procedure was applied to all model variants, the obtained scores should be interpreted as automatic comparative indicators rather than as a substitute for expert mathematical assessment.

4. RESULTS AND DISCUSSION

The evaluation results for the four analyzed configurations are presented in Table 3. In addition to final answer accuracy, the table includes the logical coherence of the generated solution process, the quality of the provided explanation, and the aggregate score, which represents an overall measure of model performance.

Table 3: Performance comparison of baseline and fine-tuned model variants.

Model	Answer Accuracy (A)	Logical Coherence (B)	Explanation Quality (C)	Aggregate Score
Mistral-7B Baseline	8.5%	21.6%	28.1%	18.3%
Mistral-7B Fine-Tuned	6.5%	19.3%	26.6%	16.4%
Mathstral-7B Baseline	22.3%	23.7%	41.2%	28.4%
Mathstral-7B Fine-Tuned	24.1%	28.0%	42.9%	30.9%

A comparison of the baseline variants already reveals a clear difference between the two models. Mathstral-7B Baseline outperforms Mistral-7B Baseline according to all three individual criteria, as well as the aggregate score. The most pronounced differences are observed in final answer accuracy and explanation quality. This result indicates that prior mathematical specialization is relevant even when the model is applied to non-standard tasks in Serbian.

The effects of fine-tuning differ between the two analyzed models. No improvement was observed for Mistral-7B. After additional training, the values of all three criteria decreased, and the aggregate score was therefore lower. This decline suggests that adapting a general-purpose model to a small dataset of complex competition tasks does not necessarily lead to better generalization to unseen examples.

A different pattern is observed for Mathstral-7B. The fine-tuned variant achieves better results according to all the criteria evaluated. The largest improvement is recorded for the logical coherence of the solution process, while final answer accuracy and explanation quality also increased. The aggregate score rises from 28.4% to 30.9%, making Mathstral-7B Fine-Tuned the best-performing configuration among the four analyzed variants.

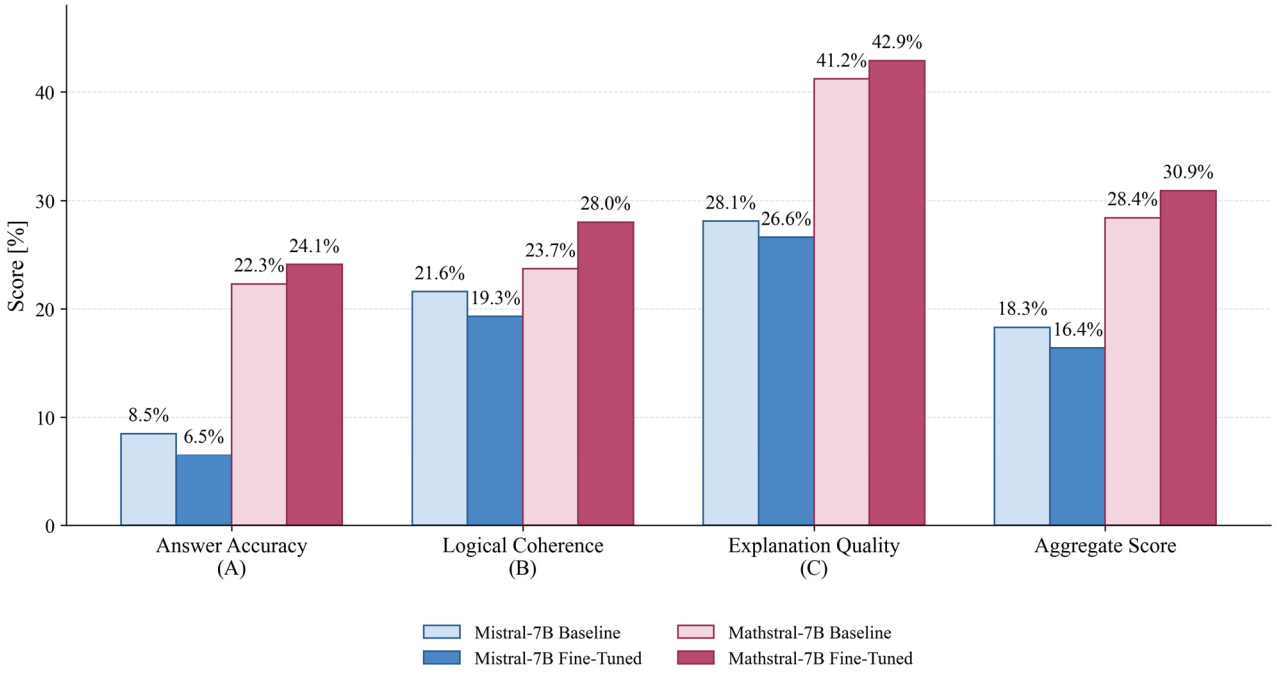


Figure 2: Performance comparison of baseline and QLoRA fine-tuned variants of Mistral-7B and Mathstral-7B across the evaluation criteria on the held-out test set of 116 tasks.

The results indicate that the effectiveness of domain-specific fine-tuning also depends on the initial characteristics of the selected model. Additional training on a relatively small corpus was not sufficient to improve the performance of the general-purpose Mistral-7B model. In contrast, Mathstral-7B benefited more from the same type of corpus, suggesting that fine-tuning may be more effective when it builds on previously developed mathematical reasoning capabilities.

The training dynamics are consistent with this finding. As shown in Figure 3, Mathstral-7B achieves a substantially lower evaluation loss throughout training, with a final value of 0.620 compared with 0.991 for Mistral-7B. At the same time, its peak evaluation mean token accuracy reaches 0.850, whereas Mistral-7B plateaus around 0.771. Notably, Mistral-7B shows a larger gap between its final training accuracy (0.804) and evaluation accuracy (0.765), suggesting a degree of overfitting, whereas Mathstral-7B displays almost no such divergence (0.851 vs. 0.849). These patterns indicate that Mathstral-7B adapted more effectively during fine-tuning, which is consistent with its stronger performance on the held-out test set.

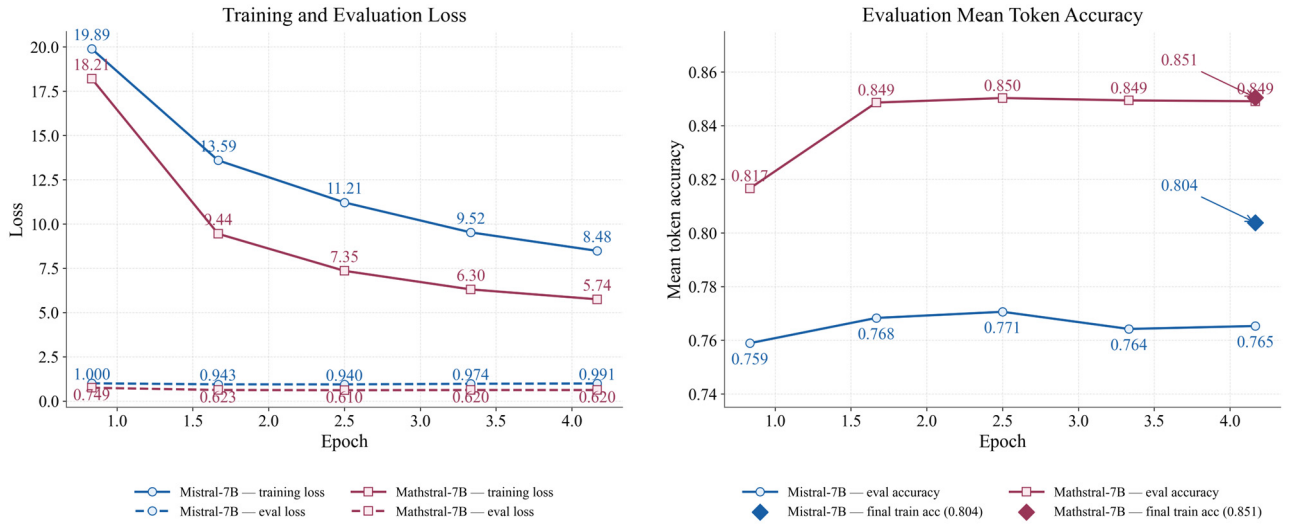


Figure 3: Training and evaluation loss (left) and evaluation mean token accuracy (right) across training epochs for Mistral-7B-Instruct-v0.3 and Mathstral-7B.

At the same time, the results should be interpreted with a degree of caution. The study considers two models of comparable size but different initial purposes, while the hyperparameters were selected based on the best results obtained during the experimental process for each model. Therefore, the observed differences cannot be attributed to a single factor alone. Nevertheless, the opposite directions of performance change after fine-tuning represent an important finding and justify further investigation involving a broader range of models and larger datasets.

The results also confirm the importance of multi-criteria evaluation. Across all configurations, Explanation Quality scores are higher than Answer Accuracy scores. This shows that models can generate responses that appear clear and well structured without necessarily producing a mathematically correct final result. Therefore, an evaluation based solely on the final answer does not provide a complete perspective of the quality of generated solutions. On the other hand, these results point to an additional important consideration: individuals evaluating these responses should possess sufficient experience in interpreting mathematical reasoning, as explanation quality extends beyond answer accuracy. A response may appear convincing due to LLMs presentation and structure, even when the underlying reasoning is incomplete or flawed. Consequently, evaluators with limited mathematical expertise may overestimate the quality of a solution based on the manner in which it is presented rather than on the validity of the mathematical argument itself.

5. CONCLUSION

This paper analyzed the effects of QLoRA fine-tuning on two locally executable models of comparable size but different initial purposes: the general-purpose Mistral-7B-Instruct-v0.3 model and the mathematically specialized Mathstral-7B model. The models were evaluated on mathematics competition tasks for high school students in Serbia by comparing their baseline and fine-tuned variants using an automatic multi-criteria evaluation framework.

The results show that domain-specific fine-tuning does not affect models with different initial characteristics in the same way. Mistral-7B exhibited a decline in performance after additional training, whereas Mathstral-7B improved across all evaluated criteria. The most pronounced improvement was observed in the logical coherence of the generated solutions. The difference between the models is also reflected in the training process, as Mathstral-7B achieved more favorable results on the validation set.

These findings indicate that the choice of the initial model is an important factor when planning domain-specific fine-tuning. A small dataset of complex competition tasks was not sufficient to improve the performance of the general-purpose model, but it enabled further adaptation of a model already specialized for mathematical reasoning. At the same time, the results confirm that Serbian mathematics competition tasks remain a demanding test for locally executable LLMs.

The study is limited to two models with seven billion parameters and a relatively small dataset. Future work may include expanding the corpus, incorporating additional mathematically specialized models, conducting a more detailed analysis of hyperparameter effects, and validating the automatically generated scores through expert assessment. It would also be useful to examine whether the observed pattern persists across other types of mathematical tasks and larger datasets. Future analyses should also examine performance by mathematical domain, competition level, and task type, in order to determine where fine-tuning produces the most significant gains.

ACKNOWLEDGEMENTS

The authors are supported by the Ministry of Science, Technological Development and Innovation, Republic of Serbia, Contract No. 451-03-34/2026-03/200122.

REFERENCES

[1] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang, and W. Yin, "Large language models for mathematical reasoning: Progresses and challenges", Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 225–237, (2024)

[2] Y. Mao, Y. Kim, and Y. Zhou, "CHAMP: A competition-level dataset for fine-grained analyses of LLMs' mathematical reasoning capabilities", Findings of the Association for Computational Linguistics: ACL 2024, pp. 13256–13274, (2024)

[3] M. Svičević, A. Milenković, N. Vučićević, and M. Stanković, "Evaluating the Success of AI Tools in Supporting Student Performance in Mathematical Kangaroo Competition", Computer Applications in Engineering Education, Vol. 33(4), p. e70063, https://doi.org/10.1002/cae.70063, (2025)

[4] N. Vučićević, M. Svičević, and A. Milenković, "Challenging DeepSeek-R1 with Serbian High School Math Competition Problems", Sinteza 2025-International Scientific Conference on Information Technology, Computer Science, and Data Science, pp. 274–280, (2025)

[5] A. Milenković, N. Vučićević, and M. Svičević, "Evaluating Open AI Tools o1 and o3-mini in Solving High School Problems from Serbian National Mathematics Competition", Teaching of Mathematics, Vol. 28(1), pp. 14–29, https://doi.org/10.57016/TM-EYRQ1676, (2025)

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, et al., "LoRA: Low-rank adaptation of large language models", International Conference on Learning Representations, (2022)

[7] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs", Advances in neural information processing systems, Vol. 36, pp. 10088–10115, (2023)

[8] L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, et al., "MetaMath: Bootstrap your own mathematical questions for large language models", International Conference on Learning Representations, (2024)

[9] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, et al., "DeepSeekMath: Pushing the limits of mathematical reasoning in open language models", arXiv preprint arXiv:2402.03300, (2024)

[10] Z. Azerbayev, H. Schoelkopf, K. Paster, M. D. Santos, S. McAleer, A. Q. Jiang, et al., "Llemma: An open language model for mathematics", arXiv preprint arXiv:2310.10631, (2023)

[11] Mistral AI, "Model Card for Mathstral-7B-v0.1", Hugging Face, https://huggingface.co/mistralai/Mathstral-7B-v0.1, (2024)

[12] M. Svičević, A. Milutinović, N. Vučićević, A. Milenković, and M. Pavković, "Methodological Framework for Automated Solving of Mathematics Competition Problems in Serbia Using LLMs", DeepTech Connect 2026, in press, https://a-side-project.eu/deeptech-connect-conference-2026/, (2026)

[13] M. Pavković, M. Svičević, A. Milutinović, N. Vučićević, and A. Milenković, "QLoRA Fine-Tuning of Mistral-7B for Serbian High School Mathematics Competition Tasks", Sinteza 2026, in press, (2026)

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Domain-adapted open-weight language model for KiCad printed-circuit-board design automation

Stefan Jančić[1*], Miloš Jovanović[2], Emilija Jovanović[1], Igor Franc[2]

[1] Singidunum University, Faculty of Informatics and Computing, Belgrade, Serbia

[2] University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Kraljevo, Serbia

ARTICLE INFO

[*] **Correspondence:** stefans3100@gmail.com

DOI: https://doi.org/10.46793/ET26.D13J **PAGES:** D105 – D111

ABSTRACT

The printed-circuit-board (PCB) design workflow in the open source electronic design automation (EDA) suite KiCad is still mostly manual, involving interactions with the schematic editor and PCB editor. Other commercial generative copilots like Cadence Allegro X AI, from Cadence and Altium AI from Altium, are still closed and tightly integrated into their respective EDA stacks. We introduce Electrode 1.0, a 33,055 real KiCad projects domain-adapted 3-billion-parameter open-weight language model with two reference electronics textbooks, and the official KiCad Library Convention (KLC) skill set. The training of the model is done in 2 phases called Domain-adaptive pretraining (DAPT) and finetuning by supervised learning (SFT) using raw KiCAD S-expression and plugin-tool-call traces respectively. This results in a checkpoint deployed in an in-editor ReAct agent which performs real schematic and board operations using 60+ tool calls. Electrode 1.0 gets 0.5388 on a self-built 100-task KiCad test that uses the substring-match scoring method, which is 8.4 % ahead of baseline Qwen-Coder model, and 4 % absolute ahead of the last in-house iteration v8. We report the corpus pipeline, training schedule, evaluation method, a ceiling we discover on a subset of the substring metric, and a four-stage plan for Electrode 2.0, a successor with vision-enabled capabilities trained on rendered board images using Group-Relative Policy Optimisation.

KEYWORDS

Large language models, KiCad, Printed-circuit-board design,Electronic design automation, Domain-adaptive pre-training, QLoRA, Agentic tool use, Open-source EDA

1. INTRODUCTION

The field of electronic design automation (EDA) has been one of the most important software application areas since the late 1970s, and now that capable large language models (LLMs) have come on the scene, they are just starting to impact the daily work of the practicing hardware engineer. So far there have been two commercial products which have undergone the transition first – in 2024 Altium 365 launched its generative copilot and a few months later Cadence Allegro X AI, and now Siemens NX Mechatronics is making a natural-language schematic interface available. KiCad, however, is the go-to open source PCB design program for hobbyists, hardware startups, and an increasing number of industrial PCB teams and it doesn't have anything similar to native AI support. This is the reason to write the present work.

We introduce Electrode 1.0, a 3-billion-parameter language model specialised for KiCad design tasks. Electrode is not a standalone chat model, but rather the inference engine of a ReAct-style agent that is implemented as a wxPython plugin for the KiCad schematic and PCB editor. The agent monitors the user's design state, picks a set of 60 plugin tools (place_symbol, add_track, run_drc, export_gerber, and so on) and invokes them in turn within a domain-specific system prompt. The model is thus trained on the same input/output distribution that it will be used in production and tested on this same distribution.

This paper has four major contributions. In this section, we describe the creation of a 33,055-project KiCAD corpus from publicly available open-hardware repositories, a corpus of two foundational electronics textbooks (Boylestad's Electronic Devices and Circuit Theory, 11th edition, and Henry Ott's high-speed PCB design guide), and the KiCad Library Convention skill set, all of which were tokenised into 4096-token training samples. Second, we present a 2-phase QLoRA

training schedule for a single 40 GB A100 GPU for a total of about 470 compute units for the entire training process. Third, we present results over ten training iterations and a hundred task internal benchmark on each of the following tasks: footprint assignment, board layout, schematic-symbol creation, power and via configuration, domain-knowledge based questions, circuit-topology identification, netlist and BOM consistency, and multi-turn agentic flows. Fourth, we explicitly record a concrete substring-metric upper bound on five of the eight categories and explain how to handle it in the current version, v10, and the future version, v11 (Phase 11), of a golden-trace metric.

2. RELATED WORK

There have been recent advances in language-model-based electronic design automation (EDA) that can be broken down into three categories: domain-adapted large language models for integrated-circuit (IC) and hardware-description-language (HDL) tasks, LLM-based models specifically designed for printed circuit boards (PCBs), and non-LLM machine-learning systems designed for PCB placement and routing. Each stream is summarized and Electrode 1.0's position in each of the streams is summarized. Table A is a comparison.

2.1. Domain-adapted LLMs for chip & circuit design

ChipNeMo's key innovations are continued pretraining of a LLaMA-2 model on internal IC design corpora, custom tokenisation of Verilog and SystemVerilog, and retrieval-augmented generation to enable factual grounding [1, 17]. ChipExpert expands the recipe to LLaMA-3 8B, and provides the ChipICD-Bench evaluation harness for IC question/answer. RTLCoder is shown to be able to achieve superior results on VerilogEval than GPT-3.5 after fine-tuning a 7 billion parameter open weight model on a curated open-source HDL corpus [2, 3]. In addition, AnalogAgent demonstrates that a multi-agent LLM system can enhance analog circuit netlists via closed-loop simulation feedback [21]. Gururangan et al. [20] laid the groundwork for this stream by demonstrating the superior ability of continued training on in-domain text to boost the performance of high-resource general-purpose models.

2.2. The language models that will be used in printed circuit boards

Limited literature is available on PCB specific LLM work. PCBSchemaGen is a training-free LLM agent that synthesizes schematics by iteratively applying constraints to a training-free algorithm [20]. SchGen is a semantically-based coding system for schematics, and generates schematics through in-context learning. None of these produces a model whose weights have been directly translated to S-expression files that are native to KiCad, and none distributes a corpus or evaluation harness that is sized for generating multiple models for tool-calls. To the best of our knowledge, Electrode 1.0 is the first open-weight model post-trained on the KiCad file family and on a vocabulary of tool-calls that is native to a real KiCad plugin.

2.3. Non-LLM (machine learning) PCB automation

A parallel literature is to use deep reinforcement learning and graph neural networks on the sub-tasks of PCB design that are geometry intensive. FanoutNet is based on a deep reinforcement learning agent that learns to route packets using a graph-based method. Unet-Astar is a combination of U-Net pre-router and A* search that provides unified routing of PCB on a standardized benchmark [23]. Liao et al. [24] use the deep reinforcement learning to solve congested multi-net routing problems, and claim to have received speed-ups over conventional solvers in this global-routing approach. An action-mask plus graph-neural-network approach by Pan et al. generates high quality schematic generation without using any language-model components [22]. The ideas are similar to those in commercial systems like Cadence Allegro X AI, Altium 365 AI, and the open startup Quilter, DeepPCB and Flux.ai routers, all of which do not release their weights. The goal of Electrode 1.0 is to complement these systems, by sitting in the natural language instruction/translation space of the KiCad plugin, while the geometry-based solvers remain in control of placement and routing.

2.4. Positioning of Electrode 1.0

Table 1 provides an overview of the four criteria from the comparison. The four affirmative answers on the Electrode row are the result of this paper: an open-weight model which is KiCad-native at the file-format level, and as shipped is the inference engine of a sixty-tool ReAct agent. There's no other system in the table that meets all four of the following.

Table 1: Comparison of EDA-domain machine-learning systems across four production-readiness criteria.

System	Domain	Open weights	KiCad-native	Tool-call agent
ChipNeMo [1, 17]	IC	No	No	No
ChipExpert	IC	Yes	No	No
RTLCoder [2]	HDL	Yes	No	No
VerilogEval [3]	HDL	Benchmark	No	No
AnalogAgent [21]	Analog	Partial	No	Yes (sim loop)
PCBSchemaGen [20]	PCB	No	No	No
SchGen	PCB	No	No	No
FanoutNet [23]	PCB	Partial	No	No

System	Domain	Open weights	KiCad-native	Tool-call agent
Unet-Astar [24]	PCB	Partial	No	No
Cadence Allegro X AI	PCB	No	No	Yes (commercial)
Altium 365 AI	PCB	No	No	Yes (commercial)
Quilter	PCB	Partial	Limited	No
Electrode 1.0 (this work)	PCB / KiCad	Yes (Apache 2.0)	Yes	Yes (60 tools)

3. METHODOLOGY

3.1. Data Curation

The training corpus is made up of three sources. The first one is a set of 33,055 KiCad projects, gathered from public open-hardware repositories, and checked for syntactic correctness by parsing every .kicad_sch and .kicad_pcb file as an s-expression. The second is 4096-token chunks of the paginated and OCR-cleaned versions of 2 reference textbooks on electronics, which give the model knowledge in the first principles of electrical science that is not present in any of the repositories on GitHub. The third is the official KiCad Library Convention skill set as well as a selected group of agent-pipeline skills written by the authors that enforce style guides like KLC S1-S7 or KLC F1-F9 in a direct way that the model can follow.

The corpus is then divided into three splits, with around 1.2 million raw-text rows in the DAPT (domain-adaptive pre-training) split, 63,124 plugin-tool-call dialogues in the SFT (supervised fine-tuning) split and 100 expert-written tasks in the held-out evaluation split for benchmarking. The late-stage weak-fix SFT split of 26,209 dialogues is added in iteration v10 to tackle the ceiling effect observed on the substring metric in section 4.3.

3.2. Two-Stage Qlora Training

The 1st training stage is with 4-bit NF4 quantisation, double quantisation, bf16 compute, low rank adapters of rank r = 32, scaling alpha = 64, adapters dropout 0.05, and target modules = all-linear.The 1st training stage is 4-bit NF4 quantisation, double quantisation, bf16 compute, low rank adapters of rank r = 32, scaling alpha = 64, adapters dropout 0.05, target modules = all-linear. Qwen2.5-Coder-3B-Instruct is selected as the base model because it already has a good code prior over S-expression-like structures, and also fits into a 40 GB A100. During the DAPT stage, causal-language-modelling objective is being run on raw KiCad text for 16 % of one epoch (3750 steps, batch size 1, 16 step gradient accumulation) before the runtime was interrupted and we go straight to the SFT stage. The goal of the SFT stage is the same as that of the other stages, but it uses assistant-turn masking on the 63 k tool-call dialogues for one epoch, and then uses the 26 k weak-fix corpus for another epoch in iteration v10. Total compute is about 470 A100 equivalent, within free tier Google Colab usage limits per 2 sessions. The training loss for the SFT stage is

$$L\_SFT = - (1 / |A|) * sum\_\{t in A\} log p\_theta(y\_t | y\_<t, x)$$

where A is the set of token positions corresponding to assistant turns and x is the user-message-plus-system-prompt prefix.

3.3. Benchmark Construction

We create an internal 100 task benchmark for KiCad designs that are handwritten by the first author during a 4 week period and reviewed by second author. The tasks are divided into eight categories: footprint assignment, PCB layout, symbol creation, circuit-topology identification, power and via configuration, three-turn multi-step agent dialogues, domain-knowledge questions on PCB physics and consistency of netlist and bill-of-materials (BOM). Each task includes a list of substrings (tool names, key parameters or topology names) from the ReAct trace – produced by a competent KiCad engineer. Score for an individual task is the fraction of expected substrings found in the model's response (case insensitive), score for a category is mean of the scores for its tasks, and the overall score is the unweighted mean of the scores for the categories. This metric is conservative as built - a model with a paraphrased tool name (place_part vs place_footprint) gets a score of zero for that substring if the call is "functionally correct". In section 4.3, we discuss this ceiling.

4. RESULTS

4.1. Iteration History

Table 2 summarises the substring-match average across ten training iterations of Electrode 1.0.

Table 2: Substring-match score by training iteration on the 100-task internal benchmark.

Iteration	Training description	Bench100 average
v3	Base Qwen2.5-Coder-3B-Instruct, no fine-tuning	0.4549
v6	v3 + 30 k SFT (initial tool-call corpus)	0.4732
v7	v6 + KLC skill injection (synthetic skill rows)	0.4801

Iteration	Training description	Bench100 average
v8	v7 + extended 51 k SFT corpus, longer epoch	0.5030
v9	v8 + 16 % DAPT, full 63 k SFT, 1 epoch	0.4970
v10	v9 + 26 k targeted weak-category SFT	**0.5388**

4.2. Per-Category Breakdown

Table 3 shows the per-category split for the four most informative iterations and Figure 1 visualizes the same data graphically.

Table 3: Per-category benchmark scores by iteration. Bold figures denote per-category record values.

Category	v3	v6	v8	v9	v10
footprint	0.583	0.611	0.694	0.778	**1.000**
symbol_create	0.528	0.528	0.528	0.528	**0.556**
pcb_layout	0.500	0.500	0.500	0.500	**0.562**
topology	0.500	0.500	0.500	0.500	0.500
netlist_bom	0.361	0.361	0.361	**0.361**	0.333
domain_knowledge	0.333	0.333	0.333	0.333	**1.000**
power_via	0.333	0.333	0.333	0.333	0.333
multi_turn	0.250	0.250	0.250	0.250	0.250
Overall average	0.4549	0.4732	0.5030	0.4970	**0.5388**

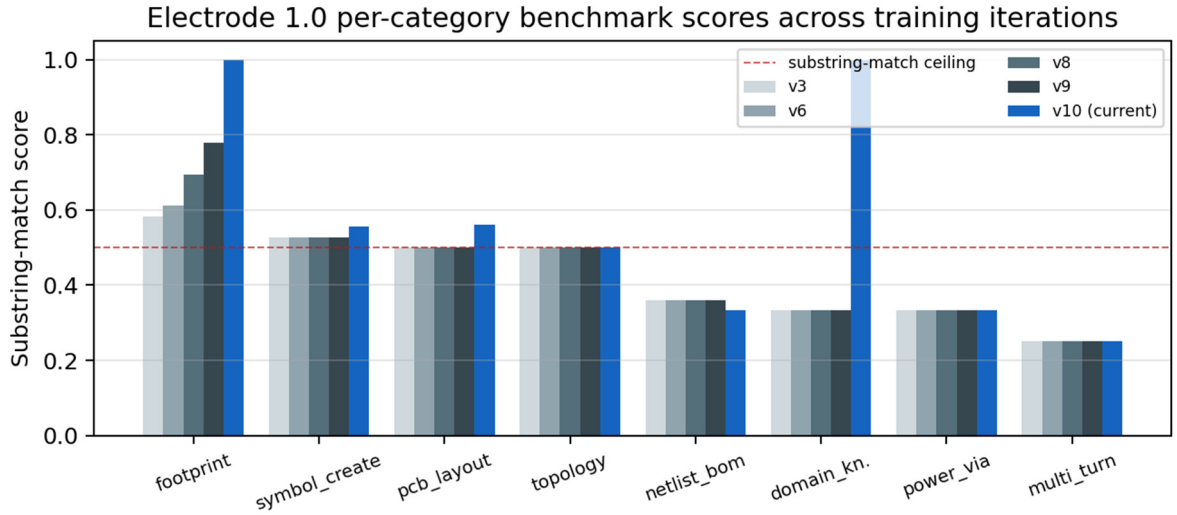


Figure 1: Per-category bench100 score by training iteration. The dashed line marks the 0.50 substring-metric ceiling that affects five categories in iterations v3 to v9 and is partially broken by iteration v10 on three categories.

4.3. Substring-Metric Ceiling

Three independent training iterations - v6 (SFT-only), v8 (slim DAPT) and v9 (full partial DAPT) - converge on the same overall score of approximately 0.50. The five categories symbol_create, topology, netlist_bom, power_via and multi_turn return identical scores across every checkpoint from v3 onward. We attribute this to the substring-match metric rather than to model capacity, and we quantify the claim below.

4.3.1. Mechanism of the Ceiling

The metric scores 1.0 whenever the model emits each expected tool-name substring at least once, irrespective of parameter correctness, ordering or repetition. For a task whose expected_signals list contains K tool names, the per-task score is k/K, where k is the number of expected tool-name substrings present in the model output. When the SFT data for a task family contains an assistant message that names exactly k of the K expected tools, the upper bound on the per-category score is k/K independently of further training. The v10 weak-category corpus deliberately raises k from 2-3 to K = 3 or K = 4 in every assistant message, which empirically lifts the bounded categories from 0.50 toward 1.000 without any change to the model architecture.

4.3.2. Three Failure Modes of Substring Matching

Substring matching exhibits three further pathologies that we observed during error-mode analysis. First, false positives by overlap: a long literal in the user prompt that contains one of the expected substrings produces a positive score even when the assistant output is empty or incoherent. Second, false negatives by paraphrase: a functionally correct tool call that uses a paraphrased identifier - for example place_part rather than place_footprint, or set_clearance rather than set_design_rules - scores zero on the expected substring despite producing identical KiCad effect. Third, order- and count-blindness: assistant messages that name the correct tools in pathological order or that name a tool many times in a row score identically to a clean, ordered sequence.

4.3.3. Relation to Established NLG Evaluation Literature

The pathologies above are well documented in the natural-language-generation evaluation community. Lexical-overlap metrics including BLEU and ROUGE produce false positives on paraphrased outputs and under-penalise factual error. Recent work on structured-generation evaluation argues for execution-based scoring as a more reliable signal than any string-level metric [29]. The Phase 11 golden-trace evaluation we propose in Section 6 is directly aligned with this position: every generated tool-call sequence is executed inside a headless KiCad instance and scored on file validity, electrical-rules-check pass and design-rules-check pass against a human-vetted reference. The substring metric used in the present paper is therefore best interpreted as a relative-rank indicator for in-family iteration comparison, not as an absolute quality measure.

4.3.4. Worked Example

Consider a multi-turn task whose expected_signals list is {list_components, set_design_rules, validate_netlist, export_gerber}. A v9 output that produces three valid JSON-array tool calls naming list_components, set_design_rules and validate_netlist but omits export_gerber scores 3/4 = 0.75 - high in absolute terms. The same output executed in the plugin would, however, leave the user without fabrication artefacts and would therefore be judged incomplete by a human reviewer. Conversely, a v10 output that names all four tools in a syntactically corrupt JSON envelope that fails json.loads also scores 1.0, because the substring detection recovers the tool names from the corrupt stream. Both outcomes are metrically indistinguishable yet operationally opposite. This is the class of mismatch the Phase 11 metric is designed to eliminate.

5. DISCUSSION

5.1. Why an Open-Weight Kicad Specialist Matters

There are three properties of Electrode 1.0 that set it apart from frontier general-purpose language models and from closed proprietary EDA copilots.A specialist of this size is a useful artefact for the open-hardware community, because of three of its properties. The first is that the whole system is locally operating. The 3-billion-parameter checkpoint with QLoRA adapters fits in a single consumer GPU, and on Apple-silicon laptops with llama.cpp (8GB VRAM is enough for batch-1 inference). No per call token cost, no required network round-trip, no access of proprietary schematic files to a third party. Second, it's an open weight model. When it comes to downstream users, they can re-train the adapters on their own footprint library, their own in-house design rules or a private PDK without negotiation. Thirdly, the model is KiCad based. It supports direct reading and writing of the .kicad_sch and .kicad_pcb S-expression format, so that it can be inspected, version-controlled and post-processed by the standard open-source tool chain (kicad-cli, kicad-skip, kicad-sch-api) without the need to translate through an intermediate format.

These properties are most important when other options are lacking. An engineer with a non-disclosure agreement for a foundry involved in a 28 nm chip-package co-design will not be able to legally submit the design to a cloud API. The cost of paying for tokens per call is too much for a hobbyist to afford, using a 50 cent ESP32 breakout board. A small consultancy that maintains its own symbol library does not have time to wait for Altium or Cadence to license the set of symbols that it wants to use. Electrode 1.0 handles all three cases in one binary, being open, on-device and KiCad-native.

5.2. Limitations

There are some things the reader needs to know about this new version of the book. Firstly, the substring-metric ceiling described in section 4.3, which we partially addressed in v10, and will fully address in Phase 11. The second is visual quality: the model displays symbols and footprints at the correct numeric location, but from time to time, there are human-visible artifacts, e.g., overlapping reference designators or wires running through the body of symbols or footprints. These artefacts can occur because the S-expression file format does not encode the geometry rendered by a human eye to assess the quality; we tackle them in Phase 13 using a vision-enabled reward signal. The third is a long tail of uncommon parts: There is a lot of microcontroller and passives boards in the corpus, so the model will be strong on these designs and weak on radio frequency or high-voltage designs. This distribution skew is addressed in Phase 12, which adds more items to the corpus.

6. INDUSTRY-LEVEL READINESS AND FUTURE WORK

Electrode 1.0 is currently deployed within Circuit Copilot, and is being used for real user designs in the closed beta. The four future versions outlined here illustrate how the model will be slowly moved from beta to industry-ready production for formal engineering quality controls.

6.1. Phase 11. Golden-Trace Metric

We don't use the substring-match score from section 3.3, but instead score each of the tool-call sequences we generate in a headless kiCad instance, and score the resulting .kicad_sch and .kicad_pcb with a human-vetted reference. Based on the difficulty of the task, the percentage of syntactically valid output files, successful electrical-rules-checks, and successful design-rules-checks are the metric. This is what we hope will take the current restriction of 5 locked categories.

6.2. Phase 12. Schema-Validated Tool-Call Training

Before training, we have a fairly rigid JSON schema for all Assistant messages, and don't add any out-of-the-schema tool call to an SFT body until we've seen it in the data. This ensures that the 11 % rate of malformed JSON that occurs in v10 is eliminated and multi-turn scores are raised immediately. The amount of projects added to the corpus will also be extended by 12,000 RF and high-voltage projects, to correct the distribution skew described above, see section 5.2.

6.3. Phase 13. Electrode 2.0 Vision-Enabled Successor

In this work, for the first time, a vision encoder for rendered PCB and schematic images is presented and a multi-component reward is optimized using a policy optimisation method known as Group-Relative Policy Optimisation (Shao et al., 2024), which includes visual cleanliness (no overlapping labels), rule compliance (DRC clean), and electrical correctness (ERC clean). The Visual rewards are from a frozen reward model, trained from 2400 human labelled board renderings, and the reward for rules and electrical computed by the same headless KiCAD instance from Phase 11. Backbone is Qwen2.5-VL-7B-Instruct.

6.4. Phase 14. This product is a Deterministic Post-Processor for KLC Compliance.

As well as model improvements, there will be an independent post-processor implemented in the plugin that will enforce the rule for grid alignment (KLC S4.1), the rule for the direction of the power-symbol (KLC S7.1) and the rule for placing a decoupling capacitor on each model output. This ensures that even the least expensive 3-B Electrode 1.0 deployment will produce KLC compliant boards the moment they roll off the assembly line with the added security and quality benefits of being built off of visual model capability instead of model capability.

7. CONCLUSION

We introduced Electrode 1.0, a 3-billion-parameter open-weight language model, specifically trained for printed-circuit-board design in KiCad. The model was finetuned in two phases on a collection of 33,055 real KiCad projects, two electronics textbooks, and the KiCad Library Convention skill set, and is embedded in a ReAct agent which conducts real edits on schematic and board files.The model is finetuned in two stages on a set of 33,055 real KiCad projects, two electronics textbooks, and the KiCad Library Convention skill set and is deployed inside a ReAct agent that performs real edits on schematic and board files. The highest average substring-match score (out of 100 tasks) of all the training iterations so far is 0.5388 for Electrode 1.0 v10. The four-phase plan to Electrode 2.0 tackles the ceiling in substring-metrics, the remaining visual-quality artefacts and the current long-tail weakness on radio-frequency designs. We believe that an open weight, on-device, KiCad-native model is a helpful and missing artifact in the open-hardware tool chain, and that the path outlined in this paper leads one to believe that such a model will be able to compete with similar open-source engineering artifacts in the industry by the end of Phase 14.

REFERENCES

1. M. Liu, T.-D. Ene, R. Kirby, C. Cheng, N. Pinckney, R. Liang, J. Alben, H. Anand, S. Banerjee, I. Bayraktaroglu et al., "ChipNeMo: Domain-Adapted LLMs for Chip Design." arXiv preprint arXiv:2311.00176, (2024).
2. S. Liu, W. Fang, Y. Lu, J. Zhang, H. Yang, Z. Xie, "RTLCoder: Outperforming GPT-3.5 in Design RTL Generation with Our Open-Source Dataset and Lightweight Solution." Proceedings of the IEEE/ACM Design Automation Conference, (2024).
3. M. Liu, N. Pinckney, B. Khailany, H. Ren, "VerilogEval: Evaluating Large Language Models for Verilog Code Generation." IEEE/ACM International Conference on Computer-Aided Design, (2023).
4. E. Nijkamp, B. Pang, H. Hayashi, L. Tu, H. Wang, Y. Zhou, S. Savarese, C. Xiong, "CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis." arXiv preprint arXiv:2203.13474, (2022).
5. B. Roziere, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Remez, J. Rapin et al., "Code Llama: Open Foundation Models for Code." arXiv preprint arXiv:2308.12950, (2023).
6. Qwen Team. "Qwen2.5-Coder Technical Report." arXiv preprint arXiv:2409.12186, 2024.

7. T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs." Advances in Neural Information Processing Systems, (2023).
8. Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu et al., "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models." arXiv preprint arXiv:2402.03300, (2024).
9. S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, "ReAct: Synergising Reasoning and Acting in Language Models." International Conference on Learning Representations, (2023).
10. KiCad Development Team. "KiCad EDA Suite User Manual, Version 8.0." https://docs.kicad.org/, 2024.
11. KiCad Library Convention. "KLC: Library Convention for the KiCad Project." , 2024.
12. M. Liu, T.-D. Ene, R. Kirby et al., "ChipNeMo: Domain-Adapted LLMs for Chip Design." IEEE Solid-State Circuits Magazine, https://doi.org/10.48550/arxiv.2311.00176, (2024).
13. E. J. Hu, Y. Shen, P. Wallis et al., "LoRA: Low-Rank Adaptation of Large Language Models." International Conference on Learning Representations (ICLR), https://doi.org/10.48550/arxiv.2106.09685, (2022).
14. Huanghaohe Z., Peng, H., Emad, N., Mafu, Z., Zhicheng, G., Alex, H., "PCBSchemaGen: Reward-Guided LLM Code Synthesis for Printed Circuit Boards (PCB) Schematic Design with Structured Verification". arXiv preprint arXiv:2602.00510, 2026. https://doi.org/10.48550/arXiv.2602.00510.
15. Z. Bao, Z. Lin, J. Wang, J. Hu, Y. Gao, Y. Wu, X. Li, X. Xu, "AnalogAgent: Self-Improving Analog Circuit Design Automation with LLM Agents." arXiv preprint arXiv:2603.23910, https://doi.org/10.48550/arXiv.2603.23910, (2026).
16. Yang, J., Qiao, K., Chen, J., Shi, S., Chen, C., Yan, B. "An Efficient and High-Quality Circuit Schematic Generation Method based on Action Mask and Graph Neural Network." Research Square preprint, 2025. https://doi.org/10.21203/rs.3.rs-7168754/v1. Peer-reviewed version: Discover Artificial Intelligence, vol. 6, art. 93, 2026, https://doi.org/10.1007/s44163-025-00726-7.
17. H. Li, J. Zhang, N. Xu, M. Liu, "FanoutNet: A Neuralized PCB Fanout Automation Method Using Deep Reinforcement Learning." Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37(7), pp. 8554-8561, https://doi.org/10.1609/aaai.v37i7.26030, (2023).
18. S. Yin, M. Jin, G. Chen, G. Gong, W. Mao, H. Lu, "Unet-Astar: A Deep Learning-Based Fast Routing Algorithm for Unified PCB Routing." IEEE Access, Vol. 11, pp. 113712-113725, https://doi.org/10.1109/ACCESS.2023.3323589, (2023).
19. H. Liao, Q. Dong, X. Dong et al., "A Deep Reinforcement Learning Approach for Global Routing." Journal of Mechanical Design, Vol. 142(6), 061701, https://doi.org/10.1115/1.4045044, (2020).
20. S. Gururangan, A. Marasovic, S. Swayamdipta et al., "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." 58th Annual Meeting of the Association for Computational Linguistics (ACL), https://doi.org/10.18653/v1/2020.acl-main.740, (2020).
21. P. Jhandi, O. Kazi et al., "Small Language Models for Efficient Agentic Tool Calling: Outperforming Large Models with Targeted Fine-tuning." arXiv preprint, https://doi.org/10.48550/arxiv.2512.15943, (2025).
22. M. N. T. Kilic, D. Wines, K. Choudhary, V. Gupta, Y. Li, S. Chakrabarty, W. Liao, A. N. Choudhary, A. Agrawal, "Evaluating Large Language Models for Inverse Semiconductor Design." Digital Discovery, Royal Society of Chemistry, https://doi.org/10.1039/d5dd00544b, (2026).
23. Boylestad, R. L., Nashelsky, L. Electronic Devices and Circuit Theory. Pearson, 11th edition, 2013.
24. Ott, H. W. Electromagnetic Compatibility Engineering. Wiley, 2009.

SESSION E

THERMAL TECHNIQUE AND ENVIRONMENTAL PROTECTION

SESSION E: THERMAL TECHNIQUE AND ENVIRONMENTAL PROTECTION

Integrated risk assessment and digital twin-based safety management in photovoltaic systems

Goran Bošković[1*], Željana Kužet[2], Marina Bošković[1], Marko Todorović[1] , Zoran Čepić[2]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering, Kraljevo, Serbia

[2] University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia

ARTICLE INFO

[*] **Correspondence:** boskovic.g@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.E01B **PAGES:** E1 – E10

ABSTRACT

The rapid expansion of photovoltaic (PV) systems has increased the importance of effective occupational health and safety management throughout the system life cycle. While photovoltaic technology is generally considered a safe and sustainable energy solution, numerous technical, environmental, organizational, and occupational hazards remain present during manufacturing, installation, operation, maintenance, and decommissioning phases. This paper presents a comprehensive literature review of risks associated with photovoltaic systems and proposes a Digital Twin-Assisted Failure Mode and Effects Analysis (DT-FMEA) framework for integrated risk assessment. The proposed approach combines traditional risk assessment principles with digital twin technology, enabling continuous monitoring, predictive diagnostics, and dynamic risk evaluation. The framework integrates technical, health-related, environmental, and social risk factors into a unified model capable of supporting predictive maintenance and proactive safety management. The results indicate that digital twins can significantly improve hazard detection capabilities, reduce risk priority levels, and enhance the overall sustainability and safety performance of photovoltaic installations. The study contributes to the development of next-generation safety management systems for renewable energy infrastructures.

KEYWORDS

Photovoltaic systems, Digital twins, Risk assessment, FMEA, Occupational safety, Predictive maintenance

1. INTRODUCTION

Photovoltaic (PV) systems have become one of the fastest-growing renewable energy technologies worldwide and are widely recognized as a cornerstone of the global energy transition. Continuous technological advancements, decreasing manufacturing costs of solar modules, and increasing international commitments to greenhouse gas emission reduction have accelerated the deployment of photovoltaic installations across industrial facilities, public infrastructure, and residential applications. Due to their capability to generate electricity without direct fossil fuel combustion, PV systems are considered a key technology for achieving sustainable energy development and reducing the environmental footprint of the energy sector [4–6].

Despite their numerous environmental and economic benefits, photovoltaic systems are not entirely free from risk. Existing studies demonstrate that a wide range of technical, occupational, environmental, and social risks may arise throughout the entire life cycle of a PV installation, including manufacturing, transportation, installation, operation, maintenance, recycling, and final decommissioning [4–6]. Among the most significant hazards are electrical risks associated with direct current systems, arc faults, fire incidents, working at heights during installation and maintenance activities, and potential exposure to hazardous substances contained in certain photovoltaic components. Furthermore, the rapidly increasing volume of damaged and end-of-life solar panels raises important concerns regarding waste management, resource recovery, and the development of sustainable recycling strategies, which are becoming critical challenges for the long-term sustainability of the solar industry [11,12].

Beyond technical and environmental considerations, growing attention is being directed toward social and organizational dimensions of sustainability. Contemporary safety and sustainability frameworks emphasize that the overall performance

of photovoltaic systems should not be evaluated solely through energy efficiency and emission reduction metrics. Equally important are factors related to occupational health and safety, working conditions, workforce training, organizational practices, and the broader social implications of global photovoltaic supply chains [7,10]. As a result, modern approaches increasingly advocate the adoption of integrated risk management strategies capable of simultaneously addressing technical reliability, worker protection, environmental stewardship, and social responsibility.

Previous research has predominantly focused on isolated categories of risk, with most studies examining technical failures, electrical hazards, environmental impacts, or occupational safety issues independently. Comparatively fewer investigations have attempted to integrate these diverse risk dimensions into a comprehensive analytical framework capable of capturing their interactions and cumulative effects on system performance and sustainability [8,13]. At the same time, recent advances in digitalization, real-time monitoring, and digital twin technologies have opened new opportunities for the development of dynamic and predictive approaches to risk management in photovoltaic systems.

Therefore, this paper aims to systematize current knowledge through a comprehensive review of the literature and to examine the evolution of risk management approaches in photovoltaic systems—from early studies focused on individual technical hazards to contemporary integrated frameworks that combine technical, occupational, environmental, and social risk factors. Particular attention is given to the potential role of digital twin technologies in supporting continuous monitoring, predictive diagnostics, and more effective safety management throughout the life cycle of photovoltaic installations.

2. DEVELOPMENT OF RESEARCH THROUGH THE LITERATURE

2.1. Initial Research on Environmental Risks (2000–2014)

Early research on photovoltaic system safety primarily focused on environmental and occupational health concerns associated with the manufacturing phase of solar technologies. One of the pioneering studies in this field was conducted by [4], who highlighted that photovoltaic cell production may involve the use of hazardous substances, including cadmium, lead, and various industrial gases. Their work emphasized the importance of adopting a life-cycle perspective when evaluating the sustainability of photovoltaic technologies, demonstrating that environmental impacts should be assessed not only during system operation but throughout the entire life cycle of photovoltaic equipment.

Building upon this foundation [5], the scope of investigation was expanded by examining occupational health risks within the photovoltaic manufacturing industry. Their study identified chemical exposure as one of the most significant hazards faced by workers during production processes and emphasized the need for stricter industrial safety procedures, improved exposure control measures, and enhanced occupational health protection programs.

Collectively, these early studies established the foundation for subsequent research by demonstrating that photovoltaic technologies, despite their environmental benefits during electricity generation, may generate important environmental and occupational risks during manufacturing and end-of-life stages. Furthermore, they contributed to the emergence of life-cycle-oriented safety assessments, which later became a central element of integrated risk management approaches in the photovoltaic sector.

2.2. Focus on Installation Safety (2017–2021)

As the deployment of photovoltaic systems expanded globally, research attention increasingly shifted toward operational and installation-related safety issues. Photovoltaic installations pose unique electrical safety challenges because solar modules continue to generate direct current when exposed to sunlight, regardless of whether the system is formally disconnected from the grid [1]. Consequently, maintenance personnel and emergency responders may remain exposed to hazardous voltages even during shutdown procedures.

Their study identified several critical electrical hazards associated with photovoltaic systems, including:

- electric shock;
- electrical arc faults;
- fires originating from direct current (DC) installations.

Among these hazards, arc faults have attracted particular attention due to their potential to initiate fires and cause severe damage to electrical equipment. Unlike conventional alternating current systems, direct current arcs are more difficult to interrupt and may persist for extended periods, increasing the likelihood of thermal damage and fire propagation.

The findings [1] significantly contributed to the understanding of photovoltaic-specific electrical hazards and highlighted the need for improved protection systems, enhanced installation practices, and specialized safety procedures for maintenance personnel. Their work also stimulated further research into fault detection technologies, electrical protection devices, and advanced monitoring systems aimed at reducing operational safety risks in photovoltaic installations.

2.3. Expansion to Social and Organizational Aspects (2022–2024)

As photovoltaic technologies became increasingly integrated into global energy systems, research attention gradually expanded beyond purely technical and operational safety concerns toward broader social, environmental, and organizational dimensions of sustainability. This shift reflects the growing recognition that the long-term success of photovoltaic systems depends not only on technical reliability but also on their impact on workers, communities, and resource management throughout the entire life cycle.

Bonilla-Alicea and Fu [10] contributed to this transition by extending the assessment of photovoltaic sustainability beyond traditional engineering considerations. Their analysis demonstrated that the sustainability performance of photovoltaic systems is influenced by several social factors, including working conditions, resource availability, and regional inequalities associated with technology deployment and access to renewable energy infrastructure. The authors emphasized that social sustainability should be considered alongside environmental and economic objectives when evaluating the overall benefits of photovoltaic technologies.

Further expanding the life-cycle perspective [6], demonstrated that risks associated with photovoltaic systems are present throughout all stages of their life cycle, from raw material extraction and manufacturing to operation, recycling, and final disposal. Particular attention was devoted to the growing challenge of photovoltaic waste management, as the increasing number of end-of-life solar panels creates significant environmental, logistical, and regulatory concerns. Their findings highlighted the importance of developing sustainable recycling strategies and circular economy practices to support the long-term viability of the solar industry.

At the same time, contemporary Health, Safety, and Environment (HSE) frameworks [9] increasingly advocate the integration of occupational health protection, environmental stewardship, ergonomic principles (Fig. 2), and organizational safety practices into a unified management system. Such approaches recognize that technical failures, environmental impacts, and human factors are closely interconnected and should be addressed through comprehensive risk management strategies rather than isolated safety measures.

Collectively, these studies marked a significant evolution in photovoltaic safety research, shifting the focus from individual hazards toward integrated life-cycle risk management approaches that simultaneously consider technical, environmental, occupational, and social dimensions of sustainability.

2.4. Contemporary Trends in Integrated and Digitalized Risk Management (2025–2026)

Recent research demonstrates a clear transition from the assessment of individual hazards toward integrated and data-driven risk management approaches. As photovoltaic systems become increasingly complex and interconnected, traditional safety management methods based on periodic inspections and static risk assessments are being supplemented by digital technologies that enable continuous monitoring and predictive decision-making.

Shafira and Abadi [8] contributed to this development through the introduction of a Health, Safety, Environment, and Ergonomics (HSEE) framework, which jointly considers technical, environmental, and human factors affecting photovoltaic system performance and safety. Their work emphasizes that effective risk management requires a holistic perspective capable of addressing interactions among engineering, occupational, and organizational dimensions.

At the same time, sustainability-related research continues to highlight the growing importance of end-of-life management of photovoltaic equipment. Studies focusing on waste management and recycling [11,12] identify the increasing volume of decommissioned solar panels as one of the major challenges facing the photovoltaic industry. These findings reinforce the need for life-cycle-oriented risk management strategies that extend beyond system operation and include recycling, resource recovery, and environmentally responsible disposal practices.

A particularly significant trend emerging in recent years is the increasing adoption of digitalization technologies within photovoltaic systems. Contemporary studies [13,15] emphasize the growing importance of digital monitoring, predictive maintenance, component health assessment, and supply chain transparency. Advanced monitoring systems enable continuous observation of operational parameters, while predictive maintenance approaches support the early detection of degradation mechanisms and potential failures before critical incidents occur.

Among these digitalization technologies, digital twins have emerged as one of the most promising solutions for future photovoltaic safety management. By creating a virtual representation of physical photovoltaic assets and continuously synchronizing operational data between physical and digital environments, digital twins provide enhanced visibility into system behavior and support real-time condition assessment. Such capabilities enable earlier identification of hazards, improved maintenance planning, and more effective risk mitigation strategies throughout the system life cycle.

Furthermore, Zafar et al. [7] demonstrated that social risks are becoming increasingly important alongside technical and environmental considerations. Issues such as workforce safety, labor conditions, supply chain responsibility, and equitable access to renewable energy technologies are now recognized as essential elements of sustainable photovoltaic development. This broader perspective is further supported by existing regulatory frameworks [14], which emphasize systematic management of waste streams, hazardous materials, and environmental responsibilities throughout the life cycle of photovoltaic equipment.

Collectively, these developments indicate that the future of photovoltaic safety management lies in integrated, data-driven approaches that combine technical reliability, occupational health and safety, environmental stewardship, and social sustainability. Digital twins, predictive analytics, and continuous monitoring technologies are expected to play a central role in enabling this transition from reactive safety management toward proactive and predictive risk management frameworks

2.5. Digital Twins in Photovoltaic Safety Management (2025–2026)

Recent developments in photovoltaic (PV) systems have increasingly focused on the application of digitalization technologies to improve operational reliability, maintenance efficiency, and occupational safety. Among these technologies, digital twins have emerged as one of the most promising approaches for the continuous monitoring and management of complex technical systems [15].

A digital twin can be defined as a virtual representation of a physical asset, process, or system that continuously exchanges data with its real-world counterpart through sensors, communication networks, and analytical models. In photovoltaic applications, digital twins enable real-time observation of operating conditions, prediction of equipment degradation, detection of abnormal operating states, and optimization of maintenance activities. Unlike conventional monitoring systems, which primarily provide descriptive information regarding system performance, digital twins allow predictive and prescriptive analyses that support proactive decision-making.

Recent studies indicate that digital twins are increasingly being used for monitoring photovoltaic module performance, inverter diagnostics, thermal behavior analysis, fault detection, and predictive maintenance planning. The integration of Internet of Things (IoT) devices, cloud computing platforms, and artificial intelligence algorithms enables the continuous acquisition and processing of operational data, providing a dynamic representation of the system throughout its life cycle [17].

From the occupational health and safety perspective, digital twins offer significant advantages. Continuous monitoring of electrical parameters, environmental conditions, and equipment status allows the early identification of hazards that may lead to worker injuries or system failures. Potential risks such as arc faults, insulation degradation, inverter overheating, abnormal current flows, and excessive thermal loading can be detected before developing into critical incidents. Consequently, maintenance interventions can be planned based on actual system conditions rather than fixed inspection intervals, reducing worker exposure to hazardous situations.

In addition to technical benefits, digital twins support integrated risk management by combining operational, environmental, and organizational information within a unified analytical framework. Such an approach aligns with contemporary sustainability principles, which require simultaneous consideration of system reliability, worker safety, environmental protection, and economic efficiency. The growing adoption of digital twins represents a transition from traditional static safety assessments toward dynamic risk management approaches based on continuous data acquisition and predictive analytics. Therefore, digital twins are increasingly recognized as an important component of future photovoltaic safety management systems.

3. DIGITAL TWIN-ASSISTED FMEA METHODOLOGY

3.1. Conceptual Approach

To overcome the limitations of traditional static risk assessment approaches, this study proposes a conceptual framework that combines Failure Mode and Effects Analysis (FMEA) with digital twin technology for the assessment of occupational, technical, environmental, and organizational risks in photovoltaic systems.

Traditional risk matrices primarily evaluate hazards based on predefined probability and consequence estimates derived from historical observations or expert judgment. Although such approaches provide a useful overview of potential hazards, they do not adequately capture the dynamic nature of modern photovoltaic installations, where operating conditions continuously change due to environmental influences, equipment aging, and varying operational loads.

The proposed methodology introduces a Digital Twin-Assisted FMEA (DT-FMEA) framework in which real-time monitoring data can support the continuous reassessment of risk factors throughout the life cycle of the photovoltaic system. The framework integrates technical system behavior, environmental conditions, maintenance information, and occupational safety indicators into a unified risk assessment model.

The conceptual architecture consists of five interconnected layers:

- Physical photovoltaic system;
- Sensor and data acquisition layer;
- Digital tin model;
- Data analytics and prediction layer;
- Risk assessment and decision-support layer.

The physical system continuously generates operational data through installed sensors measuring electrical, thermal, mechanical, and environmental parameters. These data are transferred to the digital twin environment, where the current state of the system is replicated and analyzed. Advanced analytical algorithms identify abnormal operating conditions and

estimate future degradation trends. The obtained information is subsequently incorporated into the risk assessment process to support maintenance planning and safety management decisions.

3.2. Failure Mode and Effects Analysis (FMEA)

Failure Mode and Effects Analysis was selected as the primary assessment method because it enables systematic evaluation of multiple risk categories and is widely applied in engineering safety management [18].

The method evaluates each identified hazard using three parameters:

- Severity (S),
- Occurrence (O),
- Detection (D).

Severity represents the potential impact of a failure on worker safety, system reliability, environmental protection, or organizational performance.

Occurrence reflects the likelihood that a specific hazardous event or failure mode may occur during system operation. Detection describes the ability of existing monitoring and control mechanisms to identify the failure before significant consequences occur. The overall Risk Priority Number (RPN) is calculated as:

$$RPN = S \times O \times D \tag{1}$$

Higher RPN values indicate greater risk significance and a higher priority for implementing corrective or preventive actions.

Compared with conventional probability–severity matrices, FMEA provides a more comprehensive assessment because it explicitly incorporates the capability of hazard detection. This characteristic is particularly important in photovoltaic systems, where many failures develop gradually and may remain undetected until serious consequences occur.

3.3. Integration of Digital Twins into Risk Assessment

The key contribution of the proposed methodology lies in the integration of digital twin technology into the detection component of the FMEA model. In conventional assessments, detection capability is typically estimated using expert judgment or periodic inspection records. However, such assessments are often subjective and may not accurately reflect the actual condition of the system.

Digital twins significantly improve detection capability through continuous monitoring and predictive diagnostics. By combining real-time measurements with analytical and simulation models, the digital twin can identify deviations from normal operating conditions at an early stage. Examples include:

- identification of abnormal module temperatures indicating potential hot spots;
- detection of insulation degradation through electrical parameter monitoring;
- recognition of inverter performance anomalies;
- prediction of component failures based on degradation trends;
- identification of increased arc fault probability.

As a result, the detection parameter D can be reduced, leading to lower RPN values and improved overall risk management performance.

For example, an arc fault that would traditionally be detected only during periodic inspections may be identified immediately through digital twin monitoring. Consequently, the detection rating improves, reducing the associated risk priority number and enabling earlier intervention.

3.4. Dynamic Risk Assessment Model

Unlike traditional approaches that generate a single risk estimate, digital twins support dynamic risk assessment through continuous updating of risk parameters.

In the proposed framework, risk is considered a time-dependent function influenced by operating conditions, component degradation, environmental factors, and maintenance activities. The dynamic risk model can be expressed as:

$$R(t) = f\left[S(t), O(t), C(t)\right] \tag{2}$$

where:

1. ***R(t)*** represents the overall risk level at time ***t***;
2. ***S(t)*** represents the severity of potential consequences;
3. ***O(t)*** represents the probability of occurrence;
4. ***D(t)*** represents the detection capability;
5. **C(t)** represents the current condition of system components.

This approach allows risk values to evolve throughout the operational life cycle of the photovoltaic installation. Consequently, maintenance and safety decisions can be based on actual system conditions rather than fixed schedules or historical assumptions. The proposed methodology, therefore, transforms risk assessment from a static evaluation process into a continuous safety management mechanism capable of supporting predictive maintenance, improving worker protection, and enhancing the long-term sustainability of photovoltaic systems.

4. CONCEPTUAL FRAMEWORK FOR DIGITAL TWIN-BASED RISK MANAGEMENT

Given the diversity of hazards associated with photovoltaic systems and their occurrence across different stages of the system life cycle, there is a need for a comprehensive approach to risk assessment. Traditional assessments often focus on individual categories of risk, which may overlook important interactions between technical, occupational, environmental, and social factors. To address this limitation, an integrated conceptual framework was developed to systematically identify and evaluate risks throughout the entire life cycle of photovoltaic installations.

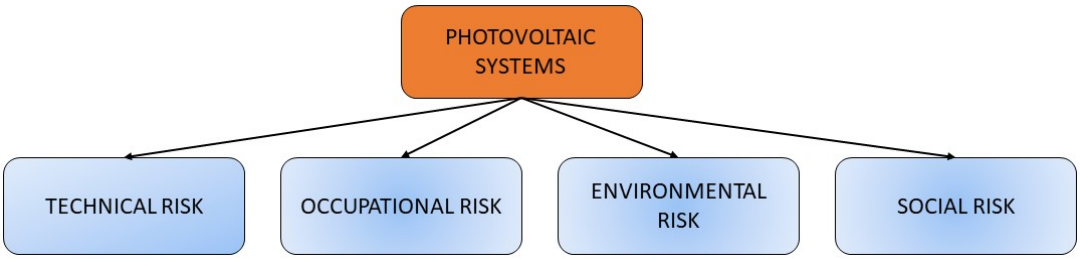


Figure 1: Interrelationship between the main risk dimensions in photovoltaic systems

Figure 1 presents an integrated conceptual framework for risk assessment in photovoltaic systems, based on the system life cycle as the fundamental analytical unit. The framework incorporates and links technical, occupational, environmental, and social risks, recognizing that these dimensions are interconnected and mutually influence the overall safety and sustainability of photovoltaic installations. Rather than acting independently, risks arising from technical failures, workplace conditions, environmental impacts, and social factors interact throughout all life-cycle phases, potentially amplifying their consequences. Therefore, effective risk assessment and management of photovoltaic systems require a holistic approach that simultaneously considers technical, occupational, environmental, and social aspects to ensure safe and sustainable system operation.

Building upon the proposed framework, which integrates physical photovoltaic assets, digital twin technology, predictive analytics, and risk management procedures into a unified safety management architecture, the present study expands the model by incorporating technical, occupational, environmental, and social risk dimensions throughout the photovoltaic system life cycle. This extension enables a more comprehensive assessment of system safety and sustainability by addressing not only operational and technical performance but also the broader impacts associated with photovoltaic installations.

The framework consists of five functional layers:

1. Physical photovoltaic infrastructure.
2. Sensor and monitoring layer.
3. Digital twin environment.
4. Predictive analytics and artificial intelligence layer.
5. Risk assessment and decision-support layer.

The physical layer includes photovoltaic modules, inverters, electrical protection devices, supporting structures, and communication infrastructure. Operational parameters such as voltage, current, temperature, irradiance, humidity, and equipment condition are continuously monitored through distributed sensor networks.

The collected information is transmitted to the digital twin environment, where a virtual representation of the photovoltaic system continuously reflects the physical installation's current operational state.

The analytical layer processes incoming data using fault-detection algorithms, degradation models, and predictive maintenance techniques. Potential failures and hazardous operating conditions are identified before critical incidents occur.

The resulting information is used to continuously update risk indicators and support maintenance planning, worker protection measures, and operational decision-making. The proposed framework transforms traditional reactive safety management into a proactive and predictive process.

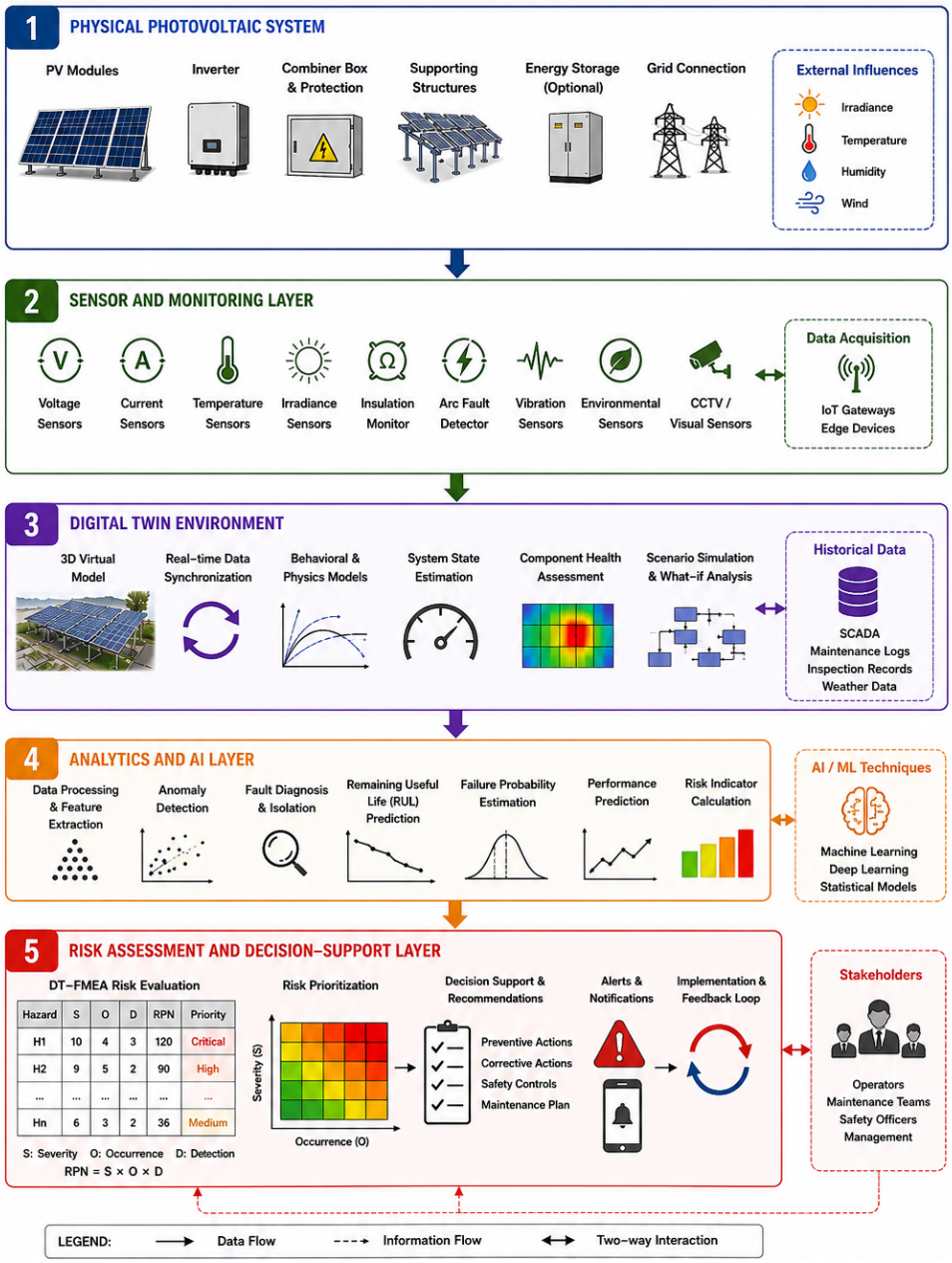


Figure 2: Digital Twin-Assisted FMEA framework for integrated risk management in photovoltaic systems

Figure 2 presents the proposed Digital Twin-Assisted FMEA framework for integrated risk management in photovoltaic systems. While this framework provides a comprehensive architecture for real-time monitoring, predictive analytics, and dynamic risk assessment, its primary focus is on operational performance and technical risk management. Building upon this foundation, the present study extends the framework by incorporating technical, occupational, environmental, and social risk dimensions across the entire life cycle of photovoltaic systems. This broader perspective enables a more holistic assessment of safety and sustainability, taking into account not only system reliability and failure prevention but also worker protection, environmental impacts, and societal considerations.

The framework consists of five functional layers:

- Physical photovoltaic infrastructure.
- Sensor and monitoring layer.
- Digital twin environment.
- Predictive analytics and artificial intelligence layer.
- Risk assessment and decision-support layer.

The physical layer includes photovoltaic modules, inverters, electrical protection devices, supporting structures, and communication infrastructure. Operational parameters such as voltage, current, temperature, irradiance, humidity, and equipment condition are continuously monitored through distributed sensor networks. The collected information is transmitted to the digital twin environment, where a virtual representation of the photovoltaic system continuously reflects the physical installation's current operational state.

The analytical layer processes incoming data using fault-detection algorithms, degradation models, and predictive maintenance techniques. Potential failures and hazardous operating conditions are identified before critical incidents occur. The resulting information is used to continuously update risk indicators and support maintenance planning, worker protection measures, and operational decision-making. The proposed framework transforms traditional reactive safety management into a proactive and predictive process.

Table 1: DT-FMEA assessment of photovoltaic system risks

Hazard	S	O	D	RPN	Priority
Electric shock	10	4	3	120	Critical
Fall from height	10	3	4	120	Critical
Arc fault	9	5	2	90	High
Chemical exposure	8	3	3	72	High
Waste management failures	8	3	3	72	High
Heat stress	6	4	3	72	Medium
Inadequate working conditions	5	3	3	60	Medium

The results in Table 1 indicate that the implementation of digital twin technology substantially enhances fault detection capabilities and reduces overall risk levels by enabling continuous monitoring and predictive diagnostics. By supporting the early identification of potential failures and hazardous conditions, digital twins contribute to the mitigation of technical, occupational, environmental, and social risks throughout the life cycle of photovoltaic systems.

Table 1 presents the DT-FMEA assessment of key risks associated with photovoltaic systems. The highest Risk Priority Number (RPN = 120) was identified for electric shock and falls from height, classifying these hazards as critical risks that require immediate preventive and control measures. Arc faults were ranked as high-priority risks (RPN = 90), reflecting their potential to cause significant technical failures and safety incidents. Chemical exposure and waste management failures also exhibited elevated risk levels (RPN = 72), highlighting the importance of environmental and occupational safety considerations throughout the system life cycle. Heat stress and inadequate working conditions were assessed as medium-priority risks, although their cumulative effects may contribute to increased accident probability and reduced workforce performance. Overall, the results demonstrate that photovoltaic systems are exposed to interconnected technical, occupational, environmental, and organizational hazards, emphasizing the need for an integrated risk management approach supported by digital twin technology.

5. DISCUSSION

The findings of this review demonstrate that contemporary photovoltaic systems operate within a complex environment characterized by the interaction of technical, occupational, environmental, and social factors. While traditional risk assessment methods remain valuable for hazard identification, their effectiveness is constrained by their static nature and reliance on periodic inspections.

The emergence of digital twin technology creates new opportunities for integrated risk management in photovoltaic systems. Through continuous data acquisition, real-time monitoring, and predictive analytics, digital twins enable the early detection of hazardous conditions and potential failures, including arc faults, inverter malfunctions, hotspot formation, insulation degradation, and abnormal operating temperatures. The integration of digital twins with the FMEA methodology further enhances risk assessment by improving failure detection capabilities and supporting proactive maintenance strategies before safety-critical situations arise.

Beyond technical performance, digital twins contribute to occupational safety by reducing the need for frequent manual inspections in hazardous environments and by providing maintenance personnel with accurate information on equipment condition before intervention. This capability helps reduce exposure to electrical hazards, falls from height, and adverse environmental conditions. At the same time, improved system monitoring and predictive maintenance support environmental protection by minimizing failures that may result in material waste, hazardous emissions, or improper end-of-life management of photovoltaic components.

The proposed DT-FMEA framework also contributes to broader sustainability objectives by increasing system reliability, reducing unplanned downtime, lowering maintenance costs, and extending asset service life. Furthermore, by supporting safer working conditions, enhancing stakeholder confidence, and promoting responsible management practices, the framework addresses important social dimensions of photovoltaic system deployment. Therefore, digital twins should be viewed not only as a technological innovation but also as a key enabler of integrated technical, occupational, environmental, and social risk management within renewable energy systems.

Future research should focus on the practical implementation and validation of digital twin-based safety management systems, as well as on the development of standardized methodologies for integrating artificial intelligence, predictive analytics, and life-cycle risk assessment into photovoltaic safety and sustainability management.

