Fakulteta za industrijski inženiring
Faculty of Industrial Engineering

Razvoj industrijskega inženiringa: Priložnosti, potenciali in izzivi za zeleni prehod z uporabo umetne inteligence

Development of Industrial Engineering: Opportunities, Potentials and Challenges for Green Transition through the Use of Artificial Intelligence

Zbornik recenziranih prispevkov
10. mednarodne znanstvene konference

Proceedings of Reviewed Papers
of 10th International Scientific Conference

Priložnosti, potenciali in izzivi za zeleni prehod z uporabo umetne inteligence

Opportunities, Potentials and Challenges for the Green Transition through the Use of Artificial Intelligence

Novo mesto, 14. maj 2026

Fakulteta za industrijski inženiring
Faculty of Industrial Engineering

Mednarodna znanstvena konferenca o razvoju industrijskega inženiringa: Priložnosti, potenciali in izzivi za zeleni prehod z uporabo umetne inteligence

International Scientific Conference on the Development of Industrial Engineering: Opportunities, Potentials, and Challenges for the Green Transition through the Use of Artificial Intelligence

Zbornik recenziranih prispevkov
10. mednarodne znanstvene konference

Proceedings of Reviewed Papers
of 10th International Scientific Conference

Priložnosti, potenciali in izzivi za zeleni prehod z uporabo umetne inteligence

Opportunities, Potentials, and Challenges for the Green Transition through the Use of Artificial Intelligence

Novo mesto, 14. maj 2026

Zbornik recenziranih prispevkov 10. mednarodne znanstvene konference | Proceedings of Reviewed Papers of 10th International Scientific Conference.

10. mednarodna znanstvena konferenca o razvoju industrijskega inženiringa: Priložnosti, potenciali in izzivi za zeleni prehod z uporabo umetne inteligence. Novo mesto, 14. maj 2026.

10th International Scientific Conference on the Development of Industrial Engineering: Opportunities, Potentials and Challenges for the Green Transition through the Use of Artificial Intelligence. Novo mesto, November 14, 2026.

Organizatorji konference | Conference Organizers:
Fakulteta za industrijski inženiring Novo mesto | Faculty of Industrial Engineering Novo mesto
Sveučilište Sjever | University of North
{Rudolfovo – Znanstveno in tehnološko središče Novo mesto | {Rudolfovo – Science and Technology Centre Novo mesto
Fakulteta za informacijske študije v Novem mestu | Faculty of Information Studies in Novo mesto
Fakulteta za organizacijske študije Novo mesto | Faculty for organisation studies Novo mesto
Visoka šola za upravljanje podeželja Grm Novo mesto | Landscape Governance College Novo mesto

Založnik zbornika | Publisher of Proceedings:
Fakulteta za industrijski inženiring Novo mesto | Faculty of Industrial Engineering Novo mesto
Kraj | Place: Šegova ulica 112, Novo mesto, Slovenija

Leto izida | Year: 2026

Tehnični urednik | Technical Editor: doc. dr. Tomaž Perme, Matic Medvešek

Tisk na zahtevo.

accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work, including, but notwithstanding language, style, data, interpretations, are appropriately investigated and resolved.

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

62-027.31:502.131.1(082)

MEDNARODNA znanstvena konferenca o razvoju industrijskega inženiringa (10 ; 2026 ; Novo mesto)
Razvoj industrijskega inženiringa: priložnosti, potenciali in izzivi za zeleni prehod z uporabo umetne inteligence = Development of industrial engineering: opportunities, potentials and challenges for green transition through the use of artificial intelligence : zbornik recenziranih prispevkov 10. mednarodne znanstvene konference = proceedings of reviewed papers of 10th International Scientific Conference : Novo mesto, 14. maj 2026 / [organizatorji konference Fakulteta za industrijski inženiring Novo mesto ... [et al.] ; uredniški odbor Tomaž Perme, Simon Muhič, Lucija Galič]. - Novo mesto : Fakulteta za industrijski inženiring = Faculty of Industrial Engineering, 2026

ISBN 978-961-7097-15-3
COBISS.SI-ID 277670147

KAZALO | TABLE OF CONTENTS

Verification of Energy Envelope Performance Through In-Situ U-Value Measurements and Their Impact on District Heating Energy Consumption……1
Đorđe Karić, Mirko Komatina, Dimitrije Manić, Simon Muhič, Goran Vorotović

Experimental investigation of the reliability of photovoltaic panel mounting systems on flat roofs using a controlled climate test chamber……9
Marko Gložinić, Matija Gložinić, Ante Čikić, Zoran Busija

Maximum Use of Renewable Energy in Low Accumulative Facilities of Intensive Vegetable Production ……15
Marko Gložinić, Matija Gložinić, Ante Čikić, Zoran Busija

Analysis of the environmental impacts of different electricity production sources in Slovenia using the Life Cycle Assessment methodology……21
Damjan Balabanič, Gregor Fijavž Kranjčan, Jelena Topić Božič

Hourly Consumption-Based Electricity Mixes for Slovenia, Croatia, and Serbia: A Time-Resolved Life Cycle Assessment Approach……27
Jelena Topič Božič, Andreja Dobrovoljc, Simon Muhič

Materials role and challenges of green transition……33
Bojan Podgornik

AI-Driven Predictive Maintenance for Sustainable Power Grid Infrastructure……40
Dunja Srpak, Srđan Skok, Tomislav Horvat, Josip Srpak

EdgeGuard-PI: Deployment-Oriented Multi-Class Prompt Injection Detection for Edge AI Systems……46
Kusai Almahamid

Towards an AI-Based Monitoring Infrastructure for Emerging Technology Discourse: Initial Mapping of Social Media Communication about Quantum……53
Wiebke Drews, Dolores Modic, Karin Dobravc Škof

An Exploratory Study of Chunking Strategies for Retrieval Augmented Generation over Scientific PDF Articles……64
Marko Rudolf, Andrej Dobrovoljc

Integrating Artificial Intelligence with Physics-Based Simulations: A Review of Methods and Applications ……72
Matic Medvešek, Lucijano Berus, Jelena Topić Božič, Simon Muhič

Z umetno inteligenco podprt organizacijski spomin: Avtomatizacija inženiringa kakovosti in preprečevanje ponavljanja napak v proizvodnji ……77
Matej Kalčič

Numerical Investigation of Sinusoidal and Arc Waveform Leading-Edge Tubercle Geometry……83
Matic Medvešek, Simon Muhič

Modeling of electric motor drives with three-phase induction machines and various types of mechanical loads……89
Pušenjak Rudolf

Experimental testing and mathematical modeling of sound insulation of recycled panels made from car tires based on statistical models……99
Anđelko Crnoja, Ivo Čolak, Aleksej Aniskin, Dalibor Kramarić

Analysis of defects in aluminum alloy castings due to thermo-mechanical cracks in die casting molds……115
Muhič Mitja

Modelling and optimisation of traction electric motor torque ……121
David Oman, Damir Vrančić

XR-Supported Workforce Training in Complex Manufacturing Systems Using Digital Twins and Artificial Intelligence……129
Gosar Žiga, Marina Panoska, Klemen Babuder, Marko Močnik

ML-Based Real-Time Monitoring System for Plasma Thin-Film Deposition in Industrial Manufacturing……..134
Žiga Gosar

Simulation Based Production Scheduling Optimization in a Digital Twin Using Genetic Algorithms ……..…....139
Vinko Longar

Razvoj modularne programske arhitekture za avtomatizirano doziranje tekočin ……..……....145
Srđan Kosić, Damir Vrančić

Verification of Energy Envelope Performance Through In-Situ U-Value Measurements and Their Impact on District Heating Energy Consumption

Đorđe Karić[1], Mirko Komatina[2], Dimitrije Manić[3], Simon Muhič[4], Goran Vorotović[5]

[1]*Faculty of Mechanical Engineering, University of Belgrade*
E-mail: djordje.karic@yahoo.com

[2]*Faculty of Mechanical Engineering, University of Belgrade*
E-mail: mkomatina@mas.bg.ac.rs

[3]*The Innovation Center of the Faculty of Mechanical Engineering, Belgrade*
E-mail: dimitrije.manic@gmail.com

[4]*Institute for Renewable Energy and Efficient Exergy Use, INOVEKS d.o.o., Cesta 2. grupe odredov 17, 1295 Ivančna Gorica, Slovenia, Faculty of Industrial Engineering Novo Mesto, Šegova ulica 112, 8000 Novo Mesto, Slovenia, Rudolfovo—Science and Technology Centre Novo Mesto, Podbreznik 15, 8000 Novo Mesto, Slovenia*
E-mail: simon.muhic@fini-unm.si

[5]*Faculty of Mechanical Engineering, University of Belgrade*
E-mail: gvorotovic@mas.bg.ac.rs

Abstract: *A reliable assessment of the actual impact of building envelope retrofits requires the integration of measured thermal performance parameters with operational energy consumption data—an approach that remains insufficiently implemented in practice. This study quantifies the relationship between in situ measured U-values of external walls and corresponding changes in operational thermal energy consumption in residential buildings. Field measurements were conducted before and after the implementation of envelope retrofit measures. The analyzed buildings are connected to the district heating system in Šabac, Serbia, enabling a direct comparison of measured heat consumption under real operating conditions prior to and following the retrofit.*

The study evaluates how changes in the thermomechanical properties of the envelope affect thermal energy demand. The results demonstrate that a significant reduction in the heat transfer coefficient after the retrofit—and consequently in transmission heat losses—leads to a measurable decrease in building-level thermal energy consumption. A direct correlation was identified between improvements in measured thermomechanical properties and achieved energy savings.

Thermal energy consumption was reduced by up to 35% in the analyzed buildings. Measured U-values and realized energy savings were compared with the design parameters specified in the energy efficiency reports for each building, allowing the determination of deviations between calculated and actual U-values, which ranged from −14.48% to +6.19% before retrofit and from −0.50% to +12.50% after retrofit. In addition, deviations between predicted and measured specific thermal energy consumption were identified, confirming the presence of discrepancies between design assumptions and actual operational performance.

In addition to facade insulation, the analysis also considered the installation of thermostatic valves, allowing the energy savings to be differentiated between envelope-related reductions in transmission heat losses and savings achieved through improved heating system control.

The findings contribute to the development of a robust framework for verifying retrofit performance by directly linking field-measured envelope properties with operational energy data, thereby enabling a reliable quantitative comparison between predicted and realized building performance.