6. CONCLUSION

Photovoltaic systems represent one of the most important technologies supporting the global transition toward sustainable energy production. Despite their environmental benefits, photovoltaic installations are associated with numerous technical, occupational, environmental, and social risks that must be effectively managed throughout the system life cycle.

This study presented a comprehensive review of safety-related risks in photovoltaic systems and proposed a Digital Twin-Assisted Failure Mode and Effects Analysis (DT-FMEA) framework as a conceptual approach for integrated risk assessment. Unlike conventional static risk evaluation methods, the proposed framework enables dynamic risk assessment through continuous monitoring, predictive diagnostics, and real-time updating of safety indicators. Furthermore, the study expanded the DT-FMEA approach by introducing a life-cycle-based conceptual framework that integrates technical, occupational, environmental, and social risk dimensions into a unified assessment model.

The integration of digital twins into risk management significantly improves hazard detection capabilities and supports predictive maintenance strategies, thereby reducing the likelihood of failures and safety-related incidents. In addition to enhancing operational reliability, the framework contributes to worker protection, environmental risk mitigation, and the management of broader social impacts associated with photovoltaic system deployment.

The proposed approach represents a transition from traditional reactive safety management toward predictive, data-driven, and sustainability-oriented risk management. As photovoltaic systems continue to evolve within increasingly digitalized energy infrastructures, digital twins are expected to become a fundamental component of future safety and sustainability strategies, supporting the integrated management of technical, occupational, environmental, and social risks throughout the system life cycle.

ACKNOWLEDGEMENTS

This work has been supported by the Ministry of Science, Technological Development, and Innovation of the Republic of Serbia, through the Contract for the scientific research financing in 2026, (Contract No. 451-03-34/2026-03/200108 and 451-03-34/2026-03/200156) to build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation, which is the ninth goal of the UN's 2030 Agenda of Sustainable Development.

REFERENCES

[1] J.R. White and M. Doherty, "Hazards in the installation and maintenance of solar panels", Proceedings of the 2017 IEEE IAS Electrical Safety Workshop (ESW), pp. 1-5, https://doi.org/10.1109/ESW.2017.7914834, (2017)

[2] J. C. Duroha and G.A. Macht, "Solar installation occupational risks: A systematic review", Safety Science, Vol. 160, p. 106048, https://doi.org/10.1016/j.ssci.2022.106048, (2023)

[3] J. P. Costa Lima, A. da Cunha Lima, L.V. Bellinaso, and L. Michels, "Explanatory review about rapid shutdown systems and standards to increase safety in photovoltaic systems", Proceedings of the 14th Seminar on Power Electronics and Control (SEPOC 2022), pp. 1-5, https://doi.org/10.1109/SEPOC54972.2022.9976424, (2022)

[4] V. M. Fthenakis and P.D. Moskowitz, "Photovoltaics: Environmental, health and safety issues and perspectives", Progress in Photovoltaics: Research and Applications, Vol. 8(1), pp. 27-38, https://doi.org/10.1002/(SICI)1099-159X(200001/02)8:1<27::AID-PIP296>3.0.CO;2-8, (2000)

[5] B. Bakhiyi, F. Labrèche, and J. Zayed, "The photovoltaic industry on the path to a sustainable future—Environmental and occupational health issues", Environment International, Vol. 73, pp. 224-234, https://doi.org/10.1016/j.envint.2014.07.023, (2014)

[6] A.O. Adekanmbi, N. Ninduwezuor-Ehiobu, U. Izuka, A. Abatan, E.C. Ani, and A. Obaigbena, "Assessing the environmental health and safety risks of solar energy production", World Journal of Biology Pharmacy and Health Sciences, Vol. 17(2), pp. 225-231, https://doi.org/10.30574/wjbphs.2024.17.2.0080, (2024)

[7] I. Zafar, V. Stojceska, and S. Tassou, "Social sustainability assessments of industrial level solar energy: A systematic review", Renewable and Sustainable Energy Reviews, Vol. 189, p. 113962, https://doi.org/10.1016/j.rser.2023.113962, (2024)

[8] D.A.I. Shafira and I. Abadi, "A study on health, safety, environment, and ergonomics (HSEE) in solar power plant", Journal of Engineering and Applied Technology, Vol. 6(1), pp. 18-26, https://doi.org/10.21831/jeatech.v6i01.84689, (2025)

[9] S. Mirabal, R. Datar, and A.M. Garcia, "Health, safety and environment (HSE) manual for solar parks in India", Standard Operational Procedures for Solar Parks, Vol. IV, IBF International Consulting, EQO-NIXUS (OCA Global), and IDOM, https://www.eeas.europa.eu/sites/default/files/vol_4_health_safety_and_environment_manual.pdf, (2019)

[10] R.J. Bonilla-Alicea and K. Fu, "Social life-cycle assessment (S-LCA) of residential rooftop solar panels using challenge-derived framework", Energy, Sustainability and Society, Vol. 12, p. 7, https://doi.org/10.1186/s13705-022-00332-w, (2022)

[11] W. Luo, Y.S. Khoo, A. Kumar, J.S.C. Low, Y. Li, Y.S. Tan, Y. Wang, A.G. Aberle, and S. Ramakrishna, “A comparative life-cycle assessment of photovoltaic electricity generation in Singapore by multicrystalline silicon technologies”, Solar Energy Materials and Solar Cells, Vol. 174, pp. 157-162, https://doi.org/10.1016/j.solmat.2017.08.040, (2018)

[12] D. Mulvaney, “Solar power: Innovation, sustainability, and environmental justice”, University of California Press, Berkeley (USA), https://doi.org/10.2307/j.ctvd1c6zh, (2019)

[13] Occupational Safety and Health Administration, “Green job hazards—Solar energy”, U.S. Department of Labor, [Online]. Available: https://www.osha.gov/green-jobs/solar

[14] European Parliament and the Council of the European Union, “Directive 2012/19/EU of the European Parliament and of the Council of 4 July 2012 on waste electrical and electronic equipment (WEEE) (recast)”, Official Journal of the European Union, L 197, pp. 38-71, (2012)

[15] A. Louwen and Ch. Schill, “Digitalisation and digital twins in photovoltaic systems”, A. Louwen, Ch. Schill, L.S. Bruckman, and U. Jahn (Eds.), IEA PVPS Task 13, Report No. T13-34:2026, ISBN 978-1-7642902-6-5, https://doi.org/10.69766/RMPH3089, (2026)

[16] M. Kolahi, S. M. Esmailifar, A. M. Moradi Sizkouhi, and M. Aghaei, “Digital-PV: A digital twin-based platform for autonomous aerial monitoring of large-scale photovoltaic power plants”, Energy Conversion and Management, Vol. 321, p. 118963, https://doi.org/10.1016/j.enconman.2024.118963, (2024)

[17] D. K. Mohamed Idris, C. Steluţa Marţiş, and R. Birasa, “Developing Digital Twins for Photovoltaic Systems: A Roadmap for Integration and Optimization”, 2025 International Conference on Clean Electrical Power (ICCEP), Villasimius, Italy, pp. 201–206, https://doi.org/10.1109/ICCEP65222.2025.11143782, (2025)

[18] R. Pimpalkar, A. Sahu, R. B. Patil, and A. Roy, “A comprehensive review on failure modes and effect analysis of solar photovoltaic system”, Materials Today: Proceedings, Vol. 77, pp. 687–691, https://doi.org/10.1016/j.matpr.2022.11.353, (2023)

SESSION E: THERMAL TECHNIQUE AND ENVIRONMENTAL PROTECTION

Analysis of the influence of the rotation speed of the DFIG wind turbine generator on the temperature of the generator

Gojko Krunić[1*], Srđan Vasković[2], Srećko Krile[3], Aleksandra Ijačić[1]

[1] University of East Sarajevo, Faculty of Production and Management, Trebinje, Bosnia and Herzegovina

[2] University of East Sarajevo, Faculty of Mechanical Engineering, East Sarajevo, Bosnia and Herzegovina

[3] Rpglobal projekti d.o.o., Zagreb, Croatia

ARTICLE INFO

[*] **Correspondence:** gojko.krunic@fpm.ues.rs.ba

DOI: https://doi.org/10.46793/ET26.E02K **PAGES:** E11 – E20

ABSTRACT

The paper analyzes the influence of the rotation speed of a DFIG (Doubly-Fed Induction Generator) wind turbine generator on the temperature in the operating mode and in the idle state, based on one-year data from the SCADA (Supervisory Control and Data Acquisition) system. The aim is to quantitatively determine the intensity and statistical significance of this relationship. The methodology includes data filtering, linear regression and correlation analysis, with the calculation of the Pearson coefficient (r) and the coefficient of determination (R^2). The accuracy of the model was assessed by RMSE and MAE, while statistical validation was carried out by ANOVA analysis and residual analysis. The results show a moderately strong positive correlation between rotation speed and temperature (r = 0.7034), with R^2 = 0.4949 indicating that speed changes explain about 49.5% of the variability. In idle state, the temperature depends mainly on external factors (ambient temperature, wind, solar radiation, humidity and surrounding objects). The proposed model enables reliable temperature prediction in the operating mode and can be used for operation optimization, monitoring and predictive maintenance of wind turbines.

KEYWORDS

Generator rotation speed, Wind turbine generator, Generator temperature, Linear regression, Residuals

1. INTRODUCTION

Wind energy is one of several renewable energy sources used in the production of electricity due to rising electricity consumption and environmental pollution [1]. Advances in wind energy technology have significantly decreased the cost of producing electricity from this renewable source. Nowadays, the generation of energy from wind sources plays a crucial role to increasing the green energy [2]. WPPs have become more popular in recent decades as they are cleaner, greener, and more sustainable [3].

The most prevalent wind power generator is the variable-speed wound-rotor doubly fed induction generator (DFIG) [4]. The DFIG technology has the advantage that it extracts maximum energy from the wind for low wind speeds by optimizing the turbine speed, also by minimizing the mechanical stresses on the turbine during sudden gusts of wind [5]. Doubly Fed Induction Generators (DFIGs) are equipped with a rotor winding that is powered through a power electronic converter. This configuration enables efficient control of both the generator's rota-tional speed and output power [6]. The stator of DFIG is directly connected to the grid, while the rotor is connected to the grid through a back-to-back power electronic converter, which is composed of a rotor-side converter (RSC) and a grid-side converter (GSC) [7]. A gearbox couples the rotor and generator of DFIG to manage the speed range differences between rotor and generator speed [8]. Bidirectionally powered induction generator of wind turbine is shown in Figure 1 [9].

Thermal analysis of a generator is a complex problem where a number of heat exchanges are involved Since per-formance and efficiency of a generator depends on the thermal condition of its windings, understanding the power losses and the thermal mechanism within a DFIG are of great importance to the industry [10]. The rotation speed of a DFIG generator is a fundamental factor that directly affects the thermal load and functional reliability of the generator. DFIG systems are

designed to operate at variable rotation speeds, which allows for maximum utilization of the kinetic energy of the wind and improves the efficiency of electricity generation. Changes in rotation speed directly affect the stator and rotor currents, causing changes in copper and iron losses, which leads to an increase in the generator temperature.

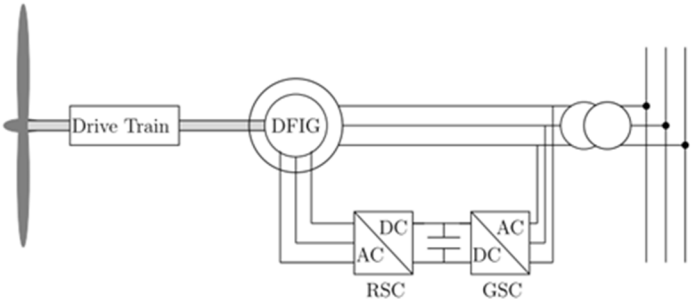


Figure 1: Bidirectionally powered induction wind turbine generator

Power losses of the generator and the temperature of stator winding increase with wind speed. As the induction generator torque is subject to the rotor current and stator voltage, therefore any unbalance of the stator voltage will cause an unstable rotor current in order to maintain an appropriate torque value rather than to maintain a constant power output. Therefore, an increase of stator voltage will cause a slight increase of stator current, and thus an increase of winding temperature [10]. The increase in power dissipation and changes in thermal conductivity lead to an increase in the temperature of the stator winding, and therefore of the entire generator.

The contribution of the work is reflected in the quantification of the influence of rotation speed on the temperature load of the generator based on measurements at one representative point, with a separate analysis of the operating mode and the idle state. The work applied linear regression analysis to determine the dependence between rotation speed and temperature, using Pearson's correlation coefficient, regression function parameters (slope and intercept), coefficient of determination, as well as statistical indicators of model accuracy and significance, including RMSE, MAE, *p*-value, *t*-statistics and residual analysis.

However, it is important to emphasize that the temperature analysis in this paper is based on measurements at a single selected point of the generator, which is considered representative for assessing the overall thermal state of the system. Although in real conditions the temperature of the generator is monitored at multiple locations, including the stator and rotor windings, as well as other critical parts, this approach implies a certain level of simplification. Accordingly, the obtained results describe local temperature characteristics and do not encompass the complete spatial distribution of temperature within the generator. The aforementioned limitation should be kept in mind when interpreting the results, while future research may be directed towards the application of multi-point measurements and more advanced thermal models.

The paper is organized as follows: after the introduction, the mathematical foundations of linear regression are presented in the second chapter, the analysis of the influence of the rotation speed on the generator temperature is given in the third chapter, while the conclusions and an overview of the literature used are presented at the end of the paper.

2. MATHEMATICAL BASIS OF LINEAR REGRESSION

The analysis of the influence of the rotation speed of a wind turbine DFIG generator on the generator temperature is based on the application of appropriate statistical methods that allow quantitative description and measurement of the degree of correlation between the observed quantities. In this context, simple linear regression represents the basic model for examining the linear relationship between the rotation speed as an independent variable and the generator temperature as a dependent variable.

In order to enable a detailed and quantitative analysis of the observed dependence, the mathematical foundations of the Pearson correlation coefficient (r), the coefficient of determination (R^2), as well as the parameters of the linear regression function, including the slope and intercept of the regression line, are presented below. In addition, the basic equation of the simple linear regression model is given, which allows for the description and prediction of the change in generator temperature depending on the change in rotation speed.

Pearson's correlation coefficient (r) is an indicator of the linear relationship between two quantitative variables. Its value reflects the direction and strength of the relationship between the variables being analyzed. In other words, it shows how much a change in one variable affects a change in the other, either in the same (positive correlation) or opposite (negative correlation) direction.

Based on the value of the Pearson correlation coefficient, the direction and strength of the linear relationship between the observed variables can be determined.

The direction and strength of the linear association of the variables depending on the value of the Pearson correlation coefficient are shown in Table 1 [11].

Table 1: The direction and strength of the linear association of the variables depending on the value of the Pearson correlation coefficient

The value of the Pearson correlation coefficient	Direction and strength of linear correlation of variables
$r = -1$	functional negative correlation
$r = 1$	functional positive correlation
$-1 < r \leq -0{,}8$	strong negative correlation
$0{,}8 \leq r < 1$	strong positive correlation
$-0{,}8 < r \leq -0{,}5$	medium strong negative correlation
$0{,}5 \leq r < 0{,}8$	medium strong positive correlation
$-0{,}5 < r < 0$	weak negative correlation
$0 < r < 0{,}5$	weak positive correlation
$r = 0$	there is no correlation

A scatter plot shows the relationship between two numerical variables using dots. A scatter diagram showing data with zero, complete positive, incomplete positive, complete negative, and incomplete negative correlations is shown in Figure 2.

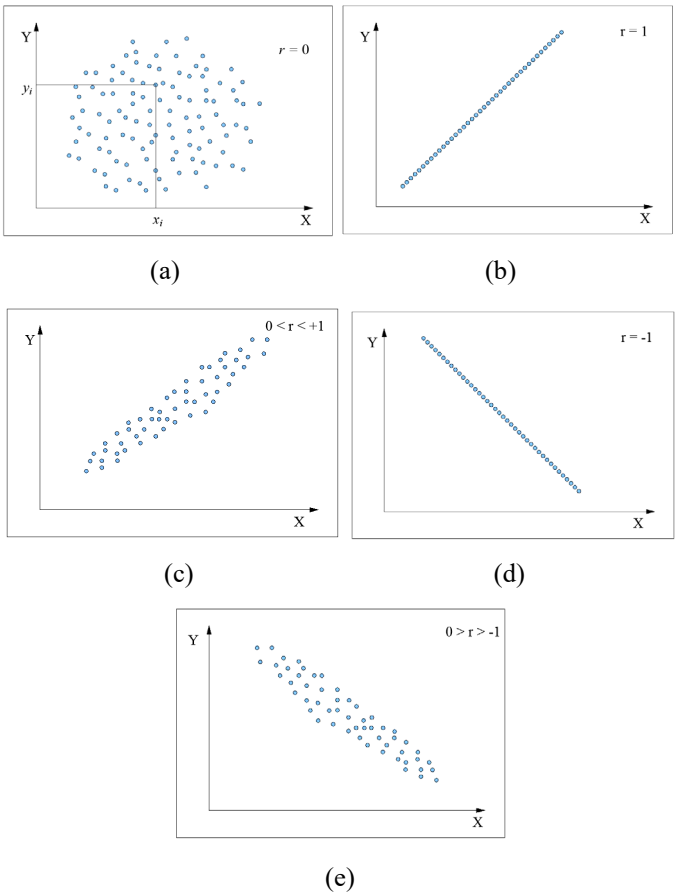


Figure 2: Scatter plot showing data with: (a) zero correlation, (b) complete positive correlation, (c) incomplete positive correlation, (d) complete negative correlation, and (e) incomplete negative correlation

The value of Pearson's correlation coefficient ranges from –1 to +1. A value of $r = 0$ indicates the absence of a linear relationship between variables, which does not necessarily mean that there is no relationship between them, but only that the relationship is not linear. A value of $r = +1$ indicates a complete positive linear relationship between the variables, while $r = -1$ indicates a complete negative linear relationship between the variables.

A complete positive linear relationship means that an increase in the value of one variable always follows an increase in the value of the other, that is, that the variables move in the same direction. Incomplete positive correlation exists when two variables tend to increase together, but the relationship is not perfectly linear. In this case, the increase in one variable usually follows the increase in the other, but the points in the scatter plot do not lie exactly on a straight line.

A perfect negative linear relationship means that an increase in the value of one variable is accompanied by a simultaneous decrease in the value of the other. An incomplete negative correlation exists when an increase in one variable often accompanies a decrease in the value of the other, but the relationship is not perfectly linear and the points in the scatterplot

do not lie exactly on a straight line, meaning that the variables change in opposite directions. The closer the value of the Pearson coefficient is to the values 1 or –1, the stronger the linear correlation.

Pearson's correlation coefficient is represented by the following equation [12]:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (1)$$

where:

- x_i, y_i - individual values of the observed variables related to the rotation speed and temperature of the wind turbine generator.
- $\bar{x}$, $\bar{y}$ - arithmetic means of data sets related to the rotation speed and temperature of the wind turbine generator, respectively.
- n - sample size.

The coefficient of determination (R^2) in linear regression indicates how well the model explains changes in the dependent variable (wind turbine generator temperature) by the independent variable (wind turbine generator rotation speed). The coefficient of determination (R^2) is defined as the square of the Pearson correlation coefficient, i.e. $R^2 = r^2$.

The coefficient of determination (R^2) in linear regression is represented by the following equation:

$$R^2 = r^2 = \left(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \right)^2 \quad (2)$$

The value of the coefficient of determination is in the interval from 0 to 1, i.e. $0 \leq R^2 \leq 1$. The value of $R^2 = 0$ means that the linear model does not explain changes in the dependent variable at all, while $R^2 = 1$ indicates that the model perfectly explains all changes in the dependent variable. A higher value of the coefficient of determination indicates a better ability of the regression model to describe changes in the dependent variable.

Equation (2) represents the basis for analyzing the quality of a linear regression model, as it allows quantifying the degree and reliability of the linear dependence between two variables.

The slope b in a linear regression model indicates how much the dependent variable changes when the independent variable increases by one unit.

The slope in linear regression is represented by the following equation [13]:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2} \quad (3)$$

where:

- x_i - individual values of the rotation speed of the wind turbine generator.
- y_i - individual temperature values of wind turbine generators.
- $\bar{x}$ - the arithmetic mean of all values of the rotation speed of the wind turbine generator.
- $\bar{y}$ - the arithmetic mean of all the temperature values of the wind turbine generator.

The slope shows the direction and strength of the linear relationship between the independent and dependent variable. If the slope is positive ($b > 0$), it means that the value of the dependent variable increases as the value of the independent variable increases. In this case, the diagram shows a straight line going up from left to right.

If the slope is negative ($b < 0$), the value of the dependent variable decreases as the independent variable increases, which is shown in the diagram by a straight line going downward from left to right. In the case where the slope is zero ($b = 0$), there is no linear trend between the independent and dependent variables. The line is horizontal, indicating that changes in the independent variable have no effect on the values of the dependent variable.

The intercept in a linear regression model indicates the value of the dependent variable when the independent variable is zero. In the linear regression model, the intercept is represented by the following equation [13]:

$$b_0 = \bar{y} - b_1 \bar{x} \quad (4)$$

The intercept indicates the point where the regression line intersects the y-axis. A simple linear regression model is represented by the following equation [13]:

$$E(Y|x) = \beta_0 + \beta_1 x \tag{5}$$

where:

- $E(Y|x)$ - predicted (expected) value of the dependent variable Y for a given level of the variable x.
- β_0 - the intercept of the regression line.
- β_1 - the slope of the regression line.

A simple linear regression model is an analytical method that describes and quantifies the degree of linear relationship between an observed dependent and independent variable. A simple linear regression model is used in this paper, since the influence of one independent variable (generator rotation speed) on a dependent variable (generator temperature) is analyzed. The purpose is to determine the appropriate regression line that best depicts changes in the dependent variable in relation to the independent variable.

3. RESULTS AND ANALYSIS OF THE DEPENDENCE OF DFIG GENERATOR TEMPERATURE ON ROTATION SPEED

This chapter analyzes the influence of wind turbine generator rotation speed on generator temperature. The data used in the analysis were collected from one wind turbine of the Rudine wind farm. The generator temperature data were taken from the SCADA system and refer to one measurement point, i.e. one built-in temperature sensor, while at the same time the rotation speed data is included. The original dataset contains 52,510 time series records, which represents a high measurement resolution, but at the same time a significant processing load in standard software tools, including Microsoft Excel.

Downsampling is subsetting data such that the timestamps occur at a lower frequency than in the original time series [14]. In order to efficiently process data while preserving relevant information, a downsampling method was applied. From the original data set, systematic sampling was performed by selecting every fourth sample, reducing the total number of data records to 13,128.

This approach is a common technique in time series analysis, which reduces the size of the data set while preserving the basic characteristics of the observed process, thereby ensuring the representativeness of the reduced set. In this way, the data set used retains the key trends in the changes in the observed variables and allows for reliable statistical analysis of the relationship between the rotation speed as an independent variable and the generator temperature as a dependent variable. The relationship between the observed variables was examined using a linear regression model, which enabled quantification of the impact of changes in rotation speed on changes in generator temperature. The dependence of temperature on rotation speed was analyzed using a scatter diagram with a linear trend line, while the statistical characterization of temperature distribution was performed using a box-plot diagram for different generator operating modes.

3.1. Generator Temperature in Operating Mode

This subsection analyzes the influence of the rotation speed of the DFIG generator of a wind turbine on the temperature of the generator, as well as the temperature distribution, which illustrates the thermal behavior in the operating mode.

The influence of the rotation speed of the DFIG wind turbine generator on the temperature of the generator is shown in Figure 3.

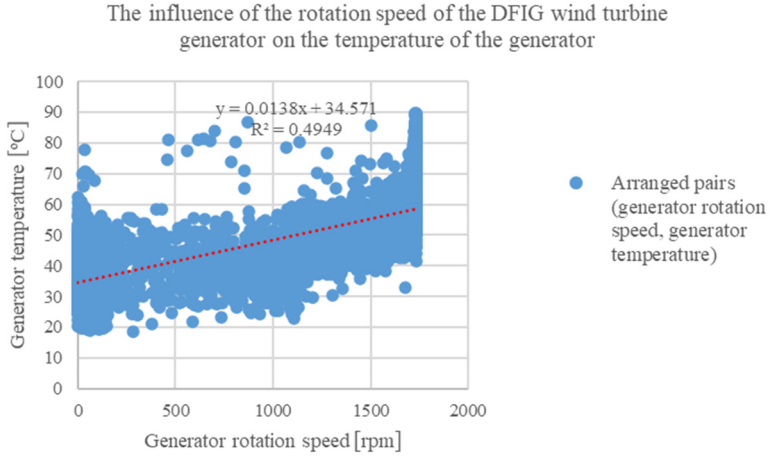


Figure 3: The influence of the rotation speed of the DFIG wind turbine generator on the temperature of the generator

Figure 3 shows a graphical representation of the dependence of the generator temperature on the rotation speed of the DFIG generator in operating mode. The blue dots represent ordered pairs of measured values, i.e. the corresponding values of the rotation speed and the generator temperature.

The dependence is approximated by a linear regression model, which confirms the obtained regression lines. However, the scattering of experimental points around the regression line indicates deviations from ideal linear behavior, which can be attributed to the influence of additional operating and ambient factors (generator load, cooling conditions, ambient temperature, etc.).

It is observed that at lower rotation speeds the temperatures are lower and relatively stable, while with increasing speed there is a slight but continuous increase in temperature. At higher rotation speeds there is more pronounced heating and greater dispersion of measured values, which indicates an increased sensitivity of the thermal regime of the generator in that range.

An approximately linear dependence with moderate deviations is observed, where the rotation speed is a significant factor in determining the thermal state of the generator, but not the only one.

Figure 3 shows a linear regression line that illustrates the trend of generator temperature versus rotation speed. The equation of the regression line (red dashed line) is given by the following expression:

$$y = 0,0138 \cdot x + 34,571 \tag{6}$$

where:

- y - temperature of the wind turbine generator (ºC).
- x - rotation speed of the wind turbine generator (rpm).

The slope of the regression line is $b_1 = 0.0138$, which indicates that for each increase in the generator rotation speed by one revolution per minute, the generator temperature increases by approximately 0.0138 °C on average.

The intercept of the regression line is $b_0 = 34.571$ and corresponds to the estimated generator temperature at zero rotation speed (0 rpm). This value serves as a reference point and allows for the estimation of the initial temperature of the generator when not in operation.

The Pearson correlation coefficient is $r = 0.7034$, which indicates a medium-strong positive correlation between the rotation speed of the wind turbine generator and its temperature. With increasing rotation speed, a tendency for the generator temperature to increase is observed. This result confirms the existence of a significant linear dependence between the observed variables.

The value of the coefficient of determination $R^2 = 0.4949$ indicates a moderate level of statistical correlation between the rotation speed and the generator temperature. The results show that the rotation speed explains approximately 49.5% of the temperature variability, while the remaining part of the variation (about 50.5%) depends on other factors, such as cooling quality, generator load, ambient temperature, etc. Therefore, the rotation speed can be seen as a significant, but not dominant parameter in modeling the thermal behavior of the generator under operating conditions.

The generator temperature distributions, which illustrate the thermal behavior in the operating mode ($n > 0$ rpm) depending on the rotation speed, are shown using box-plot diagrams in Figure 4.

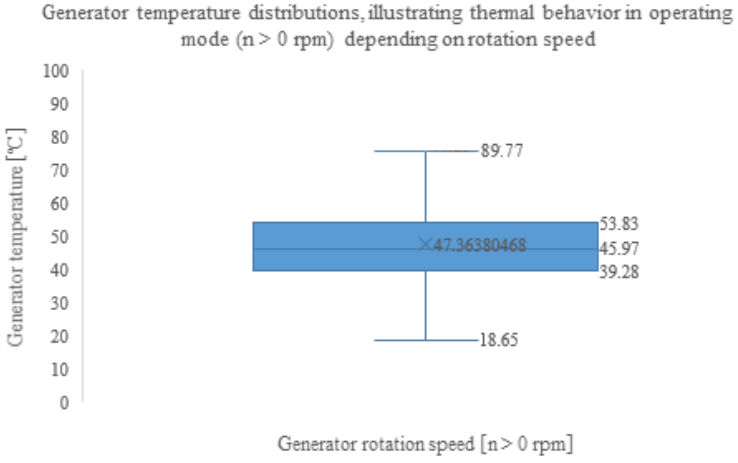


Figure 4: Generator temperature distributions illustrating thermal behavior in operating mode

The first quartile (Q1) is 39.28 °C and represents the value below which 25% of the measured temperatures are located. The median temperature is 45.97 °C and represents the central, i.e. representative, value of the thermal state of the generator in operating mode. The third quartile (Q3) is 53.83 °C and defines the value below which 75% of the measured temperatures are located, and represents the upper limit of the interquartile range. Based on the displayed parameters, it

can be concluded that most of the measured temperatures are in the interval from 39.28 °C to 53.83 °C. The arithmetic mean of the temperature is 47.36 °C and represents the average value of the measured temperatures. The minimum and maximum temperature values are 18.65 °C and 89.77 °C, respectively.

3.2. Generator Idle Temperature

The generator temperature distributions, which illustrate the thermal behavior in the idle mode (n = 0 rpm), are shown using a box-plot diagram in Figure 5.

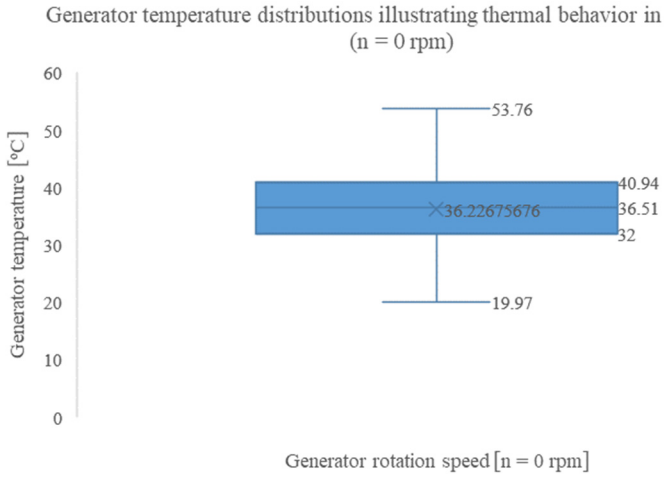


Figure 5: Generator temperature distributions illustrating thermal behavior in idle mode

The first quartile is 32 °C and indicates the value below which 25% of the measured temperatures are located. The median temperature is 36.51 °C and represents the central value of the thermal state of the generator in idle mode. The third quartile is 40.94 °C and indicates the value below which 75% of the measured temperatures are located, representing the upper limit of the interquartile range.

Based on the parameters shown, it can be concluded that the temperatures in the idle mode are relatively stable, with moderate dispersion and no pronounced extreme deviations. The arithmetic mean temperature is 36.22 °C and represents the average value of the measured temperatures, while the minimum and maximum values are 19.97 °C and 53.76 °C, respectively, which defines the overall range of temperature variation.

The analysis of the generator temperature distribution in the operating mode and the idle mode provides insight into the basic statistical parameters and differences between the observed data sets. The minimum generator temperature in the operating mode is 18.65 °C and is lower than the temperature in the idle mode, which is 19.97 °C. This phenomenon can be explained by the influence of ambient conditions, time asynchrony of the measurements and the measurement uncertainty of the sensor, where the difference of 1.32 °C belongs to the range of expected experimental variability.

In the idle mode, the generator temperatures are characterized by lower values and greater stability, with less dispersion and a more uniform distribution of measurement data. In contrast, in the operating mode, the temperature increases with increasing rotation speed, accompanied by more pronounced variability and a wider range of values, which indicates a more complex thermal behavior of the generator due to the action of additional influencing factors.

4. MODEL AND RESIDUAL ANALYSIS

This chapter presents the statistical validation of the developed model, as well as the analysis of residuals with the aim of assessing its adequacy.

4.1. Statistical Validation of the Model

ANOVA detects the presence or absence of a global effect of the independent variable on the dependent variable [15]. The ANOVA (Analysis of Variance) table of the regression model includes degrees of freedom (*df*), sum of squares (*SS*), mean squares (*MS*), *F*-statistics, and the corresponding *p*-value of the *F*-test (Significance *F*). The listed elements are systematically arranged according to the principal components of the model, including regression, residual (error), and total variance. ANOVA within the regression model was performed using the Data Analysis – Regression tool in [Microsoft Excel, and the results are shown in Table 2.

Table 2: ANOVA regression model

	df	SS	MS	F	Significance F
Regression	1	2429939429	2429939429	12027.17717	0
Residual	12275	2480008907	202037.3855		
Total	12276	4909948336			

In order to analyze the effect of the rotation speed of the wind turbine DFIG generator on its temperature, the linear regression method was applied, while the statistical significance of the model was assessed using the ANOVA approach. ANOVA analysis allows the decomposition of the total variability of the dependent variable into the part explained by the regression model and the part attributed to residual (unexplained) variation. The degrees of freedom (*df*) in the ANOVA table represent the number of independent pieces of information used to estimate the generator temperature variability, with the total degrees of freedom divided into the component associated with the regression model that includes the wind turbine DFIG generator rotation speed and the residual error component.

The resulting *F*-value indicates a high ratio between explained and unexplained variation, with the generator temperature variation explained by the DFIG generator rotation speed being significantly higher than the residual variation. The accompanying *p*-value of the *F*-test (Significance $F \approx 0$) shows that the effect of the DFIG generator rotation speed on the generator temperature is statistically highly significant, i.e. the probability that the observed relationship between these two variables arose by chance is negligibly small. This confirms that the rotation speed has a real and measurable effect on the generator temperature changes in the observed model. Based on the above results, it can be concluded that the linear regression model is statistically highly significant, i.e. that the rotation speed of the DFIG generator has a statistically significant effect on the generator temperature. The results confirm a statistically significant linear dependence between the observed variables.

The model shows high statistical significance; however, its validity is limited by the assumption of linearity of the relationship between the observed variables. Furthermore, the model does not consider additional physical influences, such as load changes, cooling system configuration and ambient conditions, which have a significant impact on the thermal processes in the generator. Based on the results obtained, ANOVA analysis confirms that the linear regression model adequately describes the statistical relationship between rotation speed and generator temperature, with the influence of rotation speed being statistically significant and dominant in explaining temperature variability. The high value of the *t*-statistic (109.67) and the negligibly small *p*-value (≈ 0) confirm that the influence of rotation speed on generator temperature is statistically highly significant, with the probability that the observed effect is the result of chance being practically negligible.

The RMSE and MAE values, which in this case are 8.28 and have identical numerical values, represent the basic indicators of the accuracy and quality of the prediction of the linear regression model that describes the influence of the rotation speed of the DFIG generator on the generator temperature. RMSE represents the root mean square error and is used as a measure that gives greater importance to larger deviations, which is why it is more sensitive to the presence of larger errors in the model prediction. MAE represents the mean absolute error, or the average absolute difference between the actual and predicted temperature values, where the obtained value shows that the model deviates on average by about 8.28 temperature units from the measured values, without taking into account the direction of the deviation. The identity of the RMSE and MAE values indicates that the model errors are relatively uniform, i.e. there are no pronounced extreme deviations that would lead to a significant increase in RMSE relative to MAE. This is an important indicator of model stability, as it indicates that the linear regression model consistently describes the relationship between rotation speed and generator temperature, without the presence of significant local deviations.

In the context of DFIG wind turbine generators, the obtained error values indicate that the model approximates the thermal behavior of the system with relatively high accuracy. The average deviation of about 8.28 temperature units can be viewed as a systematic error of the model, which is probably a consequence of neglected physical factors, such as load changes, cooling regime, wind turbulence and ambient conditions. Nevertheless, the low and stable error value indicates that the linear regression model has good predictive ability and is adequate to describe the general trend of the dependence of temperature on the generator rotation speed.

4.2. Residual Analysis

In order to verify the adequacy of the developed model for estimating the temperature of the DFIG generator, a residual analysis was performed. The residuals represent the difference between the measured and model-predicted generator temperature values, and provide insight into possible model deviations, the presence of systematic errors, and data variability. The residual diagram as a function of the generator rotation speed is shown in Figure 6.

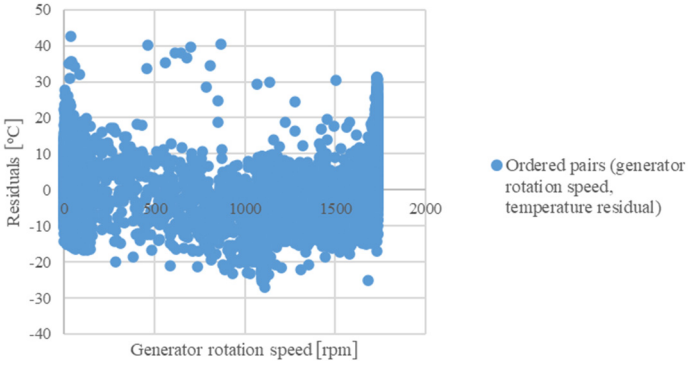


Figure 6: Residual diagram as a function of generator rotation speed

Figure 6 shows a diagram of residuals as a function of generator rotation speed, which allows a visual assessment of the existence of a dependence between these quantities. The blue points on the diagram represent ordered pairs (generator rotation speed, temperature residual), with each point corresponding to one measurement sample. Based on the diagram of residuals, it can be seen that the residuals are approximately randomly distributed around the zero value, without a clearly expressed systematic pattern, which indicates that the assumption of linearity of the model is satisfied and there is no significant systematic deviation.

However, there is a variation in the dispersion of the residuals along the generator rotation speed range, which suggests a deviation from the assumption of homoscedasticity. Also, the observed range of residuals and the presence of extreme values indicate limitations of the model in terms of the accuracy of temperature prediction in certain generator operating modes.

The observed changes in the variability of the residuals suggest that additional factors, such as load, cooling conditions and operating modes, can have a significant impact on the generator temperature, and that their inclusion in the model could contribute to improving its accuracy. Based on the overall analysis, the model shows a satisfactory level of adequacy, however, the observed deviations indicate the need for its further improvement by applying more complex model approaches, in order to increase the reliability and accuracy of the prediction.

Based on the analysis of the residuals, it can be concluded that the assumption of linearity of the model is satisfied, while the observed indications of impaired homoscedasticity indicate certain limitations of the observed model.

5. CONCLUSION

In this paper, a quantitative analysis of the dependence of the temperature of a wind turbine DFIG generator on its rotation speed was conducted, using linear regression analysis and statistical processing of SCADA data. The results of the correlation analysis confirm the existence of a medium-strong positive linear dependence between the observed quantities ($r = 0.7034$), while the coefficient of determination ($R^2 = 0.4949$) shows that the rotation speed explains almost half of the total variability of the generator temperature.

The results of the statistical validation of the model based on ANOVA analysis confirm its high statistical significance, with the *F*-test and *t*-statistics indicating a significant influence of the rotation speed on the generator temperature. The results of the error analysis (RMSE and MAE = 8.28) indicate a uniform and stable pattern of model deviations, without the presence of pronounced extreme values, which further confirms the high reliability and consistency of its predictions.

The findings of the residual analysis indicate their random distribution around the zero value, which confirms the adequacy of the linearity assumption of the model. However, the identified variability of the residual variance indicates a deviation from the homoscedasticity assumption, which indicates the presence of additional factors that affect the thermal behavior of the generator, such as load, cooling conditions and ambient influences.

Although the developed linear regression model confirms a statistically significant relationship between the rotation speed and the temperature of the DFIG generator, it is necessary to point out certain limitations of the applied approach. The developed model is based on a single independent variable, while other relevant factors, such as generator load, cooling regime and ambient conditions, are not included in the analysis. In addition, the temperature analysis is based on measurements at a single representative point, which limits the possibility of observing the spatial distribution of the thermal field inside the generator.

Although the linearity assumption is statistically confirmed on the observed data set, it may lead to a decrease in the accuracy of the model in conditions of nonlinear system behavior at extreme operating modes. The observed heteroskedasticity indicates that the application of more robust models could improve the reliability of the analysis.

Based on the obtained results, it can be concluded that the linear regression model adequately describes the basic dependence between the rotation speed and the temperature of the DFIG generator under operating conditions, and represents a reliable and efficient approach for assessing the thermal behavior of the system. However, the observed limitations in explaining the overall variability, as well as the presence of heteroscedasticity, indicate the need to improve the model by including additional input variables or applying more advanced nonlinear and multivariate approaches, in order to improve the accuracy and reliability of the generator temperature prediction.

In future research, it is recommended to expand the set of input variables, as well as the application of advanced multivariate regression models and machine learning methods, including neural networks and ensemble methods, in order to improve the accuracy, robustness and generalization ability of the model for generator temperature prediction under different operating conditions. Also, the application of multipoint measurements would enable a more detailed analysis of the spatial distribution of temperature within the DFIG system, which would improve the understanding of local thermal gradients and their impact on the generator operating characteristics and reliability.

REFERENCES

[1] C. P. Chioncel, E. Spunei and G.O. Tirian, “Wind Turbines Around CutIn Speed: Startup Optimization and Behavior Analysis Reported to MPP,” Appl. Sci., Vol. 15, p. 3026, --https://doi.org/10.3390/app15063026, (2025).

[2] A. Dannier, E. Fedele, I. Spina and G. Brando, “DoublyFed Induction Generator (DFIG) in Connected or Weak Grids for TurbineBased Wind Energy Conversion System,” Energies, Vol. 15, p. 6402, -https://doi.org/10.3390/en15176402, (2022).

[3] M. Arsal, T. Kamel, H. Asad and A. Khan, “A Systematic Review of Cyber Risk Analysis Approaches for Wind Power Plants,” Energies, Vol. 19, p. 677, https://doi.org/10.3390/en19030677, (2026)

[4] M. Shuaibu, A. S. Abubakar and A. F. Shehu, “Techniques for ensuring fault ridethrough capability of grid connected DFIGbased wind turbine systems: A review,” Niger. J. Technol. Dev., Vol. 18, pp. 39–46, --https://doi.org/10.63746/njtd.v18i1.591, (2021).

[5] R. Sharma, V. S. Chauhan and H. Shah, “Doubly Fed Induction Generator for Wind Energy System,” Int. J. Adv. Res. Electr. Electron. Instrum. Eng., Vol. 6(5), pp. 3208–3213, https://doi.org/10.15662/IJAREEIE. 2017.0605004, (2017).

[6] P. Prajzendanc and C. Kreischer, “A Review of New Technologies in the Design and Application of Wind Turbine Generators,” Energies, Vol. 18, p. 4082, https://doi.org/10.3390/en18154082, (2025).

[7] K. Ni, Y. Hu, Y. Liu and C. Gan, “Performance Analysis of a Four Switch Three Phase Grid Side Converter with Modulation Simplification in a Doubly Fed Induction Generator Based Wind Turbine (DFIG WT) with Different External Disturbances,” Energies, Vol. 10, p. 706, https://doi.org/10.3390/en10050706, (2017)

[8] H. A. Behabtu, T. Coosemans, M. Berecibar, K. A. Fante, A. A. Kebede, J. Van Mierlo and M. Messagie, “Performance Evaluation of Grid Connected Wind Turbine Generators,” Energies, Vol. 14, p. 6807, https://doi.org/10.3390/en14206807, (2021).

[9] R. K. Behara and A. K. Saha, “Optimised Neural Network Model for Wind Turbine DFIG Converter Fault Diagnosis,” Energies, Vol. 18, p. 3409, https://doi.org/10.3390/en18133409, (2025).

[10] Y. Qiu, W. Zhang, M. Cao, Y. Feng, and D. Infield, “An Electro Thermal Analysis of a Variable Speed Doubly Fed Induction Generator in a Wind Turbine,” Energies, Vol. 8(5), pp. 3386–3402, https://doi.org/10.3390/en8053386, (2015).

[11] M. Biljan-August, S. Pivac and A. Štambuk, Uporaba statistike u ekonomiji, 2. izd., Rijeka: Ekonomski fakultet Sveučilišta u Rijeci, 2009.

[12] A. Tabak, Pearsonov korelacijski koeficijent, Diplomski rad, Sveučilište J. J. Strossmayera u Osijeku, Odjel za matematiku, Diplomski studij matematike – smjer Financijska matematika i statistika, Osijek, Hrvatska, 2018.

[13] H. J. Seltman, Experimental Design and Analysis, 2018. Available: http://www.stat.cmu.edu/~hseltman /309/Book/Book.pdf, accessed: 27.04.2026.

[14] A. Nielsen, “Practical Time Series Analysis: Prediction with Statistics and Machine Learning” O’Reilly Media, Inc., Sebastopol, CA, USA, 2020. https://www.oreilly.com/

[15] J. Juarros-Basterretxea, G. Aonso-Diego, Á. Postigo, P. Montes-Álvarez, Á. Menéndez-Aller i E. García-Cueto, “Post-Hoc Tests in One-Way ANOVA: The Case for Normal Distribution,” Methodology, Vol. 20(2), pp. 84–99, https://doi.org/10.5964/meth.11721, (2024).

SESSION E: THERMAL TECHNIQUE AND ENVIRONMENTAL PROTECTION

Influence of overheating exposure time on the microstructural stability and mechanical properties of HP40-0.63Nb alloy

Milica Timotijević[1*], Olivera Erić Cekić[2,3], Dragan Rajnović[4], Petar Janjatović[4]

[1] University of Belgrade, The Academy of Applied Studies Polytechnic, Department of Mechanical Engineering, Belgrade, Serbia

[2] University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Kraljevo, Serbia

[3] University of Belgrade, Innovation Centre of the Faculty of Mechanical Engineering, Belgrade, Serbia

[4] University of Novi Sad, Faculty of Technical Science, Department of Production Engineering, Novi Sad, Serbia

ARTICLE INFO

[*] **Correspondence:** mtimotijevic@politehnika.edu.rs

DOI: https://doi.org/10.46793/ET26.E03T **PAGES:** E21 – E28

ABSTRACT

HP40-Nb heat-resistant cast steels are widely used for reformer furnace tubes operating at elevated temperatures, where accidental overheating may induce microstructural degradation and alter mechanical properties. This study investigates the influence of overheating time on the microstructure and mechanical properties of centrifugally cast HP40-0.63Nb alloy. Samples were overheated at 950 and 1150 °C for 30, 120 and 480 min. Microstructural characterization was performed using light microscopy (LM), scanning electron microscopy (SEM) and energy-dispersive X-ray spectroscopy (EDS), while tensile testing and hardness measurements were used to evaluate mechanical performance. The as-cast microstructure consisted of an austenitic dendritic matrix and eutectic chromium-rich and niobium-rich carbides. Overheating promoted the precipitation of secondary chromium-rich carbides, causing significant changes in carbide morphology and distribution. The highest yield strength (346 MPa) was achieved after overheating at 950 °C for 120 min due to intensive precipitation hardening. Prolonged overheating led to carbide coarsening and partial dissolution, reducing the strengthening effect. The highest tensile strength (582 MPa) and favorable ductility (8.76 %) were obtained after overheating at 1150 °C for 30 min. The results demonstrate that overheating time strongly influences carbide evolution and mechanical behavior of HP40-0.63Nb alloy, with 950 °C/120 min providing maximum strengthening and 1150 °C/30 min the best strength–ductility balance.

KEYWORDS

HP40-Nb alloy, Overheating time, Secondary carbides, Mechanical properties.

1. INTRODUCTION

Heat-resistant centrifugally cast austenitic steels of the HP family are extensively used in petrochemical and chemical industries for the production of reformers, cracking and pyrolysis furnace tubes operating at high temperatures [1-4]. Their widespread application is primarily associated with excellent resistance to creep, oxidation and carburization at elevated temperatures [5-8]. These properties are achieved through a combination of austenitic matrix strengthening and the presence of primary and secondary carbide phases.

Among HP-type alloys, niobium-containing grades have attracted considerable attention because niobium promotes the formation of thermally stable MC carbides, improves high-temperature strength and enhances creep resistance [9]. The microstructure of HP-Nb alloys in the as-cast condition typically consists of an austenitic dendritic matrix and a continuous network of eutectic carbides distributed along grain boundaries and interdendritic regions [10,11]. Depending on alloy composition, and thermal history, chromium-rich carbides (mainly M_7C_3 in the as-cast condition and $M_{23}C_6$ after thermal exposure) and niobium-rich MC carbides may coexist in the microstructure.

During exposure to elevated temperatures, significant microstructural changes occur. Carbon diffusion from the supersaturated austenitic matrix promotes the precipitation of secondary chromium-rich carbides, while prolonged thermal

exposure may lead to carbide coarsening, partial dissolution and changes in carbide morphology. These microstructural processes directly influence mechanical properties, particularly yield strength, tensile strength and ductility [12-14].

Previous studies have shown that carbide precipitation, transformation and coarsening significantly influence the mechanical behavior of HP-Nb alloys during thermal exposure [15-18]. However, the influence of short-term overheating duration remains insufficiently investigated.

In industrial service conditions, accidental overheating events may occur due to process disturbances, burner malfunctions or local hot spots, resulting in temperatures significantly above the design operating range. Such events can induce rapid microstructural changes and consequently alter the mechanical performance of HP-Nb alloys.

Therefore, the aim of this work was to investigate the influence of overheating time on the microstructure and mechanical properties of HP40-0.63Nb alloy. Special attention was paid to microstructural evolution, particularly changes in carbide morphology and microstructural features, and their effect on the mechanical properties of the alloy. The influence of overheating at 950 °C and 1150 °C for different exposure times on yield strength, tensile strength, elongation and hardness was systematically evaluated and interpreted in relation to the observed microstructural changes.

2. EXPERIMENTAL PROCEDURE

In this study, specimens were sectioned from a centrifugally cast HP40-Nb austenitic heat-resistant cast steel alloy in the as-cast condition. The chemical composition of the alloy was determined using an Optical Emission Spectrometer (OES, I Spark 8860, Thermo Scientific™, USA). The specimens were subjected to thermal exposure at 950 and 1150 °C for 30, 120 and 480 min, followed by cooling to room temperature. Heat treatments were carried out in a chamber-type laboratory furnace equipped with a PID temperature controller. The furnace temperature was continuously monitored using Pt–Pt13%Rh thermocouples positioned in close proximity to the specimens.

Following thermal exposure, the specimens were prepared for metallographic examination using standard grinding and polishing procedures. Grinding was performed with SiC abrasive papers ranging from 180 to 2400 grit, followed by polishing with diamond suspensions of 6, 3, 1, and 0.25 μm particle size. The polished surfaces were etched using a solution containing 15 mL HCl, 10 mL glycerol, and 5 mL HNO_3.

Microstructural characterization was performed using light microscopy (LM), scanning electron microscopy (SEM) and energy-dispersive X-ray spectroscopy (EDS). For LM a Leitz-Orthoplan microscope was used, while for SEM a JEOL JSM-6460LV was used. Phase identification and chemical characterization of the microstructural constituents were conducted by energy-dispersive X-ray spectroscopy (EDS), INCA, Oxford Instruments attached to the SEM.

Vickers hardness (HV10) was measured in accordance with SRPS EN ISO 6507-1:2023 using a load of 98.07 N (10kgf) and a dwell time of 15 s on an HPO 250 hardness tester (WPM, Markkleeberg, Germany). Five measurements were performed for each specimen, and the mean hardness value was calculated after excluding the highest and lowest readings.

Mechanical properties were evaluated by tensile testing at room temperature in accordance with SRPS EN ISO 6892-1:2020 on three specimens. Tests were carried out on a Schenck Trebel universal testing machine (Fmax = 250 kN) at a crosshead speed of 2 mm/min. The yield strength, ultimate tensile strength, and total elongation were determined from the resulting engineering stress–strain curves for specimens in the as-cast and heat-treated conditions.

3. RESULTS AND DISCUSSION

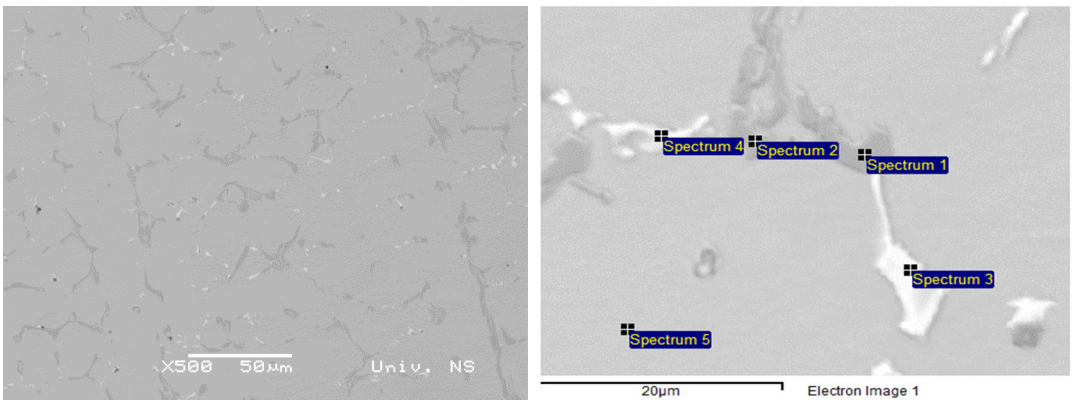
3.1. Microstructure

3.1.1. As-Cast Microstructure

The microstructure of the HP40-0.63Nb alloy in the as-cast condition is presented in Fig. 1. The typical as-cast microstructure consists of a γ-austenitic dendritic matrix surrounded by a network of primary eutectic carbides concentrated in the interdendritic regions and along grain boundaries. Figure 1b shows two morphologically distinct carbide phases within the interdendritic regions. The dark grey phase was identified as a Cr-rich M_7C_3 carbide, whereas the bright phase corresponded to Nb-rich MC-type carbides containing minor amounts of Ti and Cr.

The contrast differences observed in the BSE images (Fig. 1b) arise from variations in the average atomic number of the constituent phases, with phases containing heavier elements appearing brighter. In the present alloy, Nb-rich carbides exhibit a brighter contrast than Cr-rich carbides due to the significantly higher atomic number of niobium compared with chromium. The applicability of SEM-BSE imaging for distinguishing carbide phases in heat-resistant cast steels has been widely reported in the literature [19–21].

Results of further EDS analyses of carbides are given in Table 1. The Spectrum 1 and 2 revealed high chromium contents, confirming the presence of chromium M_7C_3 carbides. In contrast, Spectrum 3 and 4 were enriched in Nb, indicating the formation of MC-type carbides. Spectrum 5 corresponded to the γ-austenitic matrix.



a) b)

Figure 1: The microstructure of the HP40-0.63Nb alloy in the as-cast condition: a) SEM-BSE, b) SEM-BSE [10]

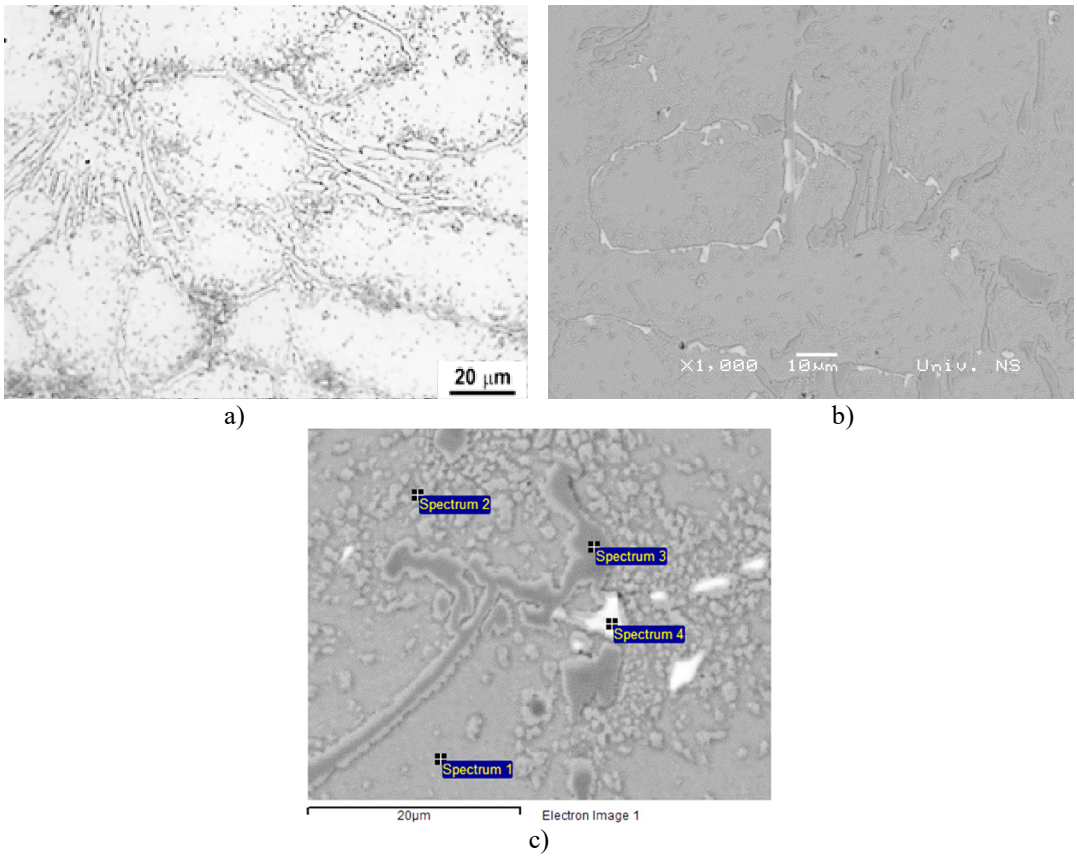
Table 1: Chemical composition (mass. %) of phases in HP40 Nb alloy, according to Fig.1b

Spectrum	C	Si	Ti	Cr	Mn	Fe	Ni	Nb
Spect. 1	36.97			54.39		6.92	1.72	
Spect. 2	15.83	3.41		28.73		27.11	24.93	
Spect. 3	44.98		4.30	4.62		2.56	1.80	41.73
Spect. 4	38.99	1.52	1.02	11.78		10.03	8.49	28.16
Spect. 5	10.76	2.56		22.12	0.94	34.03	29.58	

The presence of thermally stable MC-type carbides is particularly important for the high-temperature performance of HP40-Nb alloys. Unlike chromium-rich carbides, Nb-rich MC carbides exhibit superior thermal stability and resistance to coarsening and dissolution during long-term exposure at elevated temperatures. Consequently, these carbides play a crucial role in maintaining creep strength and microstructural stability during service [22-23].

3.1.2. Effect of Overheating Time at 950 °C on the Microstructure

Representative microstructural features observed in the HP40–0.63Nb alloy after overheating at 950 °C for different exposure times (30, 120, and 480 min) are shown in Figure 2.



a) b)

c)

Figure 2: Microstructure after overheting at 950 °C: a) for 30 min (LM), b) for 120 min (SEM-BSE), c) for 480 min (SEM-BSE)

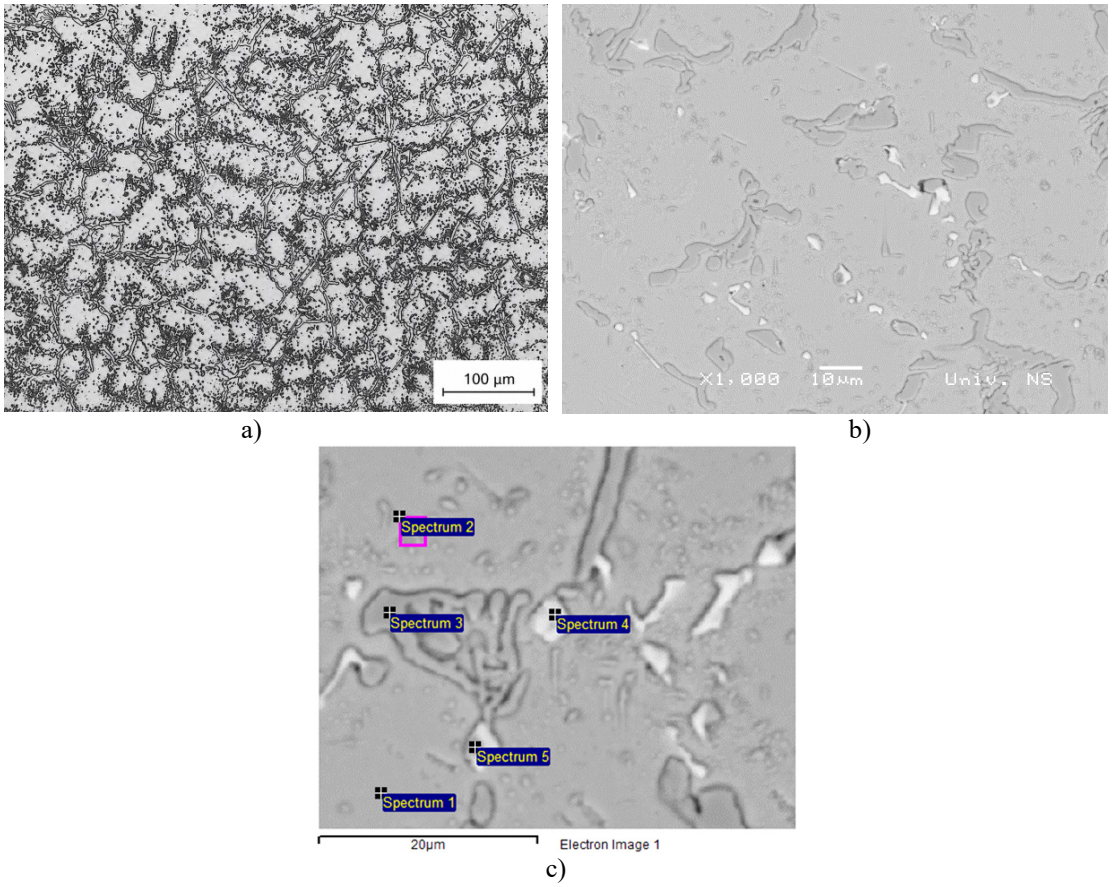
Table 2: Chemical composition (mass. %) of participating phases in HP40 Nb alloy, corresponding to Fig.2c

Spectrum	C	O	Si	Ti	V	Cr	Mn	Fe	Ni	Nb
Spectrum 1	16.17	2.62	2.09			19.26	1.13	30.71	28.01	
Spectrum 2	21.61		3.15			20.84	1.34	26.61	26.45	
Spectrum 3	26.77				0.41	61.06	1.57	7.84	2.35	
Spectrum 4	54.13		0.69	0.53	0.36	6.94		4.39	4.71	28.24

Overheating at 950 °C for 30 min resulted in the precipitation of numerous fine secondary chromium-rich carbides within the austenitic matrix as carbon diffused from the supersaturated solid solution. Thermal exposure promoted the transformation of metastable M_7C_3 carbides into the more stable $M_{23}C_6$ phase, accompanied by the precipitation of secondary chromium-rich carbides, thereby strengthening the matrix [26]. Increasing the overheating duration to 120 min led to a substantial increase in the density of secondary carbides. EDS analysis confirmed the presence of chromium-rich $M_2{}_3C_6$ carbides and thermally stable niobium-rich MC carbides. The extensive precipitation of fine $M_{23}C_6$ carbides corresponded to the highest yield strength (346 MPa), while elongation decreased to 5.06%, indicating significant precipitation hardening. Further extension of the overheating time to 480 min promoted carbide coarsening and partial dissolution. As a result, the strengthening effect diminished slightly, as evidenced by a small reduction in yield strength and a lower fraction of chromium-rich carbides. These findings suggest that carbide precipitation reached its maximum at approximately 120 min, after which coarsening became the dominant microstructural process.

3.1.3. Effect of Overheating Time at 1150 °C on the Microstructure

The microstructural changes induced by overheating at 1150 °C are shown in Figure 3, where representative micrographs corresponding to 30, 120, and 480 min of exposure are presented.



a) b) c)

Figure 3: Microstructure after overheting at 1150 °C: a) for 30 min (LM), b) for 120 min (SEM-BSE), c) for 480 min (SEM-BSE)

Table 3: Chemical composition (mass. %) of participating phases in HP40 Nb alloy, corresponding to Fig.3c

Spectrum	C	Si	Ti	V	Cr	Mn	Fe	Ni	Nb
Spectrum 1	21.47	2.81			18.69	1.11	28.13	27.79	
Spectrum 2	31.43	2.28			20.46	0.94	22.91	21.98	
Spectrum 3	36.64				48.06	0.90	9.76	4.63	
Spectrum 4	70.21		2.31	0.26	2.65		1.58	1.26	21.72
Spectrum 5	54.73		16.44	0.63	3.67		1.09	0.93	22.51

After 30 min of overheating at 1150 °C, the density of secondary carbides in the austenitic matrix decreased, while chromium-rich grain-boundary carbides became coarser and developed a blocky morphology. Although $M_{23}C_6$ carbides are considered thermodynamically stable, they undergo morphological changes during exposure to elevated temperatures [27]. At the same time, globular secondary carbides appeared in interdendritic regions. These microstructural changes resulted in the highest tensile strength recorded in this investigation (582 MPa), while elongation remained relatively high (8.76 %).

Increasing the overheating time to 120 min promoted further coarsening of primary and secondary carbides. EDS analysis confirmed the presence of chromium-rich $M_{23}C_6$ carbides and thermally stable niobium-rich MC carbides, indicating that overheating affected carbide morphology rather than carbide type.

After 480 min, the amount of secondary carbides decreased further due to coarsening and partial dissolution. Dissolution occurred both within dendritic regions and near primary grain-boundary carbides, increasing the fraction of the austenitic matrix. Consequently, the tensile strength decreased to 553 MPa, while elongation remained relatively high (7.93 %). The results indicate that prolonged exposure at 1150 °C reduces precipitation strengthening because carbide coarsening and dissolution become dominant microstructural processes.

3.2. Mechanical Properties

The effects of overheating temperature and holding time on the mechanical properties of the HP40-0.63Nb alloy are shown in Table 4 and Figures 4-6.

3.2.1. Hardness

The hardness results are presented in Table 4. The hardness of the as-cast alloy was 215 HV10. After overheating at 1150 °C, the hardness increased from 201 HV10 after 30 min to 232 HV10 480 min, which was the highest value measured. At 950 °C, the hardness increased from 197 HV10 after 30min to 226 HV10 after 480 min. Overall, increasing the overheating time from 30 to 480 min resulted in a progressive increase in hardness at both temperatures. The increase in hardness is attributed to the precipitation of secondary carbides within the austenitic matrix and the associated strengthening effect, which becomes more pronounced with increasing overheating time.

The hardness values obtained in this study are in good agreement with those reported in the literature [18]. Furthermore, Table 4 shows that the highest hardness values were achieved after 480 min of overheating, regardless of the treatment temperature.

Table 4: The hardness values of the alloy HP40-0.63Nb in its initial state and after being overheated at 950 and 1150°C for 30, 120 and 480 min, respectively.

Temperature, [°C]	Hold Time, [min]	Hardness HV10					Average	Standard Deviation
		1	2	3	4	5		
As-cast	-	228	199	222	230	197	215	16
950	30	184	205	197	198	199	197	8
	120	210	205	206	206	206	207	2
	480	224	224	230	226	228	226	3
1150	30	204	196	204	199	203	201	4
	120	204	208	208	209	208	207	2
	480	233	230	230	233	232	232	2

3.2.2. Tensile Characteristics

The influence of overheating time on the tensile strength of the alloy is presented in Fig. 4. Compared with the as-cast condition (450 MPa), all overheated specimens exhibited substantially higher tensile strength.

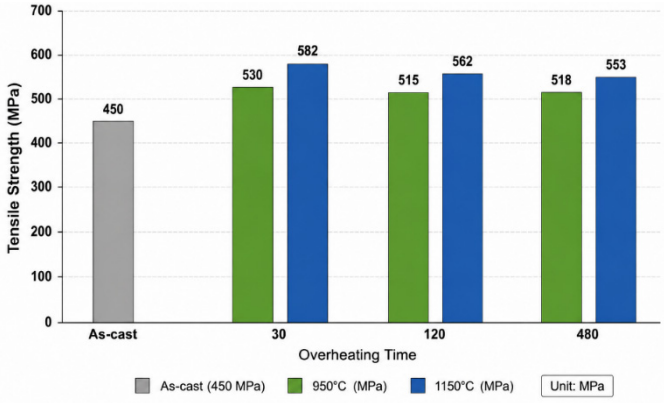


Figure 4: Influence of overheating treatment on the tensile strength of the HP40–0.63Nb alloy

The maximum value of 582 MPa was achieved after 30 min at 1150 °C, whereas overheating at 950 °C resulted in a tensile strength of 530 MPa. Increasing the overheating duration led to a gradual decline in tensile strength; however, even after 480 min, the measured values remained above that of the as-cast alloy. These results suggest that short-term overheating promotes strengthening, while prolonged exposure contributes to microstructural coarsening and a moderate reduction in strength.

The effect of overheating time on yield strength is shown in Figure 5. Relative to the as-cast alloy (240 MPa), overheating at 950 °C produced a marked increase in yield strength, reaching 346 MPa after 120 min, followed by a slight decrease to 342 MPa after 480 min. In contrast, specimens treated at 1150 °C exhibited lower yield strength values, which gradually decreased from 308 MPa to 303 MPa with increasing exposure time. Overall, overheating at 950 °C proved more effective in enhancing yield strength, whereas prolonged exposure at 1150 °C promoted softening.

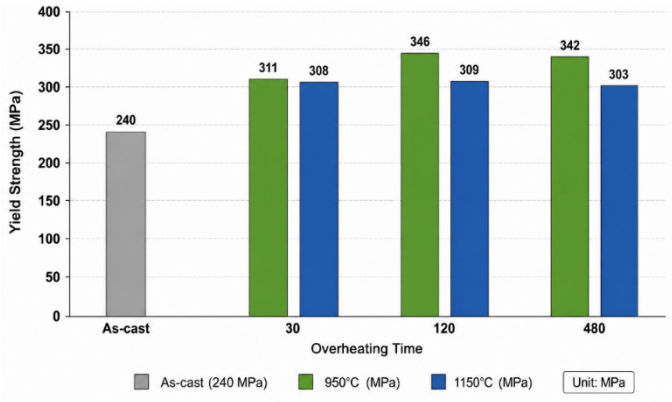


Figure 5: Influence of overheating treatment on the yield strength of the HP40–0.63Nb alloy

Changes in elongation as a function of overheating time are presented in Fig. 6. The highest ductility was recorded in the as-cast condition (9.67%). Exposure to elevated temperatures reduced elongation at both overheating temperatures, with a more pronounced decrease observed at 950 °C.

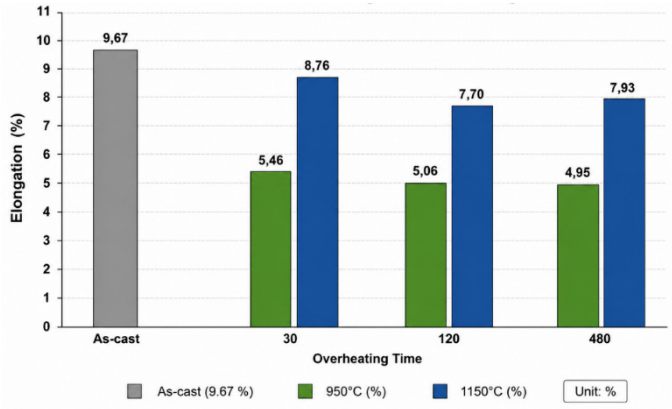


Figure 6: Effect of overheating treatment on the elongation of the HP40–0.63Nb alloy

Throughout the investigated period, specimens overheated at 1150 °C retained higher elongation values, decreasing only slightly from 8.76% after 30 min to 7.93% after 480 min. Although all heat-treated samples exhibited lower ductility than the as-cast alloy, the results indicate that overheating at 1150 °C provides a more favorable balance between strength and plasticity than overheating at 950 °C.

3.3. Correlation Between Carbide Evolution and Mechanical Properties

The evolution of mechanical properties was closely related to time-dependent changes in carbide morphology and distribution. Short-term overheating promoted the precipitation of fine secondary chromium-rich carbides within the austenitic matrix, leading to precipitation strengthening [26,28]. With increasing exposure time, carbide coarsening and partial dissolution reduced the effectiveness of precipitation strengthening, resulting in a gradual decrease in tensile strength. In contrast, niobium-rich MC carbides remained stable throughout all overheating treatments, confirming their superior thermal stability and their contribution to maintaining microstructural stability during prolonged exposure at elevated temperatures [22,23,25].

4. CONCLUSIONS

1. Analysis of variance revealed that overheating exposure time is the dominant factor affecting the mechanical properties of the HP40-0.63Nb alloy, whereas overheating temperature was found to be statistically significant only for tensile strength.