Keywords*: in-situ U-value measurement, energy retrofit, district heating, thermal transmittance*

Article Classification*: Scientific Paper*

1 INTRODUCTION

Improving the thermal performance of the building envelope remains one of the most effective approaches for reducing space heating demand, operational energy consumption, and associated greenhouse gas emissions in

the residential sector. Existing buildings account for a significant share of final energy use worldwide, and large-scale renovation strategies increasingly prioritize envelope retrofits as a key pathway toward achieving energy efficiency and decarbonization targets. Among envelope parameters, the thermal transmittance (*U*-value) of external walls represents a fundamental indicator used to quantify transmission heat losses and assess the effectiveness of insulation measures [1].

Traditionally, U-values are determined through steady-state calculations based on standardized thermal properties of construction materials and simplified boundary conditions. However, increasing evidence suggests that calculated values may not accurately reflect real thermal performance during building operation. Discrepancies between theoretical and measured U-values can be attributed to multiple factors, including construction defects, thermal bridging, moisture content, material ageing, and variations in internal and external environmental conditions. As a result, field measurements have gained importance as a more reliable method for evaluating the actual thermal behaviour of building envelope components [1,2].

The methodology for in-situ determination of thermal resistance and thermal transmittance is defined by international standards, most notably ISO 9869-1, which specifies the heat-flow meter method for assessing envelope performance under real operating conditions. The standard establishes requirements regarding sensor placement, minimum measurement duration, data convergence criteria, and environmental stability conditions necessary for obtaining representative results. Complementary standards, such as EN ISO 6946, addressing calculation methods for thermal resistance and transmittance, and EN ISO 52016-1, which provides procedures for assessing the energy performance of buildings, support the interpretation of measured thermal parameters within a broader energy analysis framework. The application of standardized procedures is therefore essential for ensuring the reliability and comparability of experimental results.

Recent research has demonstrated that in-situ measurements can provide valuable insights into the dynamic thermal response of building envelopes. O'Hegarty et al. conducted an extensive review and experimental investigation of U-value monitoring in highly insulated constructions, highlighting that measurement uncertainty is influenced by transient heat transfer effects and the duration of data acquisition [1]. Their findings emphasize the importance of carefully selecting measurement periods and applying appropriate data processing techniques to ensure convergence and accuracy. Similarly, Nicoletti et al. analyzed different in-situ measurement approaches and showed that the accuracy of thermal conductance estimation depends strongly on boundary condition variability and signal processing methods [3].

Further experimental studies have confirmed that measured U-values may vary seasonally due to climatic influences and moisture transport processes within multilayered wall assemblies. Yu et al. reported significant variations in thermal transmittance measured in cold-climate buildings, demonstrating that such variations can influence estimates of heating energy demand and consequently affect retrofit performance assessment [2]. Investigations focusing on advanced monitoring techniques have also explored the use of embedded sensors and dynamic heat-transfer models to improve the reliability of in-situ measurements and to better capture transient thermal behaviour [4].

The growing interest in field measurement methodologies is closely linked to the broader issue of the building energy performance gap, defined as the discrepancy between predicted and actual energy consumption. Numerous studies have identified limitations of design-stage modelling assumptions, uncertainties in input parameters, and behavioural factors as major contributors to this gap. In this context, integrating experimentally measured envelope properties with monitored operational energy data has been recognized as a promising approach for improving the accuracy of retrofit performance evaluation and for enhancing confidence in energy savings estimates [5].

In addition to improving building-level energy performance assessment, reliable determination of in-situ U-values has implications for large-scale renovation planning and energy system optimization. Accurate knowledge of transmission heat losses supports better forecasting of heating demand and facilitates the development of more effective energy efficiency policies. This is particularly relevant for buildings connected to district heating networks, where changes in envelope performance directly influence system load profiles and long-term infrastructure planning strategies. Recent research has emphasized the need to consider building retrofits and district heating decarbonization pathways as complementary rather than competing measures in achieving sustainable urban energy systems [6].

Furthermore, the availability of high-quality field measurement data enables more robust calibration of building energy models and supports the development of advanced performance verification methodologies. Experimental validation of retrofit outcomes can help reduce uncertainties in predicted energy savings, improve cost-effectiveness assessments, and provide valuable feedback for improving design guidelines and regulatory frameworks. As building renovation programs continue to expand globally, the role of empirical performance evaluation is expected to become increasingly important.

In this context, the present study investigates the impact of building envelope retrofits on operational thermal energy consumption by combining in-situ measurements of external wall U-values with monitored heat consumption data in residential buildings connected to a district heating system. Measurements were conducted

before and after retrofit implementation, allowing direct comparison of thermal performance and energy use under real operating conditions. The study further evaluates deviations between measured and design parameters defined in energy efficiency documentation, thereby contributing to the development of more reliable verification approaches for assessing retrofit effectiveness.

By integrating standardized in-situ measurement procedures with operational energy analysis, the research provides experimentally validated evidence of the relationship between improved envelope thermal characteristics and realized energy savings. The findings aim to support more accurate assessment of retrofit performance and to reduce uncertainties associated with building energy predictions in practical applications.

In recent years, additional research has further emphasized the importance of integrating experimental envelope performance assessment with advanced modelling and large-scale retrofit evaluation approaches. Studies based on urban building energy modelling and monitored retrofit projects have demonstrated that reliable field data significantly improve the calibration of simulation models and reduce uncertainties in predicting long-term energy savings [7]. Moreover, investigations focusing on post-retrofit performance monitoring confirm that the magnitude of realized energy savings depends strongly on the accuracy of input thermal parameters and on the interaction between building envelope improvements and system operation strategies [8]. Large-scale analyses of residential retrofit programs also highlight the need for systematic performance verification frameworks based on measured operational data rather than solely on design assumptions [9]. In addition, recent experimental research on envelope thermal characterization has explored improved methodologies for dynamic heat-transfer analysis and uncertainty reduction in *U*-value measurements, further supporting the role of standardized in-situ procedures in building performance evaluation [4].

2 METHODS

The research is based on a combined experimental and analytical approach, including the determination of the thermal transmittance of external walls and the analysis of operational thermal energy consumption before and after the implementation of energy retrofit measures.

The thermal transmittance (*U*-value) was determined using two approaches: (i) calculation based on the wall composition in accordance with relevant standards, and (ii) in-situ measurement using the heat flux method in accordance with ISO 9869-1. Measurements were conducted both before and after the retrofit, enabling a direct comparison of the thermal performance of the building envelope. The thermal transmittance was calculated according to:

$$U = \frac{1}{(R_{si} + \Sigma R_i + R_{se})} [\mathrm{W/m^2K}] \quad (1)$$

where R_{si} and R_{se} represent internal and external surface thermal resistances, and ΣR_i represents the thermal resistance of individual layers of the wall.

During in-situ measurements, temperature was recorded at four measurement points: the internal air temperature, internal surface temperature of the wall, external surface temperature of the wall, and external air temperature. This measurement setup allowed for a more accurate determination of boundary conditions and reduced the influence of local temperature variations. The in-situ thermal transmittance was determined as:

$$U = \frac{\dot{q}_{avg}}{\left(T_{i_{avg}} - T_{e_{avg}}\right)} \quad (2)$$

where $\dot{q}_{avg}$ $[\mathrm{W/m^2}]$ is the average heat flux density, and T_i and T_e represent indoor and outdoor temperatures, respectively. In order to interpret the experimental results, a parametric analysis of the external surface heat transfer resistance was performed. Calculated U-values were obtained for a range of R_{se} values and compared with measured values to identify the conditions that provided the best agreement. The external surface heat transfer coefficient was additionally determined based on measured data as:

$$\alpha_2 = \frac{\dot{q}}{t_3 - t_4} [\mathrm{W/m^2K}] \quad (3)$$

where q represents the measured heat flux density, while t_3 and t_4 correspond to the external wall surface temperature and external air temperature, respectively.

The external surface thermal resistance was then calculated as:

$$R_{se} = \frac{1}{\alpha_2} [\mathrm{m^2K/W}] \quad (4)$$

This approach enables the estimation of effective boundary conditions during the measurement period and provides insight into the influence of external heat transfer conditions on the measured thermal performance. In addition to the thermal performance of the envelope, annual thermal energy consumption at the building level was analysed before and after the implementation of retrofit measures. Two phases of intervention were considered: (i) facade insulation and (ii) facade insulation combined with thermostatic control. This enabled the separation of the individual contributions of envelope improvement and heating system regulation to the overall energy savings.

For each building, the following performance indicators were determined: normalized specific heating energy consumption before retrofit, normalized specific heating energy consumption after retrofit, absolute heating energy savings, and relative heating energy savings. Absolute savings were defined as the difference between normalized specific consumption before and after retrofit, while relative savings were expressed as the percentage reduction relative to the pre-retrofit value. Distinguishing between absolute and relative savings is important because these indicators reflect different aspects of retrofit effectiveness and may exhibit different dependencies on building characteristics and operational conditions.

In order to quantify the relationship between building characteristics and heating energy performance, a set of normalized and derived indicators was defined. Heating degree days (HDD) were calculated as the cumulative temperature difference between the reference indoor air temperature and the mean daily outdoor temperature during the heating season:

$$HDD = \sum_{i=1}^{n} max\left(0, T_{ref} - T_{out,i}\right) \tag{5}$$

where T_{ref} is the reference indoor temperature (°C), $T_{out,i}$ is the mean daily outdoor temperature on day i (°C), and n is the number of days within the heating season. Only the days for which $T_{out,i} < T_{ref}$ were included in the calculation. The normalized specific heating energy consumption was determined based on the ratio between the reference and measured Heating Degree Days (HDD), allowing for the correction of climatic differences between analyzed heating seasons. Normalized specific heating energy consumption was determined as:

$$Q = \frac{Q_{year}}{HDD \cdot A} \tag{6}$$

where Q denotes the normalized specific heating energy consumption, Q_{year}represents the measured annual heating energy consumption, A is the heated floor area of the building, and HDD denotes the seasonal heating degree days used for climate normalization. This formulation enables comparison of heating energy performance across buildings and heating seasons under equivalent climatic conditions.

Based on the above, a comparison between calculated and measured values was performed, along with an analysis of the relationship between changes in U-values and achieved energy savings.

3 RESULTS AND DISCUSSION

3.1 Comparison of calculated and measured U-values

The comparison of calculated and experimentally determined thermal transmittance values highlights the differences between predicted and actual thermal performance of the analysed buildings before and after energy retrofit. Before retrofit, the calculated U-values ranged from 0.97 to 1.78 W/m²K, depending on wall composition. Buildings constructed with solid brick walls exhibited the highest U-values, reflecting poor thermal performance. In contrast, Building 1, constructed with hollow block brick, showed slightly better thermal characteristics, while Building 4, which included an internal insulation layer, had the lowest initial U-value.