2. Overheating at 950 °C for 120 min exhibited the most favorable effect on the mechanical properties of the alloy, resulting in an increase in yield strength without a statistically significant reduction in elongation.
3. Increasing overheating exposure time promoted the evolution of the carbide phase, characterized by a reduction in the amount of secondary carbides within the austenitic matrix and the growth of $M_{23}C_6$ carbides. The highest tensile strength, 582 MPa, was achieved after overheating at 1150 °C for 30 min, which was associated with the presence of coarser and uniformly distributed secondary carbides.
4. Overheating at 1150 °C for 480 min led to the dissolution of secondary carbides and a reduction in the amount of primary carbides due to their transformation and dissolution, indicating reduced microstructural stability of the alloy at elevated temperatures.
5. The obtained results demonstrate that the changes in mechanical properties are directly related to the evolution of the carbide phase during overheating, with exposure time being the key parameter controlling the microstructural stability and performance of the HP40-0.63Nb alloy.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science, Technological Development, and Innovation of the Republic of Serbia, Contract No. 451-03-34/2026-03/200156, 451-03-34/2026-03/200108, 451-03-33/2026-03/200213; and by the Faculty of Technical Sciences, University of Novi Sad, through project “Scientific and Artistic Research Work of Researchers in Teaching and Associate Positions at the Faculty of Technical Sciences, University of Novi Sad 2026”, Contract No. 01-3609/1.

REFERENCES

[1] L. Bonaccorsi, E. Guglielmino, R. Pino, C. Servetto, and A. Sili, “Damage analysis in Fe–Cr–Ni centrifugally cast alloy tubes for reforming furnaces,” Engineering Failure Analysis, Vol. 36, pp. 65–74, https://doi.org/10.1016/j.engfailanal.2013.09.020, (2014)

[2] M. Ilman and N. Kusmono, “Analysis of material degradation and life assessment of 25Cr–38Ni–Mo–Ti wrought alloy steel (HPM) for cracking tubes in an ethylene plant,” Engineering Failure Analysis, Vol. 42, pp. 100–108, https://doi.org/10.1016/j.engfailanal.2014.03.020, (2014)

[3] W. Z. Wang, F. Z. Xuan, Z. D. Wang, and C. J. Liu, “Effect of overheating temperature on the microstructure and creep behavior of HP40Nb alloy,” Materials & Design, Vol. 32, pp. 4010–4016, https://doi.org/10.1016/j.matdes.2011.03.008, (2011)

[4] D. J. Tillack and J. E. Guthrie, "Wrought and Cast Heat-Resistant Stainless Steels and Nickel Alloys for the Refining and Petrochemical Industries," Nickel Development Institute, Toronto (Canada), (1998)

[5] L. M. Shen, J. M. Gong, Y. Jiang, and L. Y. Geng, “Effects of aging treatment on microstructure and mechanical properties of Cr25Ni35Nb and Cr35Ni45Nb furnace tube steel,” Acta Metallurgica Sinica (English Letters), Vol. 24, pp. 235–242, https://doi.org/10.11890/1006-7191-113-235, (2011)

[6] Y. Nishiyama, N. Otsuka, and T. Nishizawa, “Carburization resistance of austenitic alloys in CH4–CO2–H2 gas mixtures at elevated temperatures,” Corrosion, Vol. 59, pp. 688–700, https://doi.org/10.5006/1.3277598, (2003)

[7] M. Mobaraki, B. Afshang, M. R. Rahimpour, and M. E. Bahrololoom, “Effect of cracking feedstock on carburization mechanism of cracking furnace tubes,” Engineering Failure Analysis, Vol. 107, Art. no. 104216, https://doi.org/10.1016/j.engfailanal.2019.104216, (2020)

[8] H. M. Tawancy, “Degradation of mechanical strength of pyrolysis furnace tubes by high-temperature carburization in a petrochemical plant,” Engineering Failure Analysis, Vol. 16, pp. 2171–2178, https://doi.org/10.1016/j.engfailanal.2009.02.009, (2009)

[9] M. H. S. Lissarrague and C. A. Lanz, “NbC transformation during aging in HP40-Nb heat resistant alloy,” Acta Metallurgica Slovaca, Vol. 28, pp. 147–150, https://doi.org/10.36547/ams.28.3.1562, (2022)

[10] M. Timotijević, O. Erić Cekić, D. Rajnović, and P. Janjatović, “Insight into the microstructure analysis of a HP austenitic heat-resistant steel under short-term high-temperature exposition,” Metals, Vol. 15(5), Art. no. 486, https://doi.org/10.3390/met15050486, (2025)

[11] J. Laigo, F. Tancret, R. Le Gall, and J. Furtado, “EBSD phase identification and modeling of precipitate formation in HP alloys,” Advanced Materials Research, Vol. 15–17, pp. 702–707, https://doi.org/10.4028/www.scientific.net/AMR.15-17.702, (2007)

[12] M. Timotijević, O. Erić Cekić, D. Rajnović, and P. Janjatović, “Microstructural analysis of a HP40Nb alloy aged,” Engineering TODAY, Vol. 1, pp. 41–47, (2022)

[13] M. Timotijević, O. Erić Cekić, D. Rajnović, and P. Janjatović, “The influence of overheating temperature on the microstructure of HP40 alloy,” in Proceedings of the 15th International Scientific Conference, Novi Sad, Serbia, (2024)

[14] D. W. Hetzner and W. Van Geertruyden, “Crystallography and metallography of carbides in high alloy steels,” Materials Characterization, Vol. 59, pp. 825–841, https://doi.org/10.1016/j.matchar.2007.07.005, (2008)

[15] R. Skindaras, A. V. Valiulis, and W. L. Spychalski, “The structure and mechanical properties of the high chromium and nickel content cast alloy after long duration work in high temperature,” Mechanika, Vol. 19, pp. 706–710, https://doi.org/10.5755/J01.MECH.19.6.5986, (2014)

[16] G. N. Haidemenopoulos, A. D. Zervaki, H. Kamoutsi, and K. Polychronopoulou, “Creep rupture in HP-Nb refractory steel tubes due to short-term overheating,” European Journal of Materials, Vol. 1(1), pp. 1–22, https://doi.org/10.1080/26889277.2021.1994841, (2021)

[17] Y. Zhou, Y. Liu, X. Zhou, C. Liu, and J. Yu, “Precipitation and hot deformation behavior of austenitic heat-resistant steels: A review,” Journal of Materials Science & Technology, Vol. 33, pp. 1448–1456, https://doi.org/10.1016/j.jmst.2017.01.025, (2017)

[18] A. C. Picasso, C. A. Lanz, M. S. Lissarrague, and A. D. Garófalo, “Microstructure evolution of a nickel-base alloy resistant to high temperature during aging,” Journal of Minerals and Materials Characterization and Engineering, Vol. 4, pp. 48–61, https://doi.org/10.4236/jmmce.2016.41006, (2016)

[19] W. T. H. Honeycombe and R. W. K. Honeycombe, “Structure of centrifugally cast austenitic stainless steels: Part 1. HK40 as cast and after creep between 750 and 1000 °C,” Materials Science and Technology, Vol. 1, pp. 385–389, https://doi.org/10.1179/mst.1985.1.5.385, (1985)

[20] C. Murkin and J. Brightling, “Eighty years of steam reforming,” Johnson Matthey Technology Review, Vol. 60, pp. 263–269, https://doi.org/10.1595/205651316X692923, (2016)

[21] B. Estruch and C. Lyth, “High temperature alloys for use in reformer furnaces,” in Materials Technology in Steam Reforming Processes: Proceedings of the Materials Technology Symposium Held on October 21-22, (1964), C. Edeleanu, Ed. Oxford, U.K.: Pergamon Press, 1966, pp. 29–60, https://doi.org/10.1016/B978-0-08-011323-4.50007-7

[22] A. A. Wahab, C. R. Hutchinson, and M. V. Kral, “A three-dimensional characterization of creep void formation in hydrogen reformer tubes,” Scripta Materialia, Vol. 55, pp. 69–73, https://doi.org/10.1016/j.scriptamat.2006.02.045, (2006)

[23] A. F. Padilha, I. F. Machado, and R. L. Plaut, “Microstructures and mechanical properties of Fe–15Cr–15Ni austenitic stainless steels containing different levels of niobium addition submitted to various processing stages,” Journal of Materials Processing Technology, Vol. 170, pp. 89–96, https://doi.org/10.1016/j.jmatprotec.2005.05.002, (2005)

[24] T. Sourmail, “Precipitation in creep resistant austenitic stainless steels,” Materials Science and Technology, Vol. 17, pp. 1–14, https://doi.org/10.1179/026708301101508972, (2001)

[25] S. Shi and J. Lippold, “Microstructure evolution during service exposure of two cast heat-resisting stainless steels—HP–Nb modified and 20–32Nb,” Materials Characterization, Vol. 59(8), pp. 1029–1040, https://doi.org/10.1016/j.matchar.2007.08.029, (2008)

[26] C. J. Liu and Y. Chen, “Variations of the microstructure and mechanical properties of HP40Nb hydrogen reformer tube with time at elevated temperature,” Materials & Design, Vol. 32, pp. 2507–2512, https://doi.org/10.1016/j.matdes.2010.08.031, (2011)

[27] Y. Zhou, F. Chen, J. Gong, L. Geng, Y. Jiang, and J. Tang, “Effect of overheating on the microstructures and mechanical properties of HP40-Nb furnace tube,” Pressure Vessel Technology, Vol. 38(6), pp. 1–9, https://doi.org/10.3969/j.issn.1001-4837.2021.06.001, (2021)

[28] J. W. Martin, Precipitation Hardening: Theory and Applications. Oxford, U.K.: Elsevier Science & Technology, (1998)

SESSION E: THERMAL TECHNIQUE AND ENVIRONMENTAL PROTECTION

Assessment of the microclimate and subjective perception of the thermal environment in school

Valentina Zochinova[1*], Mihaela Ivanova[1], Tsvetelina Shalamanova[1]

[1] National Centre of Public Health and Analyses, Physical Factors Department, Sofia, Bulgaria

ARTICLE INFO

[*] **Correspondence:** zochinova@abv.bg

DOI: https://doi.org/10.46793/ET26.E04Z **PAGES:** E29 – E33

ABSTRACT

The report presents results of a study of thermal comfort as a criterion for ensuring a favorable learning environment in school buildings. By combining objective measurements of the physical factors of the environment, PMV/PPD indices and subjective surveys, the influence of the microclimate on students in high school are analyzed The study includes objective measurement and assessment of the microclimate by the classical method and indices for thermal comfort and assessment of the subjective thermal sensation of students during the measurement. It is performed during the "cold" period of the year, which is characterized by an average twenty-four-hour temperature below 10ºC and covers 300 male students (14–16 years old) in schools in Sofia. Specialized equipment is used to measure microclimate parameters: temperature, relative humidity, air velocity, radiation temperature and PMV/PPD indices. Tools for visualizing comfort data of the PMV and PPD indices were used as well. To assess the subjective thermal sensation of the environment an anonymous survey among students of the mentioned age group was conducted with a non-representative sample. Conclusions have been drawn concerning the connection between the data from the two types of research, which provides information about the relationship between the values of the studied parameters of the microclimate in the classrooms and the ability of students to concentrate at the temperatures causing discomfort. Over 70% of respondents indicate that thermal conditions affect in moderate to very strong degree affects their attention and concentration. The existing limit values for microclimate could be used as a starting point for development a new legislation for schools environment, but other approach is needed for assessment and control thermal environment.

KEYWORDS

Microclimate, Thermal comfort, School hygiene, High school students, PMV and PPD indices.

1. INTRODUCTION

Ensuring an thermal comfort in schools is a key factor for the health and performance of students. Thermal comfort significantly affects the concentration and performance of students in an educational environment. Scientific research shows that optimal thermal conditions of the learning environment consistent with thermal comfort conditions improve students' ability to concentrate, especially on tasks that require sustained attention and cognitive effort [1] and perform well academically, while discomfort can negatively affect these abilities. The impact of microclimate on learning outcomes can be observed in various aspects of student performance, including academic achievement, emotional well-being, and social development. This relationship is especially important in environments where students spend long periods, such as classrooms in schools. The present study is focused on the dynamics of thermal sensation during different seasons of the year. The aimo is to establish to what extent the objective assessment of the physical parameters of the environment corresponds to the results of a study of the subjective thermal sensation of students.

2. MATERIAL AND METOD

The study includes objective measurement and assessment of the microclimate using classical and exposure indices and assessment of the students' subjective thermal sensation during the measurements.

The applied method takes into account the specifics of microclimate in classrooms, to determine the conditions of thermal comfort in schools, both for students and for characterizing the working environment of teachers.

The present research covers 14 classrooms. The studied premises are in a five-story building with the same geographical orientation east/southeast. They are centrally heated and naturally ventilated.

A preliminary survey and analysis of the learning environment is carried out to identify the sources of heat and humidity, and the air movement. The conditions and parameters of the external environment are taken into account: external temperature and, sunlight. These data make it possible to make an accurate assessment of the parameters of the microclimate in the classroom and to search for the possible connection between the registered environmental parameters values and the subjective sensation among the students during the measurements.

A multidisciplinary approach was applied in the study, which was structured in three stages as follows:

- Objective measurements of temperature, humidity and air velocity.
- Measurements of exposure indices PMV (Predicted Mean Vote) and PPD (Predicted Percentage of Dissatisfied) according to ISO 7730 [2].
- Assessment of subjective heat perception by conducting an anonymous survey among 300 students (14-16 years old) using the 7-point Fanger scale [3].

2.1. Characterization of the Studied Group

The study was conducted among high school students aged 14 to 16. They were characterized by level of education, age, gender, thermal insulation of the clothing worn and metabolism. Thermal insulation was determined according to BNS EN ISO 7933 [2] for the thermal characteristics of clothing, and metabolic rate according to BNS EN ISO 8996 [4].

2.2. Objective Study of Microclimate Parameters in Classrooms

The measurements of temperature, humidity, air velocity and radiation temperature at the predetermined points was carried out at regular intervals of time depending on the learning process or the characteristics of the classroom. The interval between the first and last measurements for the day was not less than half of the established duration of the students' shift in the classroom, and their number was at least 3 to determine the dynamics of the environment.

The measurements were carried out during the cold period of the year which is characterized by an average twenty-four-hour temperature below 10 ºC and the day interval is chosen to cover the period of the classes. They were performed three times a day, at three levels above the ground – at the level of ankles, abdomen and head of students at sitting position. Measurement duration is determined by the time constant of the measuring equipment.

The points are indicated in a schematic field and in a study protocol, where all measured values of the microclimate components are reflected in the time range determined according to the methodology.

Due to the lack of a specific regulatory document on the requirements for the microclimate in schools for children and adolescents, a work environment regulation was used [5]. It determines the limit values of the parameters air temperature, air velocity and humidity for two periods of the year: “warm” and “cold”. For this case limit values for the “cold" period on the year, and "light physical work” (sitting without systematic physical exertion or lifting and carrying of weights). The limit values are as follows: air temperature from 18 to 25 °C, air velocity up to 0.2 m/s and relative humidity in the interval from 30% to 75%.

The results of the measurements of the microclimate parameters are calculated separately for each room determined for measurement. The results for the values of temperature (t), relative humidity (Φ) and air velocity (v) used for assessment are determined by the formulas:

$$X_{\min} = X_{cp} - S;$$

$$X_{\max} = X_{cp} + S;$$

$$X_{cp} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n};$$

$$S = \sqrt{\frac{1}{n-1}\left[\left(X_1 - X_{cp}\right)^2 + \left(X_2 - X_{cp}\right)^2 + \ldots + \left(X_n - X_{cp}\right)^2\right]},$$

where X1, X2....., Xn are the measured values of the respective parameter; Xmin – the minimum value of the parameter X; Xmax – the maximum value of the parameter X; n – the total number of measurements of the respective parameter (temperature, relative humidity and air velocity) carried out at the workplace designated for measurement for the respective period of the year; S – the standard deviation of the respective parameter. The resulting intervals (tmin; tmax), (Φmin; Φmax) and (Vmin; Vmax) are the final results for the comparison with limit values.

2.3. Microclimate Assessment with Thermal Comfort Indices PMV /PPD

The measurement of thermal comfort indices is in parallel to the measurement of microclimate components. Two main parameters are defined for calculating the heat balance:

- Metabolism (Met): Determined based on BNS EN ISO 8996 Ergonomics of thermal environment. Determination of metabolic rate [4]

For the study period, a value of 1.2 met (corresponding to 70 Wm^2) was adopted , corresponding to light sedentary activity (studying).

- Thermal insulation of the worn clothing (Clo): Determined according to BNS EN ISO 7730 [2] Table C.2 - Thermal insulation for clothing and changes in optimal working temperature

An average value of 0.7 clo (winter everyday clothing) was adopted for the study period.

For the purposes of the study, the following measuring equipment and auxiliary tools are used:

- Testo 400 combined device
- Thermal comfort meter type 1212 Bruel and K jaer
- PC software testo Data Control;
- Software tool for calculating and visualizing thermal comfort CBE Termal Comfort Tool - https://comfort.cbe.berkeley.edu/

Aiming to objectively characterizing the solar heating of the premises, measurements of the radiation temperature were also carried out with a spherical thermometer during the measurements.

2.4. Study of the Subjective Perception of Thermal Comfort Among High School Students

Subjective thermal sensation was investigated using a questionnaire method. The developed questionnaire for the study uses Fanger scale [3] and ISO 7730 [2] to assess thermal sensation. The questionnaire is applied during the measurements of the physical parameters characterizing the microclimate, so that the answers to the questions reflect the environmental conditions as much as possible. The questionnaire is used to study the sensitivity to various environmental stressors. The questions refer to the specific individual reactions of the person to environmental factors such as air temperature, air velocity; air humidity. To check the understandability of the questions, a pilot study was conducted on 75 students before the final study. Before the start of the study, written informed consent was obtained from the parents/guardians of the students, as well as the consent of the participants themselves. Participation in the survey is voluntary. No personal data (names, personal identification numbers) were collected that could identify a specific student. The data were processed in aggregate form and are used solely for the purposes of the study. In online surveys, unique codes are used to prevent repeated completion and guarantee objectivity. The results obtained are compared with the data obtained from measurements of environmental parameters and thermal comfort indices. The questionnaire study was approved by the Medical Ethics Committee of the National Center for Public Health and Analysis.

3. RESULTS AND DISCUSSION

3.1. Objective Study of Microclimate Parameters

Air temperature

- The calculated air temperatures in the studied premises are in the range from 15,6 to 26,1°C, with temperature amplitudes during the working day being within 3 – 5 °C, which is the result of sunlight.
- No vertical temperature gradient between ankle level (0.1 m) and head level (1.1 m when sitting) was detected. This prevents the feeling of discomfort in the lower extremities, which is a common problem with poorly designed heating.
- No local cooling was observed from the window surfaces, which indicates good thermal insulation of the window frames used.

Air velocity

- The air velocity in the students' living area ranges between 0.00 m/s to 0.05 m/s.
- During the cold period, strong air flow contributes to the cooling of the human body, but the established values are significantly below the limit values.

Air humidity

- The relative humidity of the air is in the range of 30% to 55%.
- The lower calculated values are close to limit values, which is typical for the winter period due to the operation of heating appliances.
- Humidity remains within limit values, preventing excessive drying of the air and mucous membranes of inhabitants.

Spatial model of a studied room

To visualize the distribution of parameters, a spatial model of the room was prepared with average values plotted by zone: next to a window, next to a door, in the middle of the studied room.

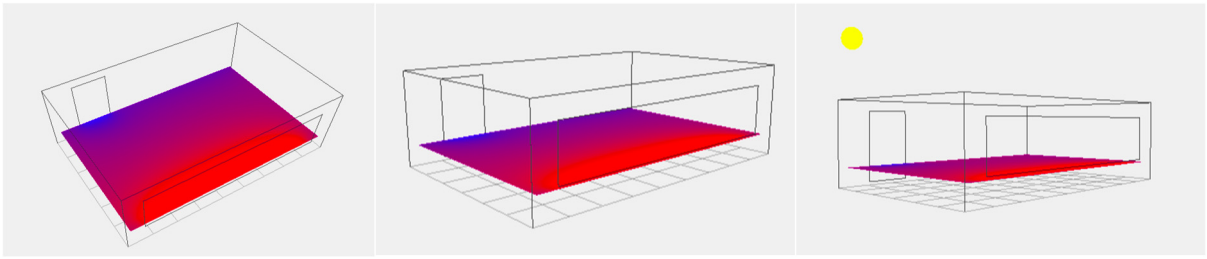


Figure 1: Spatial model

3.2. Results of PPD and PMV Indices Measurements

The use of the PMV (Predicted Mean Vote) and PPD (Predicted Percentage Dissatisfied) indices according to BNS EN ISO 7730 [2] is critically important for the winter period, as it takes into account the subjective thermal sensation of the person, influenced by clothing and metabolism.

As a result of the measurements, the following average values were obtained:

- PMV index: -1.1 to 0.6
- PPD index: 30% to 5%

The minimum value of the measured/calculated PMV values (-1.1) is found near the windows and external walls. The value indicates a feeling of “cool” and is outside the thermal comfort zone according to ISO 7730 (from -0.5 to +0.5). The maximum value (0.6) is observed near the heating appliances. This value corresponds to a feeling of “slightly warm”, which is slightly above the comfort limit. There is a pronounced thermal heterogeneity in the rooms (temperature imbalance between all measurement points), such as registered on sunny days.

The PPD index values from 30% to 5% indicate maximum dissatisfaction (30%), registered in the areas with PMV = -1.1 (around the windows). Nearly a third of the students in these places experience thermal discomfort. Comfort conditions (PPD = 5%) are registered in the central areas of the room. The reasons for the high number of dissatisfied students are cold perception due to the windows, which are the reason for low PMV values (-1.1) and a high percentage of dissatisfied students (30%). Heating devices create peaks up to 0.6 PMV, which also leads to dissatisfaction. The dynamics are due to the low morning temperature in the room and afternoon sunlight.

3.3. Results of the Subjective Study of Thermal Comfort in a Learning Environment

The present research covers 300 respondents, predominant age group is 15 years old (54.5%), followed by of 16-year-olds (33.8%) and 14-year- olds (11.7%). The sample is homogeneous, as prevail men (99.1%). The participants are relatively evenly distributed between 8th and 9th grade, and almost all respondents (99.6%) are found in classrooms type of the premises, which does the results representative basically for this one type of educational environment.

By attitude to thermal comfort, majority of the students (61.9%) determine the conditions as comfortable. Despite this, significant share reports for deviations – 18.2% feel medium cold, and 13.9% – mild heat. Although the extreme sensations (very cold or very warm) have low relative share, frequency on emergence of uncomfortable conditions is substantial: 43.5% of respondents indicate that sometimes they experience discomfort by the time on educational classes, and about 14% – often or very often. This shows that even at compliance with the limit values, there are dynamic fluctuations that influence on perception on the environment.

Key result from research is the identification of "unstable/unsteady" temperature "as main problem stated from 65.8% of participants. This significantly exceeds the rest factors as too low or too high temperature, stationary air and air flow. Therefore, not so much the absolute values of the temperature, as much as the lack on sustainability and control on the microclimate appears as a leading factor for discomfort.

The results clearly show that deviations from the comfort temperature show essential negatively impact on the students. At high temperature about 78% of respondents report different degree of discomfort, and at low temperature this share is similar. This testimonies for high sensitivity to temperature changes, regardless from the direction of their deviation from the comfort (neutral sensation).

Especially the effect is significant on cognitive processes. Over 70% of respondents indicate that thermal conditions affect in moderate to very strong degree affects their attention and concentration. From this result it is evident that there is connection between the microclimate in the classroom environment and efficiency on the school process.

Concerning the air quality - 61.3% of respondents define air quality as "moderate", and 17% - as "bad". Only 9.4 % evaluate it as excellent. This shows existence of potential ventilation problems that can strengthen the feeling for discomfort and influence negatively health and working capacity.

By the attitude on the preferred measures for improvement, greatest share from students (59.9%) supports the use on climatic systems, which reflects aspiration to better control and stability of the premises environmental parameters. The

opening of the windows also is also considered as an effective measure (28.6%), while smaller share supports the use on fans or implementation of specialized ventilation systems.

Interesting result is the lack of clear dominant preference by attitude on temperature range – 54.6% prefer warmer environment, and 45.4% – colder, which emphasizes the necessity from adaptable and flexible solutions.

The results from the study show that although in most cases the school environment is perceived as relatively comfortable, there are essential problems related to instability of temperature conditions and the insufficient ventilation. These factors are shown to have significant influence on concentration and total well-being on the students. The need from improvement of the control of microclimate, including implementation of additional criteria for comfort and methods for measurement and assessment, it is clear expressed and accordingly to be priority in the management of the school environment.

The analysis on spatial distribution of the results shows that the subjective perception is related to the position of the students, finding "to window" (30.6%), more often reported as "unstable" temperature", which could be explained by the influence of the direct heating from the sun and airmovement. Unlikely the students located in the middle on the room (34.1%), more often indicate the problem of “the immovable air "and the lack of fresh air. The lack on flexibility in the system for heating and ventilation is the reason for the high percentage of those experiencing discomfort (43.5%).

The data from the questionnaire also show that 81.3 % of the students’ experience declare strong discomfort or unpleasantness sensation during the school process.

It makes an impression that preferences for the environment are almost equally divided (54.6% for warmer against 45.4% for colder.

The data confirm the hypothesis that the microclimate in the classrooms premises often goes out outside the borders on thermal comfort. The high values on radiation temperature (from the sun) additional load thermoregulation of the students. There is a discrepancy between the normative requirements (limit values) and subjective preferences, such as the students show higher tolerance to cool, but fresh air than to warm and steady environment.

4. CONCLUSION

On the basis on the study, following basic conclusions could be made: The research confirms that the microclimate in the classroom room is a factor of the environment that requires permanent monitoring. The comparison of data from the subjective study with the data from the objective measurements allows more precisely definition of the limit values for thermal comfort, specific for the educational infrastructure. The limit values for microclimate parameters according to Regulation No. RD-07-3 [5] could be used as a starting point for development a new legislation for schools environment, but other approach is needed for assessment and control thermal environment.

The rating "Moderate" for freshness of the air (61.3%) shows that the natural ventilation through opening on windows is insufficient for maintenance of high quality on the air at full capacity on the classrooms.

The students demonstrate high readiness for acceptance on contemporary technologies, with 59.9 % preferring climatic systems, which would decide the problem with the individual regulation.

The thermal comfort it is not simple subjective feeling, but a factor for cognitive efficiency, as the large part of respondents consider that temperature discomfort affects in moderate to very strong degree their attention and concentration.

REFERENCES

[1] Z.S. Zomorodian, M. Tahsildoost, M. Hafezi, Thermal comfort in educational buildings: A review article, Renewable and Sustainable Energy Reviews, Vol. 59, pp 895-906, ISSN 1364-0321, (2016) https://doi.org/10.1016/j.rser.2016.01.033. BNS EN ISO 7730

[2] P.O. Fanger, Thermal Comfort. Danish Technical Press. Copenhagen, (1970).

[3] BNS EN ISO 8996 Ergonomics of thermal environment. Determination of metabolic rate

[4] Ordinance No. RD-07-3 on the minimum requirements for the microclimate on workplaces, State gazette No. 63/2014

[5] EN ISO 9886:2004 - Ergonomics. Assessment of heat stress by physiological measurements

[6] BNS EN ISO 7726, Ergonomics of the thermal environment — Instruments for measuring physical quantities.

[7] ISO 8996, Ergonomics of the thermal environment — Determination of metabolic rate

SESSION E: THERMAL TECHNIQUE AND ENVIRONMENTAL PROTECTION

Urban acupuncture and visual activism as foundations for environmentally responsible riverside transformation: The case of the Vrbas River in Banja Luka

Igor Kuvač[1*], Diana Stupar[1], Markus Schwai[2], Isidora Karan[3], Borjana Mrđa[4]

[1] University of Banja Luka, Faculty of Architecture, Civil Engineering and Geodesy, APLE Laboratory, Banja Luka, Bosnia and Herzegovina

[2] Norwegian University of Science and Technology - NTNU, Department of Architecture and Planning, Trondheim, Norway

[3] Centre for Spatial Research, Banja Luka, Bosnia and Herzegovina

[4] University of Banja Luka, Academy of Arts, Banja Luka, Bosnia and Herzegovina

ARTICLE INFO

[*] **Correspondence:** igor.kuvac@aggf.unibl.org

DOI: https://doi.org/10.46793/ET26.E05K **PAGES:** E35 – E45

ABSTRACT

Environmental protection of urban riparian spaces increasingly demands approaches that go beyond technical and regulatory frameworks - yet the role of cumulative small-scale interventions in establishing standards for future large-scale transformation remains poorly understood.

This paper asks: can student-led small-scale physical interventions and visual activism progressively build the spatial, pedagogical, and civic foundations necessary for environmentally responsible riverside transformation?

A qualitative analysis of three interrelated student initiatives along the Vrbas River in Banja Luka was conducted, spanning nearly a decade: the *Banjalučki mol* (2017), a wooden riverside platform realised through an international student workshop; the Sewage Pipes Festival (2021), combining spatial mapping and an urban installation to expose river pollution caused by absent wastewater treatment infrastructure; and the *"Vrbas is Neither a Construction Site - Nor a Trash Can"* campaign (2025), producing seven visual interventions with civil society organisations in response to illegal riverbed construction. Each case was examined through process documentation, material outcomes, institutional responses, and patterns of community engagement.

The *Banjalučki mol* revitalised a marginalised riverside location and demonstrated the public value of riparian space. The Sewage Pipes Festival initiated public debate on systemic ecological infrastructure. The 2025 campaign triggered inspection procedures and established civic monitoring as environmental advocacy.

Small-scale experimental projects, developed through student pedagogy over time, function not only as independent acts of environmental protection but as preparation for - and implicit criteria governing - future large-scale riverside transformation.

KEYWORDS

Environmental protection, Urban acupuncture, Visual activism, Vrbas River, Small-scale interventions, Riparian transformation, Student pedagogy.

1. INTRODUCTION

The relationship between cities and their rivers [1] has undergone profound transformation across the post-socialist world. In Banja Luka, the Vrbas River has long defined the city's landscape character and collective identity - a city traditionally described as green and dispersed, shaped by the presence of water [2, 3]. Yet despite this historical intimacy, the Vrbas today remains largely inaccessible along most of its urban course: its banks unregulated, its water polluted by direct

sewage discharge, and its riverbed subject to unauthorised construction interventions that proceed without public consultation or environmental assessment [4].

This condition reflects a broader pattern characteristic of post-transitional cities in the Western Balkans, where formal planning frameworks exist but function weakly, regulatory enforcement is inconsistent, and spatial decisions frequently reflect political rather than public interests [5]. In such contexts, environmental protection cannot rely solely on institutional mechanisms. It requires complementary approaches capable of operating outside - and in dialogue with - fragile regulatory systems.

This paper argues that small-scale experimental interventions, developed through student pedagogy over time, represent precisely such an approach. Rather than treating modest spatial and visual projects as ends in themselves, the paper proposes that their cumulative effect is the gradual construction of spatial precedents, ecological standards, and civic frameworks that define what environmentally responsible riverside transformation should look like. In this sense, small-scale interventions function not only as independent acts of environmental protection but as preparation for - and implicit criteria governing - future large-scale urban riverside transformation.

The argument is developed through qualitative analysis of three interrelated student initiatives along the Vrbas River in Banja Luka, spanning nearly a decade of practice: the *Banjalučki mol* (2017), the Sewage Pipes Festival (2021), and the *"Vrbas is Neither a Construction Site - Nor a Trash Can"* campaign (2025). Together, these cases trace a progression from physical revitalisation of marginalised riverside space, through the making visible of systemic ecological problems, to active resistance against environmental degradation - each intervention building on the last, and each contributing to a cumulative body of spatial and civic knowledge about what the Vrbas riverfront can and should become.

2. METHODOLOGY

This paper employs a qualitative research methodology based on case study analysis. The case study approach was selected as appropriate for the research question, which concerns the cumulative effects of small-scale interventions in a specific urban context over time - a phenomenon that resists quantification but is amenable to systematic qualitative examination [6]. Three cases were analysed: the *Banjalučki mol* (2017), the Sewage Pipes Festival (2021), and the *"Vrbas is Neither a Construction Site - Nor a Trash Can"* campaign (2025). The cases were selected on the basis of three criteria: direct engagement with the ecological and spatial conditions of the Vrbas River in Banja Luka; involvement of architecture students as primary agents; and the availability of documentary evidence sufficient for systematic analysis.

Each case was examined through four categories of evidence: process documentation, including workshop programmes, participant records, and organisational materials; material outcomes, including built structures, visual artefacts, mapping products, and installation documentation; institutional responses, including inspection records, media coverage, and administrative decisions; and patterns of community engagement, including social media metrics, civic network activity, and evidence of sustained participant involvement beyond individual projects.

Data were drawn primarily from internal project documentation produced by the Centre for Spatial Research and APLE Lab, supplemented by published conference papers, professional magazine articles, and direct observation by the authors as participants in all three initiatives. This participant-observer position carries both advantages and limitations: it provides access to process dimensions and contextual knowledge unavailable to external researchers, while requiring careful reflexive attention to the risk of confirmation bias in interpretation. The analysis addresses this risk by maintaining a consistent analytical framework across all three cases and by explicitly acknowledging the structural limitations of each intervention alongside its achievements.

The analytical framework applied across cases draws on four theoretical constructs developed in the preceding section: urban acupuncture and small-scale intervention logic; visual activism as environmental advocacy; experimental student pedagogy; and the cumulative "massive, small" logic of incremental spatial change. Each case is assessed against these constructs, and the cross-case analysis in the Discussion section identifies the cumulative pattern that emerges from their combination.

3. THEORETICAL FRAMEWORK

3.1. Small-Scale Interventions and Urban Acupuncture

Contemporary urban theory increasingly recognises that large-scale planning alone is insufficient to address the complexity of post-transitional cities. Urban space is better understood as a multi-scale system in which small, precisely located interventions can generate effects disproportionate to their physical dimensions [7]. This logic underpins the concept of urban acupuncture, developed by Jaime Lerner, which draws on the medical metaphor of the needle: an intervention at a specific point produces not only a local effect but a systemic response across the wider urban body [8]. Manuel de Solà-Morales extends this argument, proposing those small-scale urban projects - precisely because of their proximity to human dimension, perception, and everyday life - can serve as the foundation for larger strategic transformations [9].

Jan Gehl's research on public space confirms that modest interventions, such as the placement of benches or the creation of accessible edges along water, significantly improve urban activity and social interaction [10]. Small-scale is therefore not an aesthetic category but a social and psychological one: it operates at the scale of the human body, perception, and

the possibility of encounter [11]. Critically, small-scale interventions must always be understood as components of a larger system - their value lies not in isolation but in their capacity to catalyse further change and to accumulate into a coherent spatial narrative [7].

3.2. Visual Activism and Environmental Advocacy

Visual communication has emerged as a significant instrument of environmental advocacy, particularly in contexts where institutional channels are weak or captured by political interests. As W.J.T. Mitchell argues, images are not passive representations but active agents capable of producing effects that exceed those of textual or technical communication [12]. Their simultaneous, affective register enables them to cross linguistic, disciplinary, and social boundaries, reaching publics that professional planning discourse rarely engages [13].

In the specific context of spatial and environmental activism, humour and visual irony have proven particularly effective strategies. By combining emotional engagement with critical distance, they generate what Ann Thorpe describes as a redirection of visual production from market-driven persuasion toward political critique and public awareness [14]. This approach is especially relevant in post-socialist contexts, where distrust of official discourse is widespread and alternative communicative channels - social media, civic networks, informal circulation - carry significant weight in shaping public perception of spatial decisions.

3.3. Experimental Student Pedagogy and Social Responsibility

The pedagogical dimension of the interventions analysed in this paper draws on a tradition of experiential and critical learning in architectural education. Paulo Freire's critique of "banking education" - in which students passively receive knowledge for later professional application - remains foundational: authentic learning emerges through engagement with real problems, real consequences, and real communities [15]. Donald Schön's concept of reflective practice formalises this process, distinguishing between reflection-in-action and reflection-on-action as complementary modes through which professional judgment develops under conditions of genuine uncertainty [16].

Etienne Wenger's framework of communities of practice further clarifies how learning occurs through participation in collective work oriented toward shared objectives [17]. When students collaborate with civil society organisations, affected communities, and institutional actors, they develop capacities - strategic thinking, communicative flexibility, civic commitment - that conventional studio pedagogy cannot cultivate. The workshop format, as applied in the cases examined here, operationalises these frameworks: it positions students as co-investigators rather than recipients, and treats real spatial problems as the primary medium of education.

3.4. Cumulative Logic: Massive, Small

The three cases examined in this paper share a common logic that Katie Campbell terms "massive, small": the recognition that substantial urban change can emerge not from singular large-scale interventions but from the strategic accumulation of multiple modest actions, each individually limited but collectively consequential [18]. This logic has relevance for environmental protection in contexts of institutional fragility, where large-scale transformation is rarely achievable through formal planning alone. Small interventions establish spatial precedents, build community identification with place, and generate civic knowledge about what a given environment can and should become - gradually constructing the conditions under which responsible large-scale transformation becomes both possible and accountable.

4. CASE STUDIES

4.1. Banjalučki Mol (2017): Physical Revitalisation and the Public Value of Riparian Space

4.1.1. Context

The right bank of the Vrbas River in the Mejdan neighbourhood, directly across from the medieval Kastel Fortress, represents one of the most historically significant yet spatially neglected stretches of the urban riverfront in Banja Luka. Despite its proximity to the city centre and its exceptional landscape qualities, the location functioned for decades as an underused, inaccessible margin - neither properly maintained nor integrated into the everyday life of the city. It exemplified a broader condition of the Vrbas in Banja Luka: a river present in the city's identity and imagination, but largely absent from its functional public space [2, 3].

4.1.2. Process

The *Banjalučki mol* emerged from the small SCALE international student workshop, organised by the Centre for Spatial Research (Istraživački centar za prostor) in Banja Luka. The workshop brought together students from three universities - the Norwegian University of Science and Technology (NTNU) in Trondheim, the University of Sarajevo Faculty of Architecture, and the University of Banja Luka Faculty of Architecture, Civil Engineering and Geodesy [19]. The project was developed within the framework of the HERD Energy university cooperation programme, supported by the Norwegian Ministry of Foreign Affairs.

The workshop followed an established small SCALE methodology: an initial phase of research and critical analysis of the urban context, followed by collaborative architectural design, and concluded by a five-day on-site construction phase in which students themselves built the structure [6]. Throughout the design process, residents, city representatives, and

local associations were engaged through presentations and dialogue, introducing a participatory dimension uncommon in local planning practice.

The project thus operated simultaneously on three levels: as an educational experiment in applied, community-engaged architectural pedagogy; as a bottom-up spatial intervention in a context where formal planning processes had failed to address the riverside; and as a civic act demonstrating the possibility of public space activation through collaborative, non-institutional initiative.

4.1.3. Intervention

The result of the workshop was a wooden structure of 92 m², comprising a meandering walkway platform threading between existing trees, a small dock extending toward the river, and a series of benches - a combination of urban, landscape, artistic, and architectural design [19]. The simple form derives directly from the spatial conditions of the site, leaving room for appropriation and informal use by visitors. The structure was designed to fulfil two qualities considered essential for strong public space: a clear spatial identity and a concept of open, unprogrammed use [19].

The *Banjalučki mol* was one of the first architectural interventions in open public space along the Vrbas in Banja Luka, and as such attracted significant attention from both the professional community and the general public. Although built on a location previously considered marginal, it gradually became accessible to an increasing number of everyday users and acquired the reputation of one of the favorite places in the city [19]. The project received the BIG SEE Architecture Award in 2019, gaining regional recognition as an example of the possibilities and effectiveness of urban acupuncture.

4.1.4. Environmental and Spatial Significance

From the perspective of environmental protection, the *Banjalučki mol* performed a function that technical or regulatory interventions alone cannot achieve: it established an affective and social relationship between the urban population and the riparian environment. By making a neglected stretch of the Vrbas accessible, legible, and desirable as public space, it generated what can be understood as civic identification with the river - a precondition for any sustained commitment to its protection.

The project also initiated a chain of subsequent changes along the riverside, consistent with the acupuncture logic of systemic response to a precisely located intervention [7]. New activities, informal appropriations, and institutional attention followed, demonstrating that a single modest structure can alter the trajectory of an urban area over time. In this sense, the *Banjalučki mol* established the first spatial precedent in the cumulative argument this paper develops: that the Vrbas riverfront is a public asset of demonstrable value, deserving of protection, investment, and responsible transformation.



Figure 1: Banjalučki mol, 2017. a) The view towards the Kastel fortress and Ferhadija mosque; b) Student construction workshop. Photos by small SCALE

4.2. Sewage Pipes Festival (2021): Making Invisible Pollution Visible

4.2.1. Context

Despite its central role in the identity and landscape of Banja Luka, the Vrbas River functions today as the city's primary open sewage collector. Dozens of direct discharge points can be identified along its urban course, the consequence of a sewage system first constructed in 1912 that has never been complemented by a wastewater treatment facility [4]. The scale of this problem is significant: untreated municipal wastewater enters the river continuously, degrading water quality, producing persistent odour, and rendering the riverbanks inhospitable for much of the year. Yet despite its longevity and visibility, the problem has consistently remained outside political debate and public awareness [4]. This condition of systematic invisibility - in which an ecological crisis of considerable scale is collectively normalised and administratively ignored - constitutes the core problem that the Sewage Pipes Festival set out to address. The festival proceeded from the premise that environmental protection in such contexts requires not only technical solutions but communicative ones: the problem must first be made visible before it can become the object of civic demand and political action.

4.2.2. Process

The Sewage Pipes Festival was organised by the Centre for Spatial Research in Banja Luka, in partnership with the Faculty of Architecture, Civil Engineering and Geodesy and the Academy of Arts of the University of Banja Luka, with co-financing from the Embassy of France in Bosnia and Herzegovina [4]. The central component of the festival was a student workshop bringing together participants from both faculties, structured around the small SCALE methodology established in previous projects [6].

Over six days, the workshop combined analytical and design phases with public performance and installation. Lectures on hydrotechnics, water landscape, contemporary visual arts, and urban public space accompanied the students' analytical process, supporting an interdisciplinary engagement with the problem that crossed the boundaries of architecture, ecology, and visual communication [4]. The workshop thus continued and extended the pedagogical model developed through the *Banjalučki mol*, applying it to a problem that was ecological and political rather than primarily spatial.

4.2.3. Interventions

The festival produced a diverse set of outputs, each addressing the problem of river pollution through a different communicative register and targeting a different audience [4]:

Spatial mapping was conducted through public urban walks and boat rides along the Vrbas, during which students systematically documented sewage discharge points within the urban stretch between the Green and Rebrovac bridges. This produced the first publicly available visual record of the scale and distribution of the problem.

The *S.O.S. for Vrbas* performance deployed coloured smoke signals along the river, sending a clear visual message about the state of pollution to both riverside users and the broader public - a performative gesture rendering the invisible infrastructure momentarily and dramatically present. A festival walks through the city centre carried banners with visual messages that addressed the problem through humour and accessible visual language, engaging passers-by in informal dialogue about pollution and its causes.

A spatial installation on Krajina Square - the main public square of Banja Luka - was constructed from plastic sewage pipes of various lengths and configurations, placing users at the centre of a simulation of the city's sewage system. The installation materialised the problem at the heart of the city's public life, raising awareness of its immediate presence in everyday experience and communicating the scale of the infrastructure deficit through direct spatial encounter [4].

Exhibition panels presented the analytical results of the mapping exercise, demystifying the technical dimensions of the problem and providing factual grounding for public debate. A complementary exhibition, *Landscape of Vrbas - Conditionally*, presented eight urban projects treating the river under one shared condition: clean water.

4.2.4. Environmental and Spatial Significance

The Sewage Pipes Festival achieved measurable communicative impact: festival-related content reached approximately 10,000 views across social media platforms, and the event generated coverage across internet portals and television programmes [4]. More significantly, it initiated a public debate about the Vrbas pollution problem that had previously been absent from civic discourse and established a connection between the local activist and professional community and the broader initiative for the construction of a wastewater treatment system.

From the perspective of the cumulative argument developed in this paper, the festival performed a distinct and essential function within the sequence of interventions. Where the *Banjalučki mol* had established the public value of the riparian space - demonstrating what the Vrbas riverfront could be - the Sewage Pipes Festival established the ecological standard it must meet: a river that is clean, accessible, and integrated into the city's public life as a living environment rather than an open drain. This standard, articulated through visual and spatial means accessible to non-specialist publics, constitutes the second layer of the cumulative framework of criteria that small-scale interventions progressively construct.



Figure 2 a) A part of the S.O.S for Vrbas performance; b) Citizens interacting with the small-scale spatial installation on Krajina Square. Photos by small SCALE

The festival also confirmed and extended the pedagogical model: students moved beyond the design and construction of physical space into the design of communicative strategies, developing capacities for environmental advocacy that conventional architectural education does not cultivate. The multidisciplinary and multimedia character of the festival combining mapping, performance, installation, exhibition, and public dialogue demonstrated that the protection of urban ecological systems requires precisely this breadth of approach.

4.3. "Vrbas Is Neither a Construction Site - nor a Trash Can" (2025): Civic Monitoring as Environmental Advocacy

4.3.1. Context

On 23 June 2025, heavy machinery appeared in the active bed of the Vrbas River in Banja Luka. Without permits, environmental impact assessment, or any form of public consultation, municipal authorities initiated construction works within the riverbed: embankments were constructed, sections of the bed were paved with concrete, and the channel was narrowed at several points [20]. The ecological consequences were immediate and measurable: channel narrowing increased flood-period water velocities, embankment construction destroyed riparian habitat, and concrete paving eliminated the substrate diversity required for aquatic and semi-aquatic species.

The core problem, however, extended beyond the physical construction itself. The modifications proceeded in a condition of institutional invisibility: spatially, because construction took place in areas peripheral to regular pedestrian movement; temporally, because incremental modifications resisted dramatic media representation; and institutionally, because decisions remained confined to administrative channels inaccessible to the public [20]. This invisibility must be understood within the broader context of post-socialist Banja Luka, where environmental regulation exists formally but functions weakly, and where spatial decisions are frequently presented as political achievements rather than outcomes of democratic deliberation [5]. Without visibility, meaningful public deliberation is impossible. The campaign that followed responded directly to this condition.

4.3.2. Process

Recognition of the problem emerged organically through distributed observation: citizens posted photographs on social media, environmental activists began documenting construction, and informal conversations circulated across civic and professional networks. Students who had participated in previous APLE Lab activities recognised the construction as a continuation of a broader pattern of spatial decisions proceeding without public consultation.

By early July 2025, an informal coalition had assembled spontaneously: architecture students interested in visual communication, civic organisations documenting regulatory violations, environmental activists concerned about ecological damage, and faculty facilitating experimental engagement. This coalition operated through a Viber community of 27 participants, functioning as an informal community of practice [17] without hierarchy or institutional mandate. It adopted the name *Diverzanti* (The Subversives), signalling a deliberate stance toward institutional opacity.

An intensive collaborative workshop convened in mid-July 2025, bringing together twenty architecture students with faculty facilitators and partner organisations including the civic initiative *Za plan*, the Centre for Spatial Research, the Centre for Environment, and the Helsinki Committee for Human Rights. Students formed working groups organised around distinct visual modalities - animation, illustration, technical drawing, data visualisation, and photomontage - and engaged in structured strategic dialogue: which audiences need to understand this problem, and which visual forms will best support their comprehension [20]?

4.3.3. Interventions

On 22 August 2025, seven visual interventions were released simultaneously across multiple platforms coordinated through partner organisations. Each embodied a different representational strategy, targeting a different constituency through a distinct visual register [20]:

Fish Walk (*Ribe na suvom*) - an animated video employing a river-level perspective and tragicomic register - positioned public audiences within the affected environment through anthropomorphised fish confronting riverbed degradation, achieving 1,900 views and 156 shares.

Mad Scientist (*Ludi naučnik*) - a video using sarcasm and exaggeration - reached design communities and civically engaged audiences through a scientist celebrating riverbed destruction, generating 1,100 views and 141 shares.

Nature Remembers (*Bager Family*) - a hand-drawn illustration depicting construction equipment as destructive agents - addressed design communities and civic audiences through sharp visual critique, achieving 945 views and 94 shares.

Octopus (*Hobotnica*) - a GIF animation employing a dramatic register - generated emotional recognition of ecological consequence across social media platforms, reaching 1,300 views and 132 shares.

Promenade (*Šetalište*) - an architectural cross-section and plan drawing employing technical professional conventions - communicated the extent of riverbed modification to planning professionals and inspection authorities, achieving only 83 views but generating the most significant institutional response.

Canva Infographic - a data visualisation presenting quantitative dimensions of the construction in terms of extent, materials, and timeline - provided factual grounding for media and general audiences, reaching 826 views and 34 shares.

Bager Friendly - a photomontage employing visual irony - reached critically oriented audiences through juxtaposition of construction equipment and ecological degradation.



Figure 3: (a) Fish Walk (Ribe na suvom) (b) Mad Scientist (Ludi naučnik) (c) Nature Remembers (Bager Family) (d) Octopus (Hobotnica) (f) Promenade (Šetalište) (g) Canva Infographic (h) Bager Friendly (Bager prijatelj). Source: Diverzanti student group, APLE Lab Banja Luka, 2025.

4.3.4. Environmental and Spatial Significance

The institutional response was immediate. Municipal inspection authorities opened procedures and temporarily halted construction within days of the campaign's release - demonstrating that public documentation and visual visibility can activate regulatory mechanisms that administrative opacity had foreclosed [20]. This response confirmed a critical asymmetry: the *Promenade* technical drawing and the *Canva Infographic*, despite achieving the lowest social media engagement, generated the most significant institutional consequences, circulating through professional networks and reaching planning officials whose individual responses carried greater consequence than aggregate public attention.

However, the suspension did not last. Municipal authorities attempted retroactive legalisation through permits and environmental assessments previously omitted. When these were not approved, construction resumed on 23 September 2025, explicitly and in a state of acknowledged administrative illegality [20]. This sequence revealed the structural limits of communicative intervention: visibility is a necessary but insufficient condition for environmental protection when institutional asymmetries allow political authority to override regulatory findings.

Nevertheless, the campaign's effects extended beyond the immediate procedural outcome. The issue entered and remained active within planning circles, civil society organisations, human rights networks, and academic communities. Students sustained voluntary engagement across multiple subsequent workshops dedicated to documentation, hydromorphological analysis, and alternative intervention strategies. A second campaign addressing proposed amendments to Banja Luka's Urban Plan followed in September 2025, demonstrating that the methods and civic commitments developed through the Vrbas campaign had generated transferable capacities for ongoing environmental and spatial advocacy.

From the perspective of the cumulative framework, the 2025 campaign contributed a third and distinct layer to the progressive construction of criteria for responsible riverside transformation: not only must the Vrbas be accessible as public space and clean as an ecological system, but any intervention within its bed must be subject to transparent regulatory process, public participation, and democratic deliberation. This procedural standard, established through civic resistance rather than institutional initiative, represents the most explicitly political dimension of the cumulative argument.

5. DISCUSSION

5.1. Three Interventions, Three Layers of Environmental Criteria

The three case studies analysed in this paper are not isolated events. Together, they reveal a progressive and cumulative construction of environmental criteria along the Vrbas River in Banja Luka - criteria that were not established through formal planning instruments but through repeated acts of spatial, visual, and civic engagement over nearly a decade.

The *Banjalučki mol* established the first layer: a spatial precedent demonstrating that the Vrbas riverfront is a public asset of demonstrable value. By making a neglected stretch of the river accessible, desirable, and actively used, it generated civic identification with the riparian environment and demonstrated - through built reality rather than planning proposal - what the riverside could be. This is not a minor contribution: in a context where the river had long been treated as a margin rather than a resource, the simple act of building a wooden platform and watching it become a favourite place in the city constitutes an argument about environmental value that no regulatory document had previously made.

The Sewage Pipes Festival established the second layer: an ecological standard that the river must meet if it is to fulfil the public value the *Banjalučki mol* demonstrated. By making the sewage problem visible - spatially, performatively, and analytically - the festival transformed a normalised condition of ecological degradation into a publicly recognised problem demanding a solution. It articulated, in communicative terms accessible to non-specialist publics, that a river functioning as an open sewer is incompatible with the vision of riparian public life that the *Banjalučki mol* had embodied.

The 2025 campaign established the third layer: a procedural and democratic standard governing how any intervention in the riverbed must be conducted. By documenting illegal construction, activating inspection mechanisms, and sustaining public attention across professional and civic networks, the campaign asserted that the Vrbas riverbed is not available for arbitrary administrative or political appropriation - that any transformation of this space requires transparency, environmental assessment, and meaningful public participation. This procedural standard is the most explicitly political of the three layers, and the most directly relevant to the governance of future large-scale riverside transformation.

Together, these three layers - spatial value, ecological standard, procedural legitimacy - constitute an implicit framework of environmental criteria that any future large-scale transformation of the Vrbas riverside would need to address. They were not written into a plan or codified in a regulation. They were constructed through practice, demonstrated through built and visual interventions, and validated through public response and institutional reaction.

5.2. The Cumulative Logic and Its Limits

The cumulative logic described above - what Campbell terms "massive, small" [18] - operates through a mechanism distinct from conventional planning. Rather than proceeding from strategic vision to implementation, it proceeds from practice to precedent: each intervention establishes a spatial, ecological, or civic reference point that conditions what comes next. The sequence is not predetermined but retrospectively coherent: it is only in looking back across the decade that the progressive construction of an environmental argument becomes visible.

This retroactive coherence has both strengths and limitations. Its strength lies in its resilience: because the criteria are embedded in built reality, public memory, and civic practice rather than in planning documents, they are not easily reversed by administrative decision or political change. The *Banjalučki mol* continues to be used and valued; the sewage problem, once made visible, cannot be re-normalised with the same ease; the 2025 campaign established a documented record of institutional illegality that persists in professional and civil society networks regardless of subsequent administrative actions.

Its limitation, however, is equally significant. The cumulative construction of environmental criteria through small-scale interventions does not, by itself, produce the large-scale transformation it prepares. The Vrbas riverfront remains largely

inaccessible, polluted, and subject to arbitrary intervention. Inspection procedures were opened and closed; construction continued in administrative illegality; the wastewater treatment system remains unbuilt. The gap between the criteria established through practice and their institutional recognition and application remains wide.

This gap points to a structural condition of post-transitional urban governance that small-scale interventions alone cannot overcome: the asymmetry between the capacity of civic and pedagogical action to establish standards and the capacity of political and administrative authority to ignore them. Visual activism and urban acupuncture can make problems visible and demonstrate alternatives, but they cannot substitute for the institutional will and regulatory capacity required to implement large-scale environmental transformation.

5.3. Student Pedagogy as Environmental Practice

Across all three case studies, the student workshop emerges as the primary organisational and methodological unit. This is not incidental. The workshop format - combining research, design, construction or visual production, and public engagement within a compressed timeframe and outside conventional curricular structures - creates conditions for the kind of authentic, consequential learning that Freire, Schön, and Wenger identify as the basis for genuine professional and civic development [15, 16, 17].

Students who participated in these workshops did not merely acquire technical skills. They developed strategic communicative capacity, civic commitment, and an understanding of environmental problems as simultaneously spatial, ecological, political, and pedagogical. The sustained voluntary engagement documented across all three cases - students returning for subsequent workshops, initiating new campaigns, and transferring methods to new contexts - suggests that the pedagogy generated forms of environmental consciousness that extend beyond any individual project.

This finding has implications for architectural and engineering education more broadly. Environmental protection in post-transitional contexts requires practitioners capable of operating across institutional boundaries, communicating with diverse publics, and sustaining engagement under conditions of uncertainty and institutional resistance. These capacities are not developed through conventional technical training. They require pedagogical frameworks that position students as agents rather than recipients - frameworks of the kind that the small SCALE and APLE Lab models represent.

5.4. Towards Large-Scale Riverside Transformation

The argument developed in this paper is ultimately prospective. The three interventions analysed here are significant not only for what they achieved but for what they prepared. The Vrbas riverside in Banja Luka remains a largely unrealised urban potential - a corridor of exceptional landscape, ecological, cultural, and historical value that formal planning has consistently failed to integrate into the life of the city [2, 3, 5].

When large-scale transformation of this corridor eventually becomes a political and institutional possibility, the criteria for what such transformation should look like will not need to be constructed from scratch. They have been built, incrementally and through practice, over nearly a decade of student-led intervention. A riverside that is publicly accessible and spatially valued; a river that is ecologically clean and free of untreated sewage; a planning process that is transparent, participatory, and legally grounded - these are not abstract planning aspirations. They are standards demonstrated, demanded, and defended through the cumulative body of work this paper analyses.

In this sense, the small-scale interventions along the Vrbas function as more than independent acts of environmental protection. They function as the advance work of responsible large-scale transformation: establishing its criteria, building its constituency, and holding its future implementation accountable to a set of spatial, ecological, and democratic standards that have already been publicly tested and validated.

6. CONCLUSION

This paper set out to ask whether student-led small-scale physical interventions and visual activism can progressively build the spatial, pedagogical, and civic foundations necessary for environmentally responsible riverside transformation. The qualitative analysis of three interrelated initiatives along the Vrbas River in Banja Luka, spanning nearly a decade, suggests that the answer is affirmative - with important qualifications.

The *Banjalučki mol* (2017), the Sewage Pipes Festival (2021), and the *"Vrbas is Neither a Construction Site - Nor a Trash Can"* campaign (2025) each contributed a distinct and irreplaceable layer to a cumulative framework of environmental criteria. The first demonstrated the public value of riparian space through built reality. The second made a normalised ecological crisis visible and transformed it into a subject of civic demand. The third established the procedural and democratic standards that any future intervention in the riverbed must meet. Together, they constitute an implicit environmental argument - not codified in planning documents, but embedded in spatial practice, public memory, and civic commitment - that defines what responsible large-scale transformation of the Vrbas riverside should look like.

Several conclusions follow from this analysis. First, small-scale interventions are not merely compensatory responses to the failures of formal planning. When strategically sequenced and analytically connected, they perform a generative function: establishing precedents, articulating standards, and building the civic constituency that large-scale transformation requires. Second, visual activism and urban acupuncture are complementary rather than alternative strategies: physical interventions establish the spatial value of an environment, while visual interventions make its ecological and procedural

conditions legible to diverse publics. Their combination produces a more complete environmental argument than either could generate alone. Third, the student workshop - as an organisational form positioned outside conventional curricular structures - proves capable of generating both spatial outcomes and forms of civic and professional development that conventional technical education cannot cultivate.

At the same time, the analysis confirms a structural limitation that small-scale interventions cannot overcome alone. The cumulative construction of environmental criteria through practice does not, by itself, produce institutional transformation. The gap between the standards established through civic and pedagogical action and their recognition and application by political and administrative authority remains the central challenge of environmental governance in post-transitional contexts. Closing this gap requires not only the continued accumulation of small-scale precedents but the development of institutional mechanisms capable of translating civic standards into binding planning obligations.

The broader implication is significant for both environmental practice and architectural education. In contexts characterised by institutional fragility and weak regulatory enforcement, the protection of urban ecological systems depends on the gradual construction of civic and spatial knowledge about what those systems can and should be. Student pedagogy, when oriented toward real problems and real communities, is a primary vehicle for this construction. The cases examined here suggest that architecture and engineering education has a role in environmental protection that extends well beyond the technical - one that is spatial, communicative, civic, and ultimately political.

The Vrbas riverside in Banja Luka remains a largely unrealised urban potential. But the criteria for its responsible transformation have already been built. They are waiting.

7. ACKNOWLEDGEMENTS

The authors wish to thank the Centre for Spatial Research (Istraživački centar za prostor), the Norwegian University of Science and Technology (NTNU), the Norwegian Ministry of Foreign Affairs, the Erasmus+ APLE project, the Embassy of France in Bosnia and Herzegovina, the Faculty of Architecture, Civil Engineering and Geodesy of the University of Banja Luka, and all students whose contributions have made this long-term research possible.

REFERENCES

[1] M. Čvoro, "Priobalje i grad ", Univerzitet u Banjoj Luci, Arhitektonsko-građevinsko-geodetski fakultet, Banja Luka, (2021)

[2] D. Simonović, "Pejzažni gradovi: poređenje razvoja urbanog identiteta Banjaluke i Graca", Arhitektonsko-građevinski fakultet, Banja Luka (Bosnia and Herzegovina), (2010)

[3] A. Kirjakov, "Generalni urbanistički plan - Banjaluka", Urbanistički zavod Generalne uprave za komunalne poslove Narodne republike Bosne i Hercegovine, Sarajevo (Bosnia and Herzegovina), (1951)

[4] I. Kuvač, B. Mrđa, I. Komljenović, A. Marić, A. Simićević, V. Otašević, and B. Ostić, "Festival kanalizacionih cijevi: vizuelizacija problema zagađenja rijeke Vrbas u Banjaluci", Zbornik radova Naučno stručna konferencija Instalacije i arhitektura, Univerzitet u Beogradu, Arhitektonski fakultet, Beograd (Srbija), 2 decembar 2021, pp. 47-55, (2021)

[5] I. Kuvač and I. Karan, "Novi koncept urbanog razvoja Banjалуке: između privatnih interesa i potreba građana", in R. Perić-Romić and D. Savić (eds.), *Novi koncept urbanog razvoja Banjалуке*, Fakultet političkih nauka, Banja Luka (Bosnia and Herzegovina), pp. 33–56, (2018)

[6] R. K. Yin, "Case Study Research and Applications: Design and Methods", SAGE Publications, Thousand Oaks (USA), (2018)

[7] I. Karan, M. Stijak, and I. Kuvač, "The small-scale approach as a generator for urbanity increase of Banja Luka city", Proceedings of Places and Technologies, Sarajevo (Bosnia and Herzegovina), 8 June - 9 June 2017, pp. 126 - 134, (2017)

[8] J. Lerner, "Acupuntura urbana", Instituto de Arquitectura Avanzada de Cataluña, Barcelona (Spain), (2005)

[9] M. de Solà-Morales, "De cosas urbanas", Gustavo Gili, Barcelona (Spain), (2008)

[10] J. Gehl, "Cities for People", Island Press, Washington (USA), (2010)

[11] D. Radović and D. Boontharm, "Small Tokyo: Measuring non-measurable", Flick Studio, Tokyo (Japan), (2012)

[12] W.J.T. Mitchell, "What Do Pictures Want? The Lives and Loves of Images", University of Chicago Press, Chicago (USA), (2005)

[13] W.J.T. Mitchell (ed.), "Landscape and Power", University of Chicago Press, Chicago (USA), (1994)

[14] A. Thorpe, "Architecture and Design versus Consumerism: How Design Activism Confronts Growth", Routledge, London (England), (2012)

[15] P. Freire, "Pedagogy of the Oppressed", Herder and Herder, New York (USA), (1970)

[16] D. Schön, "The Reflective Practitioner: How Professionals Think in Action", Basic Books, New York (USA), (1983)

[17] E. Wenger, "Communities of Practice: Learning, Meaning, and Identity", Cambridge University Press, Cambridge (England), (1998)

[18] K. Campbell, "Making Massive Small Change: Building the Urban Society We Want", Chelsea Green Publishing, White River Junction (USA), (2018)

[19] I. Kuvač, S. Peulić, U. Umićević, I. Karan, M. Burazor, M. Schwai, R. Vukomanović, and D. Zrnić, "Banjalučki mol", in BIG SEE: Odličnost jugoistočne Evrope - Bosna i Hercegovina, pp. 140–143, (2019)

[20] I. Kuvač and D. Stupar, M. Schwai, "Socially Responsible Action as a Form of New Experimental Practices for Architecture Students", Proceedings of the International Conference on Architecture Pop-up Lab Exchange (APLE), Sarajevo (Bosnia and Herzegovina), 26 March - 27 March 2026, pp. 156–175, (2026)

SESSION E: THERMAL TECHNIQUE AND ENVIRONMENTAL PROTECTION

Application of robotic systems in intensive livestock farming from the aspect of environmental protection

Suzana Knežević[1*], Milena Milojević[2], Aleksandra Milošević[3]

[1] Academy of Applied Studies Šabac, Unit for Agricultural and Business Studies and Tourism, Šabac, Serbia

[2] Independent Researcher, Belgrade, Serbia

[3] University of Belgrade, Faculty of Law, Belgrade, Serbia

ARTICLE INFO

[*] **Correspondence:** suzanaknezevic0603@gmail.com

DOI: https://doi.org/10.46793/ET26.E06K **PAGES:** E47 – E53

ABSTRACT

Modern livestock production is characterized by the increasing application of automated and robotic systems aimed at improving production processes and more efficient resource utilization. Robotic systems in livestock farming are most commonly applied in milking processes, feed distribution, manure removal, and control of microclimatic conditions on farms. Their application can contribute to more rational water and energy consumption, more efficient animal waste management, and reduction of harmful gas emissions. In addition to technological and production advantages, modern robotic systems play an important role in improving the sustainability of livestock production and environmental protection.

KEYWORDS

Robotic systems, Livestock farming, Environmental protection, Sustainability, Gas emissions, Automation

1. INTRODUCTION

Intensive livestock production represents an important segment of modern agriculture and plays a significant role in ensuring the supply of animal-derived food for the growing world population. The increasing consumption of meat, milk, and eggs has led to the development of production systems based on the concentration of large numbers of animals, improved infrastructure, and more intensive use of resources [1]. Such systems enable high productivity, but at the same time create a significant burden on the environment.

One of the most significant problems associated with intensive livestock farming concerns the large quantities of animal waste generated during various production processes. In conditions where large numbers of animals are concentrated within a limited area, increased releases of nitrogen, phosphorus, and organic matter occur, which may lead to soil and water pollution if waste is not managed adequately. A particular issue is posed by Concentrated Animal Feeding Operations (CAFO systems), which have a high potential for polluting surface and groundwater due to the presence of nitrates, ammonia, and faecal microorganisms [2].

In addition to waste management problems, intensive livestock production represents a significant source of greenhouse gas emissions, primarily methane (CH_4), nitrous oxide (N_2O), and ammonia (NH_3), which negatively affect air quality and contribute to climate change. According to a report by the Food and Agriculture Organization of the United Nations, the livestock sector has a significant impact on global greenhouse gas emissions [3]. At the same time, intensive livestock production affects natural nitrogen, phosphorus, and carbon cycles, as well as soil and water quality [4].

Bibliometric analyses of the development of smart and digital agriculture indicate a strong trend toward the integration of automation, robotics, and digital technologies into modern production systems. These development directions include the application of intelligent systems, IoT technologies, and robotic solutions, whereby increased production efficiency is associated with improved sustainability and more rational use of resources in agriculture [5].

In order to mitigate environmental pressures, modern livestock farming increasingly employs robotic and automated systems that enable more precise management of production processes. These systems are designed so that management is adapted to each animal according to its actual condition (health, production, reproduction, and behaviour). Unlike conventional systems, which focused on herd-level observation, these systems enable continuous monitoring of production conditions, individualized animal management (Precision Livestock Farming – PLF), and optimization of key processes, which may contribute to reducing resource losses and negative environmental impacts. Modern robotic systems are used for feed distribution, manure removal, control of microclimatic conditions, and monitoring of gas emissions, thereby contributing to more rational resource use and improved sustainability of livestock production [6].

The aim of this paper is to analyse the application of robotic systems in intensive livestock farming from the perspective of environmental protection, with a particular focus on the potential for reducing the negative impacts of livestock production through the use of modern technological solutions. The paper examines robotic systems used in milking, feeding, manure removal, and microclimate control processes on farms, as well as their contribution to more rational use of water and energy, more efficient management of animal waste, and reduction of harmful gas emissions.

2. ENVIRONMENTAL PROBLEMS OF INTENSIVE LIVESTOCK FARMING

Intensive livestock farming is characterized by the concentration of a large number of animals in a limited area, which leads to increased emissions of harmful gases, primarily ammonia (NH_3), methane (CH_4), and nitrous oxide (N_2O), all of which have a significant impact on air quality and climate change [3]. These gas emissions are associated with digestive processes in animals, the storage and decomposition of manure, and housing conditions on farms.

In addition to greenhouse gas emissions, an important environmental issue is animal waste management. Improper storage and application of manure can lead to the eutrophication of surface waters and the contamination of groundwater with nitrates and phosphates [2]. Large quantities of organic matter and nutrients released into the environment further disrupt ecosystem stability and the quality of soil and water.

Intensive livestock production involves a high concentration of animals per unit of space and maximised production (milk, meat, eggs) within a short period. For this reason, it is highly resource-intensive, requiring significant amounts of water, energy, and feed, thereby increasing pressure on natural resources and the overall environmental footprint of production [4]. Consequently, modern approaches in livestock farming are increasingly focused on the application of technological solutions that enable more efficient resource use and reduce negative environmental impacts.

3. ROBOTIC SYSTEMS IN INTENSIVE LIVESTOCK FARMING

Modern livestock farming is characterised by rapid technological transformation driven by the application of robotic and digital systems. According to Dilaver and Dilaver (2024), robotic systems in livestock farming cover various areas of application, including automated milking, feeding, monitoring of animal health, and automated farm management [7].

Within these systems, a high degree of automation in milking enables the identification of animals via sensors and the execution of the process without direct human involvement in operational tasks, while simultaneously collecting data on production and animal condition. Similarly, robotic feeding systems enable precise feed distribution and more efficient use of feed with reduced wastage, thereby improving feeding efficiency and farm work organisation.

In addition, the use of sensors and smart devices enables continuous monitoring of animal health and the detection of behavioural changes, contributing to timely responses and better control of the production process. Robotic systems are also used for maintaining facility hygiene, thereby improving animal housing conditions, animal welfare, and the overall organisation of the farm.

According to the literature, such application of technology leads to a reduced need for labour, increased production efficiency, and better control of production processes [7].

In a broader context, the application of robotic systems in intensive livestock farming represents one of the key directions in the modernisation of livestock production within the concept of automated and digitally supported farms. The primary goal of these systems is the optimisation of production processes and the improvement of resource-use efficiency, while simultaneously reducing negative environmental impacts [8].

3.1. Robotic Milking Systems

Automated milking systems (AMS) have undergone intensive development over the past thirty years, and their application has shown steady growth, as confirmed by an expanding body of scientific literature through a continuous increase in publications in this field. These systems offer numerous advantages, among which the reduction of labour associated with milking and an increase in milk yield per cow stand out. However, in the context of growing concerns regarding animal welfare in intensive production, it has become essential to thoroughly examine the impact of AMS on animal health and welfare, as these aspects represent key determinants of the long-term sustainability of the dairy sector [9].

Modern AMS systems (Figure 1) use sensors, laser positioning, and software algorithms for animal identification, milk quality control, and monitoring of the health status of individual animals. In addition to improving efficiency, this approach contributes to reducing stress in animals and enables more stable milk production.



Figure 1: Contemporary AMS system [10]

For an AMS to function optimally, cows must regularly pass through the robotic milking system, and to facilitate this, they need to be in constant movement within the facility. It is necessary that they have sufficient time for rest and feeding, and these behaviours must be aligned with their milking schedule [11].

During each milking session, automatic sensors enable continuous monitoring of udder health and milk quality, providing detailed data for each cow, which earlier conventional systems were unable to offer. Due to these advantages, the adoption of AMS has shown significant growth worldwide, with an estimated 50,000 units in operation across 25,000 farms in 2019. The implementation of AMS technology not only facilitates the collection of data on milk yield and quality, but also opens up the possibility for systematic study of cow behaviour and welfare under loose housing conditions. The introduction of AMS has altered the daily rhythms and behavioural patterns of cows during their lactation period, which highlights the need to consider both system efficiency and animal welfare in parallel. Despite the extensive data collected on farms equipped with AMS, a significant potential for herd characterisation and management optimisation remains insufficiently explored [12].

3.2. Robotic Feeding Systems

The development of feeding robots (Figure 2) represents one of the key solutions for intensive livestock farming, as it can mitigate problems related to labour shortages, high work intensity, and untimely feeding. Within contemporary research, autonomous feed distribution systems based on sensors and automatic navigation are being developed, enabling more precise feed distribution and continuous operation within animal housing facilities. Such systems contribute to reducing feed wastage and improving the organisation of the production process [13].