The comparison between calculated and measured values before retrofit showed deviations in the range of −14.48% to +6.19%. In most cases, measured U-values were higher than calculated ones, indicating increased heat losses under real conditions. These differences can be attributed to uncertainties in material properties, construction imperfections, and the influence of real boundary conditions that are not fully accounted for in standard calculation procedures.

After retrofit, all buildings were upgraded with an additional external insulation layer of 10 cm, resulting in a significant reduction in thermal transmittance. The calculated U-values decreased to the range of 0.29–0.34 W/m²K, while measured values ranged from 0.315 to 0.360 W/m²K.
The relative reduction in measured U-values ranged from 69.46% to 81.51%, confirming the high effectiveness of facade insulation. Buildings with higher initial U-values (solid brick structures) achieved the largest relative reductions, exceeding 80%, while Building 4, which had partially improved thermal characteristics prior to retrofit, showed a lower relative improvement.

The agreement between calculated and measured U-values improved after retrofit, with deviations ranging from −0.59% to +12.50%. Very close agreement was observed in Buildings 2 and 3, while slightly larger deviations in Buildings 1 and 4 indicate the continued influence of local boundary conditions and wall-specific characteristics. Overall, the results demonstrate that while standard calculation methods provide a reasonable estimation of thermal transmittance after retrofit, their accuracy before retrofit is limited due to uncertainties in existing wall properties and boundary conditions.

Table 1. Comparison of calculated and measured U-values before and after retrofit

Building	U_calc_before (W/m²K)	U_meas_before (W/m²K)	δU_before (%)	U_calc_after (W/m²K)	U_meas_after (W/m²K)	δU_after (%)	ΔU_meas (%)
Building 1	1.45	1.24	-14.48	0.32	0.360	12.50	70.97
Building 2	1.78	1.82	2.25	0.34	0.338	-0.50	81.51
Building 3	1.78	1.83	2.81	0.34	0.348	2.33	80.99
Building 4	0.97	1.03	6.19	0.29	0.315	8.46	69.46

Note: δU denotes the deviation between measured and calculated U-values, while ΔU_meas denotes the relative reduction in experimentally determined U-values after retrofit.

3.2 Effect of external surface heat transfer resistance

The experimentally determined values of the external surface heat transfer coefficient and corresponding external surface resistance provide important insight into the actual boundary conditions during in-situ measurements. Before retrofit, the external heat transfer coefficient ranged from 2.74 to 43.35 W/m²K, corresponding to surface resistance values between 0.023 and 0.365 m²K/W. After retrofit, the values ranged from 2.08 to 7.79 W/m²K, while the corresponding resistances ranged from 0.128 to 0.481 m²K/W. These values differ significantly from the standard value of R_{se} = 0.04 m²K/W, confirming that real environmental conditions can deviate substantially from standardized assumptions.

Buildings 1 and 2 exhibited very high values of the heat transfer coefficient before retrofit (approximately 43 W/m²K), indicating strong convective heat transfer conditions, likely caused by wind exposure and increased air movement along the facade. In contrast, Buildings 3 and 4 showed considerably lower values, corresponding to higher external surface resistance and more stable external conditions.

After retrofit, a significant decrease in the external heat transfer coefficient was observed in Buildings 1 and 2, accompanied by an increase in surface resistance. This indicates a reduction in convective heat transfer effects, which may be associated with changes in facade surface characteristics, reduced air movement near the wall, or differences in measurement conditions.

For Buildings 3 and 4, the changes were less uniform, but still demonstrate that external surface resistance is strongly influenced by local environmental conditions rather than being a constant parameter.

The comparison with the results presented in Section 3.1 shows that part of the deviation between calculated and measured U-values can be attributed to the use of standardized boundary conditions. The large variation in experimentally determined R_{se} values confirms that the external surface heat transfer resistance should be treated as a variable parameter.

This finding is particularly important for the interpretation of in-situ measurements and the calibration of thermal models, as the use of experimentally determined boundary conditions leads to a more accurate assessment of actual thermal performance.

Table 2. Experimentally determined external surface heat transfer coefficient and resistance before and after retrofit

Building	α_2_before (W/m²K)	R_{se} before (m²K/W)	α_2 after (W/m²K)	R_{se} after (m²K/W)
Building 1	33.35	0.030	4.57	0.219
Building 2	43.35	0.023	7.79	0.128
Building 3	2.74	0.365	4.58	0.218
Building 4	3.63	0.275	2.08	0.481

3.3 Energy performance and contribution of retrofit measures

The normalized annual heating energy consumption was analysed for all buildings using climate-corrected data based on Heating Degree Days (HDD). The analysed heating seasons showed significant variation in climatic conditions, with average outdoor temperatures ranging from 7.48 to 9.28 °C and seasonal HDD values ranging from 1908 to 2427.

This variation confirms the necessity of climate normalization, as differences in heating demand between seasons could otherwise be attributed to weather conditions rather than the effects of implemented energy efficiency measures. By applying HDD-based normalization, a consistent and reliable comparison of energy performance before and after retrofit was achieved.

Before retrofit, the normalized specific heating energy consumption ranged from 127.03 to 140.53 kWh/m² year, indicating relatively high energy demand associated with poor thermal performance of the existing building envelope.

Following the implementation of facade insulation, a significant reduction in energy consumption was observed across all buildings. The reduction ranged from 18.85% to 28.75%, with the highest savings recorded in buildings with solid brick walls. This confirms that improvements in envelope thermal performance directly reduce transmission heat losses.

The installation of thermostatic valves resulted in additional energy savings ranging from 6.52% to 8.88%. Although smaller in magnitude compared to facade insulation, these savings highlight the importance of proper heating system control and indoor temperature regulation.

The total reduction in normalized energy consumption ranged from 25.36% to 35.52%, with the highest savings achieved in buildings with the poorest initial thermal characteristics. Buildings with partially improved wall structures prior to retrofit exhibited lower relative savings, confirming the influence of initial envelope performance.

A clear correlation between the reduction in thermal transmittance (Section 3.1) and the achieved energy savings was observed. Buildings that exhibited the largest decrease in *U*-values also achieved the highest reduction in energy consumption, demonstrating the strong link between envelope performance and operational energy demand.

Table 3. Main energy performance indicators of the analysed buildings

Building	*q*_before (kWh/m²year)	*q*_facade (kWh/m²year)	*q*_final (kWh/m²year)	Δq_facade (kWh/m²year)	*S*_facade (%)	Δ*q*_valves (kWh/m²year)	*S*_valves (%)	Δ*q*_total (kWh/m²year)	*S*_total (%)
Building 1	118.32	93.08	82.65	25.24	21.33	10.43	8.88	35.67	30.15
Building 2	140.53	105.81	94.79	34.72	24.70	11.02	7.83	45.74	32.56
Building 3	129.20	92.05	83.31	37.15	28.75	8.74	6.76	45.89	35.52
Building 4	127.03	103.09	94.81	23.94	18.85	8.28	6.52	32.22	25.36

3.4 Financial performance of retrofit measures

The financial performance of the analyzed retrofit measures was evaluated based on experimentally verified and climate-normalized energy savings, derived from in-situ measurements of envelope thermal performance and corresponding operational heat consumption. In this way, the economic analysis directly reflects the actual thermomechanical improvements of the building envelope and their impact on real energy demand.

Table 4 summarizes the main financial indicators for facade insulation, thermostatic valve installation, and their combined implementation across all four buildings.

The results consistently demonstrate that facade insulation provides the dominant contribution to total energy savings, which is in direct agreement with the measured reduction in U-values and the corresponding decrease in transmission heat losses. This confirms that improvements in envelope thermal performance, as verified through in-situ measurements, translate into measurable and economically relevant reductions in heating energy consumption.

However, due to relatively high capital costs, facade measures exhibit longer payback periods and moderate net present values when compared to system-level interventions. In contrast, thermostatic valves show significantly higher financial efficiency per unit of investment. Owing to their low installation cost and measurable influence on indoor temperature regulation, valves achieve substantially shorter payback periods, ranging from 5.6 to 10.3 years, while maintaining positive net present values in all analyzed cases.

The combined implementation of facade insulation and thermostatic valves results in the most favorable overall financial performance across all buildings. Total net present values range from 8,654 € in Building 2 to 183,187 € in Building 3, while corresponding payback periods vary between 7.7 and 14.4 years. These results indicate that the integration of envelope improvements and system-level control enables both maximization of energy savings and improved economic feasibility.

The financial attractiveness of facade insulation alone is shown to be strongly dependent on building-specific characteristics, particularly the initial thermal performance of the envelope. In Building 3, which exhibited the largest reduction in measured U-values and the highest relative decrease in normalized energy consumption, facade insulation achieves strong financial performance without the need for subsidy support. Conversely, in Buildings 1 and 2, where the reduction in heat losses is more moderate, facade measures yield lower economic returns.

Building 4 represents an intermediate case, confirming that even when facade insulation leads to measurable improvements in thermal performance, the resulting financial indicators may remain modest if the achievable reduction in energy consumption is limited. This observation highlights the importance of linking experimentally determined envelope properties with economic evaluation.

In order to achieve financial performance of facade insulation comparable to that of thermostatic valves, substantial grant support is required in certain cases. The analysis indicates that for Buildings 1 and 2, subsidy levels of approximately 59–60% of the investment cost are necessary, while for Building 4 a support level of approximately 52% is required. In contrast, no subsidy is required in Building 3, where the strong correlation between improved envelope performance and energy savings results in high financial attractiveness.

Overall, the results confirm a consistent relationship between measured thermomechanical improvements of the building envelope, realized energy savings, and financial performance. While facade insulation remains the primary driver of energy demand reduction, thermostatic valves demonstrate superior cost-effectiveness, and their combined implementation represents the most balanced retrofit strategy.

Table 4. Financial performance indicators of facade insulation, thermostatic valve installation, and combined retrofit measures

Building	Investment facade (€)	Investment valves (€)	Total investment (€)	Annual savings (€)	Payback facade (years)	Payback valves (years)	Payback total (years)	Net present value — facade insulation (€)	Net present value — thermostatic valves (€)	Net present value — combined measures (€)	Required grant for facade (%)
Building 1	86,685	14,400	101,085	8,771	14.0	5.6	11.5	37,239	38,932	76,171	60
Building 2	19,611	2,560	22,171	1,543	16.7	6.9	14.4	3,780	4,874	8,654	59
Building 3	93,335	22,400	115,735	14,970	7.7	7.9	7.7	148,654	34,533	183,187	0
Building 4	47,155	10,800	57,955	4,061	15.6	10.3	14.3	13,033	10,102	23,135	52

4 CONCLUSION

This study provides an experimental verification of the effects of building envelope retrofit by directly linking in-situ measured U-values of external walls with climate-normalized operational heating energy consumption in residential buildings connected to a district heating system.