Figure 2: Example of a feeding robot [14]

Previous research has mainly focused on the application of robotic systems in facilities for intensive rearing of poultry and other domestic animals, particularly in cattle and sheep barns. Examples of such systems include stationary automatic feeding robots, autonomous devices equipped with feed mixers, as well as transport systems operating with magnetic or laser guidance. These systems operate effectively in strictly organised and spatially defined facilities, while their application is significantly limited in unstructured conditions characteristic of semi-intensive farms. In such conditions, the number of feeding points requiring replenishment varies, and the robot's route changes with each feeding cycle. Robot path planning in this context can be considered a multi-objective optimisation problem.

In intensive livestock farming, the energy consumption of feeding robots represents a key factor, as it directly affects operational costs. Therefore, route optimisation for robots is increasingly viewed through the lens of reducing energy consumption. The literature indicates that for a smaller number of feeding points, exact algorithms can be used to obtain a path with minimal energy consumption, while for a larger number of feeding points, problem-solving methods based on practical rules and heuristic patterns are applied in order to achieve a sufficiently good solution within a reasonable

time, i.e. heuristic algorithms based on genetic operators that enable approximation of the optimal solution. The effectiveness of these approaches has been confirmed through comparative experiments [15]. Robots make decisions in changing and unpredictable conditions.

3.3. Robotic Manure Removal Systems

Manure management represents one of the key components of modern livestock management, as manure significantly affects air quality within facilities and also contributes to greenhouse gas emissions. Robots for manure removal are widely used in the mechanisation of solid manure handling. Mobile systems offer a number of advantages, among which low investment costs, easy adaptability to different types of facilities, and the possibility of precisely defining cleaning regimes stand out [16].

The traditional approach to cleaning slatted floors on dairy farms involved the use of automatic scrapers in alleyways, based on cable or shuttle drive systems, which proved effective in pushing manure through the slats. However, in recent years, an increasing number of robotic applications for slatted floor cleaning have appeared on the market. These systems operate without cables and can be programmed to selectively remove manure in areas of the facility where it accumulates most. In addition, the robots are equipped with advanced obstacle-detection functions using ultrasonic sensors (Figure 3), enabling them to detect the presence of animals or other objects and, depending on the situation, temporarily stop operation or bypass the obstacle [17].



Figure 3: ENRO-dung removal robot [18]

3.4. Robotic Systems for Microclimate Control

Microclimate in intensive livestock housing includes air temperature, relative humidity, air velocity, and gas concentrations, as well as the presence of dust and microorganisms in the area where animals are directly located, move, and stay (Animal Occupied Zone – AOZ). These parameters have a significant impact on animal health, welfare, and productivity [19]. Modern systems use sensors and automated control units for continuous monitoring and regulation of microclimatic conditions in facilities.

Automated ventilation control enables the reduction of ammonia, carbon dioxide, and humidity concentrations in facilities, thereby improving air quality and reducing the negative environmental impact of livestock production. In addition, the optimisation of ventilation and cooling systems contributes to lower energy consumption and improved energy efficiency on farms. In modern livestock facility design, CFD models (Computational Fluid Dynamics) are widely applied, enabling the simulation of airflow and the distribution of temperature and gases within buildings, with the aim of optimising microclimatic conditions and reducing emissions of harmful substances [19].

4. APPLICATION OF ROBOTIC SYSTEMS IN ENVIRONMENTAL PROTECTION

The application of robotic and automated systems in intensive livestock farming enables more precise management of production processes, thereby contributing to more efficient resource use and a reduction in negative environmental impacts. Modern systems based on digital monitoring and automation allow for the optimisation of animal feeding, herd management, and production conditions, which may contribute to the reduction of greenhouse gas emissions and the improvement of the sustainability of livestock production.

Preethichandra, Piyathilaka, and Izhar (2025) discuss the application of robotic systems in environmental monitoring, whereby their use can also be considered in the context of livestock production through the monitoring of relevant environmental parameters. These systems enable continuous and precise data collection, which provides a basis for improving farm management and gaining a better understanding of production conditions [20].

Additionally, research based on the RuFaS (Ruminant Farm System) simulation model shows that increased feed-use efficiency achieved through automated and robotic processes can lead to a reduction in total feed intake of 6–12%, while simultaneously reducing methane emissions and the negative impacts of manure storage by 4.6–9.3% [21]. These results

indicate the significant potential of modern robotic and automated systems in reducing the environmental burden of livestock production through more rational feed use, more precise feed allocation, more uniform distribution and reduced spillage during feeding, improved feeding efficiency, and more effective management of animal waste.

However, the assessment of the impact of robotic systems cannot be viewed solely through operational effects. Life cycle analysis of robotic systems in livestock farming shows that environmental impact encompasses all stages, from production to disposal of equipment. Although feeding robots contribute to reducing feed waste and improving production efficiency, the overall environmental footprint depends on energy consumption during operation, as well as the materials and technological components involved in the equipment's life cycle. The presented data indicate a trade-off between increased production efficiency and additional technological environmental burden [22].

The positive environmental impact of robotisation is particularly pronounced in milk production systems, where automated systems enable more efficient management of production processes and resources. Research based on the Life Cycle Assessment (LCA) method, in accordance with ISO 14040 and ISO 14044 standards, shows that the application of robotic systems can contribute to reducing greenhouse gas emissions and improving production sustainability [23].

A key application in this context is automated milking systems (AMS), which enable the automation of the milking process, continuous monitoring of production parameters, and improved control of animal health status. Although their use leads to increased electricity consumption due to the operation of robotic units and supporting automation, higher production efficiency and improved herd management are simultaneously achieved [24].

Higher productivity allows for a reduction in environmental impact per unit of product, i.e. per kilogram of milk produced. In addition, automated systems enable more rational use of feed, energy, and water, as well as more efficient management of animal waste. Of particular importance is the use of anaerobic digesters, through which methane generated from manure decomposition can be used for electricity production, further reducing the negative environmental impact of livestock production [23].

Martinsson et al. (2024) indicate that robotic milking systems can increase the environmental efficiency of farms by reducing labour per animal and increasing productivity. However, the same authors emphasise that such systems may also lead to an increase in herd size and higher consumption of concentrate feed, which can negatively affect greenhouse gas emissions, highlighting the need for a balanced application of technology in livestock production [25].

5. CONCLUSION

Intensive livestock production represents a significant source of environmental pressure due to greenhouse gas emissions, large quantities of animal waste, and the intensive use of natural resources. In modern production conditions, the need to increase efficiency while simultaneously reducing negative environmental impacts is directing the development of livestock farming towards the application of automated and robotic systems.

Analysis of the available literature shows that robotic systems for milking, feeding, manure removal, and microclimate control can contribute to more rational use of feed, water, and energy, improved animal waste management, and reduced emissions of harmful gases. The particular importance of these systems lies in their ability to continuously monitor production parameters and optimise processes, thereby increasing production efficiency and improving the sustainability of livestock farming.

At the same time, research findings indicate that the environmental effects of robotisation are not uniform and depend on the way the technology is applied, the organisation of production, and the overall life cycle of the equipment. Although the use of robotic systems can reduce the environmental burden per unit of product, increased energy consumption and the technological complexity of the systems are important factors that must be taken into account when assessing their overall environmental impact.

In the Republic of Serbia, there are farms that apply automated and robotic systems in various segments of production; however, their number is still relatively small and limited to individual, technologically more advanced farms. Although their application contributes to increased production efficiency, better process control, and improved animal welfare, broader implementation is still limited due to high investment costs, the need for specialised knowledge for system management, and insufficient availability of financial incentives. For this reason, automation in Serbia is still in a phase of gradual development and slower expansion compared to developed agricultural systems in the European Union. Based on the analysed data, it can be concluded that robotisation represents a significant direction in the modernisation of intensive livestock farming and has the potential to contribute to more sustainable management of production processes. Further development and application of these systems should be aimed at achieving a balance between increased productivity, economic efficiency, environmental protection, and, above all, animal welfare.

REFERENCES

[1] M. Herrero, S. Wirsenius, B. Henderson, C. Rigolot, P. Thornton, P. Havlík, I. de Boer, and P. J. Gerber, "Livestock and the Environment: What Have We Learned in the Past Decade?", Annual Review of Environment and Resources, Vol. 40, pp. 177–202, https://doi.org/10.1146/annurev-environ-031113-093503, (2015)

[2] M. A. Mallin, M. R. McIver, A. R. Robuck, and A. Kahn Dic, “Industrial Swine and Poultry Production Causes Chronic Nutrient and Fecal Microbial Stream Pollution”, Environmental Science & Technology, Vol. 49(20), pp. 12033–12039, https://doi.org/10.1021/acs.est.5b02895, (2015)

[3] P. J. Gerber, H. Steinfeld, B. Henderson, A. Mottet, C. Opio, J. Dijkman, A. Falcucci, and G. Tempio, “Tackling Climate Change through Livestock: A Global Assessment of Emissions and Mitigation Opportunities”, Food and Agriculture Organization of the United Nations (FAO), Rome, https://www.fao.org/3/i3437e/i3437e01.pdf, (2013)

[4] A. Leip, G. Billen, J. Garnier, B. Grizzetti, L. Lassaletta, S. Reis, D. Simpson, M. A. Sutton, W. de Vries, F. Weiss, and H. Westhoek,“Impacts of European livestock production: nitrogen, sulphur, phosphorus and greenhouse gas emissions, land-use, water eutrophication and biodiversity”, Environmental Research Letters, Vol. 10(11), 115004, https://doi.org/10.1088/1748-9326/10/11/115004, (2015)

[5] J. P. Pulumbarit,“Smart Farming and Digital Agriculture: A Bibliometric Perspective on Past, Present and Future Research Trends”, Indian Journal of Agricultural Research, https://doi.org/10.18805/IJARe.AF-1023, (2026)

[6] E. Tullo, A. Finzi, and M. Guarino, “Review: Environmental impact of livestock farming and Precision Livestock Farming as a mitigation strategy”, Science of the Total Environment, Vol. 650, pp. 2751–2760, https://doi.org/10.1016/j.scitotenv.2018.10.016, (2019)

[7] H. Dilaver and K. F. Dilaver,“Robotics Systems and Artificial Intelligence Applications in Livestock Farming”, Journal of Animal Science and Economics, Vol. 3(2), pp. 63–72, https://doi.org/10.5281/zenodo.12518170, (2024)

[8] T. Duckett, S. Pearson, S. Blackmore, and B. Grieve, “Agricultural Robotics: The Future of Robotic Agriculture”, arXiv preprint arXiv:1806.06762, https://arxiv.org/abs/1806.06762, (2018)

[9] G. Gislon, L. Bava, M. Zucali, A. Tamburini, and A. Sandrucci, “Unlocking insights: text mining analysis on the health, welfare, and behavior of cows in automated milking systems”, Journal of Animal Science, Vol. 102, skae159, https://doi.org/10.1093/jas/skae159, (2024)

[10] https://dairylane.ca/products/robotic-systems/gea-dairyproq/

[11] F. P. Johansen,“Understanding dairy cow behaviour to improve production and explore welfare implications in automatic milking systems”, PhD Thesis, Queen’s University Belfast, Queen’s University Belfast Research Portal, https://pureadmin.qub.ac.uk/ws/portalfiles/portal/642374092/Understanding_dairy_cow_behaviour_to_improve_production_and_explore_welfare_implications_in_automatic_milking_systems.pdf, (2025)

[12] L. Ozella, K. Brotto Rebuli, C. Forte, and M. Giacobini,“A literature review of modeling approaches applied to data collected in automatic milking systems”, Animals, Vol. 13(12), 1916, https://doi.org/10.3390/ani13121916, (2023)

[13] G. O. Uzedhe, B. O. Akinloye, and I. C. Febaide,“Development of an Animal Farm Robotic Feeding System”, Tropical Journal of Science and Technology, Vol. 4(1), pp. 14–22, https://doi.org/10.47524/tjst.v4i1.15, (2023)

[14] https://www.dairyglobal.net/industry-and-markets/smart-farming/robotic-feeding-save-time-reduce-labour-achieve-higher-sustainability/

[15] Y. Zhang, W. Sun, J. Yang, W. Wu, H. Miao, and S. Zhang, “An Approach for Autonomous Feeding Robot Path Planning in Poultry Smart Farm”, Animals, Vol. 12(22), 3089, https://doi.org/10.3390/ani12223089, (2022)

[16] S. U. Seyfi, E. Bağcali, and K. Soku, “Robotic systems used in dairy cattle farms”, Proceeding Book of the International Symposium for Environmental Science and Engineering Research (ISESER 2019), https://api.iseser.com/doc/2019/book/O7.pdf, (2019)

[17] G. Attard, “Robots in livestock management”, in Encyclopedia of Digital Agricultural Technologies, Springer International Publishing, Cham, pp. 1202–1213, https://link.springer.com/content/pdf/10.1007/978-3-031-24861-0_245.pdf, (2023)

[18] https://en.schauer-agrotronic.com/cattle/manure-removal-and-slurry-technology/enro-manure-robot

[19] B. Bjerg, “Modelling and design of the microclimate in livestock housing”, in Encyclopedia of Smart Agriculture Technologies, Q. Zhang, Ed., Springer, pp. 1–15, https://doi.org/10.1007/978-3-030-89123-7_158-1, (2023)

[20] D. M. G. Preethichandra, L. Piyathilaka, and U. Izhar, “Review on Robotic Systems for Environmental Monitoring”, IEEE Open Journal of Instrumentation and Measurement, https://doi.org/10.1109/OJIM.2024.3493875, (2025)

[21] T. L. Hansen, M. Li, J. Li, C. J. Vankerhove, M. A. Sotirova, J. M. Tricarico, V. E. Cabrera, E. Kebreab, and K. F. Reed, “The Ruminant Farm Systems Animal Module: A Biophysical Description of Animal Management”, Animals, Vol. 11(5), 1373, https://doi.org/10.3390/ani11051373, (2021)

[22] B. Hoogerbrugge, “Measuring the environmental impact of a livestock feeding robot life cycle”, MSc Thesis, Delft University of Technology, https://repository.tudelft.nl, (2023)

[23] S. Celozzi, A. Calcante, M. Zucali, M. Pavesi, S. Colombini, S. Mattiello, and M. Battini, “Assessing the effect of transitioning from tie-stalls to free-stalls with automatic milking system on dairy cattle welfare, milk production, and environmental sustainability”, Livestock Science, Vol. 303, 105878, https://doi.org/10.1016/j.livsci.2025.105878, (2026)

[24] D. Lovarelli, M. Bovo, C. Giannone, E. Santolini, P. Tassinari, and M. Guarino, “Reducing life cycle environmental impacts of milk production through precision livestock farming”, Sustainable Production and Consumption, Vol. 51, pp. 303–314, https://doi.org/10.1016/j.spc.2024.09.021, (2024)

[25] E. Martinsson, H. Hansson, K. Mittenzwei, and H. Storm, “Evaluating environmental effects of adopting automatic milking systems on Norwegian dairy farms”, European Review of Agricultural Economics, Vol. 51(1), pp. 128–156, https://doi.org/10.1093/erae/jbad041, (2024)

SESSION E: THERMAL TECHNIQUE AND ENVIRONMENTAL PROTECTION

Chemical pollutants in the soil of Novi Pazar and ecological risk assessment

Jovana Belošević[1*], Svetlana Belošević[2], Maja Đukić[3]

[1] University of Belgrade, Faculty of Pharmacy, Belgrade, Serbia

[2] University of Pristina in Kosovska Mitrovica, Faculty of Technical Sciences, Kosovska Mitrovica, Serbia

[3] State University of Novi Pazar, Department of Natural Science and Mathematics, Novi Pazar, Serbia

ARTICLE INFO

[*] **Correspondence:** jovanabelosevic2@gmail.com

DOI: https://doi.org/10.46793/ET26.E07B **PAGES:** E55 – E60

ABSTRACT

Soil pollution by chemical pollutants poses a critical threat to environmental quality and human health. This study investigates the concentration, composition and spatial distribution of selected chemical pollutants - including heavy metals (Pb, Cd, Ni, Hg, Zn) and organic matter in urban soil of the city of Novi Pazar. Its analysis is based on data collected by the Environmental Protection Agency of the Republic of Serbia (SEPA), during year of 2023. Sampling was carried out at 7 representative locations (city park, recreation center, kindergartens, water supply source, industrial zone) at a depth of 50 cm. The results obtained were compared with the normative limit values for Serbia, and it was determined that the concentrations of heavy metals in some locations exceeded the limit values. This especially applies to Pb and Ni. The research confirms that the soil in Novi Pazar is under moderate anthropogenic pressure. Risk assessment based on international soil quality guidelines indicates potential ecological and health hazards. These findings highlight the need for continuous soil monitoring and stricter regulatory controls to mitigate pollutant accumulation and safeguard soil ecosystem.

KEYWORDS

Heavy metals; Organic contaminants; Anthropogenic sources, Soil pollution, Environmental risk.

1. INTRODUCTION

Soil is an essential component of terrestrial ecosystems and plays a crucial role in maintaining environmental quality, biodiversity, and human health. Urban soils are increasingly exposed to contamination due to industrialization, urbanization, traffic emissions, waste disposal, and intensive anthropogenic activities. Heavy metals are among the most significant pollutants because of their persistence, toxicity, bioaccumulation potential, and non-biodegradable nature [1].

Urban soils are particularly vulnerable to contamination due to intensive anthropogenic activities such as vehicular traffic, industrial emissions, combustion of fossil fuels, construction activities, municipal waste disposal, and atmospheric deposition. Pollutants deposited in urban soils may enter the food chain, affect groundwater quality, and pose potential risks to human health and ecosystems [2].

Novi Pazar is an urban center located in southwestern Serbia characterized by dense traffic, rapid urban development, industrial activities, and increasing population pressure. These factors may contribute to the accumulation of pollutants in urban soils, particularly in recreational zones, residential areas, and locations near industrial facilities [3].

The assessment of soil contamination and ecological risk is important for environmental management and sustainable urban planning. Ecological risk assessment enables the evaluation of potential adverse effects of pollutants on ecosystems and human health. In recent years, numerous studies have focused on the evaluation of heavy metal contamination in urban soils using pollution indices and comparative analysis with regulatory limits [4].

The main objectives of this study were:

1. Determination of the concentrations of selected heavy metals and chemical pollutants in soils from Novi Pazar;
2. Assessment of the spatial distribution of heavy metals and chemical pollutants in different urban locations;

3. Assessment of the ecological risk associated with detected pollutant concentrations.
4. Identification of the potential anthropogenic sources of contamination.
5. Provision of baseline data for future environmental monitoring and soil protection strategies.

2. MATERIALS AND METHODS

2.1. Study Area

Novi Pazar is situated in southwestern Serbia at approximately 43°N latitude and 20°E longitude and represents an important administrative, industrial, and commercial center. The city is characterized by moderate continental climate conditions, intensive urbanization, and increasing traffic density. Sampling locations included recreational zones, educational facilities, water supply protection areas, and public green spaces [5].

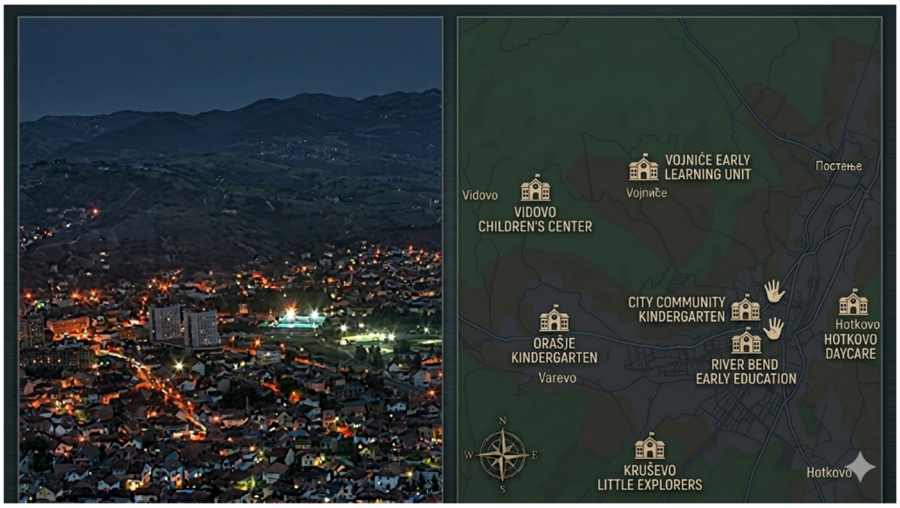


Figure 1: Novi Pazar city

2.2. Soil Sampling

Soil samples were collected during 2023 from seven representative locations within the urban area of Novi Pazar. The sites investigated included:

- City Park
- Recreational Center
- Kindergarten “Mladost”
- Pazarište – vicinity of water source
- Kindergarten “Naše dete”
- Kindergarten “Dečja radost”
- Industrial zone.

Samples were collected from the surface soil layer (sampling depth was 50 cm), according to standard environmental monitoring procedures. Geographic coordinates and physicochemical soil characteristics were recorded for each sampling point. Sampling points were selected to represent different anthropogenic influences including traffic density, recreational activities, and urban infrastructure [6].

2.2. Laboratory Analysis

The analyses were performed by the Belgrade Institute of Public Health according to standard environmental analysis procedures. The following soil parameters were determined:

- Organic matter content (%);
- Clay content (%);
- Soil moisture (%);
- Active and potential pH values;
- Sulfates, nitrates, chlorides, and fluorides;
- Heavy metals: Cu, Zn, Pb, Cd, Ni, Cr, As, and Hg.

Heavy metal concentrations were expressed in mg kg^{-1} dry soil. [6]

2.4. Ecological Risk Assessment

Ecological risk assessment was performed using descriptive statistical analysis and comparison with commonly accepted soil quality guidelines. Mean and maximum concentrations were used to evaluate contamination intensity and potential ecological effects.

The ecological significance of heavy metals was evaluated considering their toxicity, persistence, and mobility in soil ecosystems. Particular attention was focused on Pb, Cd, Ni and Hg because of their high toxicological importance.

3. RESULTS AND DISCUSSION

3.1. Physicochemical Characteristics of Soil

The investigated soils were generally alkaline, with active pH values ranging from approximately 8.0 to 8.3. Alkaline soil conditions generally decrease the mobility and bioavailability of heavy metals because higher pH values promote their adsorption onto clay minerals, organic matter, and metal oxides present in the soil matrix. Under alkaline conditions, many metal ions form stable hydroxides, carbonates, or other poorly soluble compounds that precipitate from the soil solution. As a result, the concentration of free metal ions available for plant uptake, leaching, or migration into groundwater is reduced. Consequently, heavy metals tend to remain more strongly bound within the soil structure, limiting their environmental mobility and reducing their potential toxicity to plants, microorganisms, and humans [7].

Organic matter content varied significantly among sampling locations, indicating differences in vegetation cover, land use, and anthropogenic influence. Clay content ranged from approximately 15% to 32%, suggesting moderate adsorption capacity for metallic pollutants.

3.2. Heavy Metal Concentrations in Soil

The concentrations of analyzed heavy metals at different sampling locations are presented in Table 1.

Table 1: Concentrations of heavy metals in soils of Novi Pazar

Location	Elements mg kg $^{-1}$							
	Cu	Zn	Pb	Cd	Ni	Cr	As	Hg
City Park	26.3	62.2	46.1	0.60	36.9	32.2	7.45	0.28
Recreational Center	27.9	79.8	86.5	0.79	73.5	44.1	7.06	0.20
Kindergarten “Mladost”	19.2	53.8	97.2	0.65	52.6	35.7	3.94	0.20
Pazarište-watersource vicinity	15.9	40.8	10.4	0.60	20.0	11.1	6.53	0.20
Kindergarten “Naše dete“	20.8	55.7	34.1	0.60	33.7	33.1	6.86	0.20
Kindergarten “Dečja Radost”	20.2	58.5	22.3	0.64	38.6	24.7	4.94	0.20
Industrial zone	19.1	54.3	30.9	0.60	34.9	20.0	5.86	0.20

The obtained results indicate spatial variability in heavy metal concentrations among the investigated sites. Zinc showed the highest average concentration (57.87 mg kg^{-1}), followed by lead (46.79 mg kg^{-1}) and nickel (41.46 mg kg^{-1}). Elevated concentrations of Zn are commonly associated with vehicular traffic, tire wear, galvanized materials, and urban runoff [8].

Lead concentrations showed substantial variability among locations. The highest Pb concentration was detected at the “Mladost” kindergarten location (97.2 mg kg^{-1}), suggesting strong anthropogenic influence. Elevated lead concentrations in urban soils are commonly associated with traffic emissions, combustion processes, and industrial activities. Although the use of leaded gasoline has been reduced, historical deposition may still contribute to elevated Pb levels in urban environments [9].

Nickel concentrations ranged from 20.0 to 73.5 mg kg^{-1}, with the highest value recorded at the Recreational Center. Elevated Ni concentrations may originate from the geological background as well as from anthropogenic activities (industrial emissions, fossil fuel combustion, etc.).

Cadmium and mercury concentrations were relatively low at all sampled locations, which indicates a limited industrial impact in the investigated area. Cadmium concentrations ranged between 0.60 and 0.79 mg kg $^{-1}$. In the City Park, a maximum mercury concentration of 0.28 mg kg $^{-1}$ was recorded.

Chromium and arsenic concentrations were also below commonly accepted intervention values for urban soils [10].

The lowest concentrations of most analyzed metals were observed at the Pazarište location near the water source, indicating lower anthropogenic pressure compared with densely urbanized zones.

3.3. Statistical Overview of Heavy Metal Concentrations

Table 2 shows the mean, maximum and limit values of heavy metal concentrations in the soil of Novi Pazar.

Table 2: Mean, maximum and limit value concentrations of heavy metals

Concentrations of heavy metals mg kg $^{-1}$	Mean	Minimum	Maximum	Limit value
Cu	21.34	15.90	27.90	100
Zn	57.87	40.80	79.80	300
Pb	53.49	10.40	97.20	100
Cd	0.67	0.60	0.79	3.0
Ni	46.13	20.00	73.50	50
Cr	31.03	11.10	44.10	100
As	6.09	3.94	7.45	25
Hg	0.21	0.20	0.28	1.0

The statistical overview confirms that Zn, Pb, and Ni were the dominant pollutants in the investigated soils [11,12]. Variability in pollutant concentrations among sampling locations reflects differences in local anthropogenic activities, traffic intensity, and land use.

Recreational zones and areas exposed to traffic generally exhibited higher concentrations of Pb, Zn, and Ni. In contrast, the water supply protection zone showed lower concentrations of most analyzed metals [13].

The elevated concentrations observed in urban recreational zones may be attributed to:

- Increased traffic density;
- Atmospheric deposition of particulate matter;
- Urban construction activities;
- Accumulation of contaminants in surface soils.

3.4. Ecological Risk Assessment

Ecological risk assessment showed that most analyzed metals were present at low to moderate contamination levels. However, elevated Pb and Ni concentrations at specific urban locations may represent a potential ecological concern.

Lead is particularly hazardous because of its neurotoxic effects and long-term persistence in soil. The presence of elevated Pb concentrations in areas frequented by children, such as kindergartens and recreational zones, requires special attention because children are more vulnerable to heavy metal exposure [4,14].

Nickel and chromium may affect soil microbial communities and reduce soil fertility when present in elevated concentrations. Although Cd and Hg concentrations were comparatively low, their toxicological significance should not be neglected due to their bioaccumulative behavior.

The results suggest that anthropogenic activities are the dominant source of contamination in urban soils of Novi Pazar. Traffic emissions, industrial processes, domestic heating, atmospheric deposition, and improper waste management are likely contributing factors.

The alkaline pH conditions of the investigated soils likely reduced the mobility and bioavailability of metals, thereby limiting their ecological impact. Nevertheless, continuous accumulation of pollutants may increase environmental risks in the future, especially in densely urbanized areas [15].

3.5 Environmental Implications

Urban soil plays an important role in maintaining ecological balance and environmental quality. Contaminated soils may act as secondary sources of pollution through dust generation, leaching, and uptake by plants. The obtained results emphasize the need for:

- Continuous environmental monitoring;
- Improved urban waste management;
- Reduction of traffic-related emissions;
- Development of green urban infrastructure;
- Public awareness regarding environmental protection.

Implementation of preventive measures is especially important in sensitive locations such as parks, schools, and kindergartens.

3.6 Comparison with Other Urban Areas

The concentrations measured in Novi Pazar were generally lower than values reported for heavily industrialized urban centers in Southeast Europe. The absence of major heavy industry in the region may explain the relatively moderate contamination levels.

However, the detected enrichment of Pb, Zn, and Ni confirms the influence of urbanization and traffic-related activities on soil quality. Similar trends have been reported in other Balkan cities where traffic emissions and atmospheric deposition represent dominant sources of urban soil contamination.

4. CONCLUSION

The present study evaluated the concentrations of selected chemical pollutants in urban soils of Novi Pazar and assessed their ecological risk. The results demonstrated spatial variability in heavy metal concentrations among sampling sites.

The investigated soils were predominantly alkaline, which may reduce heavy metal mobility and bioavailability.

Zinc, lead, and nickel were identified as the dominant pollutants. The highest Pb concentration was detected at the "Mladost" kindergarten site, while the highest Ni concentration was recorded at the Recreational Center. The lowest contamination levels were observed near the Pazarište water source.

Lead concentrations showed localized enrichment in urban and recreational areas influenced by anthropogenic activities.

Cadmium and mercury concentrations were relatively low at all sampling locations.

Ecological risk assessment indicated generally low to moderate ecological risk; however, elevated concentrations of Pb and Ni at several locations suggest the influence of anthropogenic activities and the necessity for regular environmental monitoring.

Although current contamination levels do not indicate severe pollution, continuous monitoring is recommended because urbanization and traffic intensity may contribute to future accumulation of pollutants.

The findings of this study provide important baseline information for environmental protection and sustainable urban planning in Novi Pazar.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education and Science of the Republic of Serbia, grants III41010. We also thank the Environmental Protection Agency of the Republic of Serbia, which enabled us to use the data.

REFERENCES

[1] B. J. Alloway, "Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability", Springer; Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability | Springer Nature Link, (2013)

[2] B. Wei, L. Yang, "A review of heavy metal contaminations in urban soils, urban road dusts and agricultural soils from China", Microchemical Journal, Vol. 94 (2), pp. 99-107, (PDF) A Review of Heavy Metal Contaminations in Urban Soils, Urban Road Dusts and Agricultural Soils From China, (2010)

[3] A. Kabata-Pendias, "Trace Elements in Soils and Plants", Taylor &Francis Group, https://doi.org/10.1201/b10158, (2010)

[4] L. Ferreira-Baptista, E. De Miguel, "Geochemistry and risk assessment of street dust in Luanda, Angola: A tropical urban environment", Atmosheric Environment, Vol. 39 (25), pp 4501-4512, https://doi.org/10.1016/j.atmosenv.2005.03.026, (2005)

[5] Grad Novi Pazar Novi Pazar – Wikipedija / Википедија

[6] Agencija za zastitu zivotne sredine, Агенција за заштиту животне средине – Министарство заштите животне средине Републике Србије

[7] B. Murray, McBride, "Environmental Chemistry of Soils", Oxford University Press, New York, 406, (1994)

[8] X. Lu, L. Wang, K. Lei, J. Huang, Y. Zhai, "Contamination assessment of copper, lead, zinc, manganese and nickel in street dust of Baoji, NW China", Journal of Hazardous Materials Vol. 161(2-3), pp. 1058-62, https://doi.org/10.1016/j.jhazmat.2008.04.052, (2008)

[9] X. Li, C. Poon, P.S. Liu. "Heavy metal contamination of urban soils and street dusts in Hong Kong", Applied Geochemistry, Vol. 16 (11-12), pp.1361–1368, https://doi.org/10.1016/S0883-2927(01)00045-2, (2001)

[10] C. Zhang, "Using multivariate analyses and GIS to identify pollutants and their spatial patterns in urban soils in Galway, Ireland", Environmental Pollution, Vol. 142 (3), pp. 501–511, https://doi.org/10.1016/j.envpol.2005.10.028, (2006)

[11] Regulation on Allowed Quantities of Hazardous and Harmful Substances in Soil and Irrigation Water and Methods of Their Testing, Official Gazette of Republic of Serbia, no. 23/94 (https://pravno-informacioni-sistem.rs/eli/rep/sgrs/ministarstva/pravilnik/1994/23/1/reg

[12] Regulation on Limit Values of Pollutants, Harmful and Dangerous Substances in Soil, Official Gazette of Republic of Serbia, no. 30/2018 and 64/2019 (https://pravno-informacioni-sistem.rs/eli/rep/sgrs/vlada/uredba/2018/30/2/reg)

[13] P. Tume, E. González, R. W. King, L. Cuitino, N. Roca, J. Bech, "Distinguishing between natural and anthropogenic sources for potentially toxic elements in urban soils of Talcahuano, Chile", Journal of Soils and Sediments , Vol. 18, pp. 2335-2349, https://doi.org/10.1007/s11368-017-1750-0, (2018)

[14] D. C. Adriano, "Trace Elements in Terrestrial Environments: Biogeochemistry, Bioavailability, and Risks of Metals. Springer New York, 2nd Edition, http://dx.doi.org/10.1007/978-0-387-21510-5 , (2001)

[15] A. Violante, V. Cozzolino, L. Perelomov, G. A. Caporale, M. Pigna, "Mobility and bioavailability of heavy metals and metalloids in soil environments, Journal of Soil Science and Plant Nutrition, Vol. 10 (3), pp. 268-292, https://doi.org/10.4067/S0718-95162010000100005, (2010)

SESSION E: THERMAL TECHNIQUE AND ENVIRONMENTAL PROTECTION

Monitoring of heavy metals and their movement through the environment of the city of Kragujevac

Jovana Belošević[1*], Svetlana Belošević[2], Maja Đukić[3]

[1] University of Belgrade, Faculty of Pharmacy, Belgrade, Serbia

[2] University of Pristina in Kosovska Mitrovica, Faculty of Technical Sciences, Kosovska Mitrovica, Serbia

[3] State University of Novi Pazar, Department of Natural Science and Mathematics, Novi Pazar, Serbia

ARTICLE INFO

[*] **Correspondence:** jovanabelosevic2@gmail.com

DOI: https://doi.org/10.46793/ET26.E08B **PAGES:** E61 – E65

ABSTRACT

This paper analyzes the presence, distribution and trends in concentrations of heavy metals (lead, cadmium, nickel, zinc, arsenic and chromium) in different environmental media - air, surface water and soil - in the territory of the city of Kragujevac over a three-year period (2022, 2023 and 2024). Special focus is placed on identifying the main sources of pollution, such as industrial activities, traffic and individual combustion plants. Data were collected through regular annual and monthly monitoring reports conducted by city departments and the Serbian Environmental Protection Agency (SEPA). Metal concentrations in suspended particulate matter (PM10) and sediments were analyzed at key measurement sites (Mala Vaga, Medical School). Although most parameters are within the legally permitted limits, continuous monitoring is necessary, especially in urban areas with high traffic volumes. The results for 2024 indicate the need for further integration of local environmental action plans to reduce emissions of carcinogenic metals and protect public health.

KEYWORDS

Heavy metals, Air pollution, Soil, Monitoring, Environment.

1. INTRODUCTION

Heavy metals represent one of the most significant groups of environmental pollutants due to their toxicity, persistence, and ability to accumulate in living organisms [1]. Unlike organic pollutants, heavy metals are non-biodegradable and may remain in environmental compartments for long periods, causing adverse ecological and health effects [2].

Industrialization, urbanization, increased traffic intensity, and fossil fuel combustion have considerably contributed to the release of heavy metals into the atmosphere, water bodies, and soils [3]. Urban environments are particularly vulnerable to heavy metal contamination because multiple anthropogenic sources operate simultaneously. Traffic emissions, industrial production, waste disposal, domestic heating, and atmospheric deposition contribute to elevated concentrations of toxic metals such as lead, cadmium, chromium, nickel, zinc, and arsenic [4].

The city of Kragujevac is one of the major industrial and urban centers in Serbia, characterized by intensive traffic, industrial activity, and population growth. Historical industrial development, particularly in the automotive and metal-processing sectors, has contributed to environmental pressures and potential contamination by hazardous substances. Continuous environmental monitoring is therefore essential for evaluating pollution levels, identifying pollution sources, and assessing ecological and public health risks [5].

The aim of this study was to evaluate the occurrence and distribution of selected heavy metals in air, surface water, and soil within the territory of Kragujevac during the period 2022–2024. The study also aimed to identify the dominant sources of contamination and assess temporal trends in heavy metal concentrations.

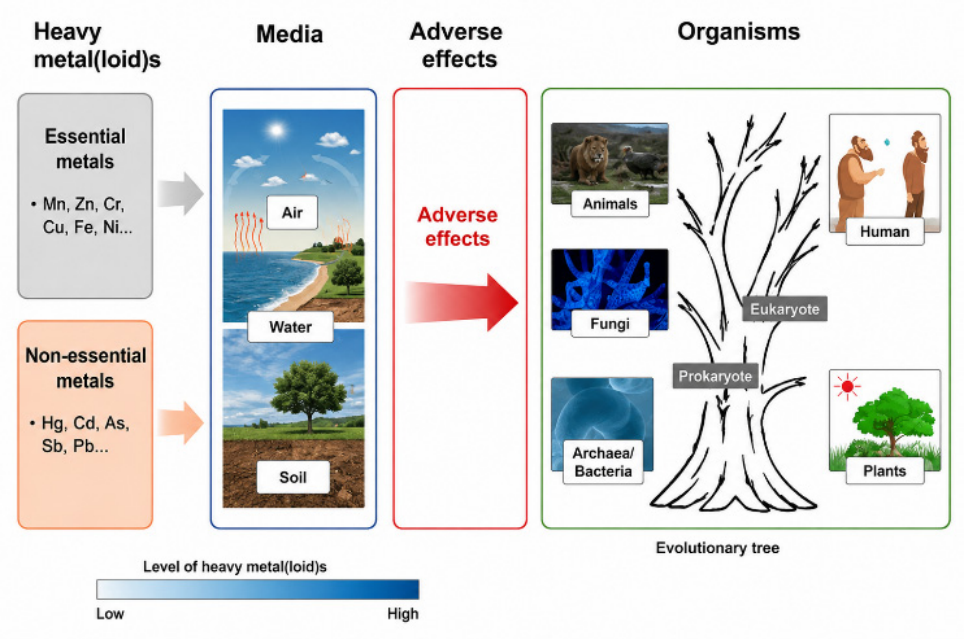


Figure 1: The cycle of heavy metals in nature

2. MATERIAL AND METHODS

2.1. Study Area

Kragujevac is located in central Serbia and represents an important administrative, industrial, and transportation center. The city is characterized by moderate continental climate conditions and a developed urban-industrial infrastructure. It is in the geographical center of the country, 145 km south of Belgrade. The city is located at an altitude of 173-220 m, with a mathematical-geographical position of + 44º 22′; +20º 56′. According to the 2022 census, there were 146.315 inhabitants, making it the fourth most populous settlement in Serbia [6].



Figure 2 a) Geographical position of Kragujevac in Serbia. b) Aerial view of the urban area of Kragujevac

2.2. Sampling and Monitoring

Environmental monitoring data were collected from official annual and monthly reports prepared by municipal environmental departments and the Serbian Environmental Protection Agency for the years 2022, 2023, and 2024 [7]. The analyzed environmental media included:

- Ambient air (PM10 suspended particles)
- Surface water and sediments
- Urban soil

Air quality monitoring was conducted at key urban locations including Mala Vaga, and the Medical School area, representing zones with intensive traffic and urban activity. In addition to the planned residential area with a large number of residential buildings, this urban area also comprises single-family housing estates equipped with individual heating units. The following heavy metals were analyzed:

- Lead (Pb)
- Cadmium (Cd)

- Nickel (Ni)
- Zinc (Zn)
- Arsenic (As)
- Chromium (Cr)

2.3 Analytical Methods

Heavy metal concentrations were determined using standard analytical procedures recommended for environmental monitoring [8]. PM10 samples were collected using automatic samplers, while soil and sediment samples were collected according to standardized environmental protocols. The collected samples were subjected to laboratory analysis using atomic absorption spectrometry (AAS) and inductively coupled plasma optical emission spectrometry (ICP-OES), depending on the monitored parameter. Measured concentrations were compared with national environmental quality standards and permissible limits prescribed by Serbian environmental legislation (Official Gazette of the Republic of Serbia, 2018) [9].

3. RESULTS AND DISCUSSION

3.1 Heavy Metals in Ambient Air

The results presented in Table 1 indicate temporal variations in the concentrations of Cd, Pb, Ni and As associated with PM10 particles during the monitoring period from 2022 to 2024 in the city of Kragujevac.

Table 1: Measurement results for Cd, Pb, Ni and As measurements by month for 2022

Month and year / maximum measured value				
Target value for the average value of the total content of suspended particles PM10	5 ng/m^3	-	20 ng/m^3	6 ng/m^3
	Cd (ng/m^3)	Pb (μg/m^3)	Ni (ng/m^3)	As (ng/m^3)
		2022.		
January	< 0.5	< 0.003	< 0.9	< 0.9
February	< 0.5	0.014 ± 0.003	15.0 ± 3.4	< 0.9
March	< 0.5	0.023 ± 0.005	< 0.9	< 0.9
April	< 0.5	<0.003	< 0.9	< 0.9
May	< 0.5	0.0187 ± 0.038	5.8 ± 1.3	< 0.9
June	< 0.5	0.0120 ± 0.025	4.3 ± 1.0	< 0.9
July	< 0.5	0.084 ± 0.017	16.5 ± 3.8	< 0.9
August	< 0.5	<0.003	< 0.9	< 0.9
September	< 0.5	0.050 ± 0.010	< 0.9	< 0.9
October	< 0.5	0.026 ± 0.005	1.4 ± 0.3	< 0.9
November	< 0.5	0.046 ± 0.009	< 0.9	< 0.9
December	< 0.5	<0.003	2.0 ± 0.5	< 0.9

Table 2: Measurement results for Cd, Pb, Ni and As measurements by month for 2023

Month and year / maximum measured value				
Target value for the average value of the total content of suspended particles PM10	5 ng/m^3	-	20 ng/m^3	6 ng/m^3
	Cd (ng/m^3)	Pb (μg/m^3)	Ni (ng/m^3)	As (ng/m^3)
		2023.		
January	< 0.5	0.025 ± 0.005	2.0 ± 0.5	< 0.9
February	< 0.5	0.018 ± 0.004	< 0.9	4.4
March	< 0.5	0.038 ± 0.008	10.6 ± 0.9	< 0.9
April	< 0.5	0.038 ± 0.008	10.6 ± 0.9	< 0.9
May	< 0.5	0.043 ± 0.009	< 0.9	< 0.9
June	3.8 ± 0.7	0.080 ± 0.016	9.5 ± 2.2	< 0.9
July	1.0 ± 0.2	0.025 ± 0.005	2.6 ± 0.6	< 0.9
August	1.0 ± 0.2	0.004 ± 0.001	< 0.9	< 0.9
September	3.3 ± 0.6	0.449 ± 0.092	< 0.9	< 0.9
October	< 0.5	<0.003	< 0.9	< 0.9
November	0.9 ± 0.2	0.026 ± 0.005	9.1 ± 2.1	< 0.9
December	< 0.5	<0.003	< 0.9	< 0.9

Table 3: Measurement results for Cd, Pb, Ni and As measurements by month for 2024

Month and year / maximum measured value				
Target value for the average value of the total content of suspended particles PM10	5 ng/m^3	-	20 ng/m^3	6 ng/m^3
	Cd (ng/m^3)	Pb (µg/m^3)	Ni (ng/m^3)	As (ng/m^3)
		2024.		
January	1.5 ± 0.3	0.054 ± 0.011	2.7 ± 0.6	3.2
February	1.7 ± 0.3	0.047 ± 0.010	3.7 ± 0.9	< 0.9
March	< 0.5	0.029 ± 0.006	2.8 ± 0.6	< 0.9
April	< 0.5	<0.003	< 0.9	< 0.9
May	< 0.5	0.007 ± 0.002	3.5 ± 0.8	< 0.9
June	0.8 ± 0.1	0.068 ± 0.014	3.3 ± 0.4	< 0.9
July	3.3 ± 0.6	0.084 ± 0.017	3.0 ± 0.7	< 0.9
August	8.7 ± 1.6	0.233 ± 0.048	51.5 ± 11.7	< 0.9
September	4.5 ± 0.8	<0.003	5.5 ± 1.3	< 0.9
October	< 0.5	0.0014 ± 0.003	< 0.9	< 0.9
November	0.9 ± 0.2	0.0039 ± 0.008	11.2 ± 2.5	< 0.9
December	< 0.5	0.012 ± 0.003	< 0.9	< 0.9

Monitoring results showed the presence of heavy metals associated with PM10 particles at all investigated locations. The highest concentrations were generally observed during the winter season due to increased emissions from domestic heating and unfavorable meteorological conditions [5].

Nickel and chromium concentrations were occasionally elevated in zones influenced by traffic and industrial emissions. Seasonal variability indicates a strong influence of anthropogenic activities in combination with meteorological conditions that favor the accumulation of pollutants. Zinc was also frequently detected, mainly originating from tire wear, brake abrasion, and industrial activities [3]. Lead concentrations remained relatively low compared to historical values due to the elimination of leaded gasoline; however, its presence in airborne particles still indicates ongoing anthropogenic influence. Cadmium and arsenic concentrations were generally below prescribed limit values but require continuous observation because of their carcinogenic potential and cumulative toxic effects [10]. The obtained results indicate that traffic emissions and fossil fuel combustion remain the dominant sources of airborne heavy metals in urban areas of Kragujevac.

3.2. Heavy Metals in Soil

Soil analysis demonstrated spatial variability in heavy metals concentrations depending on land use and proximity to pollution sources [7]. Urban soils located near roads and industrial facilities showed higher concentrations of nickel, zinc, chromium, and lead compared to suburban areas [3].

Nickel and chromium exhibited the highest concentrations among the analyzed metals, which may be associated with industrial activities, atmospheric deposition, and traffic-related emissions. Zinc concentrations were also elevated in areas with intense vehicular activity [4].

Although most concentrations remained within legally permitted values, the continuous accumulation of heavy metals in urban soils may represent a long-term ecological concern because soils act as a sink for persistent pollutants [1].

3.3. Heavy Metals in Surface Water and Sediments

Heavy metals were detected in both surface water and sediment samples. Sediments generally showed higher metal concentrations due to the accumulation of contaminants over time [2].

The presence of nickel, chromium, and zinc in sediments indicates the influence of urban runoff, industrial discharge, and atmospheric deposition. Seasonal fluctuations were observed, particularly after periods of intensive precipitation when runoff increases the transport of pollutants into aquatic ecosystems.

Although the measured concentrations did not indicate severe contamination, sediments may act as secondary pollution sources through remobilization processes under changing environmental conditions [11].

3.4. Temporal Trends (2022–2024)

Comparative analysis of monitoring data revealed relatively stable trends in most heavy metal concentrations during the investigated period. Slight increases in PM10-associated metals were observed during winter months and in highly urbanized zones.

The results for 2024 suggest moderate environmental pressure associated with urban traffic and energy consumption. Continued urbanization and industrial activity may contribute to future increases in pollutant levels if adequate mitigation measures are not implemented [5].

4. ECOLOGICAL AND HEALTH IMPLICATIONS

Heavy metals may cause significant adverse effects on ecosystems and human health even at relatively low concentrations [10]. Chronic exposure to lead and cadmium may affect the nervous system, kidneys, and cardiovascular system, while nickel, chromium, and arsenic are recognized for their carcinogenic potential [5].

Airborne particulate matter containing toxic metals represents a particular risk in densely populated urban environments because inhalation exposure directly affects the respiratory system. Soil contamination may additionally contribute to pollutant transfer into plants and food chains [1].

The obtained results indicate that current pollution levels in Kragujevac do not represent critical ecological contamination; however, preventive measures and continuous environmental monitoring remain essential for sustainable urban environmental management.

5. CONCLUSIONS

This study evaluated the occurrence and distribution of heavy metals in air, soil, surface water, and sediments in the city of Kragujevac during the period 2022–2024. The main conclusions are as follows:

1. Heavy metals including Pb, Cd, Ni, Zn, As, and Cr were detected in all investigated environmental media.
2. Most measured concentrations were within legally permitted limits.
3. Elevated concentrations were periodically observed in urban zones affected by traffic and industrial activities.
4. Traffic emissions, industrial processes, domestic heating, and atmospheric deposition were identified as the dominant pollution sources.
5. Sediments and soils act as long-term reservoirs for heavy metals and may contribute to secondary contamination.
6. Continuous environmental monitoring is necessary due to the cumulative and toxic effects of heavy metals on ecosystems and human health.
7. The results highlight the importance of strengthening local environmental protection strategies and emission reduction measures.

Future studies should include expanded spatial monitoring, bioavailability assessments, and advanced ecological risk analyses to improve understanding of heavy metal behavior in urban environments.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education and Science of the Republic of Serbia, grants III41010. We also thank the Environmental Protection Agency of the Republic of Serbia, which enabled us to use the data.

REFERENCES

[1] B. J. Alloway, “Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability”, Springer; Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability | Springer Nature Link, (2013)

[2] A. Kabata-Pendias, “Trace Elements in Soils and Plants”, Taylor &Francis Group, https://doi.org/10.1201/b10158, (2010)

[3] B . Wei, L. Yang, “A review of heavy metal contaminations in urban soils, urban road dusts and agricultural soils from China”, Microchemical Journal, Vol. 94 (2), pp. 99-107, (PDF) A Review of Heavy Metal Contaminations in Urban Soils, Urban Road Dusts and Agricultural Soils From China, (2010)

[4] H. Ali, E.Khan, I. Ilahi, “Environmental chemistry and ecotoxicology of hazardous heavy metals: Environmental persistence, toxicity, and bioaccumulation”, Journal of Chemistry, 1–14, https://doi.org/10.1155/2019/6730305, (2019)

[5] WHO (2021). Air Pollution and Human Health. World Health Organization, Air pollution, (2021) Alloway, B. J. (2013). Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability. Springer.

[6] Grad Kragujevac, Крагујевац — Википедија

[7] Agencija za zastitu zivotne sredine, Агенција за заштиту животне средине – Министарство заштите животне средине Републике Србије

[8] APHA (2017). Standard Methods for the Examination of Water and Wastewater (23rd ed). Washington DC: American Public Health Association, Apha 2017 | PDF, (2017)

[9] Regulation on Limit Values of Pollutants, Harmful and Dangerous Substances in Soil, Official Gazette of Republic of Serbia, no. 30/2018 and 64/2019 (https://pravno-informacioni-sistem.rs/eli/rep/sgrs/vlada/uredba/2018/30/2/reg)

[10] P. B. Tchounwou, Yedjou, C. G. Yedjou, A. K. Patlolla, D. J. Sutton, “Heavy metal toxicity and the environment”, Molecular, Clinical and Environmental Toxicology, Vol.101, pp. 133–164, https://doi.org/10.1007/978-3-7643-8340-4_6, (2012)

[11] UNEP, Global Environment Outlook, United Nations Environment Programme, Global Environment Outlook

SESSION F

APPLIED MECHANICS

SESSION F: APPLIED MECHANICS

Volume optimization of the worm gear using SSO algorithm

Branislav Milenković[1*], Đorđe Jovanović[2], Mladen Krstić[3]

[1] Aviation Academy, Belgrade, Serbia

[2] Mathematical Institute of SASA, Belgrade, Serbia

[3] University of Kragujevac, Faculty of Mechanical and Civil Engineering, Kraljevo, Serbia

ARTICLE INFO

*** Correspondence:** milenkovic.b@vakademija.edu.rs

DOI: https://doi.org/10.46793/ET26.F01M **PAGES:** F1 – F5

ABSTRACT

Machine design is a field of research where the process of designing a mechanical element is optimized, having certain goals and limitations in mind. In this paper, the Sperm Swarm Optimization algorithm (SSO for short) was used to optimize volume of worm gear with constant number of variables. By applying the SSO algorithm, we obtained optimal or near-optimal results. By further development of this algorithm, this method can be modified and improved in order to obtain better results.

KEYWORDS

Optimization, Worm Gear, Sperm Swarm Optimization.

1. INTRODUCTION

Optimization techniques play a significant role in modern mechanical engineering design, where the primary objective is often to improve performance while reducing material consumption, manufacturing cost, and overall system dimensions. In power transmission systems, worm gears are widely used because of their compact structure, high transmission ratio, smooth operation, and ability to provide significant speed reduction in a single stage. However, despite their advantages, worm gears are frequently associated with challenges related to efficiency, heat generation, wear, and geometric constraints. These challenges can be formulated as an optimization function, therefore making them a candidate for solving them using metaheuristics.

A worm gear is a gear consisting of a shaft with a spiral thread that engages with and drives a toothed wheel. Worm gears are an old style of gear and a version of one of the six simple machines. Basically, a worm gear is a screw butted up against what looks like a standard spur gear with slightly angled and curved teeth[8].

Finding the solution to worm gear problems in machine design is very complex and a time consuming process. Many design variables, constraints, different graphs and tables, empirical formulas are involved. As the literature reveals, not much research has been undertaken on worm gear drive optimization. Besides this certain challenging industry relevant constraints, namely, gear strength capacity, wear capacity, thermal capacity, dynamic load and face width are also not considered in the literature.

In recent decades, metaheuristic optimization algorithms inspired by natural phenomena and biological systems have become increasingly popular in solving complex engineering optimization problems. Algorithms such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), Differential Evolution (DE) [1,2,3,4,5], and many others have demonstrated strong capabilities in handling nonlinear, constrained, and multimodal optimization problems. These algorithms are particularly suitable for mechanical design optimization because they can handle constraints and explore large search spaces in an efficient manner.

Classification of optimization algorithms are shown in Figure 1.

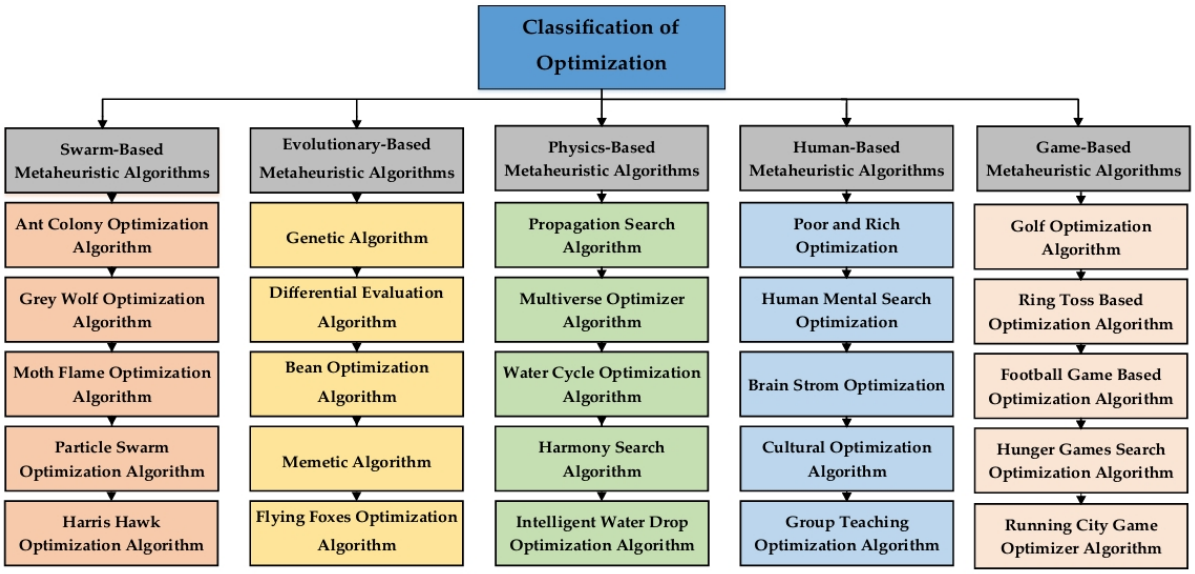


Figure 1: Classification of Optimization

The remainder of this paper is organized as follows. Section 2 presents an overview of the Sperm Swarm Optimization algorithm and its mathematical formulation. Section 3 describes the worm gear design optimization problem, including the objective function, design variables, and imposed constraints. Section 4 discusses the obtained optimization results and evaluates the effectiveness of the proposed approach. Finally, conclusions and potential future research directions are provided.

2. SPERM SWARM OPTIMIZATION

Sperm Swarm Optimization algorithm (SSO) represents a bio-inspired population-based metaheuristic algorithm, whose inspiration stems from the fertilization procedure, where a fertilized egg is surrounded by sperm cells (search agents), which compete to fertilize the egg (Figure 2). Two key biological mechanisms additionally inspire this algorithm: sperm cell's head acting as a temperature sensor, helping it locate the egg, and the female reproductive system having a differing pH value, which may be more or less harmful to the sperm cells. The algorithm goes through an initialization phase, where all search agents are initialized, and update phase, where search agents move through the search space, in order to find the best solution [6,7].

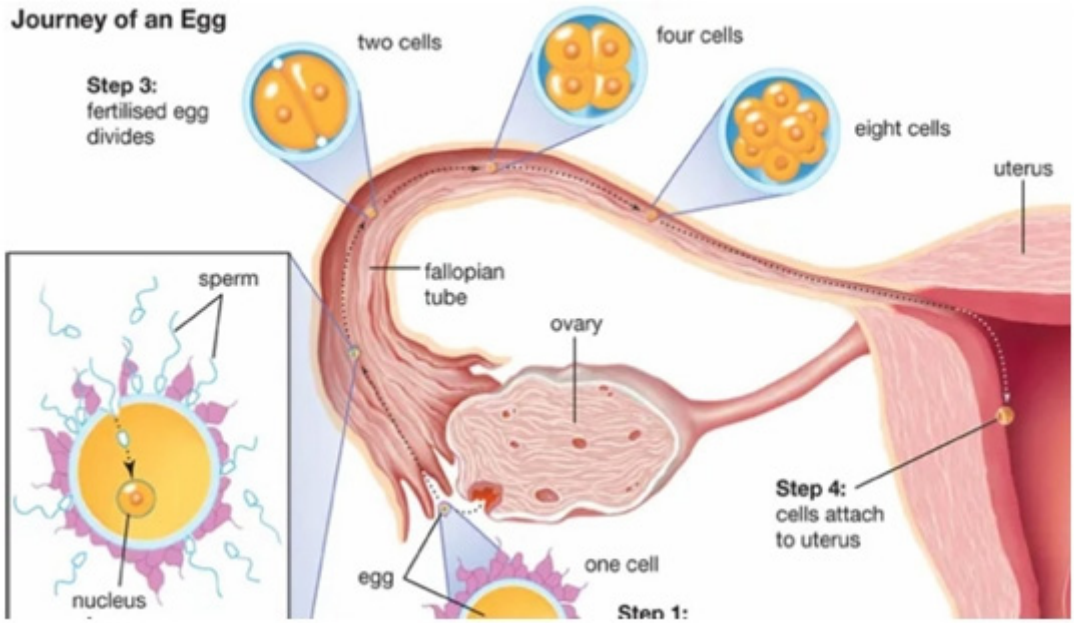


Figure 2: Mechanism of fertilization

The position of the search agents is updated by:

$$x_i(t+1) = x_i(t) + v_i(t) \tag{1}$$

Where $x_i(t+1)$ is the position of the search agent in the next iteration, $x_i(t)$ is the position of the search agent in the current iteration, and $v_i(t)$ is the velocity of the search agent in the current iteration. The search agent's velocity is calculated as follows:

$$v_i(t) = Initial\,Velocity + Current\,Best + Global\,Best \tag{2}$$

The search agent's initial velocity is calculated as follows:

$$Initial\,Velocity = D\log_{10}(PHRand)v_i \tag{3}$$

where D is the damping factor (a random number between 0 and 1), $PHRand$ is a random number between 7 and 14, and v_i is a velocity vector (randomized between 0 and maximum values of the search space).

The current best term of the velocity is calculated as follows:

$$Current\,Best = \log_{10}(PHRand)\log_{10}(TempRand)\left(xbest_i - x_i(t)\right) \tag{4}$$

where $TempRand$ is a random number between 35.1 and 38.5, and $xbest_i$ is the current best solution for the search agent.

The global best term of the velocity is calculated as follows:

$$GlobalBest = \log_{10}(PHRand1)\log_{10}(TempRand2)\left(xgbest_i - x_i(t)\right) \tag{5}$$

where $xgbest_i$ is the global best solution for the current iteration.

3. WORM GEAR DRIVE OPTIMUM DESIGN

The objective is to minimize volume of worm and worm wheel assembly. The volume equation from is:

$$f(x) = 14.310926 \cdot \frac{x_4 \cdot x_3^2}{x_1} + 455.5309x_2x_4 + 0.8382\left(\frac{256x_4^2 + x_1^2x_2^2}{x_1x_4}\right)^2 + 104634.67 \tag{6}$$

The design variables are: Gear ratio (i), Face width (b), Pitch circle diameter of worm (D_1) and Pitch circle diameter of worm wheel (D_2). The function for design variables of the worm drive can be written as $F(x) = F(i,b,D_1,D_2) = F(x_1,x_2,x_3,x_4)$

Upper and lower bounds of design variables are: $40 \le i \le 70, 100 \le b \le 200, 150 \le D_1 \le 280, 1000 \le D_2 \le 1500$.

The face width of a gear is the length of teeth in an axial plane. The face width of the gear should not exceed half the worm outside diameter.

$$g_1(x) = x_1 - 0.5(x_3 + 48) \le 0 \tag{7}$$

Mathematically, the tangent of the helix angle of the worm gear is less than the coefficient of friction between the worm and the gear. This condition is referred to as self locking of worm gears.

$$g_2(x) = \frac{22.5}{x_3} - 0.099 \le 0 \tag{8}$$

For better performance, dynamic load must be greater than strength and wear capacity.

$$g_3(x) = 6430.01 - 754.56x_2 \le 0 \tag{9}$$

$$g_4(x) = 6430.01 - 0.552x_2x_4 \le 0 \tag{10}$$

Amount of power lost in friction is converted into heat. To prevent overheating, heat generated should not exceed the rate of heat transfer.

$$g_5(x) = 120 - 406\left(0.785x_4^2 + 0.075408x_3\right) \le 0 \tag{11}$$

To achieve compact design of gear pairs, the centre distance of optimized gear pair should be less than that of actual gear pair.

$$g_6(x) = \frac{\frac{x_1x_2}{\sqrt{\frac{65x_3}{x_4}+1}} + x_3}{2} - 901 \le 0 \tag{12}$$

The condition for maximum power transmission capacity is constraint g_7 and constraint g_8.

$$g_7(x) = \frac{\left(\frac{x_1 x_2}{\sqrt{\frac{65x_3}{x_4}+1}} + x_3\right)^{0.875}}{3} - x_3 \le 0 \tag{13}$$

$$g_8(x) = x_3 - \left(\frac{\frac{x_1 x_2}{\sqrt{\frac{65x_3}{x_4}+1}} + x_3}{1.7}\right)^{0.875} \le 0 \tag{14}$$

If the worm shaft bends too much, the teeth will not mesh properly. This will result in excessive wear and early failure. To avoid this, the maximum deflection is constraint g_9 given by:

$$g_9(x) = \frac{90.887873\frac{x_4^4}{x_1}}{\left[x_2^2 - 4\frac{x_4^2}{x_1^2}\right]\cdot\left[x_3 - 2.4\frac{x_4}{x_1}\right]^4} - 0.001\cdot x_3 \le 0 \tag{15}$$

Beam strength of worm gear tooth is the maximum tangential load the worm gear tooth can take without tooth breakage. Constraint g_{10} based on Lewis equation is:

$$g_{10}(x) = \frac{66116467.3\cdot x_1}{x_4\cdot x_2\cdot x_3} - 83.33333 \le 0 \tag{16}$$

4. RESULTS AND DISCUSSION

For this algorithm, 30 search agents and 1000 iterations were chosen as input parameters. During the course of the research, it has been noted that increasing search agent and iteration count did not yield better solutions.

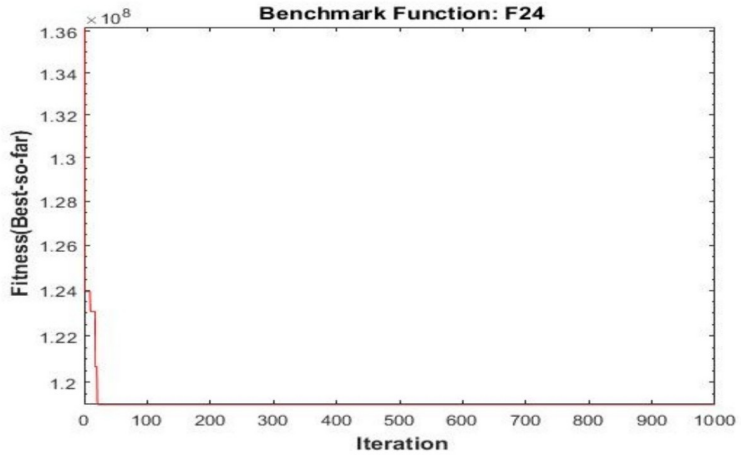


Figure 3: Convergence rate of SSO algorithm respectively for finding the best possible fitness of worm gear

In Figure 3 a convergence diagram for the problem of worm gear optimization is given. As SSO gives best results, it is compared with literature [9] in Table 1.

Table 1: Percentage change in optimal parameter values by SSO of this work (bold values) and literature

Parameters	Literature [9]	**SSO**	Change in value [%]
Gear ratio	46	**45.6**	0.87
Face width $[mm]$	197	**130.2**	33.91
Pitch circle diameter of worm $[mm]$	188	**280**	48.93

Pitch circle diameter of worm wheel $[mm]$	1000	**1000**	0
Volume of worm and worm wheel $\left[mm^3\right]$	$1.45442 \cdot 10^8$	$\mathbf{1.1822 \cdot 10^8}$	18.72

Percentage change in optimal design parameters values is shown in Table 1. It is observed that a decrease in value of 0.87 % in gear ratio, 33.91% in face width, and an increase of 48.93% in pitch circle diameter of worm. Interestingly, there is no change in value in pitch circle diameter of the worm wheel.

5. CONCLUSION

In this paper, optimum design for a worm gear drive considering industry relevant constraints using Sperm Swarm Optimization (SSO) algorithm is obtained. All constraints are well satisfied. The important findings are:

- There is 18.72 % reduction in volume of worm gear drive as compared to literature. It is worthwhile to mention that this reduction in volume is achieved after including industry relevant constraints.
- There is a decrease in design parameter value of 0.87 % in gear ratio, 33.91% in face width, and an increase of 48.93% in pitch circle diameter of worm. This research also reports the value of thermal capacity, strength capacity and wear capacity of worm gear drive which are not reported in any of the literature.
- This work is readily applicable and suitable for optimization of other similar gear drives employed in industries.

ACKNOWLEDGEMENTS

This work was supported by the Serbian Ministry of Education, Science, and Technological Development through the Mathematical Institute of the Serbian Academy of Sciences and Arts.

REFERENCES

[1] A. Alkharsan, O. Ata, “HawkFish Optimization Algorithm: A Gender -Bending Approach for Solving Complex Optimization Problems”, Electronics, Vol. 14(611), pp. 1-31, https://doi.org/10.3390/electronics14030611, (2025)

[2] M.Ghasemi, N.Aghajanpour and H.Arab, “Masterpiece Optimization Algorithm: A New Method for Solving Engineering Problems” Journal of Soft Computing in Civil Engineering, Vol.9(1), pp.1-21, https://doi.org/10.22115/scce.2024.404534.1677, (2025)

[3] M.Uzer, “New Hybrid Approaches Based on Swarm-Based Metaheuristic Algorithms and Applications to Optimization Problems” Applied Sciences, Vol.15(3), pp.1-25, https://doi.org/10.3390/app15031355, (2025)

[4] B.Milenković, Đ.Jovanović, M.Krstić, “An Application of Dingo Optimization Algorithm (DOA) for Solving Continuous Engineering Problems”, FME Transactions, Vol.50(2), pp.331-338, https://doi.org/10.5937/fme2201331M, (2022)

[5] G. Pavlović, M. Savković, “Analysis and Optimization of the Main Girder of the Bridge Crane with an Asymmetric Box Cross-Section”, Scientific Technical Review, Vol.72(1), pp.3-11, https://doi.org/10.5937/str2201003P, (2022)

[6] O.Solaiman, R.Sihwail, H.Shehadeh, I.Hashim and K.Alieyan, “Hybrid Newton-Sperm Swarm Optimization Algorithm for Nonlinear Systems”, Mathematics, Vol.11(1473), pp.1-21, https://doi.org/10.3390/math11061473, (2023)

[7] W.Mbasso, A.Harrison, P.Jangir, I.Dagal, H.Kotb, N.Alombah, R.Kumar, A.Smerat, E.Donfack, S.Gahtani and M.Elbarbary, “Sperm swarm optimization for many objective power flow problems with enhanced performance evaluation in power systems”, Scientific Reports, Vol.15(17253), pp.1-44, https://doi.org/10.1038/s41598-025-99330-z, (2025)

[8] D.Su, W.Peng, “Optimum Design of Worm Gears with Multiple Computer Aided Techniques”, Proceedings of ICCES ‘08, Vol.6(4), pp.575-581, (2008)

[9] N.Ebenezer, R.Saravanan, S.Ramabalan and S.Navaneethasanthakumar, “Worm Gear Drive Optimization Considering Industry Constraints Based on Nature Inspired Algorithms”, World Scientific News, Vol.87, pp.205-221, (2017)

[10] Sperm Swarm Optimization (SSO) - File Exchange - MATLAB Central

SESSION F: APPLIED MECHANICS

Virtual twin modeling of the Mitsubishi Rv-2frl industrial robot in the MATLAB programmable environment

Zvonko Petrović[1*], Radovan Nikolić[1], Ljiljana Brzaković[1], Milica Tufegdžić[1], Suzana Petrović Savić[2]

[1] Academy of professional studies Šumadija, Department in Trstenik, Trstenik, Serbia

[2] Faculty of Engineering University of Kragujevac , Kragujevac, Serbia

ARTICLE INFO

* **Correspondence:** zpetrovic@asss.edu.rs

DOI: https://doi.org/10.46793/ET26.F02P **PAGES:** F7 – F16

ABSTRACT

Industrial robots are complex mechatronic systems whose effective modeling plays a significant role in the development of virtual twins and advanced control systems. This paper presents the development of the kinematic and dynamic model of the Mitsubishi RV-2FRL industrial robot in the MATLAB programmable environment for the implementation of its virtual twin. The robot's kinematic model was developed based on the Denavit–Hartenberg method, through the definition of geometric parameters and transformation matrices required for determining the position and orientation of the end effector. Both forward and inverse kinematic models were implemented, enabling motion analysis and the generation of desired joint configurations. The dynamic model was derived using the Euler–Lagrange formalism, incorporating inertial, Coriolis, centrifugal, and gravitational effects acting on the robot. The developed mathematical models were implemented and validated in the MATLAB environment through appropriate simulations. The obtained results demonstrate that the models accurately describe the kinematic and dynamic behavior of the robot and provide a reliable foundation for the development of its virtual twin. Such an approach enables system performance analysis, simulation of various operating scenarios, and the development of control algorithms prior to their implementation on the physical robotic system.

KEYWORDS

Robotics, Kinematics, Dynamics, Mitsubishi RV-2FRL, Virtual Twin.

1. INTRODUCTION

The increasingly stringent requirements that technological systems must meet necessitate the use of industrial robots in assembly and packaging operations to ensure high precision, repeatability, and productivity. Industry 4.0 brings fundamental changes in the design, management, and optimization of production systems. One of the key concepts of this paradigm is the digital twin – a virtual representation of a physical system that enables two-way communication between the physical and virtual worlds [1]. In the field of industrial robotics, digital twins enable robot motion simulation, control algorithm verification, fault prediction, and production process optimization without the need for a physical prototype [2]. However, building a digital twin requires precise mathematical models of the robotic system, primarily kinematic and dynamic models, which form its mathematical core.

The subject of this paper is the Mitsubishi RV-2FRL industrial robot. It is a vertical articulated robot with six degrees of freedom, a payload capacity of 2 kg, and a reach of 649 mm [4]. Due to its precision (repeatability of ±0.02 mm) and compact design, it is widely used in assembly, palletizing, and precision manipulation tasks [4]. Although the manufacturer provides complete technical documentation, the open literature lacks detailed kinematic and dynamic models for this specific robot variant, which hinders the development of its digital twin. The main goal of this paper is to develop a complete kinematic and dynamic model of the Mitsubishi RV-2FRL robot in the MATLAB environment, as a basis for future motion simulation in the CATIA software package. MATLAB is used as a mathematical modeling platform due to its powerful numerical capabilities and wide application in robotics [11]. CATIA, as an industry standard for 3D modeling and simulation, enables visual simulation of robot motion based on calculated trajectories. The integration of these two software tools via the COM interface provides a synergistic effect: the mathematical precision of MATLAB and the visual power of CATIA.