The results confirm that facade insulation leads to a substantial improvement in thermal performance, with measured U-values reduced by 69.46% to 81.51% across the analysed buildings. The magnitude of improvement was strongly dependent on the initial wall composition, with the highest relative reductions observed in buildings with solid brick walls, while buildings with partially improved envelope structures prior to retrofit exhibited lower, but still significant, improvements.

The comparison between calculated and measured U-values showed that standard steady-state calculation methods provide a reliable estimate of post-retrofit thermal transmittance, while larger deviations may occur in the pre-retrofit condition due to uncertainties in material properties, construction characteristics, and boundary conditions. The observed deviations ranged from −14.48% to +6.19% before retrofit and from −0.50% to +12.50% after retrofit, indicating improved agreement following the implementation of insulation measures.

A key finding of the study is the significant variability of external surface heat transfer conditions. Experimentally determined external surface resistance values ranged from 0.023 to 0.365 m²K/W before retrofit and from 0.128 to 0.481 m²K/W after retrofit, which deviates considerably from the standard value commonly used in calculations.

This confirms that thermal transmittance should be interpreted as a parameter influenced not only by material properties but also by actual environmental conditions, and that the use of fixed boundary condition values may lead to inaccuracies in performance assessment.

The analysis of climate-normalized heating energy consumption demonstrated that facade insulation resulted in energy savings between 18.85% and 28.75%, while the installation of thermostatic valves provided additional savings of 6.52% to 8.88%. The total reduction in heating energy demand ranged from 25.36% to 35.52%, confirming a consistent relationship between improved envelope thermal performance and reduced operational energy consumption.

The financial analysis indicates that facade insulation represents the primary driver of energy savings, while thermostatic valves achieve higher cost-effectiveness due to lower investment costs and shorter payback periods. The combined implementation of both measures provides the most balanced retrofit strategy, enabling both significant energy savings and improved economic performance. In several analysed cases, the results indicate that achieving financial performance comparable to system-level measures requires partial grant support, highlighting the importance of integrating energy efficiency policies with financial incentives.

Overall, the study demonstrates that the integration of in-situ thermal measurements, climate-normalized energy analysis, and economic evaluation provides a robust framework for assessing the real effects of energy retrofit measures. Such an approach enables a more reliable comparison between predicted and actual performance and contributes to reducing uncertainties in building energy efficiency assessment in practical applications.

REFERENCES

[1] O'HEGARTY, R., O. KINNANE, D. LENNON and S. COLCLOUGH. In-situ U-value monitoring of highly insulated building envelopes: Review and experimental investigation. Energy and Buildings. 2021, 252, 111447. Available at: https://doi.org/10.1016/j.enbuild.2021.111447

[2] YU, Jiaqi, Wen-Shao CHANG, Ruinan ZHANG, Yu DONG, Haoyu HUANG and Tsung-Hsien WANG. In-situ U-value measurements of typical building envelopes in a severe cold region of China: U-value variations and energy implications. Energy and Buildings. 2024, 317, 114947. Available at: https://doi.org/10.1016/j.enbuild.2024.114947

[3] NICOLETTI, Francesco, Mario Antonio CUCUMO and Natale ARCURI. Evaluating the accuracy of in-situ methods for measuring wall thermal conductance: A comparative numerical study. Energy and Buildings. 2023, 290, 113095. Available at: https://doi.org/10.1016/j.enbuild.2023.113095

[4] CHOI, Doo-Sung, Ye-Ji LEE, Ji-Hoon MOON, Yong-Shik KIM and Myeong-Jin KO. Estimating In-Situ R-Value of Highly Insulated Building Walls Based on the Measurement of Temperature and Heat Flux Inside the Wall. Energies. 2023, 16(15), 5714. Available at: https://doi.org/10.3390/en16155714

[5] ZHENG, Zhihang, Jin ZHOU, Jiaqin ZHU, Ying YANG, Feng XU and Hongcheng LIU. Review of the building energy performance gap from simulation and building lifecycle perspectives: Magnitude, causes and solutions. Developments in the Built Environment. 2024, 18, 100345. Available at: https://doi.org/10.1016/j.dibe.2024.100345

[6] POPOVSKI, Eftim, Mario RAGWITZ and Heike BRUGGER. Decarbonization of district heating and deep retrofits of buildings as competing or synergetic strategies for the implementation of the efficiency first principle. Smart Energy. 2023, 10, 100096. Available at: https://doi.org/10.1016/j.segy.2023.100096

[7] JOHARI, F., O. LINDBERG, U.H. RAMADHANI, F. SHADRAM, J. MUNKHAMMAR and J. WIDÉN. Analysis of large-scale energy retrofit of residential buildings and their impact on the electricity grid using a validated UBEM. Applied Energy. 2024, 361, 122937. Available at: https://doi.org/10.1016/j.apenergy.2024.122937