1.1. Literature Review

This chapter presents a review of relevant research in the field of digital twins in industrial robotics, existing models for the Mitsubishi RV robot series, methods of kinematic and dynamic modeling, as well as methods for integrating MATLAB and CATIA software tools.

The concept of the digital twin was first systematically defined by Grieves [1] in 2014, as a virtual representation of a physical system that enables real-time simulation, analysis, and optimization. Tao and colleagues [2] further developed this concept for application in Industry 4.0, emphasizing two-way communication between the physical and virtual systems. In the field of industrial robotics, digital twins are used for robot motion simulation, control algorithm verification, fault prediction, and production process optimization [3]. However, most existing solutions are based on commercial software packages without open access to mathematical models, which limits the possibilities for further research and adaptation to specific applications.

Mitsubishi RV series industrial robots (RV-2F, RV-4F, RV-7F) are widely used in industry due to their precision, reliability, and compact design. The manufacturer provides complete technical documentation containing kinematic diagrams, joint motion ranges, and technical specifications [4]. However, detailed kinematic and dynamic models for the specific RV-2FRL model are lacking in the open literature. Existing works mainly refer to the RV-2F variant or older models [5], while no complete models have been found for the RV-2FRL. This gap represents the motivation for developing our own model tailored to the needs of building a digital twin.

Kinematic modeling of robotic systems involves the description of motion without considering the forces that cause it. The most commonly used method for modeling serial robots is the Denavit–Hartenberg (DH) convention [6], which enables the systematic definition of coordinate systems at each joint and the formation of homogeneous transformation matrices. The DH convention requires the determination of four parameters per joint: rotation angle θ, distance d, length L, and twist angle α [7]. Craig [3] provides a detailed treatment of the application of the DH convention to various types of industrial robots, including those with six degrees of freedom.

For solving the inverse kinematic problem in robots with six degrees of freedom where the last three axes intersect at a single point, the Pieper method [8] is often applied. This method enables analytical determination of joint angles for a given position and orientation of the end effector, which is computationally more efficient than numerical methods. Craig [3] and Siciliano and colleagues [5] provide a detailed description of the Pieper method and its application to anthropomorphic configuration robots.

Dynamic modeling describes the relationship between forces and torques acting on the robot and its motion. The two most commonly used approaches are the Newton-Euler and Euler-Lagrange formalisms [9]. The Euler-Lagrange approach starts from the energy of the system and arrives at explicit forms of the inertia matrix D(q), the Coriolis and centrifugal force matrix $C(q,\dot{q})$, and the gravitational force vector g(q) [10]. This approach provides a more complete insight into the structure of the dynamic model. In this paper, the Euler-Lagrange formalism is applied due to its explicitness and suitability for integration into the MATLAB environment.

2. MATERIALS AND METHODS

This chapter provides a detailed description of the materials and methods used for the development of the kinematic and dynamic model of the Mitsubishi RV-2FRL industrial robot. The methodology of kinematic modeling of the robotic system based on the Denavit–Hartenberg convention is presented, as well as the derivation of the dynamic model using the Euler-Lagrange formalism. All steps were implemented in the MATLAB software package, which represents an industry standard for numerical simulation and the development of mathematical models in the field of robotics [11]. The Mitsubishi RV-2FRL robot, as a vertical articulated manipulator with six degrees of freedom, requires precise definition of geometric and inertial characteristics in order to obtain a high-accuracy model. The data required for modeling were taken from the official technical documentation of the manufacturer [4], which includes link lengths, joint motion ranges, robot mass, payload capacity, and maximum speeds. Additionally, standard assumptions regarding mass distribution along individual segments were used, given that the manufacturer does not publish detailed data on the moments of inertia of each link. The developed algorithms were then verified through several tests, including workspace analysis, verification of the forward and inverse kinematic models, as well as comparison with the manufacturer's specifications. All results are documented and presented in appropriate tables and figures.

2.1. Description of the Mitsubishi RV-2FRL Robotic System

The robot designated as Mitsubishi RV-2FRL is a vertical articulated industrial robot manufactured by the Japanese company Mitsubishi Electric. It belongs to the MELFA FR series of robots. A detailed description of the robot is as follows:

Robot type: Vertical multiple-joint type (articulated robot).

Degrees of freedom: The robot has 6 degrees of freedom (6 axes).

Joint types: All six joints (axis designations J1 to J6) are revolute joints that enable rotational motion. Their motion ranges are:

- J1 (base): from -240° to +240° (range 480°)
- J2 (shoulder): from -117° to +120° (range 237°)
- J3 (upper arm): from 0° to +160° (range 160°)
- J4 (rotary axis): from -200° to +200° (range 400°)
- J5 (bending axis): from -120° to +120° (range 240°)
- J6 (twist axis): from -360° to +360° (range 720°)

This is a stationary manipulator. It is intended for stationary installation on the floor, ceiling, or wall. Additional characteristics:

- Payload: Maximum 2 kg
- Reach: Maximum reach radius is 649 mm
- Repeatability: ±0.02 mm
- Drive: AC servo motors with absolute encoders
- Mass: 21 kg

Controller: Connects to the CR800-02VD type controller (standalone) or others from the CR800 series depending on the application. In short, the Mitsubishi RV-2FRL is a compact, six-axis, high-precision industrial manipulator intended for tasks such as assembly, material handling, and inspection in automated production cells. This robot is delivered with controllers from the CR800 series (-D, -Q, or -R variants), but in this paper, the model is developed exclusively in the MATLAB software environment, without the need for physical connection to the actual controller for simulation purposes.

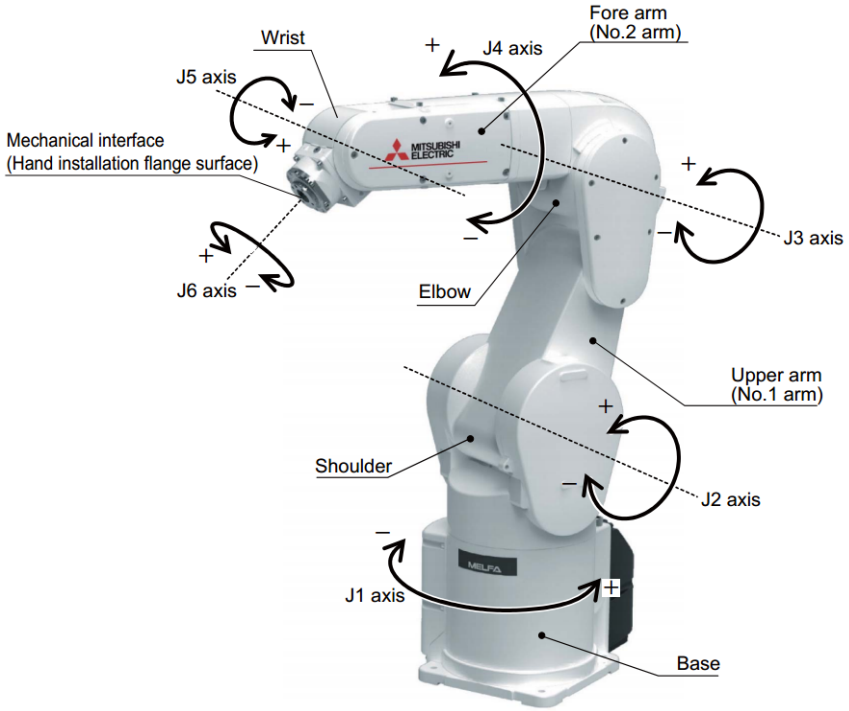


Figure 1. Robotic system [4].

2.2. Methodology for the Development of the Kinematic and Dynamic Model of the Mitsubishi RV-2FRL Robot

The methodology for the development of the kinematic and dynamic model of the Mitsubishi RV-2FRL robot consists of six main steps. These steps are carefully designed to ensure a systematic and repeatable modeling procedure, from input data to final verification.

The first step involves the collection and definition of the Denavit–Hartenberg (DH) parameters based on the complete technical documentation of the manufacturer [4]. This step is critical because the DH parameters represent the foundation for all subsequent kinematic calculations.

The second step is the implementation of the forward kinematic model in the MATLAB environment. This involves forming the homogeneous transformation matrix T_{60}, which describes the position and orientation of the end effector relative to the robot base. The forward kinematic model is necessary for determining the robot position based on given joint angles.

The third step includes the implementation of the inverse kinematic model, for which the Pieper method is applied. This method was chosen because, in the Mitsubishi RV-2FRL robot, the last three axes (J4, J5, and J6) intersect at a single point, which makes the Pieper method computationally efficient and analytically accurate.

The fourth step involves the analysis of the robot workspace through numerical simulation. For this purpose, a large number of random joint angle configurations are generated – in this paper, one million random configurations are used. For each configuration, the end effector position is calculated using the forward kinematic model, thereby obtaining a set of all reachable points. This analysis enables visualization and validation of the robot's geometric capabilities.

The fifth step encompasses the derivation of the dynamic model using the Euler-Lagrange formalism. The dynamic model describes the relationship between joint torques and robot accelerations, taking into account masses, inertias, and gravitational effects. The Lagrangian approach is chosen due to its explicitness and its ability to generate the inertia matrix D(q), the Coriolis and centrifugal matrix C(q,q̇), and the gravitational force vector g(q).

The sixth and final step is the verification of the overall model accuracy. Verification is performed by comparing simulation results (e.g., forward kinematics for a test configuration) with the manufacturer's specifications. This confirms that the DH parameters are correctly defined and that the model accurately predicts the behavior of the actual robotic system.

2.3. Mathematical Description of the Kinematic Model of the Mitsubishi RV-2FRL Robot

The Mitsubishi RV-2FRL robot is modeled as an open kinematic chain without branches, consisting of six revolute joints, designated J1 to J6. This type of model implies that each successive link is connected to the previous one, whereby the first link (the base) is fixed to the ground, and the last link (the end effector) is free. Such a representation is essential in industrial robotics because it enables a systematic approach to kinematic analysis.

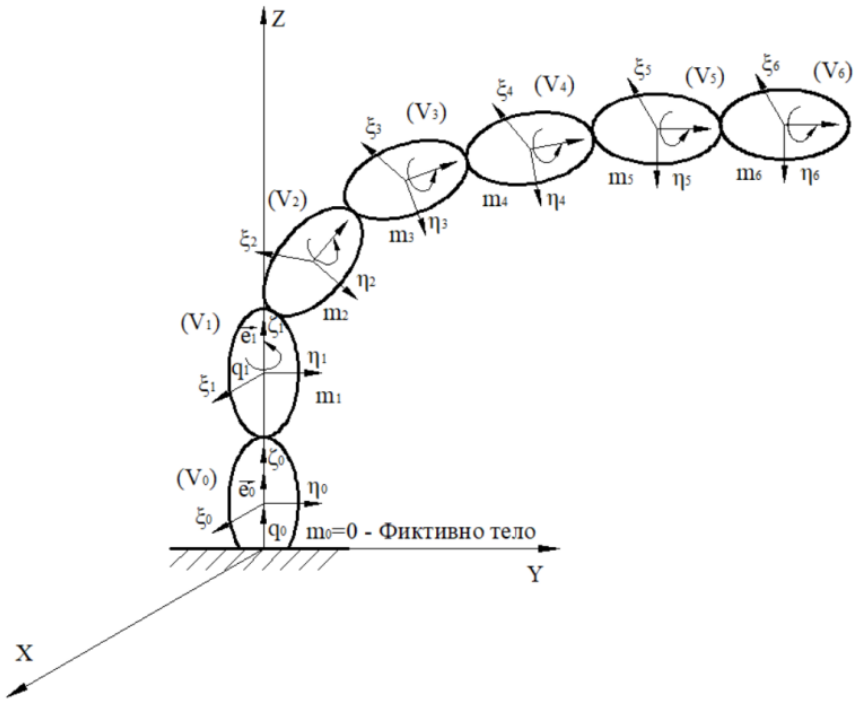


Figure 2: Model of the considered robotic system in the form of an open kinematic chain without branches

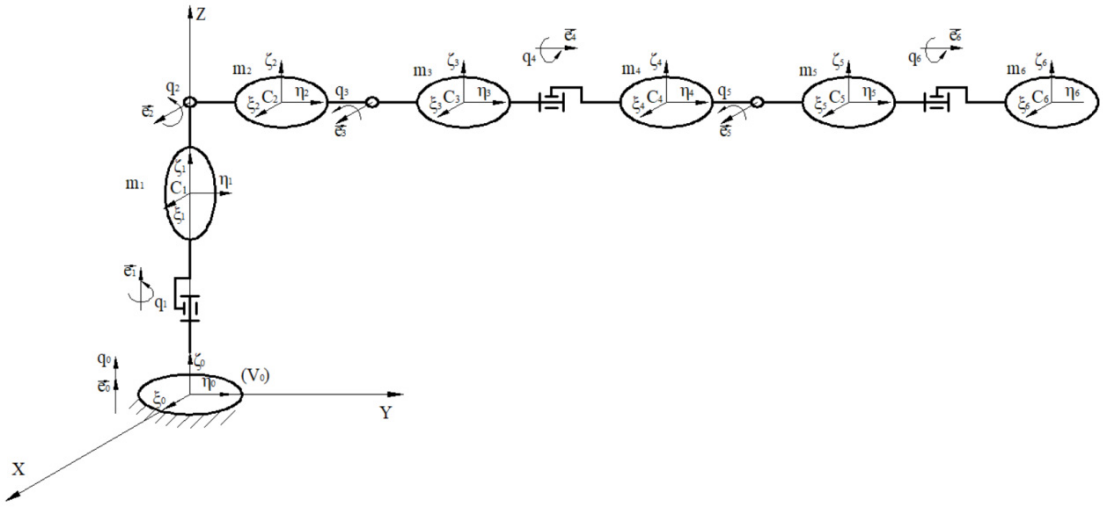


Figure 3: Model of the considered robotic system in the form of an open kinematic chain without branches

Based on the technical documentation [4] and the robot's geometry, the Denavit–Hartenberg parameters for all six joints were determined. The DH convention requires the definition of four parameters:

- Rotation angle about the axis (θ_i),
- Distance along the axis (d_i),
- Length of the common normal (L_i), and
- Twist angle about the ξ axis (α_i).

For the RV-2FRL robot, the link lengths L_2 and L_3 are 310 mm and 335 mm, respectively, while all offsets d_i are equal to zero, indicating that the joint axes intersect. The angle α_1 is -90°, $\alpha_2 = \alpha_3 = 0°$, α_4 = -90°, α_5 = 90°, and $\alpha_6 = 0°$. These parameters are summarized in Table 1.

Table 1: Denavit–Hartenberg parameters for the Mitsubishi RV-2FRL robot

Joint *i*	θ_i	d_i [mm]	a_i [mm]	α_i [°]
1	Θ_1	0	0	-90
2	Θ_2	0	310	0
3	Θ_3	0	335	0
4	Θ_4	0	0	-90
5	Θ_5	0	0	90
6	Θ_6	0	0	0

The homogeneous transformation matrix for a single joint represents a combination of rotations and translations and is given by the expression:

$$T_{ii-1} = \mathrm{Rot}(z,\theta_i)\cdot \mathrm{Trans}(z,d_i)\cdot \mathrm{Trans}(x,a_i)\cdot \mathrm{Rot}(x,\alpha_i) \tag{1}$$

The expanded form of this matrix is:

$$T_{ii-1} = \begin{bmatrix} \cos(\theta_i) & -\sin(\theta_i)\cdot\cos(\alpha_i) & \sin(\theta_i)\cdot\sin(\alpha_i) & L_i\cos(\theta_i) \\ \sin(\theta_i) & \cos(\theta_i)\cdot\cos(\alpha_i) & -\cos(\theta_i)\cdot\sin(\alpha_i) & L_i\sin(\theta_i) \\ 0 & \sin(\alpha_i) & \cos(\alpha_i) & L_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

The overall transformation matrix from the base to the end effector is obtained by multiplying the individual transformation matrices for all six joints:

$$T_{60} = T_{10}\cdot T_{21}\cdot T_{32}\cdot T_{43}\cdot T_{54}\cdot T_{65} \tag{3}$$

The final matrix T_{60} has the form $T_{60} = [R_{6x3}, p_{6x1}; 0, 1]$, where R_{6x3} is a 3×3 rotation matrix defining the orientation of the end effector, and p_{6x1} is a 3×1 position vector giving the coordinates of the end effector relative to the robot base. This model enables the calculation of the position and orientation of the end effector for a given set of joint angles $q = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6]^T$, which is known as the forward kinematic task.

For the inverse kinematic task, i.e., determining the joint angles based on a given position and orientation of the end effector, the Pieper method was applied. This method is particularly suitable for robots where the last three axes intersect at a single point, which is the case with the Mitsubishi RV-2FRL where joints J4, J5, and J6 intersect at a common point. The Pieper method enables analytical solution of the inverse kinematic problem, which is computationally more efficient than numerical methods.

The position of point J4 (intersection of axes) is determined from the desired end effector position p and rotation matrix R:

$$p_4 = p - d_6\cdot R\cdot \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T \tag{4}$$

The angle θ_1 is then determined from the position of point p_4 projected onto the base plane:

$$\theta_1 = \mathrm{atan2}\left(p_{4\gamma}, p_{4x}\right) \tag{5}$$

Angles θ_2 and θ_3 are determined using the geometric method, applying the law of cosines to the triangle formed by lengths a_2, a_3, and the distance to point p_4. After the first three angles are known, the rotation matrix for the last three joints R_{456} is calculated as $R_{456} = R_3^T \cdot R$, from which, knowing the structure of the rotation matrix for a spherical joint, the angles θ_4, θ_5, and θ_6 are obtained. This approach provides a complete solution to the inverse kinematic problem.

Based on the defined DH parameters, a kinematic model of the robot was created in MATLAB. Figure 3 shows the visualization of the Mitsubishi RV-2FRL robot in the initial position (all joint angles are 0°).

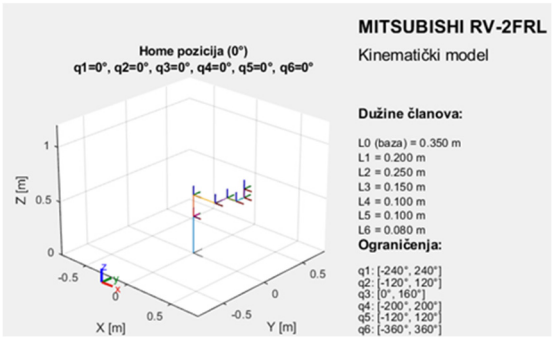


Figure 3: Visualization of the Mitsubishi RV-2FRL robot
in MATLAB – initial position (q_1=0°, q_2=0°, q_3=0°, q_4=0°, q_5=0°, q_6=0°)

2.4. Mathematical Description of the Dynamic Model of the Mitsubishi RV-2FRL Robot

he dynamic model of the robot describes the relationship between the torques generated by the actuators at the joints and the resulting robot motions. Unlike kinematics, which deals exclusively with the geometry of motion, dynamics takes into account forces, masses, inertias, and gravitational effects. In this paper, the dynamic model for the Mitsubishi RV-2FRL robot was derived using the Euler-Lagrange formalism, which represents one of the most commonly used approaches for modeling the dynamics of servo robots.

The Lagrangian of the system $L(q, \dot{q})$ is defined as the difference between the kinetic energy $K(q, \dot{q})$ and the potential energy $P(q)$ of the system:

$$L(q,\dot{q}) = K(q,\dot{q}) - P(q) \tag{6}$$

where $q = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6]^T$ is the vector of joint angles, and $\dot{q}$ is the vector of corresponding angular velocities. The kinetic energy of the system contains contributions from each robot link and depends on both masses and velocities. Potential energy originates solely from gravity and depends on the height at which the masses of individual links are located.

By applying Lagrange's equations of the second kind:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = \tau_i \tag{7}$$

the canonical form of robot dynamics is obtained:

$$D(q)\ddot{q} + C(q,\dot{q})\dot{q} + g(q) = \tau \tag{8}$$

where τ is the vector of joint torques.

The inertia matrix $D(q)$ is 6×6 in dimension, symmetric, and positive definite. For the Mitsubishi RV-2FRL robot, the total mass is approximately 21 kg, while the maximum payload is 2 kg. This means that the inertial contributions originating from the robot itself are dominant compared to the payload inertia. The elements of matrix $D(q)$ are nonlinear functions of the joint angles q and depend on the mass distribution along the robot links. The lengths of the main links are a_2 = 310 mm and a_3 = 335 mm, which directly affects the inertial characteristics of the system.

The Coriolis and centrifugal force matrix $C(q, \dot{q})$ is also 6×6 in dimension. Its elements are linear in $\dot{q}$ and can be calculated from the Christoffel symbols of the first kind. These terms describe the influence of the velocities of individual joints on the appearance of additional inertial forces. The maximum composite speed of the robot of 4200 mm/s indicates that Coriolis terms become significant when operating at high speeds. In such conditions, neglecting these terms would lead to significant errors in predicting dynamic behavior.

The gravitational force vector $g(q)$ is 6×1 in dimension and represents the torques arising from the effect of gravity on the robot masses. For the RV-2FRL robot, gravitational torques are particularly pronounced in joints J2 and J3, which correspond to the shoulder and elbow. Namely, these joints carry the largest part of the robot's mass, including the payload. On the other hand, gravity has a significantly smaller influence on joints J4, J5, and J6 due to the construction of the wrist joint, where masses are concentrated closer to the axis of rotation. The dynamic model given by equation (8) enables simulation of the robot's dynamic behavior under different loads and speeds, which is of key importance for building a digital twin. Based on this model, control algorithms can be designed, system behavior can be predicted, and production processes can be optimized.

3. VERIFICATION OF THE ACCURACY OF THE PROPOSED KINEMATIC AND DYNAMIC MODEL

Verification of the accuracy of the developed kinematic and dynamic models was conducted through four independent tests. The first test encompasses the analysis of the robot's workspace, thereby confirming that the model correctly predicts all reachable points of the end effector. The second test involves checking the forward kinematic model on a specific test configuration, where the calculated position is compared with the expected values based on technical documentation. The third test verifies the inverse kinematic model, i.e., the ability of the Pieper method to calculate correct joint angles for a given position and orientation of the end effector. The fourth test presents a tabular comparison of the robot's key characteristics according to the manufacturer's specifications and the values obtained through simulation. All results are presented in appropriate tables, graphs, and numerical outputs.

The verification was conducted in the software package MATLAB R2024a. All presented results are reproducible, and the complete source code is given in the appendix of the paper. Special attention was paid to the numerical stability of the algorithm, particularly when solving the inverse kinematic problem where the angles were calculated with an accuracy of 10^{-6} radians.

3.1. Workspace Analysis

Workspace analysis represents the first step in the verification of the kinematic model. The workspace is defined as the set of all points that the robot's end effector can reach under all possible joint angle configurations, taking into account mechanical constraints. For the Mitsubishi RV-2FRL robot, the joint motion ranges are: J1 from -240° to +240°, J2 from -117° to +120°, J3 from 0° to +160°, J4 from -200° to +200°, J5 from -120° to +120°, and J6 from -360° to +360°.

The numerical simulation was performed by generating 1,000,000 random joint angle configurations within the specified ranges. Random numbers were generated using the MATLAB function rand, which provides a uniform distribution. For each configuration, the position of the end effector in the Cartesian coordinate system was calculated using the forward kinematic model (forward_kinematics function). The resulting set of points was then visualized in three-dimensional space. Based on the generated configurations and the application of the forward kinematic model, the workspace of the end effector was obtained, as shown in Figure 2.

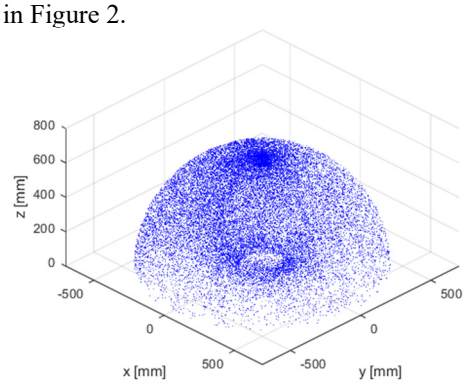


Figure 2: Workspace of the end effector of the Mitsubishi RV-2FRL robot obtained by numerical simulation in MATLAB (1,000,000 random configurations)

Based on the analysis, it was determined that the robot can reach points within a radius range of up to approximately 650 mm, which is consistent with the manufacturer's technical documentation [4] stating a maximum reach radius of 649 mm. A spherical cavity is noticeable in the central part of the workspace, which corresponds to an area inaccessible due to mechanical constraints of joints J2 and J3. This cavity is concentrated around the J1 axis and extends from approximately 50 mm to 200 mm from the base. Additionally, an asymmetrical shape of the workspace in the z-direction was observed, which is a consequence of the fact that joint J3 moves only in the positive direction (from 0° to +160°), while negative angles are not allowed. This characteristic is typical for most vertical articulated robots and directly affects the available workspace above and below the base level.

The maximum reach in the horizontal plane (x-y plane) is approximately 645 mm, while the maximum height reaches about 750 mm above the base. The minimum reachable height, due to the constraint of joint J3, is approximately 150 mm above the base. These results are of great importance for trajectory planning and robot placement in the production cell.

3.2. Verification of the Forward Kinematic Model

To verify the correctness of the forward kinematic model, a test configuration of joint angles was selected for which the end effector is expected to reach the maximum reach. Specifically, the following angles were selected: $\theta_1 = 0°$, $\theta_2 = -90°$, $\theta_3 = 90°$, $\theta_4 = 0°$, $\theta_5 = 90°$, $\theta_6 = 0°$. This configuration corresponds to the extended position of the robot along the x-axis. Using the forward_kinematics function in MATLAB, the following end effector position was obtained:

x = 644.98 mm

y = 0.00 mm

z = 0.00 mm

The calculated x-coordinate of approximately 645 mm corresponds very well with the manufacturer's specification stating a maximum reach radius of 649 mm. The small deviation of 4 mm (0.6%) can be attributed to the rounding of input parameter values (link lengths are given as 310 mm and 335 mm, while actual physical values may have small deviations) and the numerical precision of the calculations. This test confirmed that the DH parameters are correctly defined and that the homogeneous transformation matrix T_{60} correctly describes the robot's geometry.

3.3. Verification of the Inverse Kinematic Model

The inverse kinematic model, based on the Pieper method, was verified as follows: for the end effector position obtained in the previous test ($p = [645, 0, 0]^T$) and the identity rotation matrix ($R = I$, meaning the end effector is oriented so that its z-axis is vertical), the joint angles were calculated using the inverse kinematics function.

The following results were obtained:

$\theta_1 = 0.00°$
$\theta_2 = -90.00°$
$\theta_3 = 90.00°$
$\theta_4 = 0.00°$
$\theta_5 = 90.00°$
$\theta_6 = 0.00°$

It is notable that the calculated angles match the initial test configuration (noting that they are in radians and then converted to degrees for display). This confirmed that the inverse model works correctly and that the Pieper method provides analytically accurate solutions. In addition to the basic solution, the Pieper method also enables obtaining alternative solutions ("elbow up" and "elbow down" configurations), which were also tested and confirmed.

Figure 4 shows the animation of the Mitsubishi RV-2FRL robot's motion from the initial position (all angles 0°) to the test configuration of [30°, -60°, 120°, 45°, 60°, 90°]. Six characteristic moments during the interpolated trajectory are shown: 0%, 20%, 40%, 60%, 80%, and 100% of the path. In the initial position (0%), the robot is in its reference position with all joints at zero. As it moves toward the test configuration (100%), rotation of the base (J1=30°), raising of the shoulder (J2=-60°) and elbow (J3=120°) can be observed, as well as the change in wrist orientation through joints J4, J5, and J6. This animation confirms the correctness of the derived DH parameters and the Pieper method.

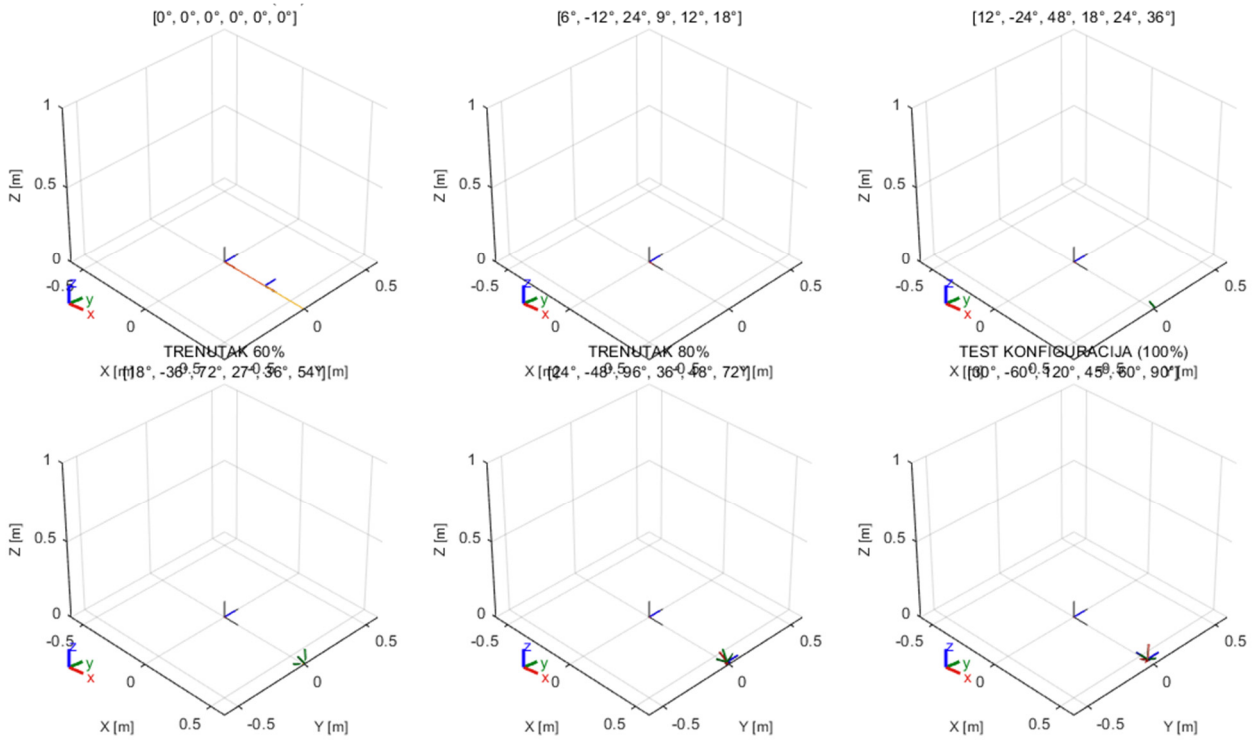


Figure 4: Animation of the Mitsubishi RV-2FRL robot motion between the initial position and the test configuration

3.4. Tabular Comparison with Manufacturer'S Specifications

The final verification was conducted through a tabular comparison of the robot's key characteristics according to the manufacturer's technical documentation [4] and the values obtained through simulation in MATLAB. The results are shown in Table 3.

Table 3: Comparison of manufacturer's specifications and simulation results

Characteristic	Manufacturer's specification [4]	Simulation result (MATLAB)
Maximum reach radius	649 mm	~645 mm
Number of degrees of freedom	6	6
Joint type	revolute (6)	revolute (6)
Robot mass	21 kg	21 kg (input data)
Maximum payload	2 kg	2 kg (input data)
Maximum composite speed	4200 mm/s	4200 mm/s (input data)
J1 range	-240° to +240°	implemented
J2 range	-117° to +120°	implemented
J3 range	0° to +160°	implemented
J4 range	-200° to +200°	implemented
J5 range	-120° to +120°	implemented
J6 range	-360° to +360°	implemented

Based on the presented results, it can be concluded that the developed kinematic and dynamic model reproduces the manufacturer's specifications with high accuracy. The small deviation in the maximum reach radius of 4 mm is within numerical tolerance and does not affect the functionality of the model. Thus, the model is verified and ready for application in the construction of the digital twin of the robotic system.

4. CONCLUSION

In this paper, a complete kinematic and dynamic model of the Mitsubishi RV-2FRL industrial robot intended for the development of its digital twin has been developed. Starting from the fact that detailed models for this specific robot variant are lacking in the open literature, while the manufacturer provides only technical documentation without mathematical models suitable for simulation, this paper represents a significant step towards filling that gap. The developed models were implemented in the MATLAB environment, which was chosen due to its powerful numerical capabilities and wide application in the field of robotics. Based on the presented material and conducted analyses, the following conclusions can be drawn.

The first significant contribution of this paper is the definition of a complete set of Denavit–Hartenberg (DH) parameters for the Mitsubishi RV-2FRL robot. Based on the manufacturer's technical documentation [4], the link lengths (a_2=310 mm and a_3=335 mm), offsets (d_i=0 for all joints), and twist angles (α_1=-90°, α_2=α_3=0°, α_4=-90°, α_5=90°, α_6=0°) were determined. These parameters form the basis for all further kinematic calculations and enable a systematic approach to modeling. Without correctly defined DH parameters, it is impossible to perform a correct kinematic analysis or build a credible digital twin.

The second contribution is the derivation of the forward kinematics matrix T_{60}, which relates the joint angles to the position and orientation of the end effector. The matrix was derived by multiplying the individual homogeneous transformation matrices for all six joints, applying the standard DH convention. This model enables the calculation of the position and orientation of the end effector in the Cartesian coordinate system for a given set of joint angles $q = [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6]^T$. The forward kinematic model is a fundamental building block of any robot control system, as it enables trajectory planning in operational space.

The third contribution is the solution of the inverse kinematic problem using the Pieper method. Due to the fact that in the Mitsubishi RV-2FRL robot the last three axes (J4, J5, and J6) intersect at a single point, the Pieper method enables analytical determination of joint angles for a given position and orientation of the end effector. This is computationally extremely efficient compared to numerical methods and enables real-time application, which is of crucial importance for systems requiring rapid inverse kinematics calculation, such as trajectory tracking systems or visually guided robots

The fourth contribution is the derivation of the dynamic model using the Euler-Lagrange formalism. The canonical form of robot dynamics was derived: $D(q)\ddot{q} + C(q,\dot{q})\dot{q} + g(q) = \tau$, where the inertia matrix $D(q)$ of dimension 6×6, the Coriolis and centrifugal force matrix $C(q,\dot{q})$, and the gravitational force vector $g(q)$ are explicitly defined. For the RV-2FRL robot, the total mass is approximately 21 kg, while the maximum payload is 2 kg, meaning that the inertial contributions originating from the robot itself are dominant compared to the payload inertia. This fact is significant for designing control algorithms, as the influence of the payload on system dynamics can be treated as a disturbance.

The maximum robot speed of 4200 mm/s indicates that Coriolis terms become significant when operating at high speeds. Under such conditions, neglecting these terms would lead to significant errors in predicting dynamic behavior. The gravitational force vector $g(q)$ is particularly significant in joints J2 and J3, which correspond to the shoulder and elbow, while gravity has a significantly smaller influence on joints J4, J5, and J6 due to the construction of the wrist joint.

The fifth contribution is the conducted verification of model accuracy. Through workspace analysis, based on 1,000,000 random configurations, it was determined that the robot can reach points within a radius range of up to approximately 650 mm, which is consistent with the manufacturer's specification of 649 mm [4]. A spherical cavity is noticeable in the central part of the workspace, corresponding to an area inaccessible due to mechanical constraints of joints J2 and J3, as well as an asymmetrical shape in the z-direction due to the fact that joint J3 moves only in the positive direction. These results are of great importance for trajectory planning and robot placement in the production cell.

By verifying the forward kinematic model on the test configuration [0°, -90°, 90°, 0°, 90°, 0°], the end effector position was obtained as x ≈ 645 mm, y = 0 mm, z = 0 mm, with a deviation of only 4 mm (0.6%) from the specification. This deviation can be attributed to the rounding of input parameter values and numerical calculation precision, which is completely acceptable for engineering applications. By verifying the inverse kinematic model, for the given position $p = [645, 0, 0]^T$ and the identity rotation matrix $R = I$, the calculated angles matched the initial test configuration, thereby confirming the correctness of the Pieper method.

It should be emphasized that the paper does not contain experimental validation of the model on a physical robot, which represents the most significant limitation. In real conditions, there are nonlinear effects such as joint friction, mechanical transmission backlash, and link elasticity, which were not taken into account in this model. Furthermore, modeling was performed under the assumption of absolute rigidity of the links, which is not the case in reality – every mechanical system possesses some elasticity that can be significant at high speeds and loads. The dynamic model does not include joint friction, which can be particularly pronounced at low movement speeds.

Future work will include several important steps. The first is the implementation of the developed models in the CATIA environment, either via the COM interface or using the Simscape Multibody tool for bidirectional real-time interaction. The second direction is experimental validation on the physical Mitsubishi RV-2FRL robot, which would quantify the accuracy of the model in real conditions and identify the dominant sources of errors. The third direction is extending the model by including joint friction and link elasticity, which would increase the accuracy of predicting robot behavior. The fourth direction is the application of the model for optimizing assembly and packaging tasks within the development of a complete digital twin, including collision-free trajectory planning.
The developed model can be used in industrial environments for simulation and optimization of tasks before physical implementation, thereby reducing production line downtime and the risk of equipment damage when testing new programs. Furthermore, the model represents a starting basis for the development of the digital twin of the Mitsubishi RV-2FRL robotic system.

REFERENCES

[1] M. Grieves, "Digital twin: manufacturing excellence through virtual factory replication", White paper, Florida Institute of Technology, (2014)

[2] F. Tao, H. Zhang, A. Liu, and A. Y. C. Nee, "Digital twin in industry: State-of-the-art", IEEE Transactions on Industrial Informatics, Vol. 15(4), pp. 2405-2415, https://doi.org/10.1109/TII.2018.2873186, (2018)

[3] J. J. Craig, "Introduction to Robotics: Mechanics and Control", 3rd ed., Pearson Prentice Hall, Upper Saddle River (USA), (2005)

[4] Mitsubishi Electric, "MELFA RV-2FRL Series Standard Specifications Manual", Mitsubishi Electric Corporation, (2020)

[5] B. Siciliano, L. Sciavicco, L. Villani, and G. Oriolo, "Robotics: Modelling, Planning and Control", Springer, London (UK), https://doi.org/10.1007/978-1-84628-642-1, (2009)

[6] J. Denavit and R. S. Hartenberg, "A kinematic notation for lower-pair mechanisms based on matrices", Journal of Applied Mechanics, Vol. 22(2), pp. 215-221, (1955)

[7] M. W. Spong, S. Hutchinson, and M. Vidyasagar, "Robot Modeling and Control", John Wiley & Sons, New York (USA), (2006)

[8] D. L. Pieper, "The kinematics of manipulators under computer control", PhD Thesis, Stanford University (USA), (1968)

[9] R. M. Murray, Z. Li, and S. S. Sastry, "A Mathematical Introduction to Robotic Manipulation", CRC Press, Boca Raton (USA), (1994)

[10] W. Khalil and E. Dombre, "Modeling, Identification and Control of Robots", Butterworth-Heinemann, Oxford (UK), https://doi.org/10.1115/1.1566397, (2002)

[11] MathWorks, "Robotics System Toolbox Documentation", The MathWorks Inc., Natick (USA), (2023)

SESSION F: APPLIED MECHANICS

Numerical investigation of a star-shaped auxetic structure under different impact velocities

Vladimir Sinđelić[1*], Aleksandar Nikolić[1], Snežana Ćirić-Kostić[1], Nebojša Bogojević[1], Olivera Cekić-Erić[1], Nenad Drvar[2]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Kraljevo, Serbia

[2] Topomatika, LLC, Sveta Nedelja, Croatia

ARTICLE INFO

*** Correspondence:** sindjelic.v@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.F03S **PAGES:** F17 – F23

ABSTRACT

Auxetic structures are a class of mechanical metamaterials characterized by a negative Poisson's ratio and enhanced energy absorption capabilities under impact loading. In this study, the influence of impact velocity on the mechanical response of a star-shaped auxetic (SSA) structure is investigated using the finite element method implemented in ANSYS Workbench LS-DYNA. A cellular structure composed of a 6×6 arrangement of SSA unit cells was subjected to compressive impact loading at velocities of 1 m/s, 10 m/s, and 100 m/s. The deformation mechanisms, force–displacement response, kinetic and internal energy histories, and energy absorption characteristics were analyzed. The results show that the deformation patterns obtained at 1 m/s and 10 m/s are nearly identical, exhibiting the characteristic elastic, plateau, and densification stages of cellular structures. Furthermore, the quasi-static condition was satisfied for these loading velocities, as the kinetic energy remained significantly lower than the internal energy throughout the simulations. In contrast, the response at 100 m/s was dominated by dynamic effects, resulting in non-uniform deformation, increased impact forces, and significantly higher energy absorption. The findings provide insight into the influence of impact velocity on the mechanical behavior of star-shaped auxetic structures and may contribute to their application in energy-absorbing systems.

KEYWORDS

Auxetic, Energy absorption, Finite element method, Quasi-static, Star-shaped structure

1. INTRODUCTION

Auxetic structures are artificially engineered materials characterized by unique mechanical properties [1–5]. They belong to the class of mechanical metamaterials, with their most distinctive feature being a negative Poisson's ratio. For isotropic materials, the Poisson's ratio typically ranges from −1 to 0.5, while in certain special cases values outside this range may be observed [6].

The fundamental building block of auxetic materials is the auxetic unit cell [3,6]. An assembly of such unit cells forms an auxetic structure. Unlike conventional materials, auxetic structures exhibit lateral contraction under compressive loading and lateral expansion under tensile loading, resulting in a negative Poisson's ratio, as illustrated in Figure 1(a). Owing to this deformation mechanism, auxetic structures demonstrate enhanced energy absorption capabilities under impact loading.

The deformation process of an auxetic structure subjected to compression can generally be divided into three stages: the elastic stage, the plateau stage, and the densification stage, as shown in Figure 1(b). Due to their ability to undergo large recoverable deformations and efficiently absorb impact energy, auxetic structures have attracted considerable attention in a wide range of engineering applications, including robotics, sports equipment, textiles, construction, composite materials, biomedicine, automotive engineering, defense systems, aerospace structures, and sensor technologies.

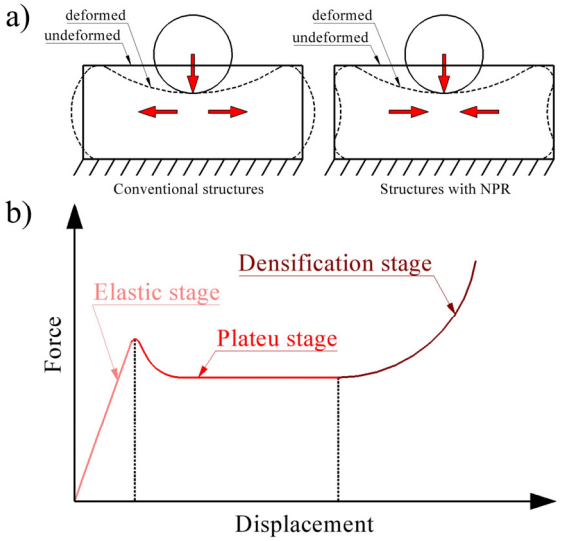


Figure 1: (a) Comparison between conventional and auxetic structures [4]
and (b) schematic representation of the deformation stages during the compression process

In this study, the influence of impact velocity on the mechanical response of a star-shaped auxetic structure is investigated using the finite element method. Section 2 presents the design of the star-shaped auxetic unit cell and the corresponding cellular structure. Section 3 describes in detail the development of the finite element model and the simulation procedure employed to analyze the structural response under different impact velocities. Finally, the numerical results obtained from the finite element simulations are presented and discussed.

2. DESIGN OF A STAR-SHAPED AUXETIC STRUCTURE

The Star-Shaped Auxetic (SSA) unit cell represents a conventional re-entrant auxetic configuration in both directions. The unit cell consists of four arrow-shaped elements, together with two vertical and two horizontal struts, as shown in Figure 2.

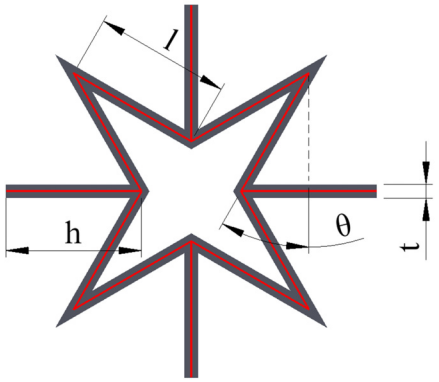


Figure 2: Geometric parameters of the SSA unit cell

The geometric parameters describing the SSA unit cell are h - the length of the horizontal and vertical struts, θ - the angle between the re-entrant and vertical struts, l - the length of the re-entrant struts, and t - the strut thickness. Based on these geometric parameters, the relative density ρ_r of the SSA structure is defined as the ratio between the volume of the unit cell V_{UC} and the volume of the space V_S occupied by that unit cell, and it is expressed by the following equation [7,8]:

$$\rho_r = \frac{V_{UC}}{V_S} = \frac{(4l+2h)t}{2\left(\sqrt{2}l\sin\left(\frac{\theta}{2}\right)+h\right)^2} \quad (1)$$

The geometric parameter values were taken from references[7–9]: $l = 4\ mm$, $h = 4\ mm$, $\theta = 30°$, and $t = 4\ mm$. A 6×6 array of unit cells was used to form the structure, which is placed between two plates with a thickness of 1 mm. The designed structure is shown in Figure 3, and its overall dimensions are: height $H = 67.57\ mm$, width $L = 65.57\ mm$, and thickness $B = 10\ mm$.

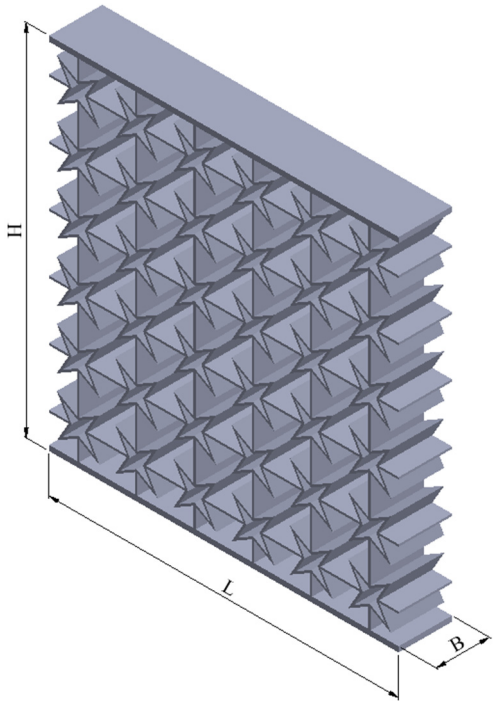


Figure 3: 6x6 structure composed of SSA unit cells

3. FINITE ELEMENT ANALYSIS OF THE SSA STRUCTURE

The mechanical properties of the SSA structure under compressive loading were investigated using the finite element method implemented in the ANSYS Workbench software package, specifically the LS-DYNA module. In this section, the structural response to the impact of a rigid plate at velocities of 1 m/s, 10 m/s, and 100 m/s is analyzed. The numerical model developed in ANSYS consists of two rigid plates and the SSA structure, as shown in Figure 4.

A constant velocity was applied to the top plate in the vertical direction, whereas the bottom plate was fixed throughout the analysis. The contact interaction between the SSA structure and the rigid plates was established using a manual contact region. Frictional contact with symmetric behavior was adopted, and both the static and dynamic coefficients of friction were assigned a value of 0.2. The planar surfaces of the SSA structure were specified as the contact bodies, while the lower surface of the top plate and the upper surface of the bottom plate were defined as the target bodies. Figure 5 presents the contact definition between the bottom plate and the SSA structure. The contact settings defined for the SSA self-contact were identical to those used for the contact interaction between the rigid plates and the SSA structure. The only difference was that both the contact and target bodies were assigned to the SSA structure, enabling the detection of potential interactions between its deformable components during compression.

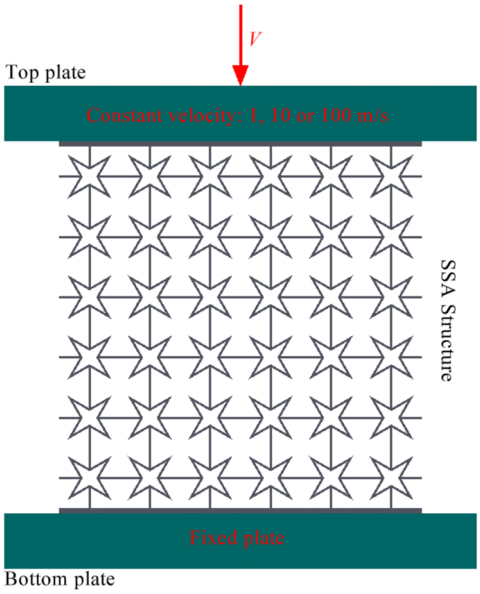


Figure 4: FEM of SSA structure under compressive loading

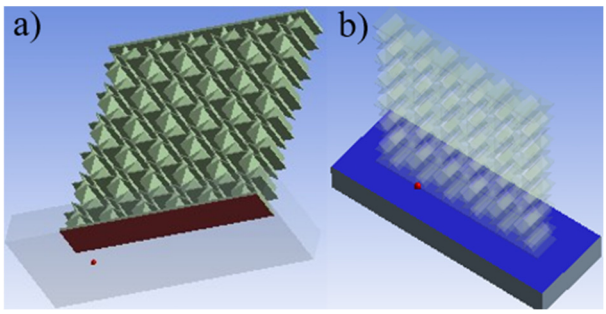


Figure 5: Contact Between the SSA Structure and the Bottom Plate: (a) Contact Body and (b) Target Body

The material used for the rigid plates in the numerical simulation was Structural Steel, with a density of ρ=7850 kg/m³ and a Young's modulus of E=200 GPa. The SSA structure was modeled using a nonlinear material, Stainless Steel NL, with the following mechanical properties: density of ρ=7750 kg/m³, Young's modulus of E=193 GPa, and yield strength of σ_{YS}=210 MPa.

In the analysis settings, the same parameters were used for all loading velocities: the time step safety factor was set to 0.9, the solver type was structural analysis only, and the Hourglass control was defined using the Flanagan–Belytschko stiffness form with a standard Hourglass coefficient of 0.1. The end time for the 1 m/s case was 0.033785 s, while for the 10 m/s and 100 m/s cases the end time was reduced by factors of 10 and 100, respectively. The selected simulation time durations correspond to a structural compression of 31.046 mm, which represents approximately 46% compression relative to the total height *H*.

In this study, no mesh convergence analysis was performed. The element type was selected as automatic with the physics preference: explicit, using a linear element order. The element size for the rigid plates was 3 mm, while for the SSA structure it was 0.4 mm.

The output results from the numerical simulation include the vertical contact force between the top plate and the SSA structure, as well as the vertical deformations of the top plate.

3.1. Numerical Simulation Results

Figure 6 illustrates the deformation behavior of the SSA structure under three different impact velocities of the rigid plate. It can be observed that, for impact velocities of 1 m/s and 10 m/s (Figure 6 (a-b)), the SSA structure exhibits nearly identical deformation patterns. During the initial stage of impact, the structure deforms uniformly until reaching the densification phase, characterized by the collapse of the unit cells. In contrast, for an impact velocity of 100 m/s (Figure 6 (c)), a significantly different deformation mode is observed. The deformation is concentrated in the upper layers of the structure, while the lower layers remain almost undeformed throughout the simulation. This behavior indicates the presence of a pronounced strain-rate effect and non-uniform deformation propagation under high-velocity impact loading.

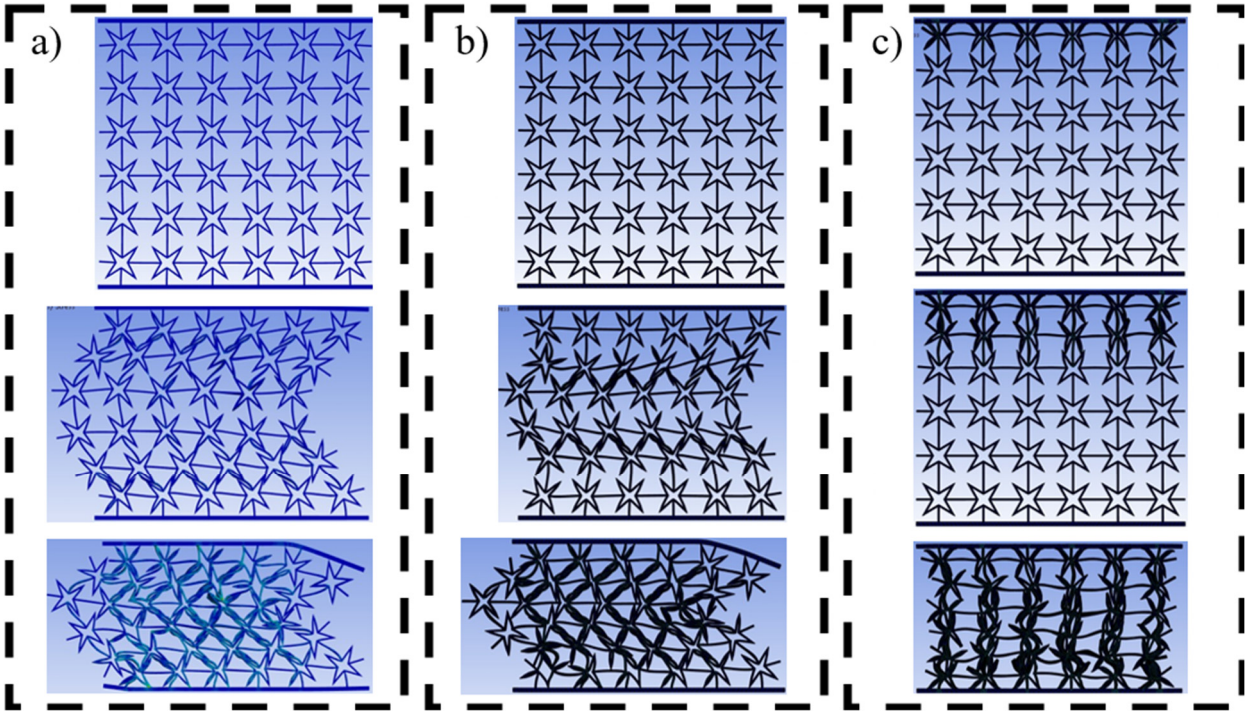


Figure 6: Deformation of the SSA structure under impact velocities of: (a) 1 m/s, (b) 10 m/s and (c) 100 m/s

To ensure quasi-static loading conditions in an explicit finite element analysis, the kinetic energy should remain significantly lower than the internal energy, typically below 5% of the internal energy throughout the simulation [10]. Figure 7 presents the histories of kinetic and internal energies for the entire duration of the simulations.

As shown in Figure 7(a), for an impact velocity of 1 m/s, the kinetic energy remains at a low level while the internal energy continuously increases during the loading process. Therefore, the quasi-static condition can be considered satisfied. A similar trend is observed for the impact velocity of 10 m/s (Figure 7(b)). Although a rapid increase in kinetic energy occurs at the beginning of the simulation, it subsequently stabilizes at a relatively low value, whereas the internal energy continues to increase. Consequently, the quasi-static condition is also satisfied in this case.

In contrast, for the impact velocity of 100 m/s (Figure 7(c)), the kinetic energy remains substantially higher than the internal energy throughout most of the simulation. As a result, the quasi-static condition is not fulfilled, indicating that the structural response is dominated by dynamic effects.

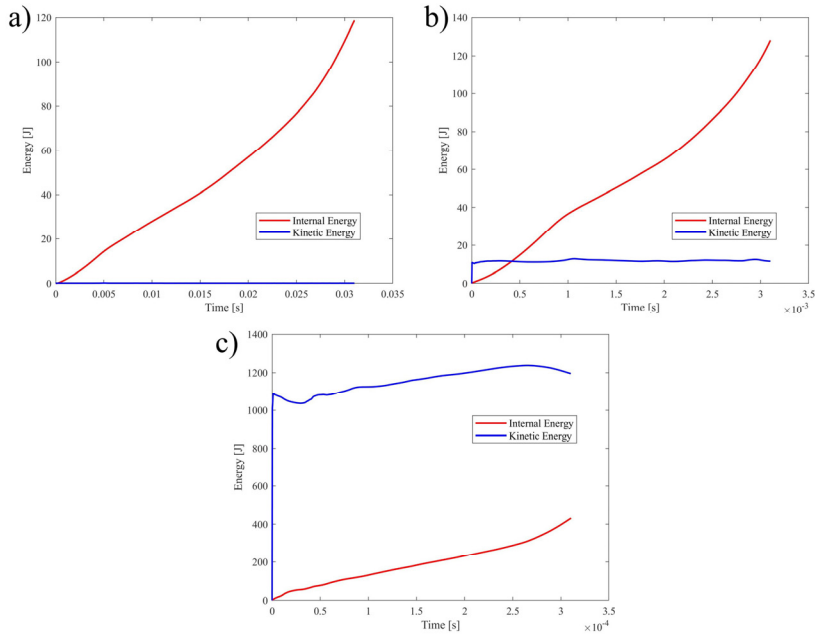


Figure 7: Internal and Kinetic energy of the SSA structure under impact velocities of: (a) 1 m/s, (b) 10 m/s and (c) 100 m/s

Figure 8 presents the force–displacement curves of the SSA structure subjected to impact velocities of 1 m/s, 10 m/s, and 100 m/s. It can be observed that, for impact velocities of 1 m/s and 10 m/s, the force–displacement response exhibits the characteristic behavior of cellular structures, consisting of three distinct stages: an initial elastic region, a plateau region associated with progressive cell collapse, and a densification region characterized by a rapid increase in force.

In contrast, the response obtained at an impact velocity of 100 m/s differs significantly. The force–displacement curve exhibits pronounced force oscillations and substantially higher peak forces, indicating the dominance of dynamic effects during impact loading. As a result, the deformation process becomes non-uniform and the characteristic quasi-static response of the structure is no longer observed.

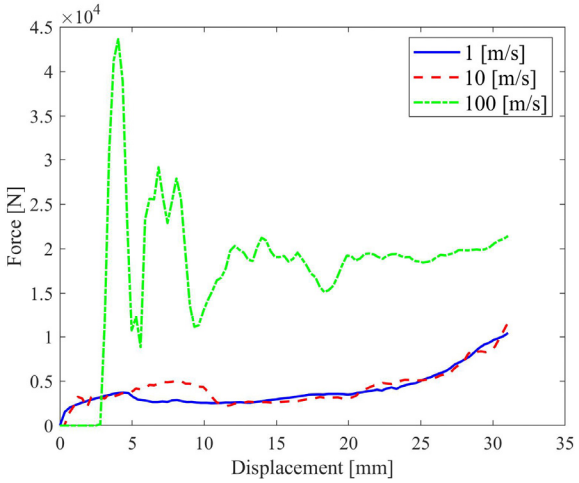


Figure 8: Force–Displacement curves of the SSA structure under impact velocities of 1 m/s, 10 m/s, and 100 m/s

Under impact loading conditions, the energy absorption (EA) capacity of the SSA structure is evaluated as the area under the force–displacement curve, which can be calculated by integrating the contact force (F(y)) over the corresponding compressive displacement (y) [11]. Accordingly, the energy absorption is expressed as:

$$EA = \int_0^y F(y)dy \tag{2}$$

For discrete data obtained from finite element simulations, the energy absorption (EA) is calculated using numerical integration based on the trapezoidal rule, which can be expressed as:

$$EA = \sum_{i=1}^{n-1} \frac{F_i + F_{i+1}}{2}(x_{i+1} - x_i) \tag{3}$$

Based on the data from Figure 8 and Equation (3), the values for all three impact velocity cases are presented graphically in Figure 9. The results show that the energy absorption (EA) values for lower impact velocities, 1 m/s and 10 m/s, are approximately identical, with a difference of less than 6%. Due to significantly higher impact forces, the EA value at an impact velocity of 100 m/s is considerably higher, which is a consequence of pronounced dynamic effects during impact loading.

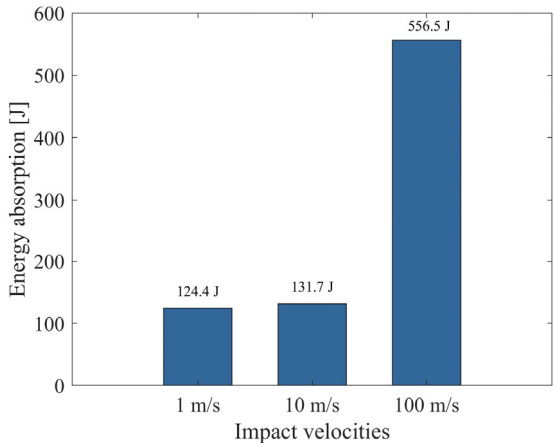


Figure 9: Energy absorption response of the SSA structure under impact velocities of 1 m/s, 10 m/s, and 100 m/s

4. CONCLUSION

This study presents a numerical investigation of the mechanical response of the SSA structure under impact loading conditions at velocities of 1 m/s, 10 m/s, and 100 m/s. Based on the results presented in this work, it can be concluded that the quasi-static condition is satisfied for impact velocities of 1 m/s and 10 m/s.

Future research should focus on the influence of different geometric parameters of the unit cell on the mechanical behavior of the SSA structure. In addition, experimental testing would be valuable to further validate the numerical results and to determine the critical impact velocity threshold required to ensure quasi-static conditions consistent with experimental observations.

Furthermore, it is of interest to investigate the evolution of the Poisson's ratio during the compression process of the SSA structure, as it may provide additional insight into its auxetic behavior under large deformations.

ACKNOWLEDGEMENTS

This research was supported under grant no. 451-03-34/2026-03/200108 by the Ministry of Education, Science and Technological Development of the Republic of Serbia. This support is gratefully acknowledged.

REFERENCES

[1] M. Periyasamy, N. Srinivasan, "A review on auxetic structures: Applications, properties, classifications, finite element analysis, manufacturing methods, and selective laser melting process parameters", Results in Engineering, Vol. 28 107660. https://doi.org/10.1016/j.rineng.2025.107660, (2025)

[2] V. Sinđelić, S. Ćirić-Kostić, N. Bogojević, A. Nikolić, "Numerical Analysis of the Modified 3D Arc-Star-Shaped Auxetic Structure", Machine and Industrial Design in Mechanical Engineering, pp. 695–705. https://doi.org/10.1007/978-3-031-80512-7_68, (2025)

[3] V. Sinđelić, A. Nikolić, G. Minak, N. Bogojević, S. Ćirić Kostić, “An improved 2D arc-star-shaped structure with negative Poisson’s ratio: In-plane analysis”, Materials Today Communications, Vol. 37, https://doi.org/10.1016/j.mtcomm.2023.107593, (2023)

[4] V. Sinđelić, A. Nikolić, N. Bogojević, S. Ćirić-Kostić, G. Minak, “Analytical and numerical analysis of the modified 2D arc-star-shaped structure with negative Poisson’s ratio”, Engineering Today Vol. 3, https://doi.org/10.5937/engtoday2400004S, (2024)

[5] Vladimir Sinđelić, Aleksandar Nikolić, Nebojša Bogojević, Olivera Erić Cekić, Snežana Ćirić Kostić, “Modified 2D arc-star-shaped structure with negative Poisson’s ratio”, Heavy Machinery-HM 2023, Vrnjačka Banja, p. E.21-E.26, (2023)

[6] V. Sinđelić, S. Ćirić-Kostić, A. Nikolić, N. Bogojević, G. Minak, “Analytical and Numerical Approach for Determining the Mechanical Properties of a Novel 3D Double Arc Star-Shaped Structure with Two Different Cross-Sections”, Mechanics of Solids, Vol. 60, https://doi.org/10.1134/S0025654425601958, (2025)

[7] G. Luo, D. Mo, C. Chai, J. Liu, Y. Chen, “Mechanical and Energy-Absorption Properties of a 3D-Printed Star-Shaped Auxetic Honeycomb under Combined Compression-Shear Loading”, Latin American Journal of Solids and Structures Vol. 20, https://doi.org/10.1590/1679-78257624, (2023)

[8] L. Wei, X. Zhao, Q. Yu, W. Zhang, G. Zhu, “In-plane compression behaviors of the auxetic star honeycomb: Experimental and numerical simulation”, Aerospace Science and Technology, Vol. 115, https://doi.org/10.1016/j.ast.2021.106797, (2021)

[9] J.P. Ren, Z.P. Gu, A.G. Zhao, C.G. Huang, X.Q. Wu, “Enhancing energy absorption of star-shaped honeycombs by utilizing negative Poisson’s ratio effect under high-velocity impact”, International Journal of Impact Engineering, Vol. 202, https://doi.org/10.1016/j.ijimpeng.2025.105297, (2025)

[10] S. Wang, H.-T. Liu, “Quasi-static compression response of a novel multi-step auxetic honeycomb with tunable transition strain”, Aerospace Science and Technology, Vol. 155, https://doi.org/10.1016/j.ast.2024.109730, (2024)

[11] Y. Hu, Y. Li, Y. Zhang, S. Ding, R. Wang, R. Xia, “Design methodology for functional gradient star-shaped honeycomb with enhanced impact resistance and energy absorption”, Materials Today Communications, Vol. 38, https://doi.org/10.1016/j.mtcomm.2023.108020, (2024)

SESSION G

CIVIL ENGINEERING

SESSION G: CIVIL ENGINEERING

Road surface wear - analysis of tests and possibilities for application of new materials

Blagoyka Paleva-Kadiyska[1*]

[1] Todor Kableshkov University of Transport, Faculty of Machinery and Construction Technologies in Transport, Department of Mechanics, Sofia, Bulgaria

ARTICLE INFO

*** Correspondence:** paleva-kadiyska.bl@abv.bg

DOI: https://doi.org/10.46793/ET26.G01PK **PAGES:** G1 – G6

ABSTRACT

Wear and tear is a process, which is related with the operation of roads, and regardless of the measures taken to reduce it, this process cannot be completely avoided. The surface layer of the road surface is in continual contact with the vehicle tires of moving vehicles and the impact of aggressive environments of various origins. The efficiency and the safety of road transport depends on several indicators, one of which is the degree of road surface wear. Now only to traffic safety, wear and tear significantly impacts to human health and the environment. In this paper has been conducted an extensive literature review of studies, which are related to road surface wear. New methods and materials to reduce this process are considered. Their results have been analyzed and systematized. Possibilities for using innovative construction materials have been examined, taking into account their specific properties and advantages that would allow their use as part of the road surface, in various facilities, and as road elements.

KEYWORDS

Road surface, wear, self-compacting concrete

1. INTRODUCTION

The road has a complex, multi-layered structure (Figure 1). The top layer is mainly subject to wear, because with this layer vehicle tires come into contact and which is directly exposed to various aggressive environments, especially in winter conditions. Asphalt and concrete pavements are mainly used for road surfaces, but other materials can also be used, such as flagstone, artificial stone, cobblestone, tiles, or bricks.

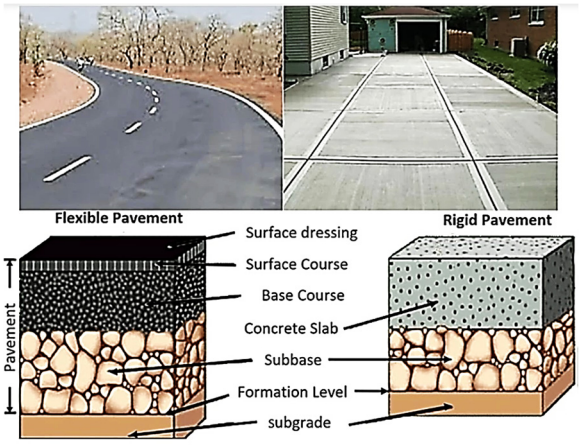


Figure 1: Road structure [1]

The main ingredients of concrete rigid pavement are coarse aggregate, sand and cement. Flexible asphalt pavements are mixtures of mineral aggregates, sand, filler, and bitumen binder. Typically, the coarse aggregate content is around 90–

95% and the bitumen binder around 5–10%. The properties of asphalt can be modified through admixtures such as adhesives, polymers, and various types of fillers. The composition and ratio of materials for two pavement types can vary widely across individual countries and regions.

The friction between the road surface and the tires of motor vehicles is crucial for traffic safety, as it provides the necessary cohesion when driving and braking. On the other hand, as a result of friction, the road surface wears out. This process is inevitable and negatively affects both traffic safety and human health, as well as the environment. In addition to the release of harmful fine particles, wear is also associated with increased noise when driving a motor vehicle. At the same time, it increases the costs of road maintenance and repair.

The number of vehicles and kilometers traveled is increasing every year. Traffic is one of the main sources of particulate matter. Emissions other than those from non-exhaust gases may arise from abrasion of motor vehicle's parts (tires, brake pads, clutch, bodywork, etc.) or from abrasion of the pavement surface, as well as from re-spraying of deposited particles.

The development includes a literature review of studies related to road pavement. The goal is to analyze the results of studying this inevitable, negative process accompanying the operation of roads. Possibilities for using alternative innovative materials, e.g., self-compacting concrete for pavement and road elements, are proposed, taking into account their specific properties.

2. ANALYSIS OF RESEARCHES RELATED TO ROAD PAVEMENT

Flexible pavements are multi-layered, often with asphalt concrete as the surface layer. They are the most common type of road pavements due to their lower initial cost and relatively easy maintenance. This kind of pavements provide elasticity and adaptability to the ground, which undergoes expansion and contraction caused by freeze-thaw cycles or minor shrinkages. In contrast, rigid pavements consist of a single layer of reinforced or unreinforced concrete. Concrete pavements are characterized by exceptional durability and have a long lifespan (often exceeding 30 years), but their placing and repair are more expensive than flexible pavements [1].

The wear of road pavements depends on many parameters such as the composition of the mixtures, traffic intensity, atmospheric changes in the environment, the composition and concentration of the aggressive environment used in winter road maintenance, and can vary significantly. The choice of materials for the mixture, as well as the technology of its placing, greatly affects the wear of the pavement and is crucial for its characteristics and efficiency. The distribution of traffic load also affects wear.

The research examined in this article have the aim to study the mechanism and degree of wear of road surfaces. Understandably, most of the researches are related to the wear of flexible pavements, due to their widest application. Some authors report that particle emissions from pavement wear are more difficult to quantify than those from the car tires wear and other vehicle undercarriage components. This is mainly due to two facts: the first one is related to the chemical composition of bitumen, which is complex, obtained as a result of recipe modeling. This fact makes it difficult to quantify its part in the overall chemical balance of the mass of released particles. The second fact is the result of the mixing of primary wear particles with fine road particles [2].

Abrasive wear of the road surface, in addition to endangering traffic safety, leads to the spread of dust pollution in the roadside area. Fine particulate matter can cause harm to people and the environment. Studies of the composition of pollutants have been conducted to determine the size and distribution of particles, and the harm to human health has been highlighted [3, 4]. Some authors [4] believe that the abrasive wear of the pavement depends primarily on several factors, such as: degree of mixing (homogeneity); dimensions of the additive materials, their properties, and the characteristics of the binding component.

In [5], based on testing three groups of samples, it is concluded that the existence of air voids affects abrasive wear by reducing the abrasion resistance of asphalt concrete. This can be explained by the separation of asphalt particles due to a lack of good adhesion between them. The process of weakening the bond between particles can be explained by an elementary mechanical model [6]. On the one hand, the presence of air voids negatively affects some properties of asphalt concrete, such as tensile strength, frost resistance, and fatigue resistance [7]. On the other hand, their reduced content leads to more intense rutting, just like poor compaction [8]. The contradictory influence of air voids in these experiments can be explained by poor compaction and poor homogenization of the asphalt concrete mixture [9]. Experiments in [5] show that the deterioration of the abrasion resistance of pavement materials depends on several processes, including the destruction and cracking of coarse aggregate particles. This imposes limitations on the laying mechanism and especially on the compaction process. It must exclude the destruction and cracking of the particles of the inert material. In [5], a relationship between the type of binder and the abrasive wear of asphalt concrete has been proven. It has been found that the simple abrasive wear model [6] can also explain the influence of the type of binder. For example, bitumen modified with styrene-butadiene-styrene polymer has good elastic properties, and its use reduces the processes of particle separation and cracking by more than 10-20% (Figure 2, c, d) [5].

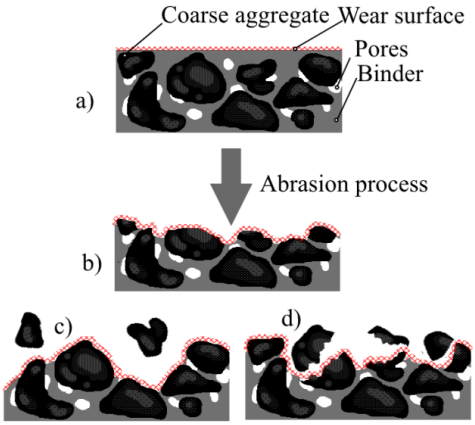


Figure 2: Process of abrasive wear of asphalt concrete:

(a) asphalt concrete before abrasion,

(b) abrasion process with striation,

(c) particle separation, (d) cracking of coarse aggregates [5]

Abrasion resistance also depends on the properties of the pavement materials. The use of coarse aggregates with a larger grain size and the use of SMA (split mastic asphalt) mixtures instead of dense-structure mixtures reduces wear [10].

In [11], to determine the abrasion of the top layer of bituminous pavements, asphalt mixtures from the wearing courses are compared in terms of mechanically separated dust particles. The particle measurements were performed in laboratory conditions, which allowed particle samples to be separated from the road surface only. Distinguishing particulate emissions from road abrasion and re-dusting is very difficult due to similar elemental compositions [12].

The results obtained in the studies in the above-mentioned works show that increased wear resistance can be achieved by increasing the size of the coarse aggregates, using a polymer-modified binder, and reducing the air void content in the asphalt concrete mixtures.

Some publications [13] investigate the noise generated by road surface wear. Noise is an environmental problem that leads to a deterioration in the lifestyle of many people and damage to their health. Outdoor noise usually has levels above 55 dB(A), with road traffic being one of its main sources. There are several options to reduce the noise: building noise barriers; insulating facades - an effective way in densely built-up areas; reducing noise in the car engine, and last but not least, taking measures to reduce the noise generated by the interaction between the tire and the road surface. It has been established that when the vehicle is moving at speeds between 30-45 and 120 km/h, this noise dominates. In [13], low-noise road surfaces are considered as an important factor in reducing road traffic noise. The surface of the pavements is a crucial factor that ensures noise reduction or amplification. The most common types of asphalt pavements with low noise levels are thin asphalt pavements and porous asphalt. Porous asphalt pavements reduce noise levels due to the high percentage of air voids in them [14]. Tests conducted in [15] found that thin-layer surfaces wore out very quickly and had a shorter life compared to porous asphalt pavements. However, the results in [16] show that porous asphalt pavements tend to lose their noise reduction characteristics after two to three years, which is assumed to be due to pore clogging. Another type of low-noise road surfaces are poroelastic pavements. They consist of granulated rubber or fibers that are combined with a binder. In this pavements the air void content is usually between 30 and 40%. The sound absorption as well as the noise reduction of poroelastic surfaces are very high. The introduction into operation of such pavements has not been successful to date [17], which is a result of: lack of adhesion between the surface and the sublayer; poor wear resistance; poor friction on a wet surface; high costs; and not good fire resistance.

Optimizing the texture of asphalt mixtures is the most sensible solution for creating low-noise pavements. The rough texture increases vibrations and noise-generating mechanisms. With the smooth surface with high porosity, there is a significant noise reduction. Finding an asphalt mixture that ensures the preservation of these characteristics in the long term is a challenge. Especially in regions with cold winters and high precipitation levels, where are used studded winter tires. Steel studs have a highly abrasive effect and cause mechanical wear on the road surface during the winter season. In such regions [18], to avoid abrasion, dense asphalt mixtures with a larger coarse aggregate size - Dmax 16 mm - are usually used. However, noise measurements in these regions show that the noise levels increase with increasing Dmax. To achieve noise reduction, it is recommended that Dmax be reduced [16]. At the same time, as Dmax decreases, pavement wear increases [19], leading to shorter pavement life and more health problems caused by dust formation due to abrasion. The challenge is to find the right balance and choose a grain size that provides good wear resistance and leads to positive acoustic characteristics of the road surface.

A significantly smaller proportion of the developments are aimed at studying rigid pavements.

The durability of the concrete pavement surface, including its wear resistance and material durability over time, is the subject of research in [20]. The wear of the concrete pavement surface is a function of both the cycles of repeated tire loading and the number of freeze/thaw cycles in the presence of de-icing chemicals. Heavy truck tires, tire chains, and studded winter tires cause significant damage even to high-quality pavements [21]. Studies have shown that ruts as deep as ¼ to ½ inch (6–12 mm) can appear in just one winter in regions where typically 30% of passenger cars are equipped with studded tires [22]. This quickly deteriorates the pavement surface, causing road safety problems.

There is ongoing research and development to create more wear-resistant concrete coatings [23], such as polymer-cement concretes, ultra-high-performance concretes, self-compacting concretes, etc. One aspect of quality is the amount of calcium hydroxide in the hardened mixture, which is directly related to the water/cement ratio – the so-called water-cement ratio (w/c) – used in the concrete [24]. This product of the cement hydration process is soft and erosive, often resulting in microscopic pits in the surface of the concrete.

Freeze/thaw resistance is also affected by the water/cement ratio, as it determines the size of pores and other voids. The total amount of pores include gel pores, capillary pores, and air voids [25]. Typically, concrete with low permeability (i.e., low water/cement ratio) is able to withstand freeze/thaw cycles better than concrete with a higher w/c. The water that enters the pores, as a result of freezing, expands and increases its volume by about 10%. This leads to the appearance of significant internal stresses, which destroy the concrete structure and lead to a violation of the integrity of the road surface [26]. After several freeze/thaw cycles, the local scaling of the surface results in direct exposure of the concrete to water penetration and aggressive salts [27].

Some of the research is aimed at studying different patterns of abrasive wear of concrete. In [28], a theoretical model based on test results of abrasion caused by the movement of detached particles transported by water is presented. The study shows that the mass loss in concrete can be expressed as a function of the work done by the abrasive mixture and a parameter related to the resistance of the concrete. Another study that extends this theoretical model, the loss of material, besides as a dependence on the composition of the concrete, is also expressed as a function of the abrasion effect caused by the rolling wheel and the total abrasion time [29].

Researchers are hampered by the fact that, in some cases, contradictory requirements are placed on them. For example: the mixture should have sufficient aggregate size to ensure good adhesion, but not too much so as not to increase noise; it should be well compacted to ensure good adhesion, but not too much so as not to damage the integrity of the aggregate; the binders should have good adhesion properties and at the same time not pollute the environment, etc.

This requires a continued search for new road surface materials.

3. SELF-COMPACTING CONCRETE - POSSIBILITIES FOR ITS USE AS AN INNOVATIVE MATERIAL FOR PARTS OF THE ROAD PAVEMENT, IN VARIOUS FACILITIES AND ROAD ELEMENTS

The search and work on creating new materials that will maximally satisfy the conditions that road pavements must meet, continues. To solve with some of the above-described problems about rigid pavements, the so-called self-compacting concrete (SCC) can be used. Invented in Japan [30], this type of concrete is suitable for use in elements with a high percentage of reinforcement and in facilities where a denser structure of the finished elements is required. Studies of its applicability show very good resistance to the effects of aggressive environments due to its denser structure. A study was conducted in [31], which showed that elements made with SCC have a longer life compared to those made of ordinary concrete when treated with aggressive anti-freeze mixtures in winter.

In rigid pavements, the increased size of the coarse aggregates is a prerequisite for the formation of cavities around the larger grains of the aggregate, which creates ideal conditions for water ingress and the destruction of the pavement according to the mechanism described above. To solve this problem, the size of the coarse aggregates can be limited to 20mm or 16mm, as recommended in the SCC formulations [32, 33].

The main advantage of rigid concrete pavements, namely their significantly greater load-bearing capacity, can be further developed by working towards increasing the adhesion between concrete and reinforcement under dynamic loading conditions. Studies in this direction have been conducted, and the results are presented in [34, 35], which prove that due to its denser structure, SCC has better bond strength to reinforcing steel than ordinary concrete, with this difference reaching up to 7%. At the same time, it has been found that when treated with an aggressive environment, SCC shows higher strength and durability compared to ordinary concrete compositions at the same water-cement ratio. The difference reaches up to 12%, which makes the SCC compositions suitable for use in constant and/or periodic exposure to aggressive environments [35]. Laboratory studies of the abrasion resistance of concrete show that it depends on the hardness of the binder, the type of coarse aggregates, and the strength of the bond between the coarse aggregates and the binder on the one hand and the bond strength between the concrete mix and the reinforcing steel on the other [35].

Optimal limits for the volume of the total amount of paste and the ratio between the additives have been established to achieve maximum resiliency and stability of the concrete mix, which guarantees its dense structure, preventing water and aggressive agents from entering [36, 37].

The above mentioned studies show that SCC has properties that make it applicable for road surfaces.

4. CONCLUSION

Researches related to road surface wear has been studied and systematized. Important characteristics and their influence by factors such as formulation composition, grain size, presence of air voids, frost resistance, noise generation, etc. have been identified.

Due to its good mechanical characteristics, SCC is increasingly entering construction. Studies related to the development of SCC with different formulation compositions have been studied. It has been established that its properties, such as a very good dense structure that prevents the entry of water and aggressive agents, which provides it with durability and a longer life, as well as its significantly greater load-bearing capacity, make it a suitable material for road surfaces.

The results of the studied tests [30-37] provide grounds to propose the use of SCC for parts of the road surface, as well as for various elements and facilities on the road. Researches should continue with a different formulation, with the emphasis on its properties essential for its application as a road surface.

Given the growing amount of household waste, when experimenting with new formulations, it is good to emphasize the use of recycled materials in their composition, especially those with low toxicity. This will lead to saving large amounts of energy, as well as reducing the consumption of natural resources for the production of new materials.

REFERENCES

[1] https://www.jouav.com/blog/road-construction.html

[2] L. Ntziachristos and P. Boulter, “Road vehicle tyre and brake wear. Road surface wear”, EMEP/EEA emission inventory guidebook, Copenhagen (Denmark), (2009)

[3] U. Olofsson, M. Tu, O. Nosko, Y. Lyu, and S. Dizdar, “A pin-on-disc study of airborne wear particle emissions from studded tyre on concrete road contacts”, Wear, Vol. 410-411, pp. 165-172, https://doi.org/10.1016/j.wear.2018.06.012, (2018)

[4] B. Snilsberg, R. G. Saba, and N. Uthus, “Asphalt pavement wear by studded tires – Effects of aggregate grading and amount of coarse aggregate”, Proceedings of the 6th Eurasphalt & Eurobitume Congress, Prague (Czech Republic), https://dx.doi.org/10.14311/EE.2016.072, (2016)

[5] R. Gayfutdinov, G. Bajmukhametov, and E. Hafizov, “Pavement wear process and abrasive wear resistance of asphalt concrete”, Proceedings of 2nd International Scientific Conference on Socio-Technical Construction and Civil Engineering (STCCE – 2021), Kazan (Russia), https://doi.org/10.1051/e3sconf/202127402008, (2021)

[6] G. Y. Lee, H. Dharan, and R. O. Ritchie, “A physically-based abrasive wear model for composite materials”, Wear, Vol. 252(3-4), pp. 322-331, https://doi.org/10.1016/S0043-1648(01)00896-1, (2002)

[7] A. Hassan, M. Mahmud, N. Adi, N. Rahmat, M. Hainin, and R. P. Jaya, “Effects of Air Voids Content on the Performance of Porous Asphalt Mixtures”, Journal of Engineering and Applied Sciences, Vol. 11(20), pp. 11884-11887, (2016)

[8] A. M. Zaltuom, “A Review Study of The Effect of Air Voids on Asphalt Pavement Life”, Proceedings of the First Conference for Engineering Sciences and Technology, Al-Khums (Libya), https://doi.org/10.21467/proceedings.4.29, (2018)

[9] J. Hu, P. Liu, and B. Steinauer, “A study on fatigue damage of asphalt mixture under different compaction using 3D-microstructural characteristics”, Frontiers of Structural and Civil Engineering, Vol. 11(3), pp. 329-337, https://doi.org/10.1007/s11709-017-0407-9, (2017)

[10] L. Chakar, “Features of the concept of an asphalt mixture for a wearing course with slag”, Proceedings of National Transport Infrastructure Conference with International Participation, Sofia (Bulgaria), pp. 74-82, (2019)

[11] D. Fullova and D. Durcanska, “Laboratory Measurements of Particulate Matter Concentrations From Asphalt Pavement Abrasion”, Civil and Environmental Engineering, Vol. 12(2), pp. 94-102, https://doi.org/10.1515/cee-2016-0013, (2016)

[12] D. Durcanska, “Analysis of Particulate Matter Composition”, Communications - Scientific Letters of the University of Zilina, Vol. 12(3A), pp. 17-22, https://doi.org/10.26552/com.C.2010.3A.17-22, (2010)

[13] D. Siebert, “How wear affects road surface texture and its impact on tire/road noise”, PhD Thesis, Norwegian University of Science and Technology, Trondheim (Norway), (2017)

[14] M. Kloth, K. Vancluysen, F. Clement, and L. Ellebjerg, “Local Noise Action Plans - Recommendations from the SILENCE project”, Forum des Laboratoires Nationaux Routiers, Bergisch Gladbach (Germany), (2008)

[15] R. Evensen, “Environmentally Friendly Pavements - Final Report”, Norwegian Public Roads Administration - Road Directorate, Oslo (Norway), (2009)

[16] T. S. Berge, F. Haukland, and A. Ustad, “Environmentally friendly pavements - Results from noise measurements 2005-2008”, SINTEF ICT, Trondheim (Norway), (2009)

[17] R. G. Saba, “Miljøvennlige vegdekker - Spesialdekker - Poroelastiske dekker”, Statens vegvesen, Oslo (Norway), (2006)

[18] M. Låg, T. Myran, M. Refsnes, P. E. Schwarze, and J. Øvrevik, “Veidekker: Svevestøv og helse”, Nasjonalt folkehelseinstitutt, Oslo (Norway), (2004)

[19] B. Snilsberg, “Pavement wear and airborne dust pollution in Norway - Characterization of the physical and chemical properties of dust particles”, PhD Thesis, Norwegian University of Science and Technology, Trondheim (Norway), (2008)

[20] C. N. Abou Sleiman, X. Shi, and D. G. Zollinger, “An Approach to Characterize the Wear ability of Concrete Pavement Surface Treatments”, Transportation Research Record, Vol. 2673(1), pp. 230–239, https://doi.org/10.1177/0361198118821668, (2019)

[21] ACI Committee 201, “Guide to Durable Concrete, ACI Manual of Concrete Practice, Part 1”, American Concrete Institute, Farmington Hills, MI (USA), (2001)

[22] P. Smith and R. Schonfeld, “Studies of Studded-Tire Damage and Performance in Ontario During the Winter 1969-1970”, Transportation Research Board, Washington D.C. (USA), (1971)

[23] K. J. Hode, “Resistance of Various Types of Bituminous Concrete and Cement Concrete to Wear Studded Tires”, Transportation Research Board, Washington D.C. (USA), (1971)

[24] B. D. Scott and M. Safiuddin, “Abrasion Resistance of Concrete – Design, Construction and Case Study”, Concrete Research Letters, Vol. 6(3), pp. 136–148, (2015)

[25] A. B. Harnik, U. Meier, and A. Rosli, “Combined Influence of Freezing and Deicing Salt on Concrete – Physical Aspects”, Durability of Building Materials and Components, West Conshohocken, PA (USA), pp. 474–484, (1980)

[26] M. A. Mohammed and Y. M. Al-Mala, “Freeze-Thaw Performance of Low-Cement Content Stabilized Soils for Containment Applications”, MASc Thesis, Dalhousie University, Halifax, NS (Canada), (2015)

[27] J. Dipayan, “Concrete Scaling – A Critical Review”, Proceedings of the 29th Conference on Cement Microscopy, Greensburg, PA (USA), (2007)

[28] E. Horszczaruk, “The Model of Abrasive Wear of Concrete in Hydraulic Structures”, Wear, Vol. 256(7–8), pp. 787–796, https://doi.org/10.1016/S0043-1648(03)00525-8, (2004)

[29] A. García, D. Castro-Fresno, J. A. Polanco, and C. Thomas, “Abrasive Wear Evolution in Concrete Pavements”, Road Materials and Pavement Design, Vol. 13(3), pp. 534–548, https://doi.org/10.1080/14680629.2012.694094, (2012)

[30] H. Okamura and M. Ouchi, “Self-compacting concrete”, Journal of Advanced Concrete Technology, Vol. 1(1), pp. 5-15, https://doi.org/10.3151/jact.1.5, (2003)

[31] G. Broceta, M. Malesev, V. Radonjanin, S. Supic, A. Savic, I. Lukic, A. Cumbo, and M. Krndija, “Effect of aggregate origin on freeze/thaw resistance of self-compacting concrete with and without a de-icing agent”, Building Materials and Structures, Vol. 66(4), pp. 21-35, https://doi.org/10.5937/GRMK2300013B, (2023)

[32] P. Tosheva, D. Bojadjieva, and B. Petrov, “Self-compacting concrete – a research of concrete mixture and hardened concrete”, Proceedings of the 12th International Scientific Conference VSU' 2012, Sofia (Bulgaria), Vol. 2, pp. 37-42, (2012)

[33] EFNARC, “Specification and Guidelines for Self-Compacting Concrete”, EFNARC, Farnham (UK), (2002)

[34] P. Stefanova, “Self-compacting concrete – a research of the bond strength with reinforcement”, Proceedings of the 13th International Scientific Conference VSU'2013, Sofia (Bulgaria), Vol. 3, pp. 162-167, (2013)

[35] P. Stefanova, B. Paleva-Kadiyska, and S. Bojadjiev, “Self-compacting concrete – application in transport facilities”, Proceedings of VII Scientific Conference Machinery and Construction Technologies in Transport, Sofia (Bulgaria), pp. 28–34, (2024)

[36] B. Petrov and P. Stefanova, “Influence of composition factors on the mobility and stability of self-compacting concrete mix”, Annual of the University of Architecture, Civil Engineering and Geodesy, Sofia (Bulgaria), (2014)

[37] B. Petrov and P. Stefanova, “Study of structural parameters of self-compacting concrete composition”, Annual of the University of Architecture, Civil Engineering and Geodesy, Sofia (Bulgaria), (2014)

SESSION G: CIVIL ENGINEERING

Comparative study of bleeding process in ordinary and self-compacting concrete – influence on the reliability of reinforced concrete elements

Petya Stefanova[1*]

[1] Todor Kableshkov University of Transport, Faculty of Machinery and Construction Technologies in Transport, Department of Mechanics, Sofia, Bulgaria

ARTICLE INFO

*** Correspondence:** p_stefanova@vtu.bg

DOI: https://doi.org/10.46793/ET26.G02S **PAGES:** G7 – G12

ABSTRACT

Worldwide there is an intensive development of the construction sector, which is invariably associated with the building of various reinforced concrete elements and facilities. Increasingly attractive projects, significant support distances imply the use of large cross-sections of reinforced concrete, which are made with a high percentage of reinforcement. Their precise technological implementation and correct composition contribute to high durability and minimizing maintenance costs. This article examines the features of the bleeding process, which invariably accompanies all concrete works. Its influence on finished reinforced concrete elements is analyzed. A comparison of the bleeding process in ordinary concrete and self-compacting concrete is made, taking into account its influence on the bond strength between the two types of concrete and reinforcing steel. Conclusions are drawn and recommendations are given for the applicability of the two types of concrete, especially in elements with great thickness and/or height.

KEYWORDS

Self-compacting concrete, bleeding effect, bond strength concrete-reinforcement, durability

1. INTRODUCTION

In recent decades, large-scale construction projects combined with design standards have led to an increase in the amount of reinforcement used in new concrete structures. As a result are produced elements with a high percentage of reinforcement and small distances between the reinforcing bars.

A mandatory process during producing of new concrete elements is the compaction of the fresh concrete mix. For this purpose, needle vibrators, external formwork vibrators and other technological approaches are used, which aim is to penetrate the concrete everywhere into the formwork and to ensure its qualitative compaction. The compaction process becomes difficult by complex formwork shapes and large amounts of reinforcement. Improper or insufficient compaction of the fresh concrete mix can lead to visible surface defects, but also to reduced bond strength between concrete and reinforcing steel inside the element.

While visible defects can be treated and reduced, deep defects appear at a later stage, when the facility is already in use, and often cannot be repaired. Additional costly processes are required to strengthen the existing structure, and if this is impossible, the construction service life is reduced significantly.

For this reason, for elements with a high percentage of reinforcement, are recommended and applied specific highly workable concretes, which kind of concrete can fill the formwork without the need for vibration. An example of such a type of construction material is self-compacting concrete \SCC\.

A research team including K. Maekawa, K. Ozawa, M. Ouchi and H. Okamura [1, 2] developed a solution to the problem of poorly compacted reinforced concrete elements by proposing the use of self-compacting concrete. Its use in the construction sector was reported for the first time in Japan in the construction of the abutments of the Akashi-Kaikyo Bridge, built in April 1998.

The wider use of SCC can lead to shorter construction times and to reduce the need for skilled workers for concrete placement [5].

The purpose of this article is to summarize the influence of bleeding effect in conventional concrete and SCC, providing recommendations for their applicability in modern reinforced concrete structures.

2. FEATURES OF SELF-COMPACTING CONCRETE

2.1. Paper Length, Size and Orientation

SCC is a highly plastic mixture that does not segregate during placement and can fill formwork without the need for additional external intervention, i.e. there is no vibration. It has been found that the use of this type of concrete reduces noise and working hours [8]. The fresh mix of SCC is more resistant to segregation [5] and is suitable for use in places with narrow openings where placing conventional concrete and its vibration are difficult.

The characteristics of SCC lead to a shorter construction schedule and to produce concrete structure with better quality, which are more homogeneous [5]. Using this type of concrete, the occurrence of deficiencies and defects in the element due to bleeding or segregation, which are caused by improper vibration activity, can be avoided.

The applicability of SCC occurs after a properly determined design mix that guarantees the "flowability" of the concrete and prevents segregation.

When determining the mix proportion for SCC, certain features must be taken into account, which are given in the EF-NARC prescriptions [3,4], as follows:

- Total powder content - (380÷600) kg/m^3 includes the amount of cement and fine mineral admixture;
- Mandatory use of fine mineral admixture – in the ratio of water to total powder content by volume within the limits W/P=(0.85÷1.1). Fine additives are considered particles with a size ≤0.125mm;
- Coarse aggregate with a limited size of up to 20 mm and according to some authors up to 16 mm maximum size – quantity (750÷1000) kg/m^3;
- Water - (150÷210) dm^3/m^3, its quantity being strictly monitored throughout the entire process of preparing the SCC mixture;
- Mandatory use of superplasticizers (SP) – chemical additives that improve the workability of SCC.

The fresh mix of SCC is required to be subjected to some more specific tests to confirm its ability to pass through obstacles, to be resistant to segregation, etc. Some of these tests are specific only to this type of concrete and have their normative characteristics according to BDS EN – spreading test, L-box test, etc. [6, 7].

3. BLEEDING AND ITS INFLUENCE ON THE BOND STRENGTH BETWEEN CONCRETE AND REINFORCEMENT

3.1. Nature of Bleeding

After the fresh concrete mix is placed in the formwork and vibrated, there is a period of time in which it is not yet hardened. During this period, the heavy particles of the concrete mix – coarse aggregate and sand, begin to settle in depth. This leads to the expulsion of free water (not included in chemical compounds) to the upper layers of the placed fresh concrete.

Total bleeding is divided into two components [10] – internal and external. External bleeding is the amount of water that reaches the surface of the concrete and this water is visible.

Part of the free water, when moving upwards, encounters obstacles – coarse aggregate grains or reinforcing bars, and the water is retained under them. This is described like internal bleeding. This process is invisible to the naked eye, but is an inevitable part of the concrete placing process. A large part of the water by internal bleeding is retained around the coarse aggregate grains and/or reinforcing bars [10]. Over time, the water evaporates, leaving voids in the interior of the finished reinforced concrete element. This leads to permanent porosity inside the material, and due to the created cavities, there is no connection between concrete and reinforcing steel, as well as between cement stone and coarse aggregates. This, according to some authors [9], leads to a decrease in the strength between the concrete components and to a general decrease in the bond strength between concrete and reinforcing steel.

The combination of bleeding, settling of heavy components in the fresh concrete mix, and segregation, according to some authors [11], are the main cause for the formation of voids under the horizontal reinforcement in reinforced concrete elements. In the absence of vibration, as in the case of SCC, some of the processes that cause or increase bleeding are missing.

The methods to determining the total bleeding are divided into two directions – direct measurement of the surface water discharged - external bleeding and a methodology for determining the internal bleeding. The above-mentioned team of authors [10] suggests that the measurement of internal bleeding should be done using a non-contact laser sensor, due to

the impossibility of direct visual inspection. The value of internal bleeding [10] can be defined as the difference between an empirically determined total bleeding and the measured external bleeding.

3.2. Top Bar Effect

The process of settling of heavy components in the fresh concrete mix and bleeding have an adverse effect on the upper horizontal reinforcing bars. As a result of these processes, voids are formed under the reinforcing bars, especially under those bars, which are located at a height of more than 30 cm from the bottom of the formwork. In scientific literature, this phenomenon is known as the "top bar effect". The formed voids reduce the bond strength between concrete and reinforcement, which leads to an overall decrease in the load-bearing capacity of the element. An additional undesirable effect is when the already formed cavities are filled with a solution from an aggressive environment, which causes corrosion in the concrete-reinforcement contact area. After conducting researches, it was found that in ordinary concrete, bleeding process and the top bar effect are significant.

The influencing factors are mostly dependent on the composition of the concrete, the bar position and the shape of the reinforcing bars. To minimize the influence of the "top bar effect", mainly caused by bleeding, Eurocode 2 introduces a coefficient that increases the overlapping and anchoring length of the reinforcing bars for elements with a height greater than 30 cm. The length is increased by about 30%.

3.3. Factors That Influence the Bleeding Process

Previous studies [13] have shown that the main factor influencing bleeding process and hence the bond strength between concrete and reinforcement, is the water-cement ratio. At a W/C ratio of 0.65, the difference in the bond strength between SCC and ordinary concrete reaches 30%. When reducing W/C=0.45, the difference drops to 10% between the two types of concrete. At both W/C ratios studied above, SCC is less affected by bleeding, which can be explained by the increased amount of fine mineral admixture in its composition. This admixture accumulates water and prevents or minimizes the bleeding process.

The above-mentioned authors [10] also confirmed that a higher W/C ratio leads to a significant increase in internal bleeding. Their exams are provided only with ordinary concrete compositions. Fig. 1 shows the degree of bleeding /external and internal/, expressed in percentages for different W/C ratios. Composition 2 has W/C=0.469, composition 3 – W/C=0.469 and composition 5 – W/C=0.550.

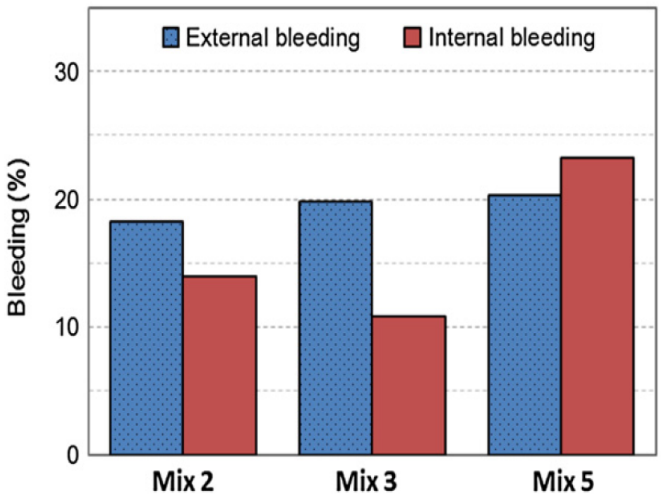


Figure 1: External and internal bleeding, reported in % for ordinary concrete compositions with different W/C ratios and different coarse aggregates grain shapes [10]

The shape of the coarse aggregates also has a significant influence on the internal bleeding. Concrete mix 2 and 3 have the same W/C ratio, but in composition 3 the coarse aggregates grains have a regular spherical shape. As can be seen from Fig. 1, composition 3 has lower internal bleeding values compared to all other tested compositions of ordinary concrete.

3.4. Bond Strength Between Concrete and Reinforcement by Bleeding

Studies have been conducted and they show that the decrease in bond strength between concrete and reinforcement along the height of the element can vary from 30% to 55%. The loss in bond strength is being greater in ordinary concrete, which is vibrated after being placed in the formwork.

The effect of bleeding on the bond strength was investigated for different diameters of reinforcing steel [8]. For some of the most commonly used cross-sections, namely Ø 10, Ø 12 and Ø 16 mm, the dependence has been confirmed that with increasing height from the bottom of the formwork, the bond strength decreases - the reason is cavities resulting from

bleeding. For reinforcement diameter of Ø 16 mm, the drop in the bond strength reaches, according to some authors [8], up to 106%, and for a diameter of Ø 12 mm – up to 63% in ordinary concrete. These values are 20% and 9% respectively for the same diameters, but for SCC. As expected, the loss in the bond strength is the smallest for the smallest diameter. This pattern can be explained by the fact that a larger diameter represents a greater obstacle that free water may encounter as it moves toward the surface.

To study the change in the bond strength caused by bleeding, some authors [8, 12] created formwork forms for a column, with reinforcing bars with the same diameter anchored along the height. In the formwork are created the so-called lower, middle and upper zones, in which the change in the bond strength is studied – Fig. 2. In both experimental setups, the exact contact length between concrete and reinforcing steel was ensured using plastic tubes that were placed on the reinforcing bar. Additionally, the holes are filled to prevent mortar to entering in the hole between the reinforcement and the plastic pipe. A tensile force is applied to the free end of the reinforcement emerging from the formwork to determine the bond strength.

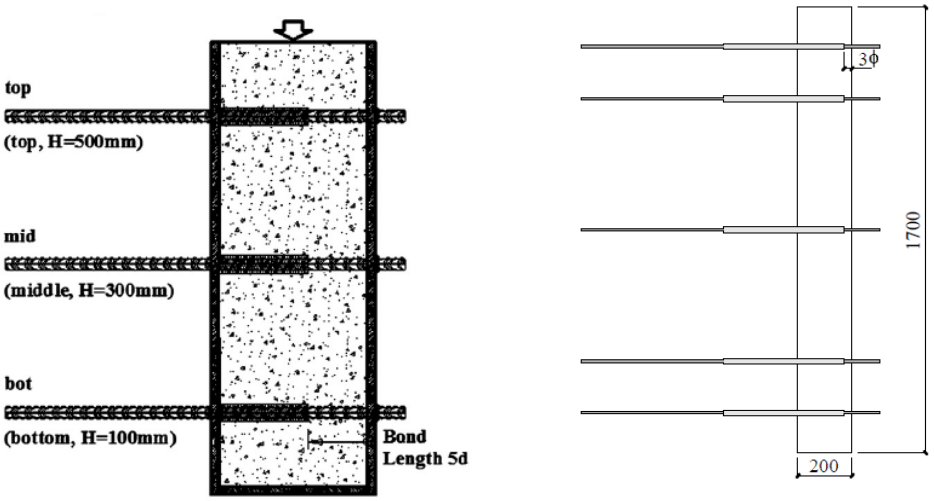


Figure2:Experimental setup A [12] and B [8] for determining the variation in bondstrength inheight affected by bleeding

After conducting the pullout test, all literature sources reviewed in this article observe a similar pattern – with increasing the height, ordinary concrete compositions are more strongly influenced by bleeding. Accordingly, the decrease in bond strength is greater by ordinary concrete. In Fig. 3, the authors [12] observed that the ratio of the bond strength for upper/lower reinforcing bars at different W/C ratios in the SCC compositions is close to unity /1/. In SCC mixes with increasing W/C ratio, this ratio /upper-lower reinforcement/ remains around unity /1/, which proves that decrease in bond strength along the height of the column is insignificant. For the SCC compositions, the ratio of the bond strength between the upper and lower reinforcement does not fall below 0.7 at all used W/C ratios. In ordinary concrete, even at the lowest W/C ratio, the bond strength between the upper and lower reinforcement reaches 0.7 once – see Fig.3.

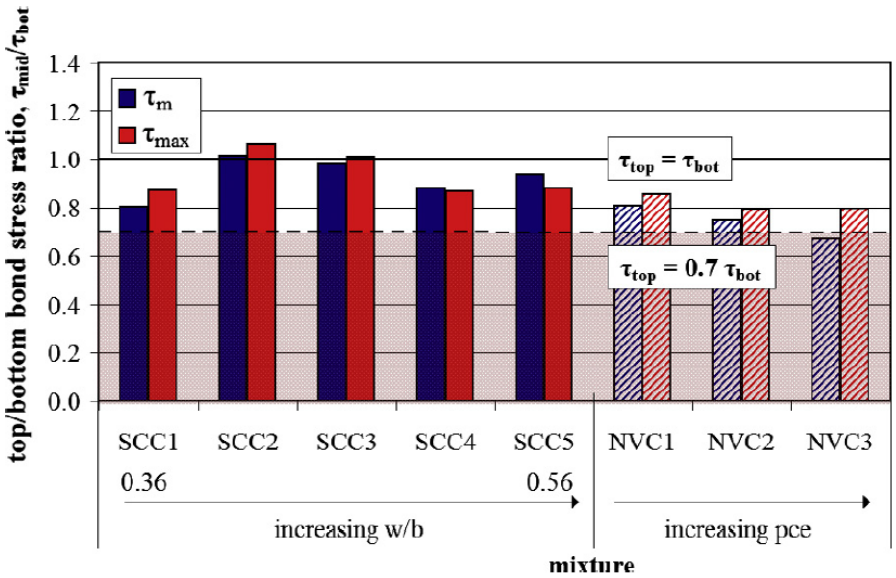


Figure 3:Relationship between the bond strength of top and bottom reinforcement for SCCand normal vibratedconcrete (NVC) [12]. “PCE” in figure means superplasticizer

3.5. Corrosion of Concrete and Reinforcing Steel Due to Bleeding

Bleeding, as mentioned above, leads to the so-called “top bar effect”. It contributes to create voids under the reinforcing steel at a height of its placement more than 30 cm above the bottom of the formwork [11]. Studies have shown that such bars corrode much faster in the presence of aggressive environments than others located lower in the formwork. The presence of voids caused by bleeding can initiate corrosion in reinforcing bars at significantly lower levels of aggressive agents. In bars affected by the "top bar effect" is observed significant corrosion situated on the underside of the bar, while their top surface remains intact - Fig. 4.

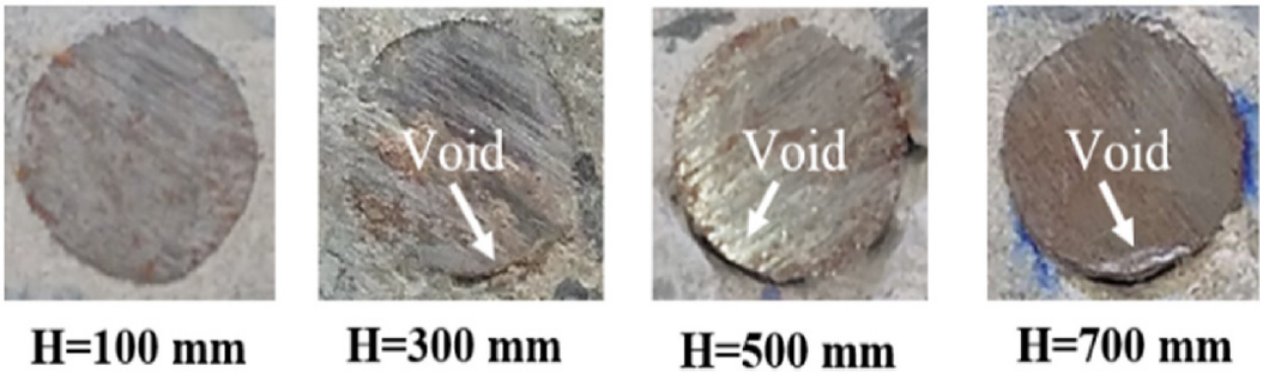


Figure 4:Cross-section of bars located at different heights of 100 mm, 300 mm, 500 mm and 700 mm f rom the bottom of the formwork [11]

4. CONCLUSIONS

As a result of the literature review and comparative analysis of bleeding process in conventional and self-compacting concrete, the following conclusions can be made:

- The ratio of water to total powder content W/P affects bleeding, as when the W/P ratio decreases, the free water, which is the basis of bleeding, decreases. As a result of the above, the total bleeidng in the fresh concrete mix decreases;
- The bleeding strongly influences the formation of voids in the concrete structure. This defect is noticeably observed in bars that are located more than 30 cm above the bottom of the formwork. As the height at which the longitudinal reinforcement is placed increases, the amount of voids under it, caused by bleeding, increases;
- Due to bleeidng and the so-called “top bar effect”, as the diameter of the reinforcing steel increases, the bond strength between reinforcement and the concrete decreases in height. This is more noticeable in conventional concrete, while in SCC the influence of the diameter is less pronounced;
- The reinforcing steel, which is located higher in the reinforced concrete element, is subject to a stronger influence of aggressive agents from the environment, due to the presence of more voids around it and connected cavities through which the aggressive environment can enter the interior;
- As the height of the reinforcement increases in both types of concrete, a decrease in the bond strength is observed. This effect is more noticeable in ordinary concrete;
- The need to increase the length of overlapping and anchoring of the reinforcement by 30% is confirmed. This is necessary if the reinforcement is placed at a height of more than 30 cm from the bottom of the formwork. This percentage increase is completely suitable for ordinary concrete compositions, and it can be further increased if the expected height of the concreted element and the diameter of the reinforcement used are more than 20mm;
- To minimize the effects of bleeding, it is recommended that for tall elements or those with a thickness of over 30 cm, bars with a smaller diameter be used for upper reinforcement, if the structural requirements allow.

REFERENCES

[1] H. Okamura, M. Ouchi, “Self-compacting concrete”, Journal of Advanced Concrete Technology, Vol.1, No.1, 5-15, (2003), https://doi.org/10.3151/jact.1.5

[2] H. Okamura, M. Ouchi, „Self-compacting concrete. Development, present and future“, 1st International RILEM Symposium on SCC, (1999)

[3] Specification and Guidelines for Self-Compacting Concrete, EFNARC, (2002)

[4] The European Guidelines for Self-Compacting Concrete, (2005)

[5] R. Eshafani, M. Reza Kianoush, M. Lachemi, „Top-bar effect of steel bars in self-consolidating concrete (SCC)." Cement and Concrete Composites, Vol. 30. No.1, 52-60, (2008), https://doi.org/10.1016/j.cemconcomp.2007.05.012

[6] БДС EN 12350-8 – Testing concrete mix - Part 8: Self-compacting concrete – Slump flow test

[7] БДС EN 12350-10 – Testing concrete mix – Part 10: Self-compacting concrete –L-box test;

[8] P. Desnerck, G. De Schutter, and L. Taerwe, “Influence of Bar Diameter on Top-Bar Effect in Self-Compacting Concrete”, International RILEM Conference on Advances in Construction Materials Through Science and Engineering, ISBN: 978-2-35158-116-2, (2011)

[9] J. Kim, H. Yim, S. Kwon, “Quantitive measurement of the external and internal bleeding of conventional concrete and SCC”, Cement and Concrete Composites, Vol.54, 34-39, (2014), https://doi.org/10.1016/j.cemconcomp.2014.05.006

[10] H. Yim, J. Kim, H. Kwak, J. Kim, “Evaluation of internal bleeding in concrete using a self-weight bleeding test”, Cement and Concrete Research, Vol.53, 18-24, (2013), https://doi.org/10.1016/j.cemconres.2013.05.015

[11] Y. Wang, A. Mukherjee, A. Castel, “Ultasonic guided waves for monitoring incipient in reinforced concrete with top-bar defect”, Cement and Concrete Composites, Vol.141, (2023), https://doi.org/10.1016/j.cemconcomp.2023.105116

[12] K. Trezos, I. Sfikas, Ch. Pasios, “Influence of water-to-binder ratio on top-bar effect and on bond variation across length in Self-Compacting Concrete specimens”, Cement and Concrete Composites, Vol.48, 127-139, (2013), https://doi.org/10.1016/j.cemconcomp.2013.11.012

[13] M. Valcuende, C. Parra, (2009), “Bond behavior of reinforcement in self-compacting concretes”, Construction Building Materials, Vol.23, No.1, 162-170, (2009), https://doi.org/10.1016/j.conbuildmat.2008.01.007

SESSION G: CIVIL ENGINEERING

Biocomposite based on corn cobs as an innovative thermal insulation material

Jovana Bojković[1], Nenad Stojić[1*], Vesna Bulatović[2], Iva Despotović[1], Stefan Mihajlović[1]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Kraljevo, Serbia

[2] University of Novi Sad, Faculty of Technical Sciences, Department of Civil Engineering and Geodesy, Novi Sad, Serbia

ARTICLE INFO

*** Correspondence:** stojic.n@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.G03B **PAGES:** G13 – G18

ABSTRACT

Under contemporary conditions of increasingly intensive climate change and growing energy demands, the role of renewable energy sources has become increasingly important. However, their availability and efficiency largely depend on variable meteorological conditions, which directly affect the stability and reliability of energy systems based on renewable energy sources. At the same time, climate change is altering energy consumption patterns, particularly through the growing need for adequate thermal insulation of buildings in order to maintain energy efficiency. Based on these facts, this research investigates the possibility of using ground corn cobs as an aggregate in the production of thermal insulation materials. The aim is to develop a composite material with satisfactory thermal insulation properties that could be applied in construction practice. This study contributes to the development of innovative materials that support the improvement of building energy efficiency.

KEYWORDS

Biocomposite; Thermal conductivity; Thermal insulator; Energy efficiency

1. INTRODUCTION

Increasing energy consumption in the building sector represents one of the key challenges of modern society, considering that the construction industry accounts for a significant share of global energy consumption and greenhouse gas emissions. Consequently, growing attention is being directed toward the development of materials that, in addition to possessing satisfactory technical characteristics, also exhibit a favorable environmental profile throughout their entire life cycle.

Materials produced from agricultural waste are of particular interest because they enable the efficient utilization of locally available resources while reducing the amount of waste generated in the environment. In this context, biocomposite thermal insulation materials represent an important direction in the development of sustainable construction, with significant potential for application in energy-efficient and environmentally friendly buildings. This research is specifically focused on investigating the feasibility of using corn cobs as a component of an innovative thermal insulation biocomposite.

2. THERMAL INSULATION MATERIALS

Thermal insulation is one of the key factors affecting the energy efficiency of buildings, as it directly contributes to reducing heat losses through the building envelope and lowering the overall energy consumption required for heating and cooling. Despite this, construction practice is still dominated by the use of synthetic thermal insulation materials such as expanded polystyrene (EPS), extruded polystyrene (XPS), polyurethane (PU), and mineral wool (Figure 1).

These materials exhibit excellent thermal insulation properties, with thermal conductivity coefficients typically ranging from 0.025 to 0.048 W/mK. However, their production requires substantial amounts of energy and the use of synthetic chemicals. Furthermore, at the end of their service life, they pose significant environmental challenges due to their limited biodegradability and the accumulation of waste.



Figure 1: Synthetic Thermal Insulation Materials

In addition, certain synthetic insulation materials may release harmful and toxic gases when exposed to high temperatures, posing risks to both human health and the environment. Although mineral wool demonstrates good fire resistance, its moisture absorption capacity may negatively affect the durability and thermal performance of building components. Due to these disadvantages, research in the field of construction materials is increasingly focused on the development of sustainable alternatives based on natural and renewable raw materials.

2.1. Natural Biocomposite Materials

Biomass-based materials represent one of the most promising groups of sustainable thermal insulation materials. Their main advantages include the renewability of raw materials, biodegradability, and lower energy requirements during the manufacturing process. Furthermore, the use of these materials contributes to waste reduction and lower carbon dioxide emissions, thereby supporting the principles of sustainable development in the construction industry.

Natural thermal insulation materials include reed boards, cork, sheep wool, straw bales, wood fibers, and corn-based materials (Figure 2). Although their thermal conductivity coefficients are generally higher than those of synthetic insulation materials, typically ranging from 0.037 to 0.145 W/mK, these materials offer significant environmental and health-related benefits. Their application is still relatively limited due to insufficient practical experience in the design and construction of such systems. However, recent research indicates a growing potential for their wider use in the construction sector.



Figure 2: Natural Thermal Insulation Materials

A particularly important area of development is represented by biocomposite materials, which are produced by combining natural fibers and natural binders. Biocomposites are defined as materials composed of at least two components of natural origin, where the fibers act as aggregates, while the binders enable the formation of a compact material structure. The analysis of heat transfer processes through such materials is of crucial importance for determining their thermal insulation properties and their potential application in building structures.

Numerous studies have confirmed the feasibility of using biocomposites for thermal insulation purposes. Cerezo et al. [1] analyzed the thermal and mechanical properties of hemp–lime biocomposites and demonstrated that increased porosity improves thermal insulation performance but may simultaneously reduce the mechanical strength of the material. Domenico Curto et al. [2] investigated the acoustic and thermal properties of the same type of biocomposite, confirming its suitability for application in energy-efficient and acoustically optimized buildings.

Similar studies were conducted by Maris Sinka and Genādijs Šahmenko [3], who examined the possibility of producing thermal insulation biocomposites from locally available raw materials, with particular emphasis on reducing CO_2 emissions throughout the material life cycle. Sophie Walker et al. [4] highlighted the significant influence of manufacturing processes and compaction levels on the internal structure and thermal insulation characteristics of hemp–lime composites. Bojković et al. [5] demonstrated that wood sawdust-based biocomposites can represent a sustainable alternative to conventional thermal insulation materials, offering satisfactory thermophysical properties and considerable environmental benefits. Florence Collet and Sylvie Pretot [6] further analyzed the effects of density and moisture content on thermal insulation performance and developed predictive models describing material behavior under different service conditions. Černý et al. [7] investigated the hygrothermal and physico-mechanical properties of flax shiv-based composite materials,

showing that agricultural waste can have significant potential for use in sustainable thermal insulation systems. Marín-Calvo et al. [8] explored the production of thermal insulation materials based on recycled cellulose and rice husks, concluding that agro-industrial waste can provide a sustainable alternative to conventional insulation materials while reducing environmental impacts.

Based on the available research, it can be concluded that natural and biocomposite thermal insulation materials represent a significant opportunity for the advancement of sustainable construction. Although certain limitations regarding their practical application still exist, ongoing research and technological development continue to improve their properties and expand their potential use in the energy-efficient buildings of the future.

3. METHODS FOR DETERMINING THE THERMAL TRANSMITTANCE COEFFICIENT

In accordance with modern requirements for sustainable and energy-efficient construction, building design must comply with current regulations that define minimum energy performance requirements and permissible levels of energy consumption. Compliance with energy-efficiency requirements is achieved through the application of relevant regulations that define the energy performance characteristics of buildings, methodologies for calculating the thermal properties of building components, and maximum allowable energy consumption values for heating, cooling, ventilation, domestic hot water preparation, and lighting.

In addition to requirements related to building energy consumption, special attention is given to the analysis of water vapor transmission through external building elements and structures adjacent to unheated spaces. Building assemblies must be designed to prevent both surface and interstitial condensation, since the presence of moisture can significantly affect structural durability and reduce the thermal insulation performance of materials. In cases where condensation occurs, it is necessary to determine the drying period required for the structure to regain its functionality and long-term stability.

Within this research, double-wall assemblies incorporating a biocomposite thermal insulation material were analyzed. The overall thermal transmittance coefficient (U-value) was determined using the URSA software package. As the thermal insulation layer, a biocomposite panel designated as K was used, produced from ground corn cobs, an agricultural waste material. In order to improve the thermal insulation properties of the material, expanded polystyrene (EPS) granules were incorporated into the biocomposite mixture, while a lime–gypsum paste in a 4:1 ratio was used as the binding component. In this way, a composite material with enhanced thermophysical characteristics and potential application in energy-efficient building structures was developed.

The percentage composition of the individual components of biocomposite K was as follows:

1. ground corn cobs – 53.5%
2. granules of expanded polystyrene (EPS) – 13.5%
3. lime-gypsum paste – 33%

Table 1 presents the particle size distribution of the aggregate used for the production of biocomposite K. The largest proportion consists of particles in the 4–8 mm size range (approximately 62%), while the 2–4 mm fraction accounts for about 28.8%. The particle size of the expanded polystyrene (EPS) granules was 0.3 cm. Figure 3 shows the appearance of the materials used in the composition of biocomposite K.

Table 1: Average Particle Size Distribution of the Aggregate

Biocomposite/size (mm)	**<0,4**	**0,4-0,5**	**0,5-1**	**1-2**	**2-4**	**4-8**	**>8**
Corn-cobs	3,1	0,5	9,4	3,6	28,8	62,3	8,9



Figure 3: Components of Biocomposite K

Methods for determining the physical and thermophysical characteristics of biocomposites were thoroughly analyzed in the study by Bojković et al. [9], where experimental testing on samples of identical composition was used to determine

the thermal conductivity coefficient of the biocomposite, K, which was then applied in the analysis of the energy performance of wall assemblies. To assess the energy efficiency of the analyzed constructions, the overall thermal transmittance coefficient (U-value) (1) was determined using the standard expression [10]:

$$U = \frac{1}{\frac{1}{\alpha_i} + \sum \frac{d_m}{\lambda_m} + \frac{1}{\alpha_e}} \quad (1)$$

where $R_{si} = 1/\alpha_i$ and $R_{se} = 1/\alpha_e$ are the thermal resistance coefficients for the inner and outer wall sides, λ_m is the coefficient of thermal conductivity, and d_m is the thickness of the m-th layer of the wall.

4. RESULTS AND DISCUSSION

The overall thermal transmittance coefficient (U-value) was determined for two types of wall assemblies: the first consisting of a ceramic block, and the second using a concrete panel as the primary structural layer. In both cases, a biocomposite K-based board was used as the thermal insulation layer. In addition to the calculation of the thermal transmittance coefficient, a water vapor diffusion analysis was also performed to assess the possibility of condensation occurring within the structure.

4.1. Wall Assembly 1 – Biocomposite K / Ceramic Block

The characteristics of the wall assembly incorporating a 50 mm thick thermal insulation layer made of biocomposite K and a ceramic block as the main structural material are presented in Table 2. The values of the thermal conductivity coefficient and the water vapor diffusion resistance factor for the remaining layers of the structure were taken from the relevant national regulation, while for biocomposite K, a thermal conductivity value of λ = 0.125 W/mK was used [10]. The water vapor diffusion resistance factor was determined using the dry cup method, as presented in the doctoral dissertation [11].

Table 2: Wall Assembly with Thermal Insulation Biocomposite K and Ceramic Block as the Base Material

	Layers	d (m)	λ (W/mK)	μ	R (m^2K/W)
	i				0.125
1	Extension plaster	0.02	0.85	15	0.024
2a	Ceramic block	0.19	0.61	6	0.311
2b	Cement–lime plaster	0.19	0.87	15	0.218
3	TI biocomposite K	0.05	0.125	7.8	0.400
4	External finishing plaster	0.03	0.87	15	0.034
	e				0.04

The obtained value of the overall thermal transmittance coefficient for the analyzed assembly is:

U = 1.082 W/m²K

Compared with conventional thermal insulation materials, it was determined that when EPS of the same thickness is used, the U-value would be 0.572 W/m²K, while for mineral (rock) wool it would be 0.490 W/m²K. Based on these results, it can be concluded that the wall assembly incorporating biocomposite K exhibits approximately twice the poorer thermal insulation performance compared to assemblies with EPS and rock wool of the same thickness.

In order to achieve equivalent thermal transmittance values, it is necessary to increase the thickness of the biocomposite layer to approximately 0.16 m relative to EPS, or 0.19 m relative to rock wool.

Table 3: Water Vapor Diffusion Resistance Factor for Assembly 1 with Ceramic Block

	Layers	d (m)	λ (W/mK)	μ	R (m^2K/W)	r (m)	t (°C)	p'(kPa)	p(kPa)
	i				0.125		20.00	2.337	1.285
							15.36	1.749	1.285
1	Extension plaster	0.02	0.85	15	0.024	0.30	15.36	1.749	1.285
							14.48	1.650	1.136
2	**Ceramic block**	0.19	0.61	6	0.311	1.14	14.48	1.650	1.136
							2.92	0.752	0.570
3	**TI biocomposite K**	0.05	0.125	7.8	0.400	0.39	2.92	0.752	0.570
							-11.93	**0.219**	**0.376**
4	External finishing plaster	0.03	0.87	15	0.034	0.45	-11.93	**0.219**	**0.376**
							-13.21	0.195	0.152
	e				0.04		-13.21	0.195	0.152
							-14.70	0.169	0.152

The vapor diffusion analysis was performed under design conditions: indoor temperature ti = 20°C, outdoor temperature te = −14.7°C, indoor relative humidity φi = 55%, and outdoor relative humidity φe = 90%. Based on the results presented in Table 3, condensation was found to occur between the third and fourth layers of the construction. However, the structure's drying period is 15 days, which is acceptable under the applicable regulations.

4.2. Wall Assembly 2 – Biocomposite K / Concrete Panel

The characteristics of the second wall assembly, in which the primary structural element is a concrete panel, are presented in Table 4. The thermophysical parameters of the remaining layers were adopted in the same manner as for the previous assembly, while the thickness of the thermal insulation layer was kept unchanged.

Table 4: Wall Assembly 2 with Thermal Insulation Biocomposite K and Concrete Panel as the Base Material

	Layers	d (m)	λ (W/mK)	μ	R (m^2K/W)
	i				0.125
1	Extension plaster	0.02	0.85	15	0.024
2	**Concrete panel**	0.1	2.04	60	0.049
3	**TI biocomposite K**	0.05	0.125	7.8	0.400
4	External finishing plaster	0.03	0.87	15	0.034
	e				0.04

The obtained value of the overall thermal transmittance coefficient is:

$$\mathbf{U = 1.488\ W/m^2K}$$

In the case of EPS with the same thickness, the U-value would be 0.670 W/m^2K, while for rock wool it would be 0.559 W/m^2K. The results indicate that the wall assembly consisting of a concrete panel and biocomposite K exhibits approximately two and a half times poorer thermal insulation performance compared to assemblies with conventional thermal insulation materials. In order to achieve adequate thermal transmittance values, it would be necessary to increase the thickness of the biocomposite thermal insulation layer.

Table 5: Water Vapor Diffusion Resistance Factor for Assembly 2 with Concrete Panel

	Layers	d (m)	λ (W/mK)	μ	R (m^2K/W)	r (m)	t(°C)	p'(kPa)	p(kPa)
	i						20.00	2.337	1.285
					0.125		13.55	1.557	1.285
1	Extension plaster	0.02	0.85	15	0.024	0.30	13.55	1.557	1.285
							12.33	1.430	1.238
2	**Concrete panel**	0.1	2.04	60	0.049	6.00	12.33	1.430	1.238
							9.80	1.212	0.285
3	**TI biocomposite K**	0.05	0.125	7.8	0.400	0.39	9.80	1.212	0.285
							-10.85	0.240	0.224
4	External finishing plaster	0.03	0.87	15	0.034	0.45	-10.85	0.240	0.224
							-12.63	0.205	0.152
	e				0.04		-12.63	0.205	0.152
							-14.70	0.169	0.152

The vapor diffusion analysis showed that no condensation occurs within this assembly. However, despite the more favorable moisture behavior, the assembly with the concrete panel exhibited less favorable thermal transmittance values compared to the assembly with the ceramic block.

5. CONCLUSION

Based on a comparison with reference cases using EPS and rock wool, it can be concluded that wall assemblies incorporating biocomposite thermal insulation material exhibit up to twice as high U-values, indicating poorer thermal insulation performance compared to conventional materials of the same thickness.

Nevertheless, despite these thermal limitations, the obtained results confirm that biocomposites based on corn cobs have potential for application in sustainable construction, particularly in buildings where environmental considerations and the use of renewable raw materials are of primary importance. Although their thermal insulation performance currently cannot fully compete with conventional insulation materials, opportunities for improving composition, density, and internal structure indicate significant potential for further development and application of these biocomposites in energy-efficient buildings of the future.

ACKNOWLEDGEMENTS

This study was supported by the Serbian Ministry of Science, Technological Development and Innovation from the grant 451-03-34/2026-03/200108.

REFERENCES

[1] V. Cerezo, F. Collet, S. Pretot et al., "Thermal and structural properties of a hemp–lime biocomposite", Construction and Building Materials, Vol. 48, pp. 745–754, https://doi.org/10.1016/j.conbuildmat.2013.07.096 , (2013)

[2] D. Curto, A. Guercio and V. Franzitta, "Investigation on a Bio-Composite Material as Acoustic Absorber and Thermal Insulation", Energies, Vol. 13(14), Article 3699, https://doi.org/10.3390/en13143699, (2020)

[3] M. Sinka and G. Šahmenko, "Sustainable Thermal Insulation Biocomposites from Locally Available Hemp and Lime", Environment. Technology. Resources. Proceedings of the International Scientific and Practical Conference, Vol. 1, pp. 73–77, https://doi.org/10.17770/etr2013vol1.828 , (2013)

[4] S. Walker, P. Walker et al., "The influence of the casting process on the internal structure and physical properties of hemp-lime", Materials and Structures, Vol. 50(2), Article 108, https://doi.org/10.1617/s11527-016-0976-4 , (2017)

[5] J. Bojković, N. Stojić and M. Marašević, "Determination of thermal conductivity coefficient of a biocomposite based on wood sawdust", Thermal Science, https://doi.org/10.2298/TSCI250412143S, (2025)

[6] F. Collet and S. Pretot, "Thermal conductivity of hemp concretes: Variation with formulation, density and water content", Construction and Building Materials, Vol. 65, pp. 612–619, https://doi.org/10.1016/j.conbuildmat.2014.05.039 , (2014)

[7] P. Cerny, M. Babenko, P. Bartos, Y. Kononets, P. Kriz, R. Rabenseifer and F. Spalek, "Complex Study of the Composite Building Material Based on Flax Straw Organic Waste: Hygrothermal and Physicochemical Properties", Waste and Biomass Valorization, Vol. 15, pp. 2231–2247, https://doi.org/10.1007/s12649-023-02273-7, (2024)

[8] N. Marín-Calvo, S. González-Serrud and A. James-Rivas, "Thermal insulation material produced from recycled materials for building applications: cellulose and rice husk-based material", Frontiers in Built Environment, Vol. 9, Article 1271317, https://doi.org/10.3389/fbuil.2023.1271317, (2023)

[9] J. Bojković, M. Marašević, N. Stojić, V. Bulatović and B. Radičević, "Thermal and Sound Characterization of a New Biocomposite Material", Materials, Vol. 16, Article 4209, https://doi.org/10.3390/ma16124209, (2023)

[10] "Pravilnik o energetskoj efikasnosti zgrada", Službeni glasnik Republike Srbije, br. 61/11, Available: https://www.mgsi.gov.rs/lat/dokumenti/pravilnik-o-energetskoj-efikasnosti-zgrada [Accessed: 10-Apr-2023], (2011)

[11] J. Bojković, "Development of Biomass-Based Composites for the Improvement of Thermal and Sound Properties of Building Objects", Faculty of Technical Sciences, Novi Sad (Serbia), PhD Thesis, (2024)

SESSION G: CIVIL ENGINEERING

Recycled steel in contemporary architecture: Sustainable potentials and applications through case studies

Snežana Bajić[1*], Sara Milošević[1]

[1] Academy of Technical and Art Applied Studies, School of Civil Engineering and Geodesy, Belgrade, Serbia

ARTICLE INFO

*** Correspondence:** mitrozs73@gmail.com

DOI: https://doi.org/10.46793/ET26.G04B **PAGES:** G19 – G26

ABSTRACT

Rapid urbanization and growing environmental challenges demand the use of sustainable materials, which has become a key aspect of contemporary architectural and engineering design. Recycled steel stands out due to its ability to retain mechanical properties through multiple recycling cycles while significantly reducing energy consumption and carbon emissions compared to primary production. This paper examines the potential of recycled steel in architecture and construction, focusing on its role in structural systems and spatial design. Selected case studies, including Hearst Tower, Olympic Stadium, and Empire State Building, illustrate diverse approaches to its application, from advanced structural concepts to adaptive reuse strategies. The research is based on a comparative analysis of structural systems, levels of recycled material use, and their impact on energy efficiency and overall sustainability. Particular attention is given to the relationship between structural logic, economic considerations, and architectural expression. The findings suggest that recycled steel not only reduces environmental impact but also enables the development of flexible, adaptable, and resource-efficient structures, positioning it as a key material in sustainable and circular architecture.

KEYWORDS

Recycled steel, Structural engineering, Structural systems, Material performance, Energy efficiency, Low-carbon construction, Case studies

1. INTRODUCTION

Contemporary construction and architecture face increasing demands to reduce environmental impact, particularly regarding resource consumption and greenhouse gas emissions. The building sector accounts for a significant share of global energy use and CO_2 emissions, making it a key area for implementing sustainable development principles [1, 2]. In this context, the selection of construction materials and optimization of structural systems play a crucial role in minimizing the overall environmental footprint of buildings [3]. Recent studies have also highlighted the importance of recycled materials in concrete production, demonstrating that recycled aggregates can maintain satisfactory mechanical and flexural performance while reducing construction waste and resource consumption [4, 5].

Recycling of steel scrap enables energy savings of up to 74% compared to conventional steel production from iron ore. In the case of aluminium, these savings are even more pronounced, recycling requires approximately 95% less energy than production from bauxite. Such reductions in energy consumption directly contribute to lower greenhouse gas emissions, given that industrial energy is still largely derived from fossil fuels.

In addition to its environmental benefits, metal recycling has significant economic advantages. Reduced consumption of energy and raw materials leads to lower production costs per unit, thereby increasing the overall efficiency of industrial processes.

At the same time, steel recycling contributes to the conservation of natural resources. Estimates indicate that recycling one tonne of steel saves approximately 1.5 tonnes of iron ore and about 0.5 tonnes of coal that would otherwise be required in primary production. In this way, the burden on mining activities and their environmental impacts is reduced,

Steel, as one of the most widely used structural materials, offers excellent mechanical properties; however, its primary production is energy-intensive and associated with high carbon emissions [6]. Its capacity for multiple recycling cycles

without loss of performance makes recycled steel particularly suitable for circular economy approaches [7]. The use of recycled steel significantly reduces energy consumption and emissions compared to virgin production while maintaining structural integrity [8].

Most recycled steel is processed in electric arc furnaces (EAF - Electric Arc Furnace). Within a large furnace, a powerful electric arc raises the temperature to the melting point of steel (above 1500°C). Steel scrap is melted and transformed into liquid steel. At this stage, alloying elements or additional raw materials are often introduced as needed to achieve the desired chemical composition. For example, in the production of stainless steel, chromium, nickel, and other elements are added, often sourced from recycled alloys. A key advantage of the electric arc furnace is that it allows the charge to consist of up to 100% scrap steel, unlike the blast furnace, which relies on iron ore. The process is more energy-efficient and flexible, particularly for smaller batch production. During melting, impurities are removed through refining processes-materials such as lime are added to bind undesirable elements (e.g., sulphur and phosphorus), forming slag that is subsequently removed. The resulting molten metal is then cast (typically using continuous casting technology) into semi-finished products such as billets, slabs, or blooms.

In contemporary engineering practice, recycled steel is increasingly integrated into load-bearing system designs, both through the use of secondary raw materials and strategies for reusing existing structural components [9, 10]. This approach not only reduces construction waste but also enables flexible, adaptable, and resource-efficient structural solutions.

The aim of this paper is to analyze the potential of recycled steel in modern structural systems through selected case studies, with a focus on its impact on sustainability, energy efficiency, and structural optimization.

2. METHODOLOGY

Interest in “green” building has increased significantly, with a growing emphasis on the evaluation of entire buildings as well as individual materials. The LEED rating system of the U.S. Green Building Council, along with other product classification systems, raises numerous questions that are directly or indirectly related to the selection of metals.

Some of these considerations include the recycled content of materials, the potential for product reuse, impacts on energy and water savings, the likelihood of harmful emissions, effects on indoor air quality, thermal comfort, maintenance requirements, and the influence on interior lighting conditions.

Indeed, engineers require a comprehensive database that provides objective life-cycle environmental assessments for all construction materials. Unfortunately, such a resource is not fully available. Existing data for different materials are often not directly comparable, making it essential to critically examine and question available information. Common questions and answers regarding the environmental performance of steel are discussed in the following section. [11]

This study adopts a mixed-methods approach to evaluate the potential of recycled steel in contemporary structural systems, combining literature analysis, comparative case studies, and performance assessment. The methodology is structured in three main phases:

2.1. Literature Review

The literature review presented in this chapter summarizes the most significant global experiences in the field of recycled steel management. The published studies are primarily based on mathematical and theoretical models. An overview of developments across all continents provides a comprehensive picture of the current state of recycled steel use in construction worldwide. On this basis, it is possible to compare these findings with the current situation and level of implementation in the Republic of Serbia.

An extensive review of recent publications (2015-2024) on recycled steel, circular economy practices, and sustainable construction materials was conducted. This phase identifies the environmental, structural, and economic benefits of using recycled steel, as well as the main barriers to its implementation in modern construction [6-9]. The review also informs the selection of case studies by highlighting projects with documented use of recycled steel.

When addressing demolition and construction waste arising from demolition, construction, renovation, adaptation, and reconstruction activities, the regulatory framework in the Republic of Serbia is defined through several key legal instruments. These include the Law on Planning and Construction, the Law on Waste Management, incorporating the Waste Catalogue and the Regulation on Categories, Testing, and Classification of Waste, as well as the Law on Environmental Protection. However, this regulatory framework remains insufficiently defined, particularly with regard to construction waste, and especially steel waste that could be recycled. As a result, clear procedures for the management and treatment of construction waste cannot be adequately established, despite the fact that such waste represents one of the major environmental protection challenges in the Republic of Serbia. Although numerous laws and regulations in the field of environmental protection have been adopted, they lack clearly unified objectives, and the outcomes of their implementation remain unsatisfactory.

2.2. Case Study Selection and Data Collection

Selected case studies include high-profile construction projects where recycled steel has been significantly applied, such as Hearst Tower, Olympic Stadium in London, and Empire State Building. Data on structural systems, steel sourcing, percentage of recycled content, and environmental performance indicators were collected from technical reports, journals, and official project documentation [1-3, 10].

2.3. Comparative Analysis and Performance Assessment

The study conducts a comparative analysis of the case studies to evaluate:

- The proportion of recycled steel in structural elements
- Energy savings and reductions in embodied carbon [7,9]
- Structural efficiency and design adaptability
- Cost implications and potential barriers for widespread adoption

Quantitative indicators such as embodied energy, CO_2 footprint, and steel reuse percentages are analyzed to assess the sustainability performance of recycled steel relative to conventional steel.

This methodology allows a systematic evaluation of recycled steel's technical and environmental potential, providing actionable insights for architects, structural engineers, and policymakers aiming to integrate sustainable materials into construction projects.

3. RECYCLED STEEL IN CONSTRUCTION

Recycled steel has gained increasing attention in the construction industry due to its technical reliability, environmental benefits, and economic potential [1-3]. Multiple studies demonstrate that steel can be recycled indefinitely without significant loss of mechanical properties, making it one of the most sustainable structural materials for modern construction [6, 7] (see Figure 1 for an example of recycled steel ready for use).

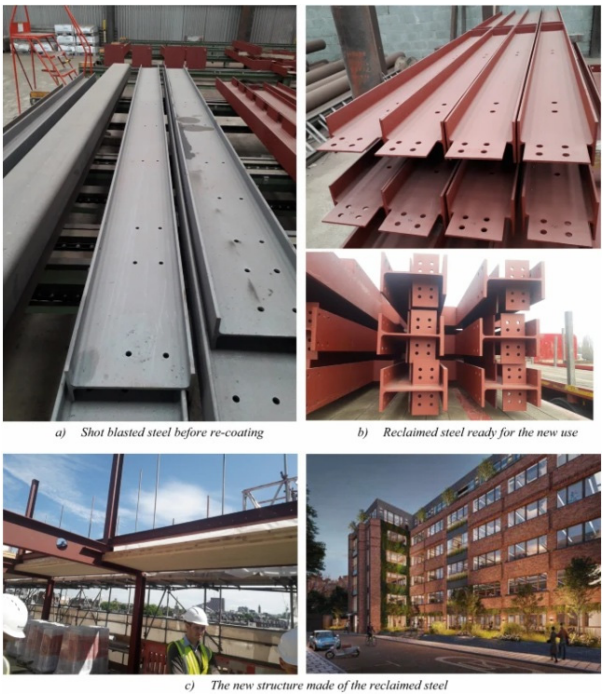


Figure 1: Recycled steel ready for construction use. [F1]

3.1. Environmental Impacts

A key driver for the adoption of recycled steel is the reduction of the carbon footprint and embodied energy of construction projects. Drewniok et al. [7] quantified energy and emission savings, showing that using recycled steel in structural elements can reduce embodied carbon by up to 60% compared to virgin steel production. Similarly, Göswein et al. [9] emphasize that lifecycle assessments (LCA) are critical for evaluating the true sustainability of steel in buildings, particularly when reuse and recycling are integrated into the design phase (illustrated in Figure 2).



Figure 2: Reduction in embodied carbon and energy savings using recycled steel. [F2]

3.2. Technical and Structural Considerations

From a structural perspective, recycled steel retains its load-bearing capacity, ductility, and fatigue resistance [6, 14]. Pombo et al. [14] analyzed different reuse strategies for structural steel elements, highlighting the importance of design for deconstruction (DfD) to maximize reuse potential without compromising structural performance. Addis and Schouten [10] further support that adaptive structural solutions, such as modular frames and prefabricated steel components, facilitate both reuse and recycling (as shown in Figure 3).



Figure 3: Modular and prefabricated steel systems facilitating reuse. [F3]

3.3. Economic and Practical Barriers

Despite the clear benefits, several barriers to widespread adoption exist. Dunant et al. [6] identify logistical, regulatory, and market-related constraints, including limited supply chains for reclaimed steel, inconsistent quality standards, and additional labor costs associated with deconstruction and sorting. The European Commission [8] stresses the importance of policy incentives and standardized guidelines to support the integration of recycled steel into mainstream construction practices (Figure 4 illustrates these challenges visually).



a) Steel from the demolished building *b) New building with reclaimed steel*

Figure 4: Logistical and market challenges in recycled steel adoption. [F4]

3.4. Integration into Circular Economy and Sustainable Design

Recent research emphasizes that recycled steel is a cornerstone of circular economy approaches in construction [8, 12, 15]. By designing buildings for reuse, recovery, and material traceability, architects and engineers can minimize resource depletion and reduce environmental impacts while maintaining structural performance [12, 15]. Ürge-Vorsatz et al. [15] highlight pathways to decarbonize the building sector, showing that recycled steel combined with energy-efficient design can significantly contribute to global climate targets.

The reviewed literature confirms that recycled steel provides high structural performance, significant environmental benefits, and potential cost savings, but its implementation requires careful planning, supply chain coordination, and policy support. Integration into circular economy frameworks and design-for-reuse strategies emerges as a central recommendation for future construction projects [1-3, 6-15].

4. CASE STUDY: SELECTION AND ANALYSIS OF RECYCLED STEEL PROJECTS

The selection of case studies for this research was guided by the need to analyze projects that exemplify the integration of recycled steel in different types of structural systems, while providing sufficient data to assess environmental and technical performance. High-profile projects were chosen for their innovative structural designs, significant use of recycled steel, and availability of detailed sustainability reports, enabling a comprehensive evaluation of both engineering and environmental outcomes.

One of the primary examples is the Hearst Tower in New York, designed by Norman Foster, which utilizes a steel diagrid frame in which approximately 85-90% of the structural steel is recycled [7,16]. The diagrid system not only optimizes material usage but also allows for substantial integration of recycled steel without compromising load-bearing performance. This project demonstrates the synergy between advanced structural design and sustainable material selection, highlighting the potential for high-rise constructions to significantly reduce embodied carbon while maintaining architectural expression [3, 7, 16, 12].

The Olympic Stadium in London, constructed for the 2012 Games, represents a contrasting typology where a lightweight modular steel frame was used. Approximately 70-75% of the steel employed was recycled [14, 17]. The modular prefabricated components facilitated design for deconstruction, allowing steel elements to be reused or recycled after the stadium's temporary function. This case illustrates how temporary large-scale structures can embrace circular economy principles, demonstrating both environmental benefits and economic feasibility [10, 12, 14, 17].

Finally, the Empire State Building renovation project in New York provides insight into how historic or existing structures can integrate recycled steel. During selective retrofitting, roughly 60% of the steel components were sourced from recycled material [13, 15, 18]. Lifecycle assessments indicate measurable reductions in embodied energy and CO_2 emissions, showing that even older buildings can contribute to sustainable construction when targeted interventions are implemented. This case highlights the importance of planning and coordination to incorporate recycled steel without affecting structural integrity or operational performance [13, 15, 18].

The analysis of these projects shows that structural typology, project scale, and design strategy influence the achievable proportion of recycled steel and its associated sustainability benefits. Innovative high-rise and modular systems tend to accommodate higher recycled content, while renovation projects demonstrate the potential for incremental improvements in material sustainability. Across all cases, it is evident that design for deconstruction, prefabrication, and lifecycle planning are critical to maximizing the environmental and economic advantages of recycled steel [7-15, 18, 19].

4.1. Comparative Analysis of Case Studies

In the Republic of Serbia, there are no buildings explicitly designed as being constructed from recycled steel, yet recycled steel is widely used through standard industrial production flows of structural steel. Most steel profiles and elements installed in industrial halls, warehouses, canopies, commercial buildings, and bridges contain a significant proportion of recycled material, as steel is produced by melting scrap steel in electric arc furnaces. Although the use of recycled steel in Serbia is not formally documented or promoted through LEED or other certification systems, it is implicitly incorporated in practice. The lack of regulation, clear guidelines, and a secondary materials market prevents buildings from being classified as "100% recycled," posing a challenge for the further development of circular construction in the country.Therefore, the further comparative analysis will be conducted on three well-known buildings worldwide.

The three selected projects - Hearst Tower, Olympic Stadium, and the Empire State Building renovation - exhibit different structural typologies, scales, and strategies for integrating recycled steel. Comparing these cases allows us to evaluate how structural design, project scope, and lifecycle planning influence the proportion of recycled steel used and the environmental benefits achieved.

The Hearst Tower demonstrates the potential of high-rise constructions to achieve the highest recycled steel content (~85-90%) due to its optimized diagrid frame and material-efficient geometry. The design not only reduces the amount of steel required but also facilitates the use of recycled material without compromising structural integrity, resulting in approximately 26% reduction in embodied carbon [7, 16, 12].

The Olympic Stadium employs a modular, lightweight steel frame where 70-75% of steel is recycled. Prefabricated elements and design for deconstruction allow material recovery after the event, yielding a CO_2 reduction of approximately 30% while providing economic benefits through reusable components [10, 12, 14, 17]. This project exemplifies circular economy principles applied to temporary or semi-permanent large-scale structures.

The Empire State Building renovation illustrates the integration of recycled steel in the retrofitting of an existing structure. With ~60% recycled steel in replaced components, the project achieves a 22% reduction in embodied carbon [13, 15, 18]. Although the overall recycled content is lower than in new constructions, targeted reuse and careful planning demonstrate how even historic buildings can contribute to sustainable material management.

A summary of these projects and their key sustainability indicators is presented in Table 1, which highlights the variation in recycled steel percentages, CO_2 reduction, and structural strategies. The comparative analysis shows that higher percentages of recycled steel are achievable in new, purpose-designed structures that incorporate modularity or innovative geometry, while renovation projects can still benefit from selective reuse strategies. The key factors influencing successful implementation include structural typology, design-for-reuse strategies, and early integration of lifecycle assessment [7-15, 18, 19]. These insights are essential for engineers and architects aiming to maximize the environmental benefits of recycled steel in various project types.

Table 1: Comparative Analysis of Case Studies

Project	Structural System	Recycled Steel (%)	CO_2 Reduction (%)	Key Sustainability Feature
Hearst Tower, NY	Diagrid high-rise frame	85-90	26	Optimized geometry, material efficiency, high-rise integration
Olympic Stadium, London	Modular lightweight frame	70-75	30	Prefabrication, design for deconstruction, circular economy approach
Empire State Building Renovation, NY	Reinforced steel retrofitting	60	22	Targeted retrofitting, selective reuse, adaptation of existing structure

4.2. Results & Discussion

The comparative analysis of the selected case studies (Table 1) demonstrates clear trends regarding the use of recycled steel in contemporary construction. Projects designed as new structures, such as the Hearst Tower and Olympic Stadium, achieve higher percentages of recycled steel (85-90% and 70-75%, respectively) compared to the Empire State Building renovation (~60%). This indicates that structural design and planning from the outset significantly influence the potential for incorporating recycled steel [7, 16, 12, 17]. High-rise diagrid systems, like that of the Hearst Tower, allow material efficiency through geometry optimization, while modular systems, as in the Olympic Stadium, facilitate reuse and disassembly, aligning with circular economy principles [10, 12, 14, 17].

The analysis also highlights the relationship between recycled steel content and CO_2 reduction. Although the Olympic Stadium achieved slightly higher CO_2 savings (30%) than the Hearst Tower (26%), the difference can be attributed to the temporary nature and modular design of the stadium, which allowed optimized steel use and efficient post-use recovery [14, 17]. The Empire State Building renovation, despite being a retrofit, demonstrates that even existing structures can significantly reduce embodied carbon through targeted replacement with recycled steel (22% reduction), emphasizing the potential of incremental interventions in older buildings [13, 15, 18].

From a structural engineering perspective, all case studies confirm that recycled steel retains full load-bearing capacity and mechanical performance, supporting its application in both high-rise and modular designs [6, 7, 14]. Additionally, the integration of design-for-deconstruction (DfD) strategies in the Olympic Stadium and the lifecycle planning in all projects demonstrates that early consideration of sustainability measures is essential to maximize environmental benefits [10, 12, 14, 19].

The findings suggest several practical recommendations for engineers and architects aiming to enhance the sustainability of steel structures:

1. Incorporate recycled steel at the design stage, particularly in high-rise and modular projects, to maximize environmental gains.
2. Apply design-for-deconstruction principles to allow easy reuse or recycling of steel elements after the building's lifecycle or functional adaptation.
3. Perform lifecycle assessments (LCA) to quantify embodied energy and CO_2 savings and guide material selection and structural optimization [9, 13, 15].
4. Consider selective retrofitting for existing structures, enabling significant reductions in carbon footprint without major structural modifications [13, 18].

Overall, this analysis confirms that recycled steel is a technically viable, environmentally beneficial, and economically feasible material for modern construction. When combined with innovative structural design, modularity, and lifecycle planning, recycled steel can substantially reduce the environmental impact of buildings while maintaining structural performance and adaptability [7-19].

5. CONCLUSION

This study demonstrates the technical, environmental, and practical benefits of using recycled steel in contemporary construction projects.

The use of recycled steel in contemporary construction projects offers a range of technical, environmental, and practical benefits. Technically, recycled steel maintains the same structural strength, durability, and versatility as primary steel, allowing it to be used in a wide variety of applications, from industrial halls and bridges to commercial buildings. Environmentally, it significantly reduces energy consumption, up to 74% less for steel compared to production from iron ore and lowers greenhouse gas emissions, while also conserving natural resources such as iron ore and coal. Practically, recycled steel supports cost efficiency, as lower energy and raw material requirements reduce production costs, and its inherent recyclability minimizes construction waste and landfill burden. Together, these advantages make recycled steel a key material in sustainable and circular construction practices.

The analysis of Hearst Tower, Olympic Stadium, and the Empire State Building renovation shows that recycled steel can be successfully integrated in high-rise, modular, and retrofit projects, achieving reduced embodied carbon, energy savings, and material efficiency [7-19]. High-rise and modular designs allow for higher percentages of recycled steel, reaching up to 90%, while renovation projects benefit from targeted retrofitting and selective reuse strategies [7, 13, 15, 19].

The comparative analysis confirms that structural system design, early planning, and lifecycle considerations are key factors in maximizing the environmental and economic advantages of recycled steel. Approaches such as design-for-deconstruction, prefabrication, and modularity enhance recyclability and support circular economy principles. Lifecycle assessment ensures that material choices lead to measurable reductions in CO_2 emissions, guiding sustainable engineering decisions [9-12, 14, 15, 17, 19].

The findings highlight that incorporating recycled steel from the earliest stages of design is essential for optimizing both structural efficiency and environmental performance. Even existing buildings can contribute significantly to sustainability goals through careful retrofitting and selective material reuse. Furthermore, policies and standards that support the use of recycled steel, along with the development of supply chains for reclaimed materials, are important to facilitate broader adoption.

In conclusion, recycled steel represents a viable and effective material for sustainable construction, bridging structural performance with environmental responsibility. Its integration into building design supports the circular economy and enables resilient, adaptable, and resource-efficient structures. By combining innovative structural solutions with material reuse strategies and comprehensive lifecycle planning, recycled steel is positioned as a cornerstone of future sustainable engineering practices.

REFERENCES

[1] International Energy Agency, “Iron and Steel Technology Roadmap: Towards More Sustainable Steelmaking,” IEA, Paris, (2020)

[2] United Nations Environment Programme, “2023 Global Status Report for Buildings and Construction,” UNEP, Nairobi, (2023)

[3] World Steel Association, “Climate Change and the Production of Iron and Steel,” World Steel Association, Brussels, (2022)

[4] B. Milošević, Ž. Petrović, M. Mijalković and S. Ranković, “Mechanical characteristics of self-compacting concrete made with coarse aggregate obtained from concrete prefabricated elements recycling,” Revista Română de Materiale / Romanian Journal of Materials, Vol. 46, No. 2, pp. 167-174, ISSN 1583-3186, (2016)

[5] Ž. Petrović, B. Milošević, M. Spasojević Šurdilović, A. Zorić and D. Turnić, “Experimental Investigation of the Flexural Performance of Continuous Self-Compacting Concrete Beams with Natural and Recycled Aggregates,” Materials, Vol. 19, p. 264, https://doi.org/10.3390/ma19020264, (2026)

[6] C.F. Dunant, M.P. Drewniok, R.C. Sansom, S.Corbey, J.M. Allwood, J. M. Cullen, “Real and perceived barriers to steel reuse across the UK construction value chain,” Resources, Conservation and Recycling, Vol. 126, pp. 118–131, https://doi.org/10.1016/j.resconrec.2017.07.036, (2017)

[7] C.F.Dunant, M.P. Drewniok, M.Sansom, S. Corbey, J.M.Cullen, J.M. Allwood, “Options to make steel reuse profitable: An analysis of cost and risk distribution across the UK construction value chain“, Journal of Cleaner Production, Vol. 183, pp. 102–111. https://doi.org/10.1016/j.jclepro.2018.02.141, (2018)

[8] European Commission, “Circular Economy Action Plan - For a Cleaner and More Competitive Europe,” Brussels, (2020)

[9] H. Feng , J. Zhao , H. Zhang , S.Zhu , D. Li , N.Thurairajah, “Uncertainties in whole-building life cycle assessment: A systematic review” Journal of Building Engineering, Vol. 50, art.104191, https://doi.org/10.1016/j.jobe.2022.104191 , (2022)

[10] C. Morgan, F. Stevenson, “Design for Deconstruction,” SEDA Design Guides for Scotland : No. 1, https://www.researchgate.net/publication/303231874_Design_for_Deconstruction, (2005)

[11] S. Bajić and M. Nikolić Topalović, "Analysis of the Possibility of Using Recycled or Reused Steel in the New Building Construction," Proceedings of the 16th International Scientific Conference "Integration, Novelty, Design, Interdisciplinarity, Sustainability" – iNDiS 2023, Novi Sad (Serbia), 16–17 November 2023, pp. 37–47, ISBN 978-86-6022-616-9, (2023)

[12] F. Pomponi and A. Moncaster, "Circular economy for the built environment: A research framework," Journal of Cleaner Production, Vol. 143, pp. 710-718, https://doi.org/10.1016/j.jclepro.2016.12.055, (2017)

[13] K. Asif, M. Muneer and R. Kelley, "Life cycle assessment: Life cycle assessment: A case study of a dwelling home in Scotland," Building and Environment, Vol. 42, No. 3, pp. 1391-1394, https://doi.org/10.1016/j.buildenv.2005.11.023, (2007)

[14] S.Vares, P.Hradil, M.Sansom, V.Ungureanu, "Economic potential and environmental impacts of reused steel structures" Structure and Infrastructure Engineering Vol. 16, pp. 1–12, https://doi.org/10.1080/15732479.2019.1662064, (2019)

[15] S. Chatterjee, B. Kiss, D.Ürge-Vorsatz, S.Teske, "Decarbonisation Pathways for Buildings," Achieving the Paris Climate Agreement Goals, Springer, https://doi.org/10.1007/978-3-030-99177-7_7, (2022)

[16] J. Allwood and M. Cullen, Sustainable Materials: With Both Eyes Open, Cambridge University Press, Cambridge, UK, (2020)

[17] J. Kim, "Olympic Structures and their Effect on the Host City's Environment, Economy, and Infrastructure ", International Journal of Civil and Structural Engineering Research ISSN 2348-7607 (Online) Vol. 11, Issue 1, pp 52-61, https://doi.org/10.5281/zenodo.7961040 , (2023)

[18] S.K.Selvaraj, T-M.Chan, "Reuse limit state: A multi-factor framework for circular steel construction: Foundational concepts and research needs" Journal of Constructional Steel Research, Vol. 245, (110471), https://doi.org/10.1016/j.jcsr.2026.110471 , (2026)

[19] H. Bartsch, "Reuse of reclaimed steel components in construction: A systematic review of potential, challenges and future directions", Structures, vol. 80 (110057), https://doi.org/10.1016/j.istruc.2025.110057 , (2025)

FIGURES REFERENCES

[F1] Springer Nature, Recycled steel ready for construction use, 2023. Available online: https://media.springernature.com/lw1200/springer-static/image/art%3A10.1007%2Fs13296-023-00778-4/MediaObjects/13296_2023_778_Fig1_HTML.jpg

[F2] BMB Steel, Construction of prefabricated steel hall, 2023. Available online: https://bmbsteel.com.vn/storage/2023/07/7633/pre-engineered-steel-building-project-by-bmb-steel.png

[F3] Cadyapi, Modular steel structural components, 2023. Available online: https://cadyapi.com.tr/uploads/images/VDrkrFRrlq2hUpWJa0201707916295.jpg

[F4] Springer Nature, Demolished and recycled steel beams, 2023. Available online: https://media.springernature.com/lw1200/springer-static/image/art%3A10.1007%2Fs13296-023-00778-4/MediaObjects/13296_2023_778_Fig2_HTML.jpg

SESSION G: CIVIL ENGINEERING

Pushover analysis with various material nonlinearity models

Marijana Janićijević Milačak[1*], Stefan Mihajlović[1], Bojan Milošević[1], Jovana Bojković[1]

[1] University of Kragujevac, Faculty of Mechanical and Civil Engineering, Kraljevo, Serbia

ARTICLE INFO

*** Correspondence:** janicijevic.m@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.G05JM **PAGES:** G27 – G34

ABSTRACT

In this paper, the nonlinear behavior of a reinforced concrete five-story building exposed to seismic action designed according to EN 1992-1-1 and EN 1998-1 is analyzed. In addition to the required load capacity, the local ductility conditions are also met in the critical sections of the structure. Nonlinear static (pushover) analysis of the structure subjected to different lateral force distributions (uniform and modal) was performed using four material nonlinearity models (infrmFB, infrmFBPH, infrmDB, and infrmDBPH) for two orthogonal directions. The analysis was carried out using the SeismoStruct software package for seismic analysis of structures. The seismic performance assessment of the structure was based on the pushover analysis results for each applied nonlinear material model.

KEYWORDS

Pushover analysis, Material nonlinearity, Seismic performance

1. INTRODUCTION

Significant global attention has been devoted to seismic protection, as experience has shown that destruction during strong earthquakes can have catastrophic proportions, with the real possibility of a large number of human victims. Given that the seismic risk in urban areas increases over time, a number of experimental and theoretical studies have been carried out in the world in order to study the behavior of structures during earthquakes [1]. Proper calculation of the seismic resistance of structures is based on the concept of designing structures to achieve acceptable behavior due to seismic effects. The main goal is for buildings to resist weak earthquakes without damage, moderate earthquakes with controlled damage to non-structural elements, and for strong earthquakes, it is necessary to prevent building collapse through controlled damage (plasticization)[2].

Due to its mechanical properties, reinforced concrete is a material that can ensure the favorable behavior of structures exposed to seismic forces [3]. If properly designed and executed, a reinforced concrete structure can develop the necessary ductility and protect against local or global collapse. Skeletal structures are the most common application of construction. They are characterized by relatively high flexibility due to the development of relatively small seismic forces. For normal loading, the structure remains in the linear-elastic range of material operation, and the relationship between load capacity and deformations is simple. In the event of an earthquake, and depending on its strength, some parts of the structure move into the non-linear area of material operation.

The main goal of modern regulations and norms for the calculation of structures in seismic areas is to ensure a sufficient capacity for dissipating seismic energy during the design seismic effect, but without a significant reduction in the local and global bearing capacity for horizontal and vertical loads, i.e., without reaching the level and distribution of damage that would cause the collapse of the building [4]. This goal is achieved by applying programmed behavior in the design of structures. In EN1998-1 [4], the concept of seismic-resistant design is based on the controlled reduction of loads through the reduction of forces, which is achieved by applying a behavior factor that practically reduces the values of the elastic response spectrum. The required bearing capacity is determined for the design level of seismic impact, which is several times lower than the values that would appear in the elastic response. This allows for a non-linear response for a true seismic effect, where the structure during an earthquake is exposed to seismic forces approximately equal to its nominal load.

The seismic analysis consists of the calculation of the dynamic characteristics of the structure, the determination of seismic forces based on the mechanical properties of the structure of the object and the specified ground displacement, and the calculation of the impact on the structure due to the action of induced seismic forces [5]. When the influences in the structure due to the action of seismic forces are determined, the dimensioning of all critical sections for the relevant combination of seismic action and other loads is carried out. After that, with the application of adequate structural solutions and the elaboration of details specific to aseismic design, the necessary bearing capacity and sufficient deformation capacity of the considered structure are ensured.

Nonlinear static analysis is a simplified method for evaluating seismic response that, along with its ease of application (compared to the nonlinear dynamic method), provides an accurate overall assessment of structural behavior [2]. Nonlinear static (pushover) analysis was employed in this study. This technique is widely recognized and utilized in numerous studies focused on evaluating the seismic performance of structures [6], [7], [8], [9], [10]. The analysis was conducted using the seismic analysis software SeismoStruct [11].

2. NONLINEAR STATIC ANALYSIS

Nonlinear static analyses are often used in seismic calculations and evaluations of structures as indicators of structural yielding and possible failure mechanisms. The main advantage of such procedures compared to linear ones is that employing nonlinear analyses, the nonlinear characteristics of the material are directly taken into account, resulting in more realistic values of internal forces and deformations of the structure [2].

The capacity of the frame structure is determined by a non-linear static ("pushover") analysis. The pushover method represents an acceptable alternative to the complex nonlinear dynamic calculation method, which uses the equations of motion [12]. With this analysis, in addition to the gravity load, the structure is also exposed to the lateral forces from the earthquake, and in this way, a capacity curve known as a "pushover curve" is obtained [5]. The capacity curve is usually determined to show the first form of structural response based on the assumption that the structural response is governed primarily by the fundamental mode of vibration. The goal of this analysis is to define the dependence of the displacement of the control node (usually a point on the last floor or roof) and the shear force at the base that causes the first yielding in the structure. Plastic hinges form when the sectional capacity of individual members is reached. The incremental lateral load continues to be applied until the "target displacement" is reached and a non-linear static capacity curve is formed. To obtain the capacity curve, which represents the horizontal displacement of the top of the structure, pushover analysis is applied up to the control displacement value corresponding to 150% of the target displacement value. Given that the structure is a system with several degrees of freedom, the pushover curve of such a system is converted into the pushover curve of an equivalent system with one degree of freedom so that the capacity of the structure can be compared with the given seismic demand [5], [13].

The force-displacement relationship can be established for any force and displacement, but is usually established for the total shear force and displacement of the top of the building. According to EN 1998-1 [4], the calculation is carried out for uniform and modal distribution of lateral forces. In the case of uniform load distribution, seismic forces in individual levels are proportional to floor masses, while in the case of modal distribution, seismic forces are distributed according to the height of the building according to the same law adopted in the elastic analysis.

With the design of the structure and adequate construction, it is necessary to prevent excessive damage to the supporting structure due to the impact of the design earthquake. Also, an appropriate degree of protection must be provided in the event of damage and limitations in the functioning of the facility. Pushover analysis evaluates the seismic performance of the object and combines the behavior of load-bearing and non-load-bearing elements, thus forming a description of the overall degree of structural damage for different levels of seismic action, which is a significant advantage of the pushover method [14].

The analysis to assess the seismic performance of the object in this work was carried out in the structural analysis program SeismoStruct [11]. SeismoStruct is a finite element-based program that has the ability to use various nonlinear modeling techniques. While calculating the behavior of spatial frames during analysis with static and/or dynamic effects, the program can analyze structural elements with the help of material and geometrically nonlinear models and calculate the response of the structure and its elements.

3. NUMERICAL ANALYSIS

A five-story reinforced concrete building (Figure 1) was analyzed in accordance with the provisions given in EN 1992-1-1 [15] and EN 1998-1 [4]. The building meets the requirements of regularity in terms of base and height. The frame construction of the building consists of six frames perpendicular to the longer side and four frames perpendicular to the shorter side of the building, with a frame spacing of 5 m in both directions. The basic dimensions of the building are 15 x 25 m, and the floor height is constant at 2.8 m. The columns are square-shaped with dimensions of 45 x 45 cm, and the beams are T and L cross-sections with dimensions: flange width 185 cm, flange thickness 16 cm, rib width 45 cm, and rib height 19 cm. The cross-sections as well as the adopted layout of the element reinforcement are shown in Figure 2.



Figure 1: Model of structure

During the analysis, the effects of seismic action and gravity (constant and variable) loads were considered. Seismic impacts were determined by multimodal analysis with a design spectrum for the horizontal direction. EN 1998-1 [4] prescribes the elastic response spectrum for horizontal seismic action with the following parameters: type 1 elastic response spectrum, soil category C (S=1.15, Tb=0.2s, Tc=0.6s, and Td=2s), maximum ground acceleration ag=0.2g, correlation factor due to damping η=1, and building significance factor γ=1. For the analyzed construction, a high ductility class DCH was adopted, for which a total behavior factor of q=5.85 was determined.

The dimensioning of the structural elements was carried out based on the authoritative combinations in the software package TOWER 8. For the dimensioning of the supporting elements, concrete C30/37 and reinforcements B500B were used. The columns are reinforced with 12Ø28 and are confined with stirrups UØ12/10cm over the length of the critical zone and over the remaining length UØ12/15cm. The beams at the support are reinforced in the upper zone with 5Ø20 and in the lower zone with 3Ø20, and in the field in the lower zone with 3Ø20 and in the upper zone with 2Ø20. Beam flanges are reinforced in the upper zone with Ø12/10cm and in the lower zone with Ø12/20cm in both orthogonal directions.

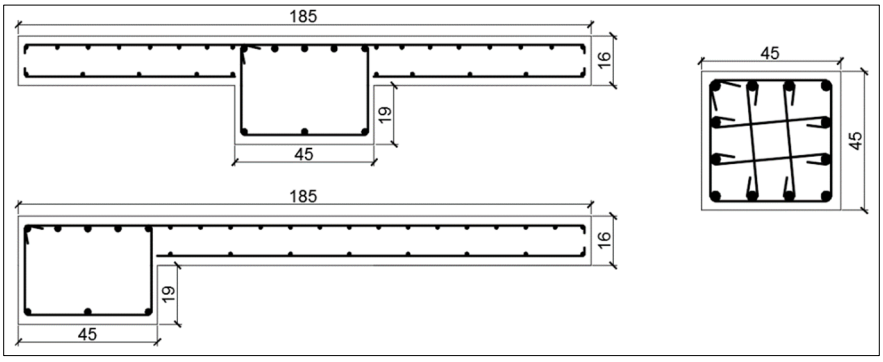


Figure 2: Cross sections of elements with adopted reinforcement

The steel and concrete models used in the SeismoStruct program are shown in Figure 3. The concrete model is based on the constitutive law for confined concrete proposed by Mander et al. [16], while the reinforcement steel is modeled using the Menegotto-Pinto [17] stress-strain relationship.

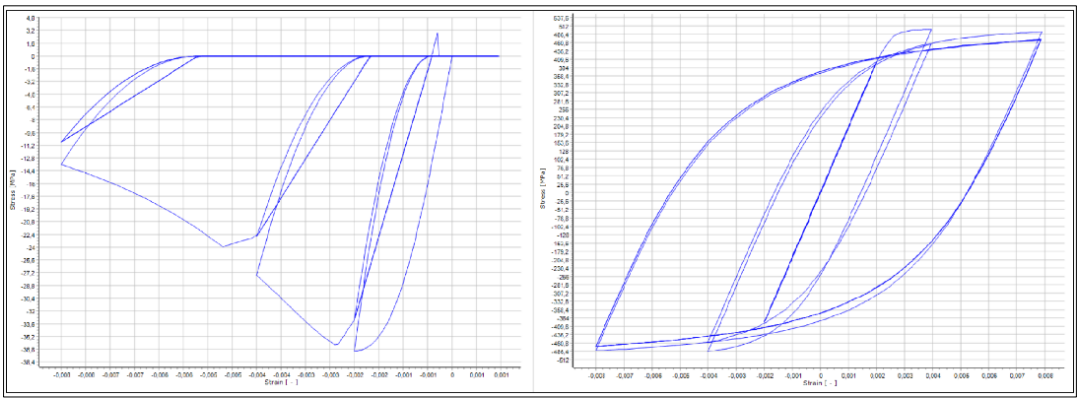


Figure 3: a) concrete and b) steel models from SeismoStruct

In addition to the required load capacity, the conditions of local ductility are also met in the critical sections (Table 1) [4], [15].

Table 1: Satisfying the condition of local ductility (EN 1998-1)

Beams	Columns
Reinforcement coefficient of the tensioned zone ρ=0,7486 % ≤ ρ_{max}=0,752 % Compression zone reinforcement coefficient ρ'=0,4489 % ≥$\rho_{tens}/2$ Length of the critical zone l_{cr}= 1 m ≥ $l_{cr,des}$ = 0,652 m	Total reinforcement coefficient ρ_{min}= 0,01 ≤ ρ = 0,0334 ≤ ρ_{max} = 0,04 Length of the critical zone l_{cl}/h_c = 2,45/0,45 = 5,44 > 3 →l_{cr} = 0,675m Fulfillment of requirements for adequate section confinement $\alpha \cdot \omega_{wd} \geq 30 \cdot \mu_\phi \cdot \nu_d \cdot \varepsilon_{sy,d} \cdot \frac{b_c}{b_0} 0{,}035$ (0,333>0,2406)

After defining all the geometric and mechanical characteristics of the structure, a non-linear static "pushover" analysis was carried out (Figures 4 and 5) [11]. The procedure was carried out for two orthogonal directions (X and Y) and two distributions of lateral forces - uniform and modal distribution. Four different models were created in the seismic analysis program to show the different effects of material nonlinearity: inelastic frame element (*infrmFB* and *infrmDB*) and inelastic plastic-hinge frame element (*infrmFBPH* and *infrmDBPH*).

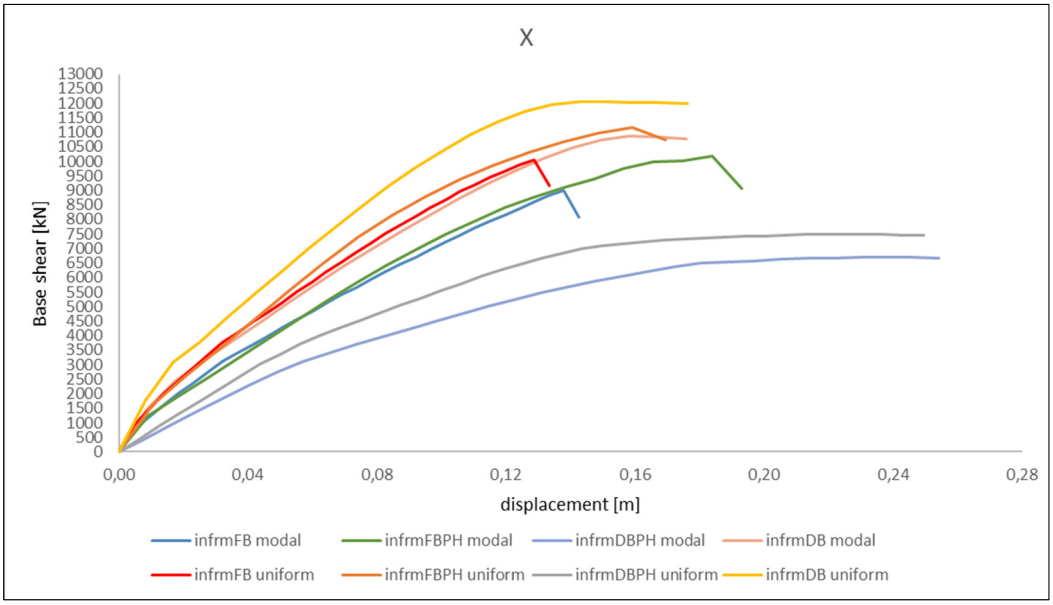


Figure 4: Pushover curve for X direction (SeismoStruct)

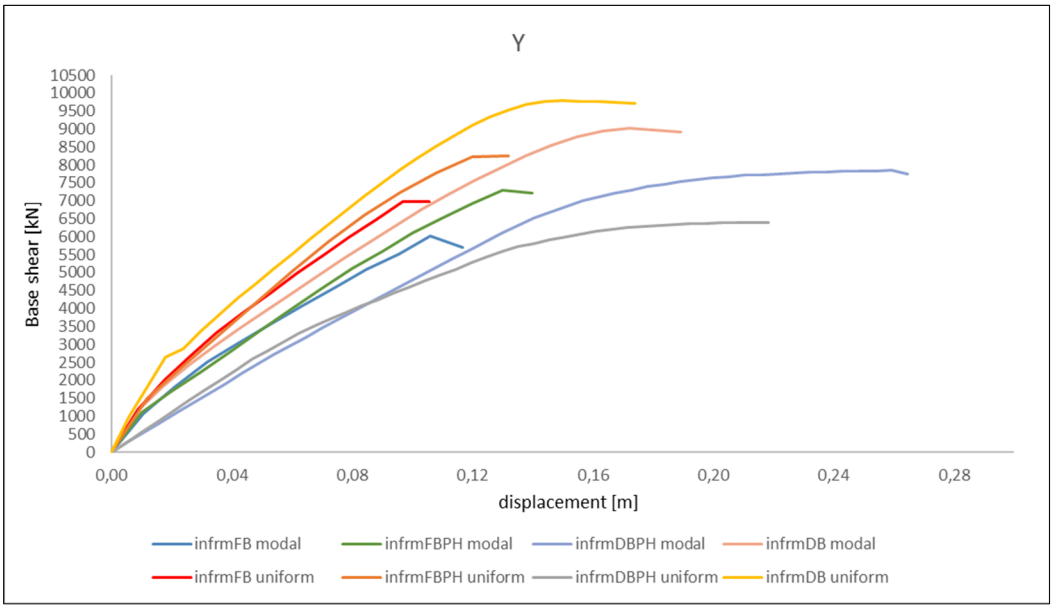


Figure 5: Pushover curve for Y direction (SeismoStruct)

InfrmFB- To simulate the nonlinear response of reinforced concrete frame structures, it is common to use simplified models such as two-node beam-column elements. Among these, force-based elements (Flexibility-FB) are especially popular because they predict longitudinal plasticity propagation, represent axial-flexural interactions, and satisfy force balance and strain compatibility [18]. The stress and deformation states in beam-column element sections are obtained by integrating the nonlinear uniaxial response of the material fibers. The section is divided into these fibers to capture inelasticity along the element's length and depth. Among SeismoStruct's four inelastic frame element types, the *infrmFB* element is the most accurate. It can represent inelastic behavior along the entire length of a structural member, even with only one element per member. The main advantage of this material nonlinearity model is its ability to use a single element to simulate a nonlinear response [11]. This reduces the number of degrees of freedom compared to using several displacement-based elements.

InfrmFBPH- The plastic hinge integration method overcomes problems with biased responses caused by deformations in force-based beam-column finite elements [19]. The method integrates deformations over certain plastic hinge lengths at the ends of the element. A positive feature of this method is that it gives exact solutions for linear problems. Incorporating the plastic length of the hinge into the integration method provides an objective response of elements and sections, which is important for the softening behavior of reinforced concrete structures. Plastic rotations are defined consistently and related to deformations in plastic hinges [19]. The advantages of such a formulation are not only a shortened analysis time (fiber integration is performed only for the two-membered end part) but also a complete control of the length of the plastic hinge (expansion of inelasticity), which allows overcoming the localization problem.

InfrmDB- This formulation is based on element stiffness. The displacement-based approach follows standard finite element method procedures where cross-sectional strains are interpolated from an approximate displacement field [20]. To approximate the response of a nonlinear element, a constant axial strain and a linear curvature distribution are applied along the entire length of the element (correct only for prismatic linear elements). Consequently, *infrmDB* elements should be used with small length members, leading to the need for mesh refinement, to achieve good accuracy in the case of higher-order strain distributions. It is usually necessary to define 4 to 6 elements per structural element [20].

InfrmDBPH- The plastic hinge integration method integrates deformations over specified lengths of plastic hinges at the member ends [19], [20]. All non-linear deformations of the element are collected in these rotational springs, while the rest of the element remains elastic. The *infrmDBPH* element consists of 3 sub-elements, two joints at the edges of the member, which model plastic rotational deformations around the second and third local axes, and an elastic element in the middle, which models the part of the element that remains elastic.

Based on the obtained calculation results for two different lateral load distributions, a very similar form of force-displacement dependence is noticeable, as well as the fact that the differences between the base force values at the same horizontal displacement value do not exceed approximately 10%. This indicates the regularity of the construction and the achieved uniformity of stiffness and load capacity for the height of the building. A manual verification of the results was performed. The calculation of the target displacement, as well as the comparison with the results from the software, is shown in Table 2 for the modal lateral load distribution and Table 3 for the uniform lateral load distribution. Figure 6 shows the relative floor displacement of the object in the X and Y directions.

Table 2: Calculation of target displacement for modal lateral load distribution

	Γ	F_y*	d_m*	E_m*	d_y*	T*	$S_{e(T^*)}$	d_{et}*	d_t*	d_t	SeismoStruct
DB X	1.06	10856.61	0.15	395.66	0.23	0.65	5.18	0.06	0.06	0.06	0.06
DB Y	1.06	9011.74	0.17	399.84	0.26	0.75	4.51	0.06	0.06	0.07	0.07
DBPH X	1.06	6693.19	0.24	623.58	0.29	0.93	3.65	0.08	0.08	0.08	0.08
DBPH Y	1.06	7843.11	0.24	723.00	0.29	0.86	3.92	0.07	0.07	0.08	0.08
FB X	1.06	9942.93	0.23	888.81	0.28	0.75	4.51	0.06	0.06	0.07	0.07
FB Y	1.06	7652.32	0.23	703.92	0.28	0.86	3.94	0.07	0.07	0.08	0.08
FBPH X	1.06	10190.58	0.13	448.66	0.17	0.58	5.83	0.05	0.05	0.05	0.06
FBPH Y	1.06	8126.6	0.13	432.64	0.15	0.61	5.52	0.05	0.05	0.06	0.07

Table 3: Calculation of target displacement for uniform lateral load distribution

	Γ	F_y*	d_m*	E_m*	d_y*	T*	$S_{e(T^*)}$	d_{et}*	d_t*	d_t	SeismoStruct
DB X	1.04	12042.04	0.16	558.24	0.23	0.61	5.54	0.05	0.05	0.05	0.05
DB Y	1.04	9796.84	0.15	392.32	0.22	0.66	5.11	0.06	0.06	0.06	0.06
DBPH X	1.04	7500.9	0.22	680.41	0.27	0.84	4.05	0.07	0.07	0.07	0.07
DBPH Y	1.04	6397.81	0.21	473.55	0.28	0.92	3.67	0.08	0.08	0.08	0.08
FB X	1.04	10802.06	0.21	882.08	0.25	0.67	5.02	0.06	0.06	0.06	0.06
FB Y	1.04	8527.14	0.19	594.68	0.25	0.75	4.49	0.06	0.06	0.07	0.07
FBPH X	1.04	11351	0.24	1037.85	0.30	0.71	4.74	0.06	0.06	0.06	0.06
FBPH Y	1.04	9228.28	0.12	562.3	0.13	0.52	6.50	0.04	0.04	0.05	0.06

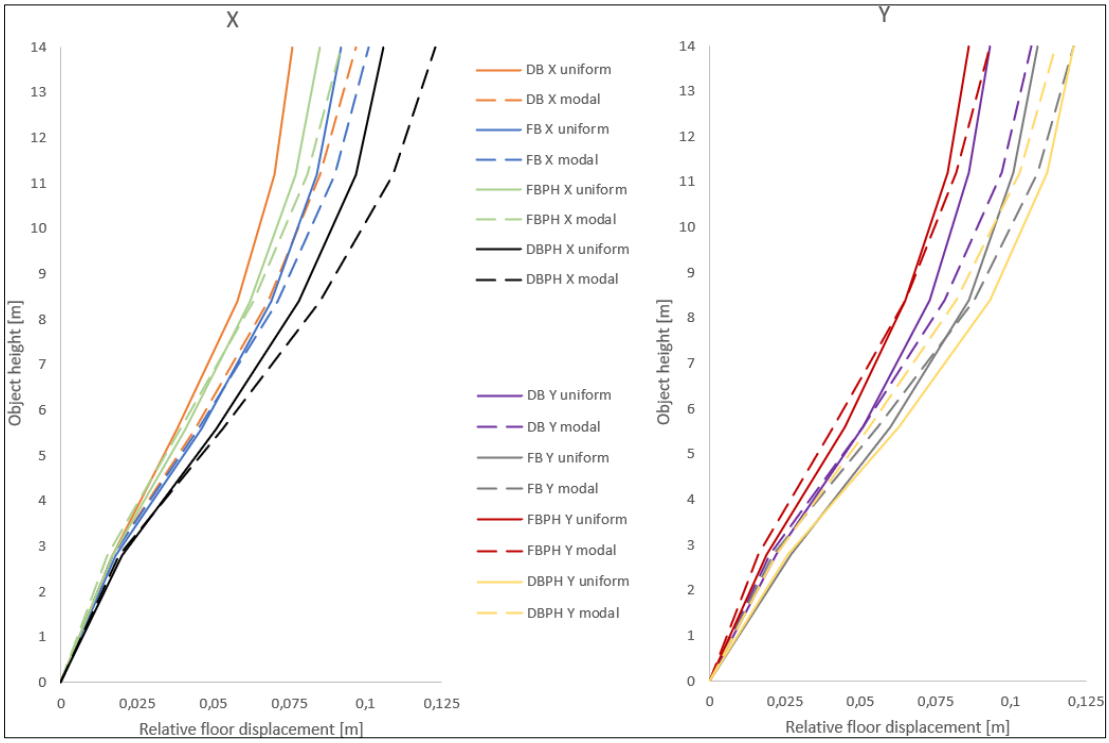


Figure 6: Relative floor displacement

The schedule for the appearance of plastic joints when reaching 150% of the target displacement is shown in Figure 7. The considered criteria for the appearance of plastic joints are concrete strain, yielding, fracture, chord rotation, and shear capacity. As can be seen in the picture, plastic hinges occur when shear capacity and yielding conditions are exceeded.

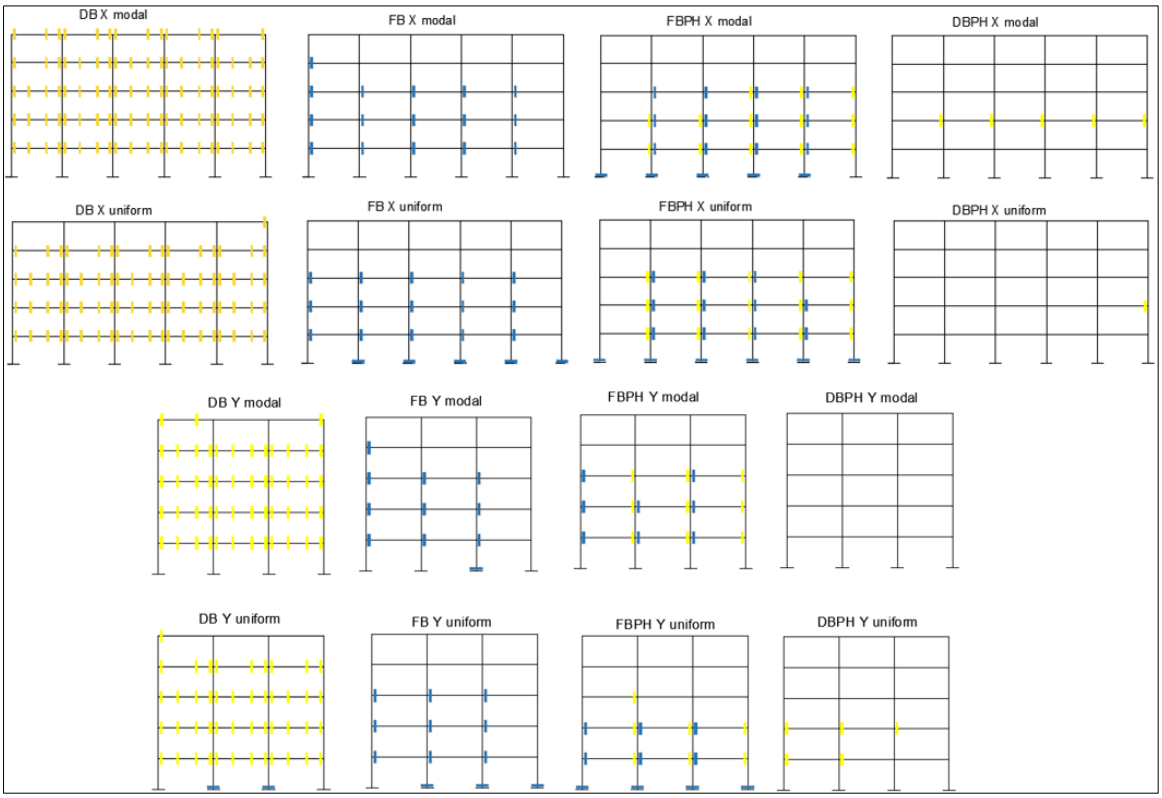


Figure 7: Schedule of appearance of plastic hinges for X and Y direction (yellow-shear capacity; blue- yielding)

Total chord rotation capacity θ_u is determined according to the expression (A.1) from EN 1998-3 [21], based on the formulations proposed by Panagiotakos and Fardis [22] and Biskinis and Fardis [23]. As shown in Table 4, the limit value of chord rotation at 150% of the target displacement was not exceeded anywhere.

Table 4: Calculation of chord rotation capacity

	Modal		Uniform	
	demand	limit	demand	limit
DB X	0.00106	0.04088	0.00117	0.04145
DB Y	0.00128	0.04360	0.00145	0.04333
DBPH X	0.00691	0.07303	0.00698	0.07259
DBPH Y	0.00788	0.08114	0.00916	0.08272
FB X	0.00609	0.06834	0.00651	0.06296
FB Y	0.00748	0.07440	0.00881	0.07097
FBPH X	0.00364	0.06932	0.00435	0.06319
FBPH Y	0.00383	0.07214	0.00464	0.07026

4. CONCLUSION

Pushover analysis of a five-story reinforced concrete building subjected to different lateral load patterns (uniform and modal arrangement) was performed using four different material nonlinearity models for two orthogonal directions. By observing the results of the seismic performance assessment of this object, the following conclusions are reached:

- The uniform loading scheme provides greater initial stiffness and shearing capacity of the base compared to the modal load distribution.
- *infrmFB* model- When reaching 150% of the target displacement, plastic hinges appear due to exceeding yielding in the beams on the 1st, 2nd, and 3rd floors; the maximum chord rotation is 0.0075 rad. On the pushover curve graphs, this model has the smallest displacement values.
- *infrmFBPH* model- When reaching 150% of the target displacement, plastic hinges appear due to exceeding the shear capacity and yielding in the beams on the 1st, 2nd, and 3rd floors and yielding in the columns at the supports; the maximum rotation of the chord is 0.00383 rad. This model has minimum relative floor displacement values for the Y direction.
- *infrmDB* model- When reaching 150% of the target displacement, plastic hinges appear due to exceeding the shear capacity in the beams on all floors. The chord rotation is the smallest in this model and is 0.00128 rad. This model has minimum relative floor displacement values for the X direction.
- *infrmDBPH* model- When reaching 150% of the target displacement, plastic hinges appear due to exceeding the shear capacity in the beams on the 1st and 2nd floors. The rotation of the chord is the largest in this model and is 0.00788 rad. This model has maximum relative floor displacement values. On the pushover curve graphs, this model has the highest displacement values.

Based on the analysis of the results, the models with the mean values of the considered criteria (shear capacity, yielding, relative floor displacement, chord rotation capacity) are *infrmFB* and *infrmFBPH* models. The *infrmDB* model gives the smallest values of chord rotation and relative story displacement, while the *infrmDBPH* model gives the highest values.

ACKNOWLEDGEMENTS

This study was supported by the Serbian Ministry of Science, Technological Development and Innovation from the grant 451-03-34/2026-03/200108.

REFERENCES

[1] P. Fajfar, "Analysis in seismic provisions for buildings: past, present and future", Bulletin of Earthquake Engineering, Vol. 16(7), pp. 2567–2608, https://doi.org/10.1007/s10518-017-0290-8 , (2018)

[2] H. Krawinkler and G. D. P. K. Seneviratna, "Pros and cons of a pushover analysis of seismic performance evaluation", Engineering Structures, Vol. 20(4–6), pp. 452–464, https://doi.org/10.1016/S0141-0296(97)00092-8, (1998)

[3] M. N. Fardis, E. C. Carvalho, P. Fajfar and A. Pecker, "Seismic Design of Concrete Buildings to Eurocode 8", CRC Press, Boca Raton (USA), (2015)

[4] EN 1998-1, "Design of structures for earthquake resistance – Part 1: General rules, seismic actions and rules for buildings", European Committee for Standardization, Brussels (Belgium), (2004)

[5] P. Fajfar, "A nonlinear analysis method for performance-based seismic design", Earthquake Spectra, Vol. 16(3), pp. 573–592, https://doi.org/10.1193/1.1586128, (2000)

[6] M. F. Haque, M. M. H. Rahat and M. M. Hoque, "Seismic performance evaluation of multi-storied RC structures by nonlinear static analysis", Results in Engineering, Vol. 25, p.104105, https://doi.org/10.1016/j.rineng.2025.104105, (2025)

[7] G. Dong, R. Garcia, K. Pilakoutas and I. Hajirasouliha, "A review of optimum seismic design of RC frames: State-of-the-art, challenges and future directions", Engineering Structures, Vol. 318, p.118455, https://doi.org/10.1016/j.engstruct.2024.118455 , (2024)

[8] M. Ferraioli, A. Lavino and A. Mandara, "An adaptive capacity spectrum method for estimating seismic demand of steel moment-resisting frames", Ingegneria Sismica, Vol. 33(1–2), pp. 47–60, (2016)

[9] R. Pinho, C. Casarotti and S. Antoniou, "A comparison of single-run pushover analysis techniques for seismic assessment of bridges", Earthquake Engineering & Structural Dynamics, Vol. 36(10), pp. 1347–1362, https://doi.org/10.1002/eqe.684, (2007)

[10] S. Marinković, B. Milošević, Ž. Petrović, D. Turina and D. Penava, "Pushover analysis for upgrading of existing residential masonry building", Engineering Today, Vol. 1(3), pp. 31–40, https://doi.org/10.5937/engtoday2203031M, (2022)

[11] Seismosoft, "SeismoStruct – A computer program for static and dynamic nonlinear analysis of framed structures, User Manual", Seismosoft Ltd., Pavia (Italy), (2023)

[12] S. Antoniou and R. Pinho, "Advantages and limitations of adaptive and non-adaptive force-based pushover procedures", Journal of Earthquake Engineering, Vol. 8(4), pp. 497–522, https://doi.org/10.1080/13632460409350498, (2004)

[13] A. K. Chopra and R. K. Goel, "A modal pushover analysis procedure for estimating seismic demands for buildings", Earthquake Engineering & Structural Dynamics, Vol. 31(3), pp. 561–582, https://doi.org/10.1002/eqe.144, (2002)

[14] ASCE/SEI 41-17, "Seismic evaluation and retrofit of existing buildings", American Society of Civil Engineers, Reston, VA (USA), (2017)

[15] EN 1992-1-1, "Design of concrete structures, Part 1-1: General rules and rules for buildings", European Committee for Standardization, Brussels (Belgium), (2004)

[16] J. B. Mander, M. J. N. Priestley and R. Park, "Theoretical stress-strain model for confined concrete", Journal of Structural Engineering ASCE, Vol. 114(8), pp. 1804–1826, https://doi.org/10.1061/(ASCE)0733-9445(1988)114:8(1804), (1988)

[17] M. Menegotto and P. E. Pinto, "Method of analysis for cyclically loaded reinforced concrete plane frames including changes in geometry and non-elastic behaviour of elements under combined normal force and bending", Proceedings of IABSE Symposium on Resistance and Ultimate Deformability of Structures Acted on by Well-Defined Repeated Loads, Lisbon (Portugal), pp. 15–22, (1973)

[18] E. Spacone, F. C. Filippou and F. F. Taucer, "Fibre beam-column model for non-linear analysis of R/C frames: Part I. Formulation", Earthquake Engineering & Structural Dynamics , Vol. 25(7), pp. 711–725, https://doi.org/10.1002/(SICI)1096-9845(199607)25:7%3C711::AID-EQE576%3E3.0.CO;2-9, (1996)

[19] M. H. Scott and G. L. Fenves, "Plastic hinge integration methods for force-based beam-column elements", Journal of Structural Engineering ASCE, Vol. 132(2), pp. 244–252, https://doi.org/10.1061/(ASCE)0733-9445(2006)132:2(244), (2006)

[20] D. Tarquini, J. P. Almeida and K. Beyer, "An enhanced displacement-based element to account for tension shift effects", Proceedings of the 16th World Conference on Earthquake Engineering, Santiago (Chile), Paper No. 3167, (2017)

[21] EN 1998-3, "Design of structures for earthquake resistance – Part 3: Assessment and retrofitting of buildings", European Committee for Standardization, Brussels (Belgium), (2005)

[22] T. B. Panagiotakos and M. N. Fardis, "Deformations of reinforced concrete members at yielding and ultimate", ACI Structural Journal, Vol. 98(2), pp. 135–148, (2001)

[23] D. Biskinis and M. N. Fardis, "Deformations at flexural yielding of members with continuous or lap-spliced bars", Structural Concrete, Vol. 11(3), pp. 127–138, https://doi.org/10.1680/stco.2010.11.3.127 , (2010)

SESSION G: CIVIL ENGINEERING

Evolution of nonlinear analysis of RC sections: From iterative computations to rapid moment–curvature prediction using ANN models

Bojan Milošević [1*], Nenad Kojić [2], Marijana Janicijević Milačak [1]

[1] University of Kragujevac, Faculty of Mechanical Engineering and Civil Engineering in Kraljevo, Kraljevo, Serbia

[2] Academy of technical and art applied studies, Department of Information and Communication Technologies, Belgrade, Serbia

ARTICLE INFO

* **Correspondence:** milosevic.b@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.G06M **PAGES:** G35 – G43

ABSTRACT

Nonlinear moment–curvature (M–φ) analysis is essential for evaluating the strength and ductility of reinforced concrete (RC) structures. Traditional numerical methods, such as iterative fiber analysis, provide high accuracy but are computationally intensive. This paper presents the evolution of approaches – from empirical and classical numerical methods to modern artificial intelligence (AI)-based models. Examples of applying artificial neural networks (ANN) and hybrid ANFIS models for predicting the nonlinear response of beams, columns, and beam–column joints are analyzed. Special attention is given to model architecture, input–output parameters, and prediction accuracy. The paper discusses differences between pointwise and parametric models, as well as recent approaches enhancing ANN interpretability. The study provides a critical review of ANN applications in defining moment–curvature relationships and identifies directions for future research in structural engineering.

KEYWORDS

Moment-curvature relationship, Reinforced concrete, Artificial Neural Networks, ANFIS, Pointwise models, parametric models

1. INTRODUCTION

Nonlinear moment–curvature (M–φ) analysis represents a fundamental step in the accurate assessment of the load-bearing capacity, stiffness, and ductility of reinforced concrete (RC) members subjected to bending. Defining the moment–curvature relationship for reinforced concrete sections is a key problem in bending analysis and is the basis for calculating the deformation capacity of individual sections and entire structural elements. The moment–curvature relationship describes the behavior of the cross-section under monotonically increasing bending moment, where the curvature increases depending on the nonlinear material response. This relationship is particularly important for understanding the behavior of structures subjected to seismic loading, where ductility and energy dissipation capacity are crucial design parameters [1].

In the traditional numerical approach, the moment–curvature relationship is obtained using methods such as iterative section analysis, fiber analysis, and nonlinear finite element methods. These methods provide high accuracy but can be computationally intensive [2] and difficult to generalize for different cross-section configurations and material parameter variations [3]. While methods such as the Newton-Raphson procedure or numerical integration over fiber cross-sections (e.g., in OpenSees) represent the gold standard for structural analysis, their application in real-time design or optimization of a large number of cross-sections becomes impractical. Precisely for this reason, in recent decades, there has been a growing interest in civil engineering in applying ANN models as fast surrogate models capable of approximating non-linear relationships between input parameters and cross-sectional responses without repeating the complete iterative procedure for each new case and eliminating the need for complex experimental or numerical analyses, which is also confirmed by contemporary review papers indicating a widespread and increasing application of ANN methods in structural analysis and design [4, 5].

In civil engineering practice, ANNs have already been successfully used for predicting various performance characteristics of concrete members, including cross-sectional capacity, moment–curvature prediction, deformations, strength, and behavior under damage conditions [1]. The current application of ANNs encompasses a wide range of RC elements, from rectangular beams strengthened with polymers [3] to columns subjected to biaxial bending [4] and complex beam-column joints [5]. The diversity of the applied architectures, ranging from standard multilayer feedforward networks with backpropagation to hybrid ANFIS models, indicates a continuous evolution of the methodology aiming to achieve a correlation higher than 95% relative to numerical and experimental data.

Given the wide range of methods and approaches used in the literature – from classical numerical models to modern ANN and neuro-fuzzy systems – there is a need for a review paper that will consolidate existing findings, systematize types of networks and their applications, and identify the advantages and limitations of different approaches. The objective of this paper is to present a systematic overview and comparative analysis of the application of Artificial Neural Networks (ANN) for predicting the nonlinear moment–curvature relationship of reinforced concrete cross-sections, with the aim of identifying key factors (network type, database, input variables) that affect model accuracy and practical applicability in structural engineering.

2. DEVELOPMENT OF COMPUTATION AND DEFINITION OF THE MOMENT–CURVATURE RELATIONSHIP

The moment–curvature relationship represents a fundamental concept in the bending analysis of reinforced concrete sections, as it describes the relationship between the bending moment and the curvature of the section under loading. The theoretical development of numerical determination of the moment–curvature (M–φ) relationship represents an evolution from simple analytical expressions to complex computational methods and, most recently, to data-driven models based on artificial neural networks.

The basic classical theory is based on the Euler–Bernoulli hypothesis of plane sections. It represents the foundation of every calculation, including those using neural networks, and enables the determination of the moment–curvature relationship in the elastic phase. This analysis neglects bond-slip between reinforcement and concrete and uses simplified constitutive laws (linear-elastic material response, or bilinear models with limited plasticity) [9]. Despite its simplicity, it defines the basic terminology (neutral axis, curvature, tensile zone, compressive zone) that remains the basis for all subsequent methods.

With the development of modern strength theory and the improvement of concrete and reinforcement models, approaches have incorporated plastic and nonlinear effects. The key challenge in this phase became the selection of an adequate constitutive material behavior law (e.g., Kent-Park, Mander, or Scott-Park models), as subtle differences in the concrete stress-strain curve (especially in the post-peak softening zone) can lead to significantly different values of curvature and ultimate moment capacity. The development of these constitutive models enabled the accurate capturing of concrete cracking, yielding of reinforcement, and ultimate failure of the section, which is crucial for seismic analysis and assessment of existing structures [9].

In engineering calculation practice, numerical methods such as iterative procedures and the finite element method have become the standard for solving the moment–curvature relationship for complex cross-sections and various loading conditions. The discretization phase (Fiber Method) represents the "gold standard" in the analysis of reinforced concrete sections. The section is divided into a large number of fibers – small sub-areas of concrete and reinforcement – where each fiber is assigned a constitutive law, and the section equilibrium is achieved iteratively (most commonly using the Newton-Raphson method). Software implementations such as OpenSees (open-source) and ABAQUS (commercial FEA) enable modeling of practically any cross-section geometry and material [10].

However, classical numerical approaches often face challenges when it comes to generalizing the model to different types of cross-sections or variations in material parameters – especially when dealing with preliminary dimensioning of structures or real-time analyses. Modern numerical developments focus on reducing the "iterative cycle" through model reduction and surrogate models, which directly opens the door to the application of machine learning techniques [1].

The transition to ANN has proven particularly promising due to their ability to model complex nonlinear functions based on datasets obtained from experiments or numerical simulations. ANN is theoretically viewed as a function approximator that "maps" input parameters (geometry and material) directly to output parameters (yield moment, ultimate curvature, ductility) in milliseconds instead of minutes, thereby bypassing the iterative cycle altogether. The practical implication is that ANN does not replace previous numerical methods but rather compresses their results into a rapid interface: the network is trained on data generated by the most precise methods (fiber analysis, FEM) and then enables almost instantaneous prediction for new parameter combinations.

Numerical development has progressed from simple equations, through iterative integration over fibers (OpenSees, ABAQUS), to today's hybrid models where ANN serves as a rapid interface that utilizes knowledge generated by the most precise numerical methods of the past. Modern development of numerical methods clearly demonstrates the tendency of combining the accuracy of the fiber method with the speed of ANN, representing a significant step forward in the real-time application of nonlinear analysis in structural engineering practice.

3. DEVELOPMENT AND APPLICATION OF ANN IN STRUCTURAL ENGINEERING

ANNs are increasingly used in structural engineering as data-driven models for predicting and approximating complex nonlinear relationships in the behavior of concrete structures. These models transform input data into output predictions through layers of interconnected neurons, using mathematical optimization algorithms for learning from the training database [1]. The key characteristic of ANN models is their ability to capture complex nonlinear patterns that are difficult to describe with conventional analytical expressions, making them particularly suitable for predicting mechanical characteristics and load-bearing capacity of concrete elements and structures.

In the early period of application in civil engineering (1989–1994), ANNs primarily served to demonstrate that it was possible to learn nonlinear material behavior rules, i.e., material mechanics, without explicit differential equations. The potential of ANN in civil engineering was originally presented by Adeli (1989), while a comprehensive overview of ANN applications in civil engineering, including numerous references, is provided by Adeli (2001) [11]. One of the earliest applications, Ghaboussi et al. (1991), demonstrated an ANN model for approximating the constitutive behavior of materials under monotonic loading [12]. These papers laid the foundation by showing that ANN can capture the basic nonlinearity of materials without explicitly specifying constitutive equations.

With the development of concrete technology (1995–2005), the focus of ANN application shifted from the general concept to specific predictions of material properties, when they were also used for predicting mechanical characteristics of concrete elements, including the strength and load-bearing capacity of columns and beams. That ANN models are superior compared to regression models in predicting the compressive strength of concrete was demonstrated by Yeh (1998) [13]. Hadi (2003) [14] applied ANN for predicting the load-bearing capacity of concrete columns with and without confinement. A comprehensive overview of ANN application in civil engineering, including numerous references, is provided by Adeli (2001) [11]. During this period, several studies confirmed that ANN can predict the mechanical characteristics of concrete elements with high accuracy ($R^2 > 0.95$).

Later (2005–2015), the focus shifted toward specific structural problems in addition to defining load-bearing capacity, ANNs were also utilized to estimate deformability and ductility, which has been successfully applied in more recent studies to predict failure loads and other performances under extreme loading [15], [16]. Studies have shown that ANNs can recognize complex non-linear characteristics in experimental data without assuming a functional form, making them suitable for modeling the behavior of concrete elements with various input parameters [17]. Research has demonstrated that ANNs can simultaneously process the effects of concrete cover and stirrup spacing, which are parameters that directly affect the confinement level of the concrete core—a factor that is extremely difficult to accurately quantify in traditional fiber models without losing convergence [18]. The trend in the literature also points to the potential application of ANNs in modeling complex relationships such as moment–curvature, where ANNs can serve as an efficient approximation method alongside numerical or experimental approaches [1], [14].

ANNs enable rapid and efficient prediction, which is useful when large datasets from laboratory tests or numerical simulations of concrete behavior are available. These models reduce computation time and improve design parameters, especially when traditional analytical methods become too complex for practical application [17]. During this period, the focus shifted from simply "whether ANN can predict" to "how reliably it generalizes and whether it can replace or complement established engineering models".

In the most recent phase, attention is increasingly directed toward model interpretability, i.e., toward reducing the "black box" effect, which is particularly evident in hybrid ANFIS systems. While early MLP models provided accurate results without explaining the logic, the most recent works (Mangir et al., 2025 [3]) introduce fuzzy rules that enable tracing the decision-making process, thus increasing the transparency and trust of structural engineers in data-driven models.

4. ANALYSIS OF PUBLISHED RESEARCH

In order to examine contemporary approaches to the application of neural networks in moment–curvature relationship analysis, a systematic review of relevant research from available scientific databases was conducted. The review covered relevant databases (Scopus, Web of Science, Google Scholar) with a focus on papers that combine the application of neural networks with the analysis of reinforced concrete cross-sections. Based on the review and analysis of papers, four characteristic studies that represent different methodological approaches were selected and comparatively analyzed.

4.1. Methodological Framework and Review of Relevant Research

Analysis of the four studies in which the process of defining the moment–curvature relationship using ANN is presented reveals that they do not constitute a homogeneous group. Jadid & Fairbairn (1996) [8] deal with an experimental beam-column connection system; Bağcı (2010) [1] and Sakcalı & Yüksel (2022) [7] are predominantly parametric models trained on numerically generated databases; while Mangir et al. (2025) [3] introduce a hybrid ANFIS model that treats curvature as an input variable. In the following text, the results of the four analyzed studies are described.

M. N. Jadid and D. R. Fairbairn (1996) [8] use ANN for solving complex nonlinear mapping of beam-column connections, predicting moment-curvature parameters based on experimental data. The significance of this work is primarily

methodological: it was demonstrated that neural networks can learn the relationship between six geometric and material parameters and the characteristic values of yield curvature (φ_y) and yield moment (My). The MLP architecture with two hidden layers (6-40-30-2), trained on 272 numerical patterns and 34 experimental samples, achieves an R.M.S. error of only 0.02%, which confirms the extraordinary approximation capability of ANN. Although the model is limited to yield point parameters of beam-column connections (which is a narrow domain), this work was the first to demonstrate the potential of neural networks for predicting moment-curvature parameters and opened the way for all subsequent research in this field.

M. Bağcı (2010) [1] uses ANN for predicting the moment-curvature relationship of reinforced concrete sections with emphasis on nonlinear material behavior. For generating training data for a multi-layer feed-forward network with backpropagation algorithm, an analytical method (filament model) was used. This work belongs to the category of parametric models where the network simultaneously predicts six characteristic parameters of the M-φ diagram (Ty, Th, Cv,C, Co, C, ε). The accuracy of the model is exceptionally high (average training error < 0.2%), which is expected since the data were generated by a deterministic mathematical model without noise. The contribution of this work lies in the demonstration that ANN can reproduce the complete moment-curvature curve through several key control points, with practically negligible error.

G. B. Sakcalı and İ. Yüksel (2022) [7] analyze reinforced concrete rectangular columns subjected to eccentric compression and biaxial bending. Training data for the network were obtained using software for cross-section analysis based on the fiber method (fiber section analysis). This work is significant because it extends the domain of ANN from simple bending to complex three-dimensional stress states. The model predicts three key parameters (yield curvature φy, ultimate curvature φu, and yield moment My) with a coefficient of determination R^2 in the range of 0.954 to 0.990. Although the database consists of only 100 samples, the model successfully captures the behavior of columns under biaxial bending.

A. Mangir, V. Okumus, and Z. Şen (2025) [3] introduce a hybrid ANFIS model (Adaptive Neuro-Fuzzy Inference System) for precise prediction of the nonlinear moment-curvature relationship of RC beams. Unlike previous works, this model treats curvature as an input variable, which enables the generation and reconstruction of the entire M-φ curve. The advantage of the ANFIS model lies in the combination of fuzzy logic and neural networks, which provides better interpretability of results – one of the biggest weaknesses of traditional ANN models. The mean relative error (MRE) is 3.55%, which is higher than in parametric models but justified by the complexity of the task – predicting the entire nonlinear curve instead of just a few characteristic points. This model represents the most modern approach in the analyzed set and demonstrates the tendency toward more transparent and interpretable models.

In order to systematically compare these approaches, and for a clearer understanding of the similarities and differences between the analyzed approaches, their key characteristics are summarized in Table 1, which serves as the basis for further comparative analysis.

Table 1: Comparative overview of ANN models in the analyzed studies

Characteristic	M. N. Jadid and D. R. Fairbairn [8]	G. B. Sakcalı and İ. Yüksel [7]	M. Bağcı [1]	A. Mangir et al. [3]
Structural element	Beam-column conection (biaxial bending)	RC columns (biaxial bending)	RC cross-sections (general model)	RC beams with rectangular cross-section
Network type	MLP Backpropagation	Forward-fed Backpropagation	Multi-layer Feed-forward	ANFIS(Neuro-Fuzzy)
Input variables	6 variables	7 variables	7 variables	6 variables
Architecture	2 hidden layers (6-40-30-2)	1 hidden layer (7-5-3)	2 hidden layers (7-12-13-6)	6 layers (Fuzzy rules)
Activation functions (between layers)	Sigmoid and Hyperbolic tangent	Sigmoid function	Sigmoid function	Gaussian membership functions (Gaussian MF)
Output variables	φ_y, My (characteristic points)	φ_y, φ_u, My (characteristic points)	6 parameters, characteristic points of TY, TH, CvC, CoC, ε, M,	Moment (M) for each diagram point, entire MC diagram
Approach type	Parametric / experimental	Parametric	Parametric	Pointwise
Database	272 valid patterns (FORTRAN + 34 experimental samples)	100 samples (MATLAB fiber analysis)	52 data sets (Filament model)	95 cross-sections (OpenSees)
Statistical data	R.M.S. error: 0.02%; Correlation: 1.0	Coefficient of determination (R^2) ranges from 0.954 to 0.990 for three parameters	Average training error: < 0.2%	Mean relative error (MRE): 3.55%

4.2. Databases and Factors of ANN Model Success

Based on the presented models, it can be observed that the success of ANN application in predicting the nonlinear response of cross-sections directly depends on several key factors. First and foremost, the quality and volume of the database represent the foundation for reliable model learning, along with proper selection of input variables and optimization of the training algorithm.

Most researchers use numerically generated databases to overcome the high costs and practical limitations of experimental testing. The dominant method for generating "ground truth" data is fiber (filament) analysis, which discretizes the section into a large number of parallel elements and performs iterative equilibrium through the Newton-Raphson procedure. The accuracy of these numerical data directly determines the upper bound of ANN model accuracy – if the database contains noise (experimental scatter) or systematic bias (simplified material models), the network cannot achieve better performance than the data on which it was trained.

4.3. Model Architecture and Activation Functions

In addition to database quality, the choice of neural network architecture and associated learning algorithms has a decisive influence on the model's ability to approximate complex nonlinear dependencies. In terms of architecture, the most commonly used architecture is the Multi-layer Perceptron (MLP), with one or two hidden layers. MLP with two hidden layers (e.g., 6-40-30-2 in [8] or 7-12-13-6 in [1]) has proven particularly suitable for parametric models where the outputs are characteristic points of the M-φ curve. In contrast, simpler architectures with one hidden layer and fewer neurons (7-5-3 in [7]) are sufficient when the goal is prediction of a small number of output parameters.

On the other hand, the application of the hybrid Adaptive Neuro-Fuzzy Inference System (ANFIS) model combines the logical reasoning of fuzzy sets with ANN learning. Research shows that ANFIS can achieve better generalization and lower error compared to standard MLP, particularly in modeling the entire nonlinear M-φ curve. As demonstrated in [3], ANFIS uses Gaussian membership functions and hybrid optimization that combines least squares and gradient descent methods.

The choice of activation function determines the network's ability to model sudden changes in stiffness (e.g., concrete cracking or reinforcement yielding). In the analyzed papers, Sigmoid and Tanh functions were applied as standard for MLP models due to their smooth differentiability. Therefore, the choice of architecture and activation function is not arbitrary – it directly reflects the physical nature of the problem: a smooth but strongly nonlinear relationship with several characteristic points of sudden change (cracking, yielding, crushing).

4.4. Analysis of Input Variables

In addition to model architecture, one of the key aspects of every ANN approach is the selection of input variables that representatively describe the physical behavior of the system. Through comparative analysis of input variables, the main cross-section comprises concrete strength, amount of main reinforcement, and cross-section geometry, which are common to all analyzed models (Figure 1). These variables directly control the internal force equilibrium in the section and represent necessary parameters for any analytical or numerical calculation.

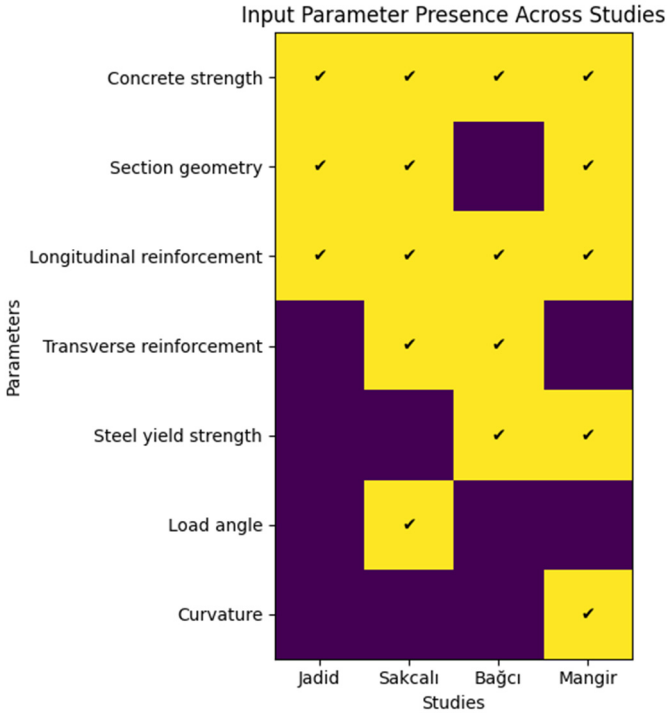


Figure 1: Comparative overview of input variables

Differences arise depending on whether beams are analyzed (where the focus is on the yield strength of steel and the cross-sectional area of reinforcement) or columns (where the influence of normal force dominates). Input parameters vary depending on the complexity of the problem; thus, geometric characteristics, width (b) and effective depth (d), are present in all models, while additional parameters such as concrete cover, confinement, axial force, or curvature value appear depending on the specific problem being analyzed. Models that include a larger number of input parameters generally show greater adaptability but also require larger databases for adequate training.

4.5. Model Typology: Pointwise and Parametric Approach

Based on the previous analysis, it can be concluded that the reviewed works are divided into two methodologically distinct groups. The first group consists of pointwise models, in which curvature is an input variable and the output is the moment for a specific point on the curve. The second group consists of parametric models, in which the network does not reconstruct the curve point by point but rather predicts several key parameters that define the shape of the M-φ diagram.

Pointwise models predict the moment for a given curvature value and are represented in the work of Mangir et al. (2025) [3], where inputs are cross-section geometry, mechanical material properties, reinforcement area, and the curvature itself, while the output is the moment. The advantage of this approach is that it enables reconstruction of the entire M-φ curve with arbitrary resolution, but requires a larger training database and can be sensitive to extrapolation outside the range of training data.

In a professional sense, neither of these two approaches can be declared "better" a priori without defining the objective of the analysis. Pointwise models are more suitable when a detailed shape of the entire M-φ curve is required, whereas parametric models are more suitable when the objective is to rapidly obtain several key indicators for structural analysis, design, or ductility assessment. The most detailed insight into the state of the material within the cross-section, predicting both the crushing points of the concrete cover and core, is presented in the research of M. Bağcı [1]. The parametric approach dominates in the works of Bağcı (2010) [1] and Sakcalı & Yüksel (2022) [7], which focus on deriving key engineering indicators such as stiffness, ductility, and capacity. These models are particularly suitable for seismic design and bilinear idealization due to their compactness, although they do not provide a detailed insight into the complete course of the curve. A similar focus on experimental parameters and the response classification of the beam-column joint is also present in the pioneering work of Jadid & Fairbairn (1996) [8]. Their main advantage is compactness and immediate usability in seismic design and bilinear idealization, while the main limitation is that they do not provide the detailed course of the complete curve. In contrast to them, the research of Mangir et al. (2025) [3] represents a strict pointwise model, aimed at the direct mapping of points and a continuous description of the structural response. The choice between these two approaches depends on engineering needs, i.e., whether the priority is a detailed reconstruction of the response or the efficient determination of key parameters.

4.6. Statistical Indicators and Model Accuracy

In order to evaluate model performance, various statistical indicators were analyzed, including correlation coefficients, R.M.S. error, and mean relative error. Statistical indicators confirm the high reliability of the models and indicate that ANN-based models achieve a level of accuracy that is practically applicable in engineering design. However, direct comparison of accuracy between models requires caution, as each model uses a different database, different input parameters, and different levels of task complexity.

Models that focus on defining characteristic points report exceptionally low errors [1,8]. The highest accuracy is recorded in the research of Bağcı [1]. The reason for this precision lies in the application of purely analytical data (generated via a "filament model") because when an ANN learns mathematically precise functions without noise, it can replicate them almost perfectly. The research of Jadid [8] is a pioneering attempt to demonstrate that an ANN can replace an experiment. Within the analyzed dataset, a very high accuracy was reported, which indicates the ability of the network to correctly classify failure modes and trends based on a limited number of laboratory tests. For columns subjected to biaxial bending in the research of G. B. Sakcalı and İ. Yüksel, the accuracy remains high, but a certain variance among the parameters is observed [7]. It is noted that the ultimate limit state (failure) is more difficult to predict than the yielding state, as the behavior of concrete before failure is highly non-linear and sensitive to small variations in the inputs. It can be said that the ANN is superior when formed on the basis of "clean" data, free of noise, and from fixed mathematical models [1, 7, 8]. It is necessary to emphasize that very high correlation values and very low errors should not be automatically interpreted as a universal superiority of the ANN approach, but always within the context of the database size, the method of splitting into training/test sets, and the fact that many databases are numerically generated, so the ANN primarily approximates the reference numerical model from which it learns.

For models that predict continuous nonlinear dependencies, such as the ANFIS model in the study by A. Mangir et al. [3], the mean relative error (MRE) is approximately 3.55%, while most results fall within the limits of ±5%, which suggests that the error is practically acceptable for engineering application, as it is smaller than the typical uncertainty of the constitutive models for concrete and reinforcement materials themselves. However, the interpretation of these indicators must be conducted with caution, taking into account the origin and characteristics of the database on which the model was trained.

4.7. Interpretability, Advantages, and Limitations of Models

In addition to accuracy, contemporary research increasingly emphasizes the importance of interpretability and practical applicability of ANN models in engineering practice. One of the greatest advantages of applying ANNs is the speed of obtaining results. While numerical simulations in software such as ABAQUS are slow and expensive, neural networks provide a non-linear response instantaneously. The practical value of these studies is further highlighted through the development of user tools. Although ANN models demonstrate high accuracy, one of the challenges remains their "black box" nature. Modern papers are beginning to introduce techniques such as Shapley Additive exPlanations (XAI - Explainable AI) to interpret the influence of input parameters, such as cross-sectional depth and reinforcement, on the final capacity prediction. Among the analyzed papers, only one study (Mangir et al., 2025) [3] is observed to have made a significant step toward eliminating the "black box" effect by using fuzzy sets and expert opinions to reduce the rule base, thereby making the decision-making process within the model transparent and verifiable for engineers. It should be emphasized that the speed of ANN / ANFIS models represents a clear practical advantage, but the cost of that speed can be reduced interpretability. Precisely for this reason, the significance of the work by Mangir et al. (2025) [3] is greater than the error percentage itself – that work demonstrates the direction of moving from purely black-box models toward approaches that are more engineering-verifiable and explainable.

Comparative analysis shows a clear tendency: while early models [8] aimed to prove the very possibility of learning, modern approaches [3] are focused on model transparency and speed (results in milliseconds) without loss of physical foundation of the calculation. Statistical indicators of all analyzed works confirm that ANN is no longer just a tool for scientific investigation but a practically applicable tool in everyday engineering practice. An overview of the chronological development of works also reveals an increase in the complexity of the problem being solved – from simple connections [8], through individual sections [1], to complex biaxial bending of columns [7] and complete M-φ curves via ANFIS [3].

5. RECOMMENDATIONS FOR FUTURE RESEARCH

Based on the detailed analysis of the papers, it is evident that science has traversed a path from proving that ANN can predict moment-curvature at all [8], through optimization of speed [1], to attempts to make the process logically transparent [3]. Also, based on the analysis, shortcomings as well as possible directions for future research can be identified. In the following text, recommendations for further research are briefly described.

Since all proposed studies primarily deal with the moment-curvature relationship under monotonic loading, it is necessary to investigate the behavior of real structures subjected to cyclic loading. Developing an ANN model that predicts hysteretic loops of the moment-curvature (M-φ) relationship, taking into account degradation of stiffness, strength reduction, and energy dissipation through cycles, remains important for application in seismic engineering.

The analyzed models are based on ideal cross-sections, where it is necessary to include parameters such as the degree of corrosion and time as input variables. Reinforcement corrosion drastically changes the moment-curvature (M-φ) relationship over time (reduction of steel area, loss of concrete-steel bond, cracking of the protective concrete layer). Therefore, ANN models that predict the moment-curvature relationship should be extended to include time-dependent effects.

A classical cross-section was analyzed, without the possible application of high-strength concrete, concrete with recycled aggregate, or fiber-reinforced polymer (FRP) reinforcement. Analyzing the moment-curvature relationship for hybrid sections (e.g., combination of steel profiles and concrete – SRC) or sections strengthened with FRP strips/sheets would expand the domain of ANN application from conventional RC to modern composite and rehabilitated structures.

In current research, the moment-curvature relationship is treated as an isolated problem, where integration into BIM and real-time simulations is needed. Developing an ANN model that can be integrated as a plugin for BIM software (such as Revit or Tekla) or software for nonlinear analysis (OpenSees, SAP2000) would enable engineers to have an M-φ response for any section in real time during the design process. This would transform ANN from a scientific tool into a practical design instrument.

6. CONCLUSION

The application of ANN in defining the nonlinear moment-curvature relationship represents a significant advancement in structural engineering, as it enables bridging the gap between high numerical accuracy and the demands for rapid engineering application. Through analysis of model evolution – from pioneering MLP networks to modern ANFIS systems – it can be concluded that the development direction is clearly defined: toward faster, more accurate, and more transparent predictive models.

Key findings of the paper indicate that:

- Computation speed in milliseconds enables optimization of thousands of cross-sections in real time, which is computationally too expensive with classical numerical methods (such as the Newton-Raphson procedure);
- ANN models have proven superior to linear regression because they naturally model the pronounced nonlinearity of concrete and steel, particularly in the yielding and ultimate failure zones;

- Database quality remains the primary factor of accuracy, where numerically generated data (e.g., via OpenSees) are used as "ground truth" for training robust predictive models.

Viewed as a whole, the analyzed works demonstrate that development in this field progresses from proving the very possibility of ANN approximation, through optimization of architecture and database, to modern models with a greater degree of interpretability. Thus, ANN and ANFIS are not imposed merely as faster tools but as a bridge between the precision of classical numerical methods and the requirements of modern engineering practice for speed and transparency.

Future research directions must focus on eliminating the "black box" effect through XAI (Explainable AI) techniques, integration into BIM workflows, and introduction of time and corrosion parameters into input variables. This will enable neural networks to cease being merely an alternative tool and become a standard part of the engineering design process, thereby contributing to safer, more efficient, and more economical concrete structures.

ACKNOWLEDGEMENTS

This research was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, under the Agreement on Financing the Scientific Research Work of Teaching Staff at the Faculty of Mechanical Engineering and Civil Engineering in Kraljevo, University of Kragujevac - Registration number: 451-03-34/2026-03/200108 dated 5 February 2026. The conducted research is related to the goal Industry, innovation and infra-structure in accordance with the universal strategy of the United Nations (Agenda 2030).

REFERENCES

[1] M. Bağcı, “Neural Network Model for Moment Curvature Relationship of Reinforced Concrete Sections,” Math. Comput. Appl., Vol. 15(1), pp. 66–78, https://doi.org/10.3390/mca15010066, (2010)

[2] Ž. Petrović, S. Ranković, M. Mijalković and B. Milošević, “Analytical Study of the Section of the RC Beams Strengthened for Flexure with FRP Materials,” Facta Univ. Ser. Archit. Civ. Eng., Vol. 19(1), pp. 15-30, https://doi.org/10.2298/FUACE210719003P, (2021)

[3] A. Mangir, V. Okumus and Z. Şen, “Moment–Curvature Relationship Prediction of Reinforced Concrete Beams by Adaptive Neuro Fuzzy Inference System (ANFIS) Modelling,” Int. J. Fuzzy Syst., https://doi.org/10.1007/s40815-025-01981-7, (2025)

[4] Ž. Petrović, B. Milošević, N. Kojić and I. Kuvač, “Artificial Neural Networks in the Design and Stability Assessment of Masonry Structures: A Comprehensive Review (1999–2024),” Facta Univ. Ser. Archit. Civ. Eng., Vol. 23(2), pp. 177-196, https://doi.org/10.2298/FUACE250610013P, (2025)

[5] B. Milošević, N. Kojić and Ž. Petrović, “Overview of the application of Artificial Neural Networks in the determination of some mechanical characteristics of masonry walls,” Engineering Today, Vol. 4(1), pp. 41-52, https://doi.org/10.5937/engtoday2500006M, (2025)

[6] T. T. L. Hoang, H. Masuya, Y. Kurihashi and T. M. Tan, “An artificial neural network for predicting the ultimate bending moments in reinforced concrete beams with fiber-reinforced polymer strengthening,” Asian J. Civ. Eng., Vol. 24(7), pp. 2295-2305, https://doi.org/10.1007/s42107-023-00641-1, (2023)

[7] G. B. Sakcalı and İ. Yüksel, “Eğik Eğilme Etkisindeki Betonarme Kolonların Moment-Eğrilik İlişkisi için Yapay Sinir Ağ Modeli,” Politeknik Dergisi, Vol. 25(3), pp. 1167-1183, https://doi.org/10.2339/politeknik.764745, (2022)

[8] M. N. Jadid and D. R. Fairbairn, “Neural-network applications in predicting moment-curvature parameters from experimental data,” Eng. Appl. Artif. Intell., Vol. 9(3), pp. 309-319, https://doi.org/10.1016/0952-1976(96)00021-8, (1996)

[9] A. H. Nilson, D. Darwin and C. W. Dolan, “Design of Concrete Structures,” McGraw Hill, New York (USA), (2010)

[10] I. M. Smith and D. V. Griffiths, “Programming the Finite Element Method,” John Wiley & Sons, West Sussex (England), (2004)

[11] H. Adeli, “Neural Networks in Civil Engineering: 1989–2000,” Comput.-Aided Civ. Infrastruct. Eng., Vol. 16(2), pp. 126-142, https://doi.org/10.1111/0885-9507.00219, (2001)

[12] J. Ghaboussi, J. H. Garrett Jr. and X. Wu, “Knowledge-Based Modeling of Material Behavior with Neural Networks,” J. Eng. Mech., Vol. 117(1), pp. 132-153, https://doi.org/10.1061/(ASCE)0733-9399(1991)117:1(132), (1991)

[13] I.-C. Yeh, “Modeling of strength of high-performance concrete using artificial neural networks,” Cem. Concr. Res., Vol. 28(12), pp. 1797-1808, https://doi.org/10.1016/S0008-8846(98)00165-3, (1998)

[14] M. N. S. Hadi, “Neural networks applications in concrete structures,” Comput. Struct., Vol. 81(6), pp. 373-381, https://doi.org/10.1016/S0045-7949(02)00451-0, (2003)

[15] Q. H. Nguyen, H.-B. Ly, T.-A. Nguyen, V. H. Phan, L. K. Nguyen and V. Q. Tran, "Investigation of ANN architecture for predicting shear strength of fiber reinforcement bars concrete beams," PLOS ONE, Vol. 16(4), p. e0247391, https://doi.org/10.1371/journal.pone.0247391, (2021)

[16] M. M. Ahmad, A. Elahi and S. Barbhuiya, "Comparative Analysis of Reinforced Concrete Beam Behaviour: Conventional Model vs. Artificial Neural Network Predictions," Materials, Vol. 16(24), p. 7642, https://doi.org/10.3390/ma16247642, (2023)

[17] P. Chopra, R. K. Sharma and M. Kumar, "Prediction of Compressive Strength of Concrete Using Artificial Neural Network and Genetic Programming," Adv. Mater. Sci. Eng., Vol. 2016, p. 7648467, https://doi.org/10.1155/2016/7648467, (2016)

[18] I. Flood, "Towards the next generation of artificial neural networks for civil engineering," Adv. Eng. Inform., Vol. 22(1), pp. 4-14, https://doi.org/10.1016/j.aei.2007.07.001, (2008)

SESSION G: CIVIL ENGINEERING

Methodology for measuring pressure at the contact surface between a column footing model and a soil layer limited by a rigid base using Earth pressure cells

Stefan Mihajlović[1*], Milan Bižić[1], Iva Despotović[1], Jovana Bojković[1], Marijana Janićijević Milačak[1]

[1] University of Kragujevac, Faculty of mechanical and civil engineering in Kraljevo, Serbia

ARTICLE INFO

* **Correspondence:** mihajlovic.s@mfkv.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.G07M **PAGES:** G45 – G54

ABSTRACT

The design of all building structures largely depends on the foundation support system, i.e., the relevant actions for structural design depend on soil–structure interaction. This paper describes a methodology for measuring pressures at the contact surface between a column footing model and a soil layer limited by a rigid base using earth pressure cells. The purpose of the research is to assess the reliability of the obtained results representing contact stress distributions between the interacting components under laboratory conditions. The concept of a reduced-scale footing model is presented, along with the procedure for preparing and installing earth pressure cells in the soil, with special emphasis on their position and data acquisition method. The main factors affecting measurement accuracy are also discussed, including soil characteristics, geometric and material properties of the tested footing model. The aim of this paper is to define an experimental procedure for a footing model and a soil layer limited by a rigid base that can serve as a basis for further experimental research and validation of numerical models. The reliability control of the measured contact stresses using earth pressure cells is based on comparison with analytical results. This experimental study examined contact pressures between a footing model and compacted sand using earth pressure cells. Measured pressures deviated significantly from calculated uniform values in the first test (48.3%) due to installation-induced soil disturbance, but improved to an acceptable 9.2% in the third test with proper surface preparation. The load–pressure relationship remained nearly linear across all measurements, consistent with sensor specifications.

KEYWORDS

Foundation, Geomechanics, Contact stresses, Column footing, Earth pressure cells, Reduced-scale footing model, Experimental methodology, Soil layer limited by a rigid base

1. INTRODUCTION

Design of buildings entails the dimensioning of all structural elements based on regulations and codes, whereby the basic design principles relating to stability, durability, and serviceability of the designed structure must be satisfied. This is a demanding engineering process based on known data and data that must be assumed. The dimensioning of structural elements depends on the method of foundation of the building. It is necessary to transfer all loads from the structure to the soil, and in the analysis of the foundation structure–soil interaction [1], as many parameters as possible should be taken so that the behaviour of the adopted model is as close as possible to the actual behaviour of the structure. It is of engineering importance to consider all existing models and apply the best one for the analysed structure and its characteristics.

Foundation of structure depends on many parameters, so prior to design it is necessary to provide all the required data. Data on the physical and mechanical characteristics of the soil are obtained through geotechnical investigation works, which include field and laboratory testing [2]. In addition, it is necessary to investigate existing data for the location where the designed structure will be built. These geotechnical investigation works define the spatial position of soil layers, their thickness and extent, and the depth of the groundwater table. Besides these data, it is necessary to provide data on the

engineering properties of the soil, such as its strength and deformability, as well as all other necessary indicators depending on the purpose and importance of the structure being built. The design of the foundation structure involves determining the dimensions of the foundation based on the bearing capacity of the foundation soil, which is defined by a geomechanical study according to the provisions of Eurocode 7 (EN1997-1:2004), and settlement control in order to satisfy the condition related to the serviceability and functionality of the designed structure.

Due to the influence of the intensity and distribution of contact stresses between the foundation structure and the soil on the dimensioning of structural elements of the building, this paper will present a measurement methodology with special emphasis on the measuring equipment. The aim of the paper is to define the method of placing earth pressure cells intended for measuring soil pressure on a footing model under laboratory conditions. Recommendations will be given regarding the setup of the measuring device and the method of collecting experimental data, based on the manufacturer's instructions, technical characteristics, and observations made during the conducted tests.

2. ANALYSIS OF SOIL-STRUCTURE INTERACTION

For the purpose of analyzing structure-soil interaction, it is first necessary to know the material and geometric characteristics of the interacting elements. When an external load is applied to a body, it induces changes in the internal forces within that body, i.e., the development of stresses, with the system's behavior varying significantly depending on the material properties of the loaded body. In the analysis of foundation structure–soil interaction, the problem is even more pronounced because soil constitutes a heterogeneous and anisotropic material whose mechanical behavior is complex to describe and model [3].

The behavior of soil is most commonly defined by constitutive models that establish the relationship between the stress and strain states. Due to the large variations in the physical and mechanical characteristics of different soil types, there is no single model that can describe all soil types and all forms of their behavior with sufficient accuracy. For this reason, a large number of models of varying complexity have been developed, whereby the choice of model directly affects the results of the foundation–soil–structure interaction analysis [4].

The relationship between stress and strain in soil is predominantly nonlinear. At lower load levels, elastic behavior is dominant, whereas with increasing loads, elasto-plastic deformations occur. The accuracy of the computation depends on the adopted constitutive model and the input parameters describing the soil, which makes their selection one of the key steps in the analysis of foundation–soil interaction.

The simplest model of elastic soil behavior is based on Hooke's law, while the most well-known idealized model of foundation–soil interaction is the Winkler model. The basic assumption of this model is that the soil deflection w (settlement) at any point is directly proportional to the applied load q, whereby this deflection is independent of the loads applied at other points, which can be expressed as (1):

$$q(x,y) = kw(x,y) \tag{1}$$

where: k – is the modulus of subgrade reaction, the unit of which is stress per unit length. In physical terms, the soil is represented by a system of mutually independent springs, due to which deformations occur only in the loaded zone, as shown in Figure 1. Despite the simplifications, the Winkler model still represents the basis of numerous foundation–soil–structure interaction analyses today.

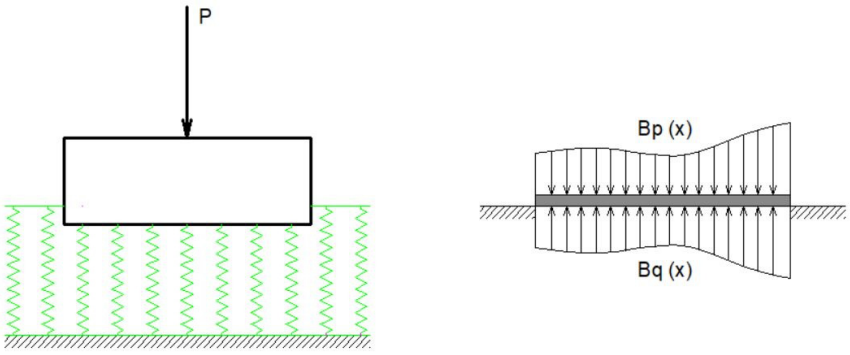


Figure 1: Winkler's soil behavior model

Due to the limitations of elastic models, numerous constitutive models have been developed to describe the nonlinear and elasto-plastic behavior of soil. Among the most significant are the Cam Clay (CC) and Modified Cam Clay (MCC) models, the Soft Soil (SS) model, the Hardening Soil (HS) and Hardening Soil Small (HSS) models, as well as models based on the 'Bubble' and 'Brick' concepts [5]. Contemporary numerical analyses most commonly use these advanced models because they enable a more realistic representation of soil behavior under different load levels and foundation conditions.

Many authors have investigated the distribution of contact stresses depending on the structure and soil type. Based on experimental research, they have established the patterns of distribution and intensity of contact stresses as a function of the applied load [6]. By analyzing the contact between rigid and elastic bodies, it can be observed that the stress distribution over the contact surface is not uniform. Its shape is influenced by a large number of factors, among which the most significant are the geometry and stiffness of the structure, the material characteristics from which it is made, as well as the composition and mechanical properties of the soil onto which the loads are transferred. The dependence of the behavior of the structure–soil system on the stiffness of the foundation and the structure, the shape and dimensions of the contact surface, the depth of foundation, the soil type, the type and eccentricity of loading, as well as other relevant parameters, has been the subject of numerous studies by authors Selvadurai [7], Olson [8], Hrubesova [9], and Dash [10].

In interaction analyses, cohesive and non-cohesive soils are considered. Figure 2 shows characteristic diagrams of contact stresses for different soil types, as well as the idealized stress distribution most commonly adopted in engineering practice. The diagram in Figure 2.a) refers to the case of a rigid foundation resting on cohesive soil, while Figure 2.b) shows the stress distribution for a flexible foundation on non-cohesive soil. Figure 2.c) represents the idealized linear distribution of contact stresses, which is often used in determining the relevant effects for structural dimensioning.

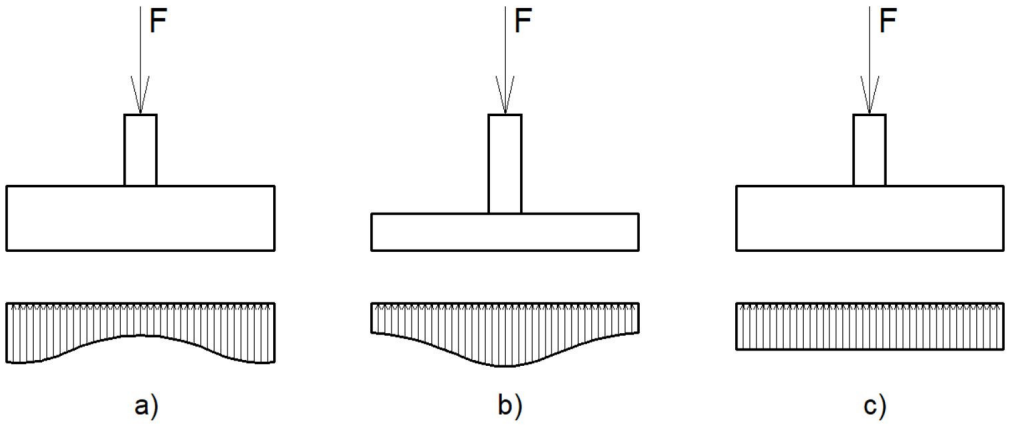


Figure 2: Distribution of contact pressure for different types of soil: a) cohesive soil (φ=0); b) non-cohesive soil (c=0); c) idealization of contact pressure

In addition to soil characteristics, the stiffness of the foundation structure also has a significant influence on the distribution of contact stresses. Figure 3 shows possible stress distributions for different combinations of foundation and soil stiffness. Depending on the intensity of loading and the stiffness ratio of the system, the diagrams may have a saddle-shaped, parabolic, or bell-shaped form.

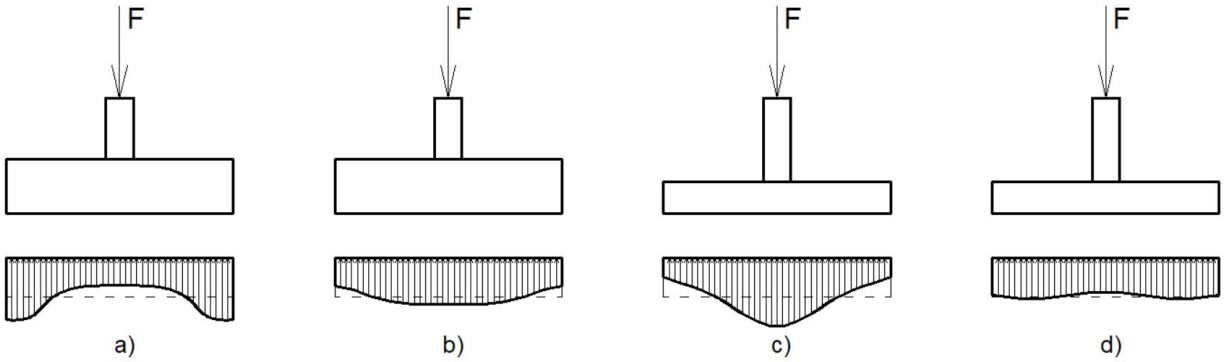


Figure 3: Distribution of contact pressure for different stiffnesses of the footing and soil: a) rigid footing, rigid soil; b) rigid footing, soft soil; c) flexible footing, rigid soil; d) flexible footing, soft soil

The theoretical foundations for the analysis of stress distribution in soil were established by the research of Boussinesq, and were later expanded through numerous experimental and numerical studies [11]. Poulos and Davis, analyzing circular plates of varying stiffness subjected to a concentrated force, showed that the shape of the contact stress diagram depends on the plate stiffness [12]. For an absolutely rigid plate, the distribution has a concave shape, whereas with decreasing stiffness, the maximum stress shifts toward the point of load application, with the stress value increasing as the foundation flexibility increases. Experimental investigations on circular foundations of different stiffness were also conducted by Barnes, and a modified diagram of contact stress distribution compared to Borowicka's research is given in Figure 4 [13]..

Research on strip foundations conducted by Terzaghi has shown a significant influence of soil type on the shape of contact stress diagrams. For cohesive soils, the distribution is predominantly saddle-shaped, while non-cohesive soils exhibit a parabolic distribution. Similar conclusions were confirmed by experimental investigations of rigid rectangular single footings carried out by Boldirev [14]. Further testing of foundations resting on compacted sand and gravel subgrades, conducted by Hegger, Djakov, Bonić and associates [15], showed that contact stresses most often have a parabolic or bell-shaped form.

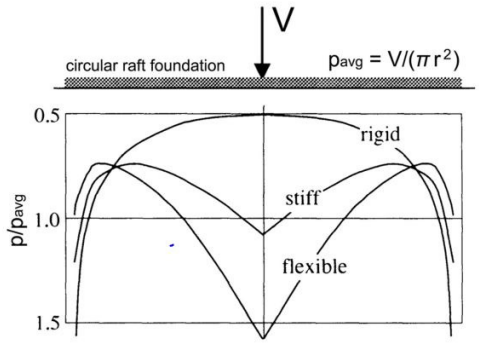


Figure 4: Shapes of contact pressure diagrams (modified from Barnes after Borowicka [13])

3. CONCEPT OF EXPERIMENTAL RESEARCH ON A FOOTING MODEL

3.1. Setup of the Experimental Research

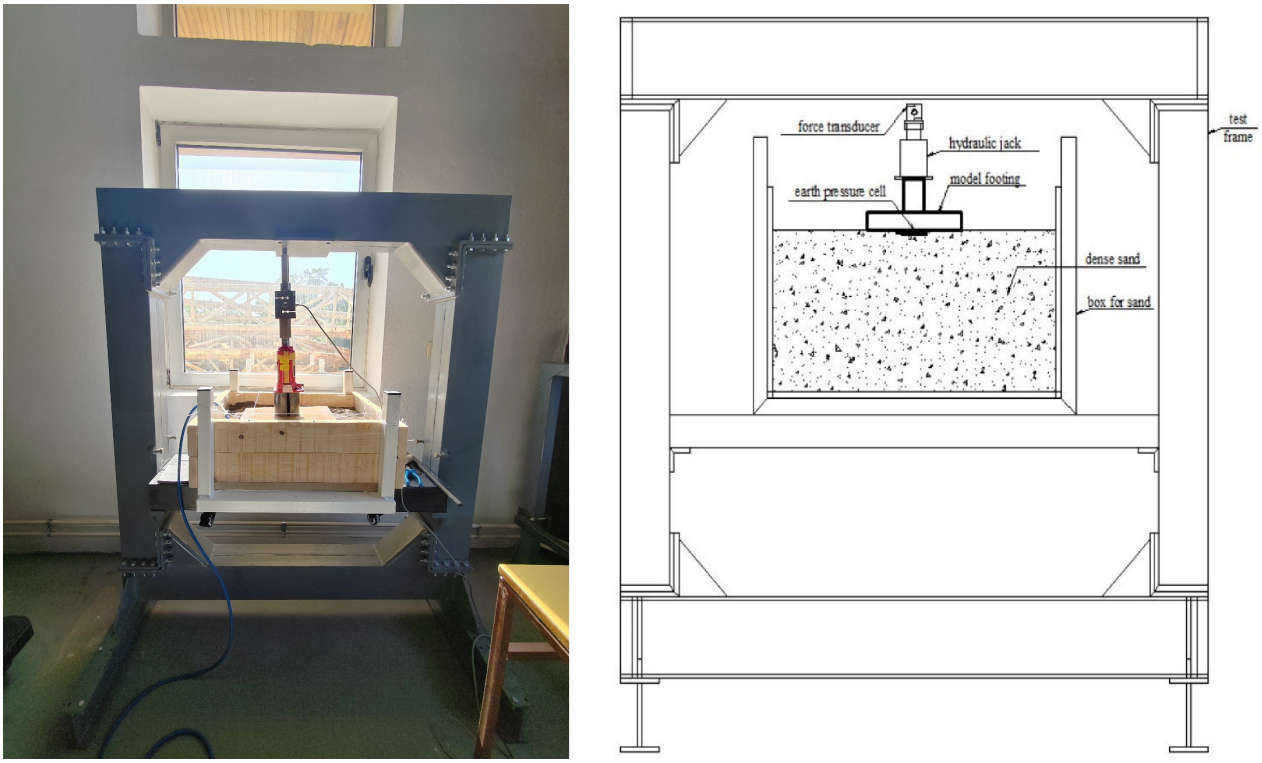


Figure 5: Experimental setup and instrumentation

The setup of experimental research on a footing model under laboratory conditions is shown in Figure 5. On the supports of the testing frame, a soil box is positioned, in which the soil was placed and compacted to a specified density. Subsequently, a earth pressure cell was installed in the soil, with the tested footing model placed above it. The load is applied to the footing model by means of a hydraulic jack, while the applied load is measured using a force transducer (load cell). Based on the measured values of contact stresses at defined time intervals and the applied load, the relationships between these quantities can be established, and thus the soil behavior can be defined.

3.2. Technical Characteristics of the Footing Model and Soil

The experimental research was conducted on a square footing model, as shown in Figure 6. The width (B) and length (B) of the footing model are 30 cm, and the thickness (h) is 6 cm. The model is made of concrete grade C20/25. The load is applied to the contact surface through a column of square cross-section, with side dimensions (b) of 10 cm and height (h1) of 10 cm.

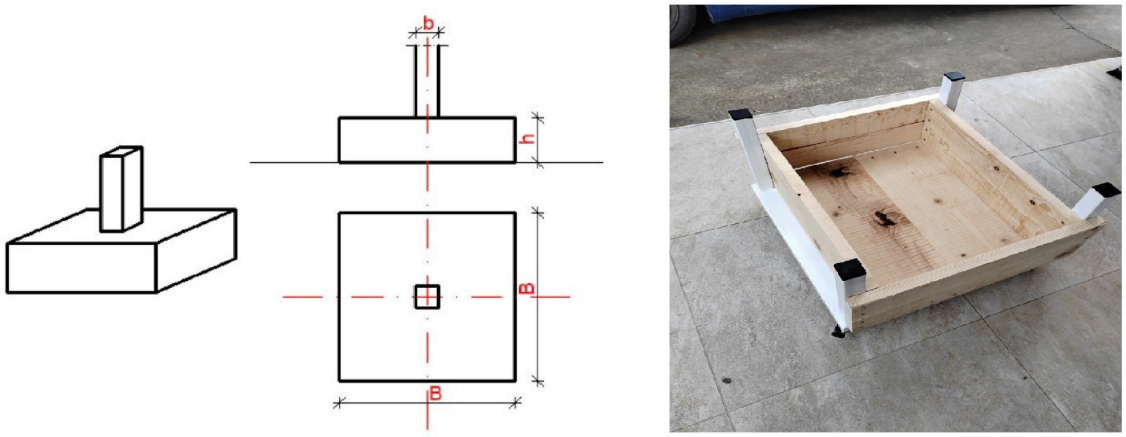


Figure 6: Dimensions of the footing and box for sand

The footing model rests on fine-grained soil – sand with a grain size of up to 2 mm, which was compacted to a specified density prior to testing. The box shown in Figure 6 was used for the experiment, with the sand placed inside and lateral expansion of the soil sample prevented. Since the stress distribution results were compared with the theoretical distribution corresponding to an rigid base, vertical displacement of the base (the bottom of the box) was also prevented in the model [16]. The sand box has a rectangular cross-section with dimensions of 60 × 47.2 cm, and the soil layer thickness is 15 cm, which corresponds to half the width of the footing. The cross-sectional dimensions of the box were determined such that approximately half the width of the footing base was provided on each side between the sidewalls of the box and the footing. For the purposes of this experimental research, the base can be considered incompressible since the vertical displacement of the bottom of the box is prevented.

3.3. Application and Technical Characteristics of Earth Pressure Cells in Structure–Soil Interaction

For measuring the contact pressure between the footing model and the soil layer, earth pressure cells (Earth Pressure Cells 4930 – EPC) manufactured by ACE Instrument Co., Ltd., shown in Figure 7, were used. These are sensors based on foil strain gauge (FSG) technology, intended for determining total stresses in geotechnical structures and experimental models. The strain gauges are placed on the active side of the cell, which is also the cell membrane, while the other side of the cell is the cover.

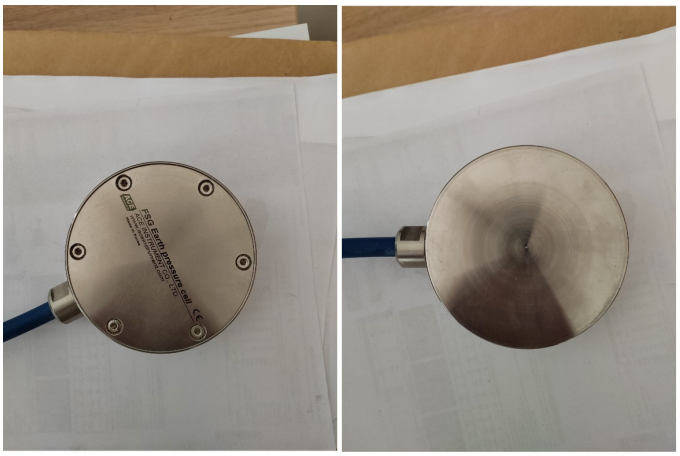


Figure 7: Earth pressure cell ACE Instument 4930, a) cover EPC; b) membrane EPC

The measuring principle is based on the deformation of the sensor's elastic membrane under the action of external pressure. The change in deformation is recorded using strain gauges connected in a Wheatstone bridge, whereby an electrical signal proportional to the applied pressure is generated.

The cells used in this research have an outer diameter of 100 mm and a total thickness of 20 mm. The sensor housing is made of stainless steel with high-quality epoxy protection, which ensures moisture resistance and long-term measurement stability. The sensor has a waterproof construction and is designed for use in demanding field and laboratory conditions.

The characteristics of the used measuring cell are shown in Table 1. The measurement range of the sensor is from 2 to 35 kg/cm², while the declared accuracy is ±0.5 % of the full scale range (FSR). The non-linearity of the sensor does not exceed ±1.0 % FSR, and the allowable overload is 120 % FSR. The output signal of the sensor is 1 mV/V at nominal load, while the nominal electrical resistance of the Wheatstone bridge is 350 Ω.

Table 1: Specification earth pressure cell

Characteristics	Model: 4930
Sensor Element	FSG sensor (foil strain gage)
Range	~23.36 kg/cm^2
Accuracy	±0.5% FSR
Nonlinearity	±1.0% FSR
Over range capacity	120% FSR
Rating output	1 mV/V (1,000x10^{-6})
Resistance	350Ω
Insulation resistance	More than 100 MΩ/500V
Exciting voltage recommended	Less than 5 VDC
Exciting voltage allowable	Less than 10 VDC
Operating temperature	-30~80°C
Water proof	300m H_2O
Weight	1.2kg
Material	Stainless steel, high grade epoxy potting
Signal cable	Ø10mm, 0.5mm^2 x 5C shielded PVC sheath cable

The sensor is powered by a DC voltage of less than 5 V, while the maximum permissible excitation voltage is 10 V. The declared operating temperature range is from −30 °C to +80 °C, which enables application in a wide range of experimental and field conditions.

The pressure cell can be installed in any position, as shown in Figure 8, and in accordance with the manufacturer's recommendations [17]. During the installation of the pressure cell, special attention was paid to achieving uniform contact between the pressure cell, the footing model, and the soil, in order to ensure the most uniform transfer of contact pressure to the active surface of the sensor (the cell membrane) and to reduce the influence of local stress concentrations.

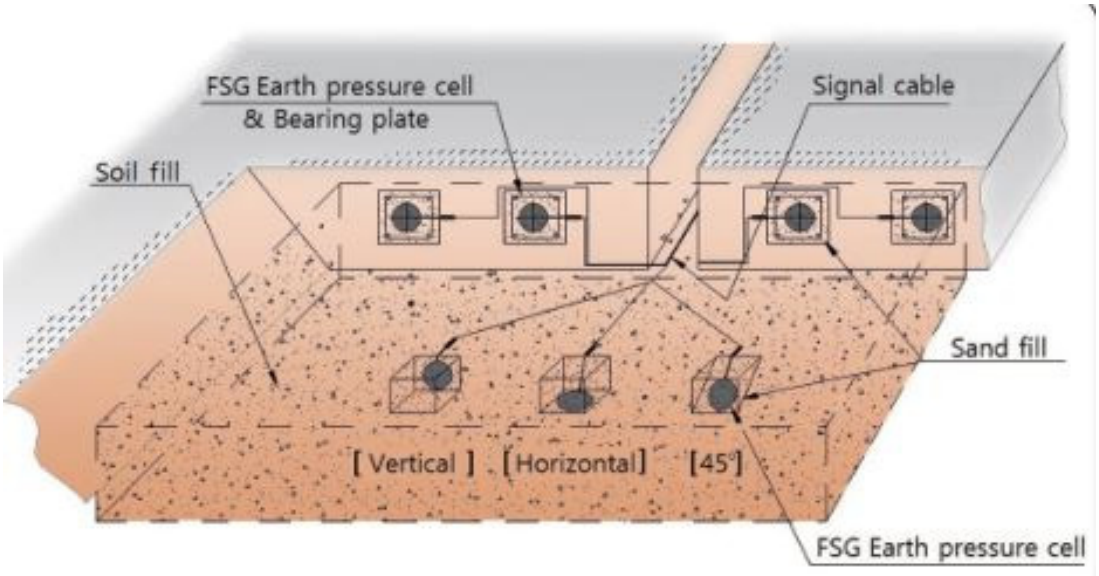


Figure 8: Installation of standard type FSG earth pressure cell

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Experimental Results

The results given in Table 2 were obtained using the measuring equipment specified in Chapter 3 with the given experimental setup. The load was applied using a hydraulic pump, and the load intensity was measured using the HBM RSCC 3C/5t force transducer. The loading results provided the load intensity values over a time interval, while the results obtained from the earth pressure cell defined the pressure on the contact surface between the single footing model and the soil, at the position where the sensor was placed, also over a time interval. Based on these results, the relationship between the applied load and the soil pressure can be established for any moment during the measurement period. Data acquisition was performed using the HBM QuantumX MX840A data acquisition system and Catman Easy data acquisition software. An example of the acquired data from one of the measurements is shown in Figures 9 and 10.

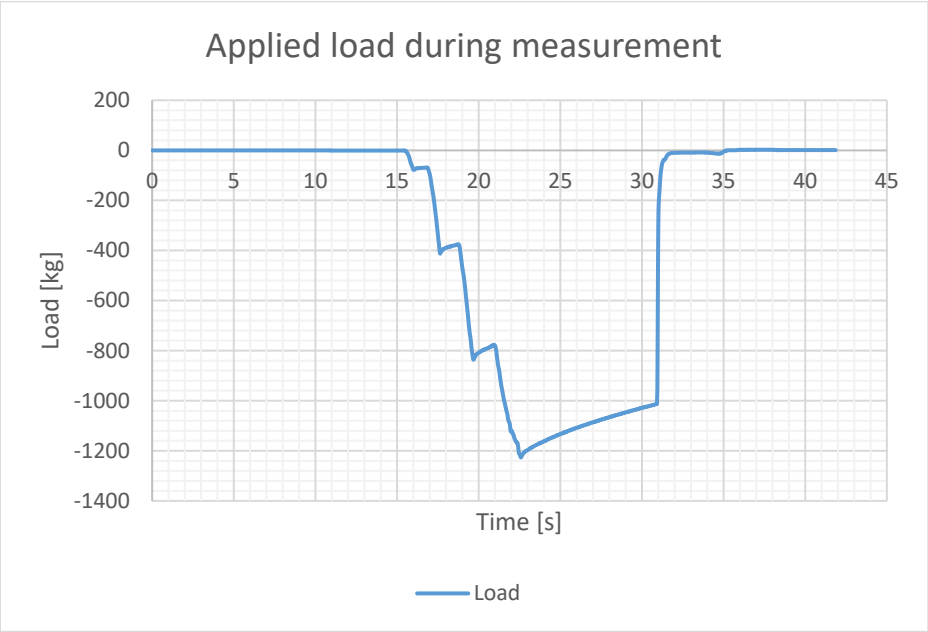


Figure 9: Applied load during measurement

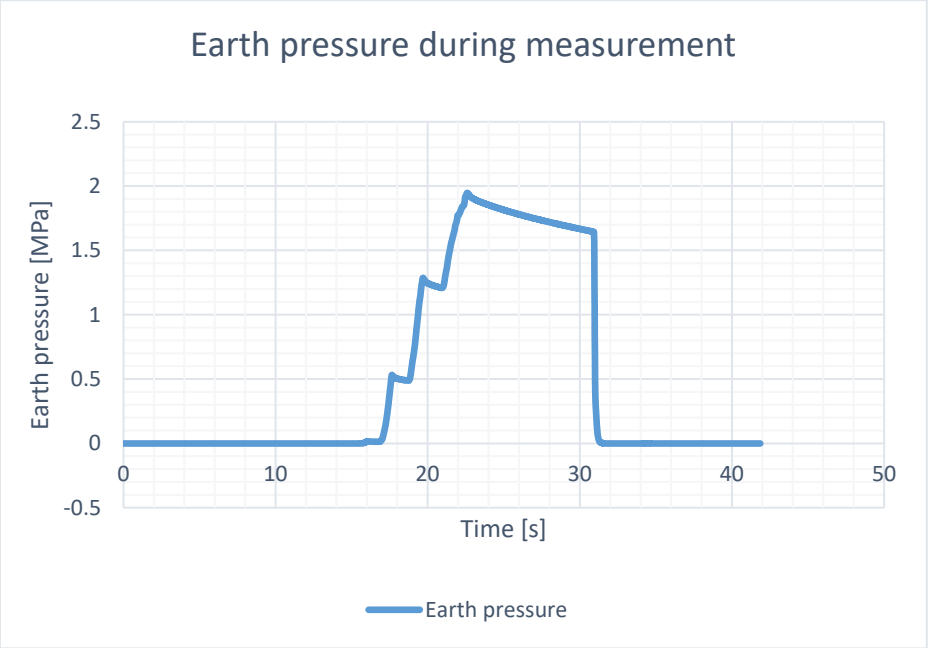


Figure 10: Earth pressure during measurement

The measurement results shown in Table 2 represent the stress values on the contact surface when the cell was embedded in the soil at the center of the footing. Three measurements were performed, in each of which the load was applied in five steps of different intensities. The first applied value was 750 kg, followed by 1250 kg, 500 kg, 1000 kg, and 1500 kg. Prior to embedding the cell into the soil, a 5 mm thick layer of fine-grained sand with a maximum grain diameter of 0.5 mm was placed between the compacted sand and the cell, in order to transfer the reactive load from the soil as evenly as possible to the cell membrane. During the test, the footing model was directly supported on the cell cover, as shown in Figure 11.a). One measurement was also performed with a 5 mm thick rubber layer between the pressure cell cover and the footing model, as shown in Figure 11.b), in which the pressure values on the contact surface were considerably lower than the expected uniformly distributed load values, so these values were not considered in further analysis. Horizontal stresses were also measured by placing the cell at the contact between the sand and the box walls, as shown in Figure 11.c); the measured pressure values amounted to 0.0 MPa.

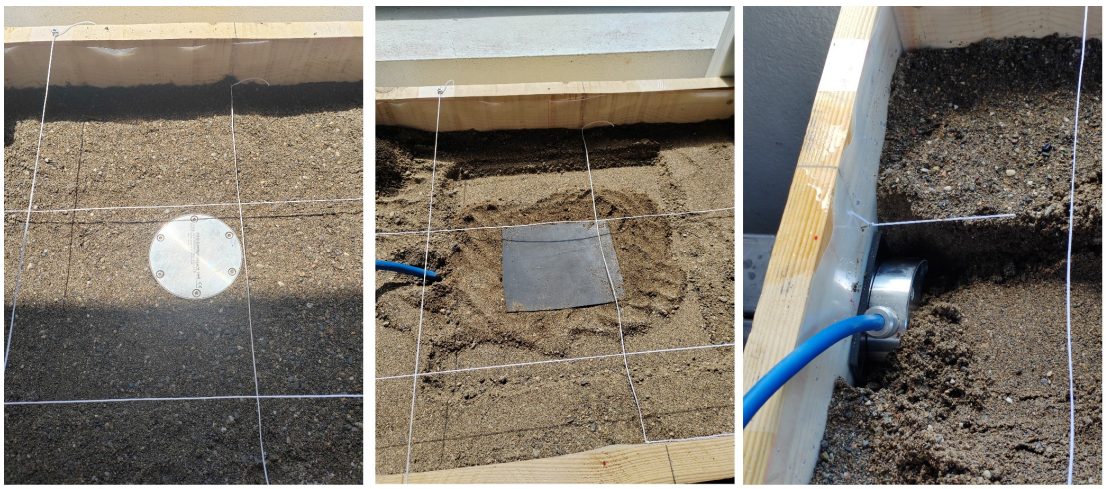


Figure 11: Different earth pressure cell placement configurations

Table 2: Applied loads and measured contact pressures during testing

Measurement ID	Load [kg]	Load [kN]	Measured contact pressures [MPa]	Calculated contact pressures [MPa]	Relative deviation [%]
1.1	526.77	5.17	0.026	0.057	54.7
2.1	769.92	7.55	0.031	0.084	63.1
3.1	1087.81	10.67	0.066	0.119	44.4
4.1	1273.84	12.50	0.071	0.139	48.9
5.1	1638.68	16.08	0.124	0.179	30.6
1.2	552.46	5.42	0.045	0.060	25.3
2.2	744.60	7.30	0.054	0.081	33.4
3.2	1022.94	10.04	0.087	0.112	22.0
4.2	1253.32	12.30	0.106	0.137	22.5
5.2	1591.70	15.61	0.146	0.174	15.8
1.3	531.46	5.21	0.055	0.058	5.0
2.3	775.53	7.61	0.076	0.085	10.2
3.3	1029.69	10.10	0.100	0.112	10.9
4.3	1248.80	12.25	0.122	0.136	10.4
5.3	1505.13	14.77	0.148	0.164	9.8

4.2. Analysis of Results

The results obtained from the testing and the relationship between the results for the first, second, and third measurements are shown in Figures 12, 13, and 14.

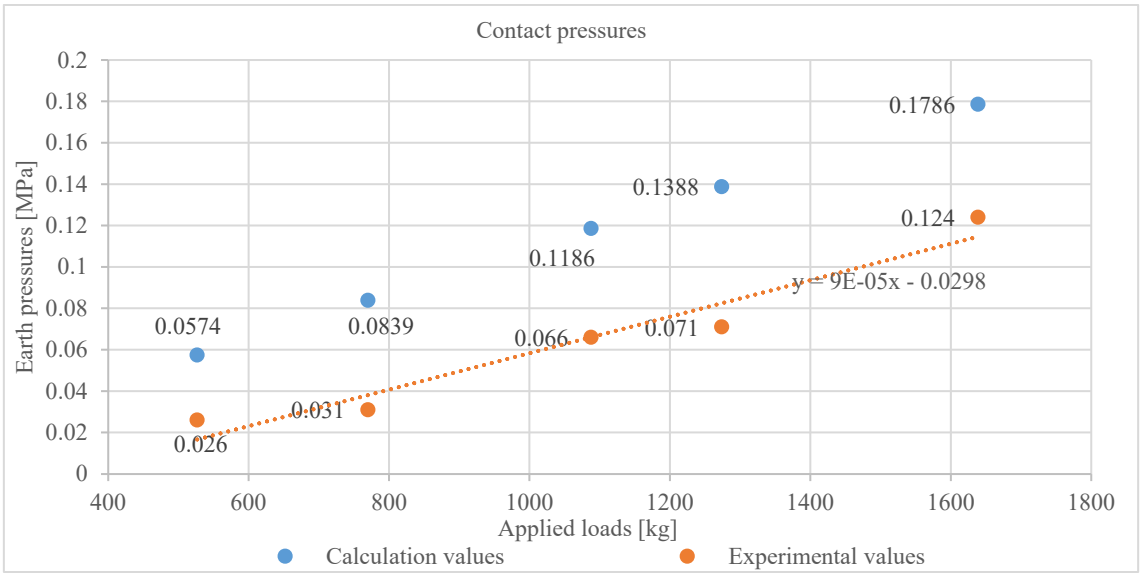


Figure 12: Experimental and calculated contact pressures as a function of applied load – measurement 1

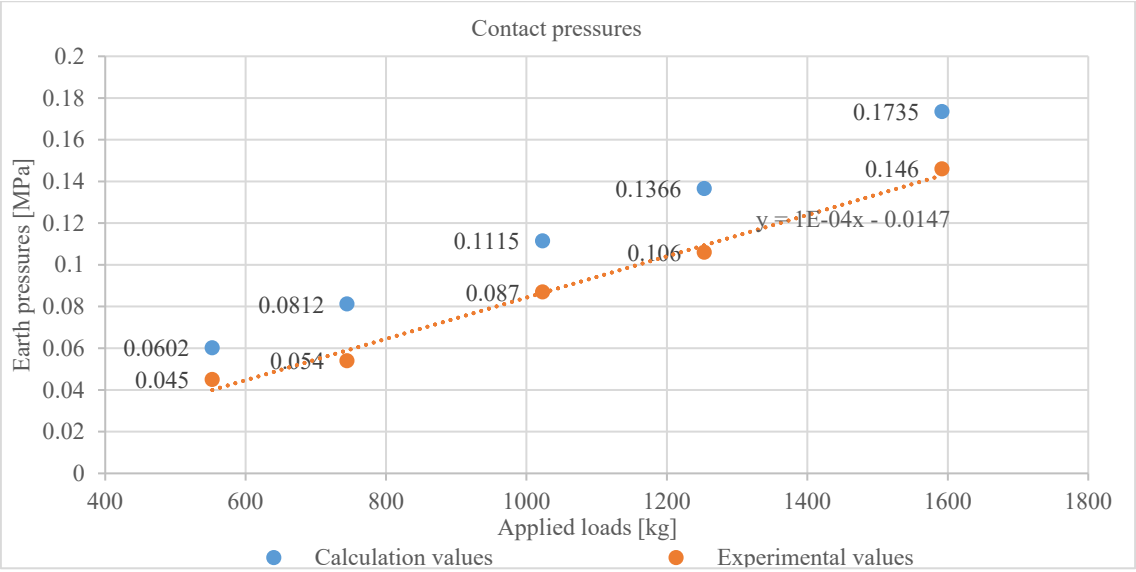


Figure 13: Experimental and calculated contact pressures as a function of applied load – measurement 2

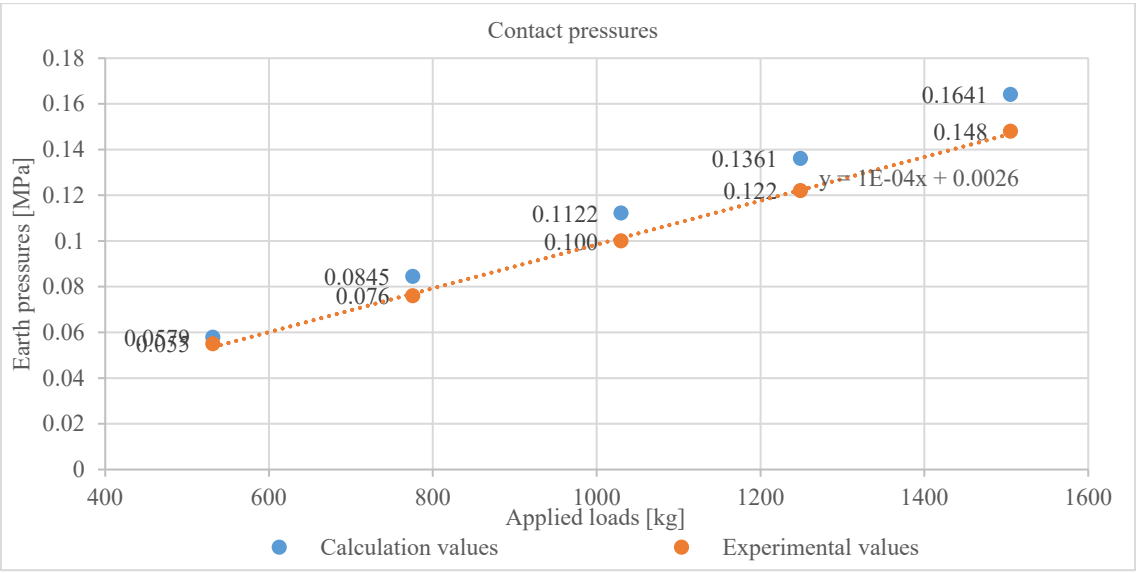


Figure 14: Experimental and calculated contact pressures as a function of applied load – measurement 3

5. CONCLUSION

Based on the presented results obtained through three measurements, it can be concluded that there is a significant deviation between the calculated uniformly distributed contact pressure and the pressure measured during the experimental research. The largest deviation can be observed in the first measurement, with an average deviation of the compared values of 48.3%. This deviation occurred due to the unevenly compacted sand layer around and beneath the cell, because after compacting the sand layer, the compaction was disturbed during the installation of the earth pressure cell. Additionally, it is difficult to estimate the required space for cell installation, so in each test, before the first load application, it happens that the footing model rests either only on the cell or only on the soil around the cell. To avoid such uneven support of the footing model, it is necessary, prior to the start of measurements and after the installation of the pressure cell into the soil, to prepare the contact surface on which the footing model will rest. This preparation entails applying the maximum load to the cell and the surrounding soil using a rigid plate with a contact area not smaller than the area of the footing model, in order to bring the ground level and the cell surface to the same horizontal plane.

This error introduced during the measurement is significantly reduced already in the second measurement, where the average deviation between the compared values is 23.8%; in the third measurement, the average deviation amounts to 9.2%, which is acceptable for the given experimental setup and its geometric characteristics. In all three measurements, the experimental values are lower than the calculated values of the uniformly distributed load on the contact surface.

Regarding the relationship between the contact pressure and the applied load obtained experimentally, in all three measurements, minimal deviations from a linear relationship are observed, which is in accordance with the technical characteristics of the earth pressure cell.

In the process of analyzing the results, the imprecise load application should also be taken into account, which could not be avoided in this experimental setup, so the occurrence of small eccentricities of the applied load was possible. Also, the small thickness of the soil layer should not be neglected, so the difference in the soil layer thickness beneath the cell and beneath the footing model relative to the cell thickness is significant. This influence could be eliminated by conducting research in the field, where there would be no limitations at least regarding the thickness of the tested soil sample.

All presented and analyzed results refer only to the contact pressures at the center of the footing model, so it would be necessary to determine multiple positions of the pressure cells in order to obtain the stress distribution over the contact surface. Due to the small dimensions of the footing model, this was not possible in the present experimental research.

ACKNOWLEDGEMENTS

This paper is a result of the research conducted within the project grant number 451-03-34/2026-03/200108 supported by the Ministry of Science, Technology and Innovation of the Republic of Serbia.

REFERENCES

[1] D. Milovic and M. Djogo, “Problemi interakcije tlo-temelj-konstrukcija”, Fakultet tehničkih nauka, Novi Sad, (Serbia), (2009)

[2] M. Vasic, “Inženjerska geologija”, Fakultet tehničkih nauka, Novi Sad, (Serbia), (2001)

[3] D. Milovic and M. Djogo, “Greške u fundiranju”, Fakultet tehničkih nauka, Novi Sad, (Serbia), (2005)

[4] I. F. Collins, “Elastic/plastic models for soils and sands”, Int. J. Mech. Sci., Vol. 47 (4–5), pp. 493–508, https://doi.org/10.1016/j.ijmecsci.2004.12.016, (2005)

[5] J. T. H. Wu and S. C.-Y. Tung, “Determination of Model Parameters for the Hardening Soil Model”, Transportation Infrastructure Geotechnology, Vol. 7 (1), pp. 55–68, https://doi.org/10.1007/s40515-019-00085-8, (2020)

[6] D. F. A. Santos, A. F. Lima Neto and M. Ferreira, “Punching shear resistance of reinforced concrete footings: evaluation of design codes”, Revista IBRACON de Estruturas e Materiais, Vol. 11 (2), pp. 432–454, https://doi.org/10.1590/S1983-41952018000200011 , (2018)

[7] A. P. S. Selvadurai, “Elastic Analysis of Soil-foundation Interaction”, Elsevier Scientific Publishing Company, Department of Civil Engineering, Carleton University, Ottawa, (Canada), (1979)

[8] R. E. Olson, “Advanced Soil Mechanics, Stress Distribution”, Chaoyang University of Technology, (Taiwan), (2003)

[9] E. Hrubesova, H. Lahuta, M. Mohyla, T. B. Quang and N. D. Phi, “Experimental sensitivity analysis of subsoil–slab behaviour regarding degree of fibre-concrete slab reinforcement”, InIOP Conference Series: Earth and Environmental Science, Vol. 143 (1), p.012005, https://doi.org/10.1088/1755-1315/143/1/012005, (2018)

[10] S. K. Dash, R. Saikia, and S. Nimbalkar, “Contact pressure distribution on subgrade soil underlying geocell reinforced foundation beds”, Frontiers in Built Environment, Vol. 5, p.137, https://doi.org/10.3389/fbuil.2019.00137, (2019)

[11] M. J. Boussinesq, “Applications des potentiels a l’etude de l’equilibre et du mouvement des solides elastiques”, FR: Gauthier–Villard, Paris, (Fance), (1885)

[12] E. H. Davis and H. G. Poulos, “Discussion of'An analysis of consolidation theories”, Journal of the Soil Mechanics and Foundations, Vol. 96 (1), pp. 334-336, https://doi.org/10.1061/JSFEAQ.0008018, (1970)

[13] H. Borowicka, “Druckverteilung unter elastischen Platten”, Ingenieur-Archiv, Vol. 10, pp. 113-125, https://doi.org/10.1007/BF02084278, (1939)

[14] G. Boldirev, S. Boldirev, O. Hrjanina, “Resistance of Wall Footings Foundations with Soil–Structure Interaction”, Penzenski University of Architecture and Civil Engineering, Penza, (Russia), (2016)

[15] Z. Bonić, E. Zlatanović, S. Brčić, P. Blagojević, N. Romić, and D. Cvetković, “Theoretical and experimental research on the punching shear capacity of reinforced concrete column footings rested on the ground”, Engineering Structures, Vol. 249, p. 113154, https://doi.org/10.1016/j.engstruct.2021.113154, (2021)

[16] S. Mihajlović, I. Despotović, J. Bojković, M. Janićijević Milačak, and V. Mandić, “Settlement analysis of isolated footings using two soil models”, Engineering Today, Vol. 4 (4), pp. 23-31, https://doi.org/10.5937/engtoday2600006M, (2025)

[17] https://svai-lab.ru/upload/docs/4910-4920-4930-4940.pdf

INDEX OF AUTHORS

A	ORCID
Aleksić, V., C13	0000-0003-2337-1288
Atmadzhova, D., B17	0000-0002-9241-6171
Avramović, Ž., B35	

B	ORCID
Bajić, S., G19	0009-0005-2052-8499
Baralić, J., C67	0000-0002-8023-7942
Bebeselea, B., A1	
Belošević, J., E55, E61	0009-0007-6326-4117
Belošević, S., E55, E61	0000-0003-3555-054X
Bernardi, E., D41	0000-0001-5248-9352
Biga, N., D57	
Bižić, M., B51, G45	0000-0002-3815-9784
Bjelić, M., C25	0000-0002-6148-1982
Bogojević, N., F17	0000-0003-3951-3519
Bojković, J., G13, G27, G45	0000-0002-0773-0087
Bošković, G., E1	0000-0002-4597-922X
Bošković, M., E1	0000-0002-3637-2741
Brzaković, L., F7	0000-0003-0583-0583
Bulatović, V., G13	0000-0002-6843-103X
Burdin, B., A17	0000-0001-9162-4545

Č	ORCID
Čepić, Z., E1	0000-0001-5853-4767

Ć	ORCID
Ćirić-Kostić, S., F17	0000-0002-9684-0850
Ćojbašić, Ž., P1, A95	0000-0002-4581-1048

D	ORCID
Defoort, M., D31	0000-0001-6346-0774
Despotović, I., G13, G45	0000-0002-4452-1372
Drvar, N., F17	0000-0001-6848-9824
Dubonjic, L., D23, D31	0000-0002-4945-6194
Dučić, N., C13	0000-0001-7351-5898
Dulović, T., C25	0000-0003-3265-5661

Đ	ORCID
Đokić, R., A25, A85	0000-0002-1578-0543
Đorđević, V., D49	0000-0002-4927-7225
Đukić, M., E55, E61	0000-0001-9382-0443
Đurić, A., C53	0000-0002-0251-6364

E	ORCID
Erić Cekić, O., E21, F17	0000-0002-4702-1728

F	ORCID
Filipović, V., D49	0000-0002-9601-5332
Franc, I., D57, D105	0009-0000-4609-1081
Francu, C., A1	0000-0002-5086-8751

G	ORCID
Gavrilović, B., B35	
Gergova, N. I., C77	0000-0001-7153-8017
Geroski, T., D87	0000-0003-1417-0521
Glišović, J., A41, B41	0000-0002-8251-1722
Grković, V., C37, C45	0000-0003-1229-4678

H	ORCID
Hozdić, E., C1	0000-0002-4725-5444
Hozdić, Em., C1	

I	ORCID
Ijačić, A., E11	
Ilanković, N., A25, A59, A69	0000-0002-9110-4926
Iontchev, E., B29	0009-0005-3585-7084
Ivanova, M., E29	
Ivanović, M., C25	0000-0002-1576-7613

J	ORCID
Janićijević Milačak, M., G27, G35, G45	0000-0003-1215-9070
Jančić, S., D105	0009-0008-2552-6275
Janjatović, P., E21	0000-0002-3779-9341

Jojić, T., A25, A85	0000-0002-2174-3394
Jordović Pavlović, M., D87	0000-0003-4038-6537
Jovanović, Đ., F1	0000-0003-1222-1292
Jovanović, E., D105	0009-0003-7007-6537
Jovanović, M., D57, D105	0009-0008-9032-8195
Jovanović, V., A33, C61	0000-0001-9252-7894
Jovičić, A., C67	0000-0001-7344-6466

K	ORCID
Karan, I., E35	0000-0001-7557-2482
Katipoğlu, G., D69	0000-0002-6774-5707
Katona, M., A25	0000-0002-2365-7975
Kelić, L., D1	0009-0009-9972-9849
Kirilov, F., B1	0009-0006-0558-5265
Klobčar, D., C53	0000-0002-6130-0328
Knežević, S., E47	
Kojić, N., G35	0000-0002-0447-7426
Kokalanov, V., D81	0009-0005-7950-2904
Kostić Miličić, A., B11	0000-0002-5654-757X
Krastanov, D., A17	
Krile, S., E11	
Krstić, M., F1	0000-0002-3544-1522
Krunić, G., E11	
Kukuseva Paneva, M., D81	0009-0001-2556-3796
Kuvač, I., E35	0000-0001-7086-2843
Kužet, Ž., E1	0009-0004-0958-517X

L	ORCID
Li, X., P11	
Liu, X., A9	0009-0002-9652-0491
Luan, X., D41	0000-0002-4805-1726
Lukić, J., B41	0000-0002-5893-3976

M	ORCID
Madić, M., C19	0000-0002-2310-2590
Marković, G., A51	0000-0002-0957-0718
Marković, S., C61	0009-0003-6334-7415
Mihajlović, J., C61	0000-0002-9787-0226
Mihajlović, S., G13, G27, G45	0000-0001-5811-5204
Mihaylov, E. M., B29	0009-0006-8815-8313
Mijailović, N., C13	
Milčić, D., C53	0000-0002-3936-7462
Milčić, M., C53	0000-0002-1089-8390
Milenković, A., D97	0000-0001-6699-8772
Milenković, B., F1	0000-0003-4606-0095
Milić, P., A95	0000-0001-7778-6097
Milićević, I., C13	0000-0003-0476-4991
Milićević, V., D57, D81	0000-0002-5587-2717
Milković, D., B11	0000-0002-3739-4600
Milojević, M., E47	
Milojković, M., C19	0000-0001-7623-1495
Miloradović, D., B41	0000-0003-1427-9789
Miloradović, N., A41, B41	0000-0001-6846-6091
Milošević, A., E47	
Milošević, B., G27, G35	0000-0003-0182-7993
Milošević, S., G19	0009-0005-1781-8268
Milutinović, A., D97	0009-0004-1300-3232
Milutinović, M., A79	0000-0002-1642-951X
Mitrović, A., C67	
Mladenović, P., A51	0000-0002-3315-4642
Mrđa, B., E35	

N	ORCID
Nadzinski, G., D11	0000-0002-0708-4812
Naydenov, V., A17	0009-0008-2211-534X
Nikolić, A., F17	0000-0003-2548-4083
Nikolić, B., A33, C61	0000-0002-9694-6719
Nikolić, M., D87	0000-0002-8537-7944
Nikolić, R., F7	0009-0008-1291-0059
Nikolić, S. S., C19	0000-0003-2745-3862
Nikolov, V., B1	0009-0004-1252-1599

O	ORCID
Obradović, O., B11	
Ojleska Latkoska, V., D11	0000-0002-6078-1916

P	ORCID
Pajović, S., C37, C45	0000-0001-9697-2266
Paleva-Kadiyska, B., G1	0000-0002-8514-4542

Paszke, W., D41	0000-0001-5060-0993
Pavković, M., D97	0000-0001-7776-6045
Pavlov, V., B59, B65	0009-0003-5829-2667
Pavlović, G. V., A51	0000-0002-7230-1908
Pavlović, J., A33	0000-0002-8825-7307
Pavlović, Z. G., B75, D75	0000-0001-6076-1811
Perić, J., C25, C37, C45	0000-0001-8747-3350
Petrova, I., B69	0009-0001-6321-7210
Petrović, A., C37, C45	0000-0002-5732-3578
Petrović, D., B51	0000-0002-1140-0662
Petrović, G., P1, A95	0000-0002-3562-9576
Petrović, N., A33, C61	0000-0002-9166-1263
Petrović, Z., F7	
Petrović Savić, S., F7	0000-0002-5629-6221
Piskulić, M., A41	0009-0005-7199-1280
Popović, M., C13	0000-0003-0318-7133
Prodanovic, S., D41	0000-0003-2682-6022
Prsic, D., D23, D31	0000-0001-7934-0865

R	ORCID
Radičević, B., C37	0000-0002-4182-4095
Radičević, V., B75, D75	
Radić, N., A79	0000-0001-6859-5774
Radulović, S., B11	0000-0001-6337-0699
Rajnović, D., E21	0000-0002-5303-4402
Ralev, V., B17, B93	
Rasinac, M., C25, C45	0000-0003-4933-3765
Ružić, D., A69	0000-0002-3320-4750

S	ORCID
Samardžić, S., A79	
Savaniu, M., A1	
Savić, A., B75, D75	0000-0002-8099-1136
Savković, M., A51	0000-0002-4501-9149
Schwai, M., E35	
Sescu-Gal, C., A1	0009-0004-5805-8368
Shalamanova, T., E29	0009-0006-9331-4739
Simić, G., B11	
Sinđelić, V., F17	0000-0001-5728-8218
Song, S., D31	0000-0002-4780-0967
Song, X., D31	0000-0001-8476-5112
Stanković, A., A95	0000-0001-8841-1258
Stanković, M., A95	
Stavrov, D., D11	0009-0004-3437-5932
Stefanova, P., G7	0009-0001-6539-8342
Stjepanović, D., B35	
Stjepanović, Z., B35	
Stojanovic, V., P11, D23, D31, D41	0000-0002-6005-2086
Stojić, N., G13	0000-0002-6228-2889
Stojkovikj, N., D81	0000-0003-2770-0771
Stoyanova, E., B29	
Stupar, D., E35	0000-0002-7801-7801
Svičević, M., D97	0000-0003-2791-3849

T	ORCID
Tao, H., P11	0000-0001-5279-2458
Timotijević, M., E21	0009-0004-6735-090X
Todorović, M., A51, E1	0000-0003-3684-2819
Tonciu, O., A1	
Tosheska Radović, S., D11	
Trajković, A., A95	0009-0006-1390-8167
Trifković, S., A79	0000-0003-0656-5718
Trifonov, K., B81	0000-0002-9130-483X
Trifunović, M., C19	0000-0002-4653-6957
Tufegdžić, M., F7	0000-0003-3856-1498

U	ORCID
Urošević, M., A9	0000-0003-2622-3694
Utku, S., D69	0000-0002-8786-560X

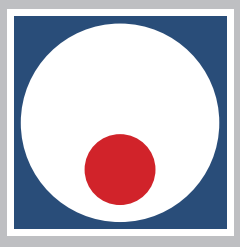
V	ORCID
Vasiljević, R., B75, D75	0000-0003-0458-8545
Vasiljević, S., B41	0000-0002-2235-1309
Vasković, S., E11	0000-0001-7953-4053
Vassilev, V., C77	0000-0002-1952-4096
Veljković, D., C67	
Vladić, J., A85	0000-0003-1518-1935
Vučićević, N., D97	0000-0002-4903-7280
Vuherer, T., C53	0000-0002-1866-8266
Vujanac, R., A41	0000-0002-8215-5668

Vujičić, V., C13	0000-0002-7037-3545

X	ORCID
Xie, X., D23	0000-0003-4822-3134

Z	ORCID
Zdravković, N., C53	0000-0001-6085-7522
Zdravković, N. B., A51	0000-0001-6387-2816
Zelić, A., A25	0000-0002-9903-1581
Zochinova, V., B59, E29	0009-0004-0697-4620
Zomer Machado, J. P., D23	0000-0002-9053-7463
Zrnić, N., A9	0000-0002-0078-7275

Ž	ORCID
Žerajić, V., A59, A69	
Živanić, D., A25, A59, A69	0000-0003-2570-0727
Živanović, M. M., D57	0009-0005-9264-3314



FACULTY OF MECHANICAL AND
CIVIL ENGINEERING IN KRALJEVO

UNIVERSITY OF KRAGUJEVAC
Dositejeva 19, 36000 Kraljevo, Serbia