[8] WANG, Zhichao, Hui SHEN, Gaofeng DENG, Xiaoyu LIU and Dong WANG. Measured performance of energy efficiency measures for zero-energy retrofitting in residential buildings. Journal of Building Engineering. 2024, 95, 109545. Available at: https://doi.org/10.1016/j.jobe.2024.109545

[9] KUIVJÕGI, Helena, Sofia VASMAN, Eduard PETLENKOV, Martin THALFELDT and Jarek KURNITSKI. Data-driven baseline generation for post-retrofit energy saving assessment, a comparison of statistical and machine learning methods. Journal of Building Engineering. 2024, 98, 111016. Available at: https://doi.org/10.1016/j.jobe.2024.111016

[10] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 9869-1:2014. Thermal insulation — Building elements — In-situ measurement of thermal resistance and thermal transmittance — Part 1: Heat flow meter method. Geneva: ISO, 2014. Available at: https://www.iso.org/standard/59697.html

[11] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 6946:2017. Building components and building elements — Thermal resistance and thermal transmittance — Calculation methods. Geneva: ISO, 2017. Available at: https://www.iso.org/standard/65708.html

[12] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 52016-1:2017. Energy performance of buildings — Energy needs for heating and cooling, internal temperatures and sensible and latent heat loads — Part 1: Calculation procedures. Geneva: ISO, 2017. Available at: https://www.iso.org/standard/65696.html

Experimental investigation of the reliability of photovoltaic panel mounting systems on flat roofs using a controlled climate test chamber

Marko Gložinić[1], Matija Gložinić[1], Ante Čikić[1], Zoran Busija[1]

[1]University North, Department of Mechatronics, 104. brigade 3, 42000 Varaždin, Croatia
E-mail: maglozinic@unin.hr, mglozinic@unin.hr, acikic@unin.hr, zbusija@unin.hr

Abstract: *Long term reliability of photovoltaic systems installed on industrial flat roofs strongly depends on the interaction between mounting structures, waterproofing systems and environmental loads such as wind, precipitation, temperature and humidity. This research presents the development and application of an experimental testing system designed to evaluate the reliability of photovoltaic panel mounting systems under controlled environmental conditions. A specialized climate test chamber was designed and constructed to simulate realistic environmental loads acting on rooftop PV installations. The chamber enables controlled regulation of air temperature, relative humidity, wind intensity of 13,5 - 18,5 m/s and precipitation levels. The experimental setup allows detailed investigation of the interaction between PV mounting structures and roof waterproofing layers. Particular attention is given to the measurement and analysis of water permeability through the waterproofing system beneath the mounting structures. Environmental parameters are monitored using an meteorological station, anemometer, temperature sensors and capacitive moisture probes embedded within the roofing layers. All signals are continuously recorded using a multi-channel data acquisition system. The results showed that wind-induced vibrations led to loosening of fastening elements, which resulted in water intrusion beneath the waterproofing layer. Water leakage was observed in all configurations where insufficient tightening torque and low structural stiffness were present. Comparative analysis indicated that configurations with panel inclination angles up to 30° exhibited stable performance, while higher inclinations resulted in increased vibration amplitudes and a consequently higher risk of leakage. The research contributes to the development of more reliable mounting solutions and improved testing methodologies for photovoltaic systems.*

Keywords*:* photovoltaic panels, mounting systems, flat roofs, waterproofing systems

Article Classification*: Scientific Paper*

1 INTRODUCTION

Among renewable energy sources, solar photovoltaics (PV) represent one of the most important technologies for mitigating climate change and supporting sustainable energy production [1,2]. Continuous technological development and decreasing manufacturing costs have made photovoltaic systems increasingly competitive in the global energy market [3]. The rapid development of the building sector increases the demand for renewable energy integration in order to meet rising energy consumption while reducing environmental pollution and greenhouse gas emissions [4]. Flat roofs of industrial and commercial buildings represent especially suitable locations for photovoltaic installations because of their large available surface area and the possibility of optimizing panel orientation and tilt angle through dedicated mounting structures [5]. One of the most important factors affecting the reliability of rooftop photovoltaic systems is wind loading. PV modules mounted on flat roofs are exposed to significant uplift and pressure forces on mounting systems [6]. Experimental studies have shown that wind loads on rooftop photovoltaic arrays strongly depend on parameters such as panel inclination, array configuration and building geometry [7]. Strong corner vortices generated near roof edges and corners can produce high uplift forces acting on solar panels installed on flat roofs [8]. In addition, architectural features such as parapets may significantly modify airflow patterns and affect the magnitude of wind loads acting on rooftop solar arrays [9]. Several investigations have analysed the influence of building geometry and array configuration on pressure distribution and aerodynamic loads acting on photovoltaic systems installed on flat roofs [10]. More recent experimental and numerical studies have further improved the understanding of complex airflow patterns and turbulence structures generated around rooftop solar arrays exposed to atmospheric winds [11]. In addition to aerodynamic loads, the durability of rooftop photovoltaic systems strongly depends on the performance of the roof waterproofing system beneath the mounting structures. Flat roof waterproofing membranes are exposed to environmental degradation, mechanical stresses and installation related damage that may lead to deterioration over time [12]. Various waterproofing technologies are commonly used in flat roof systems, including bituminous membranes and polymer based waterproofing coatings [13]. Experimental measurements conducted on buildings partially covered with photovoltaic modules have shown that PV panels can modify heat transfer processes within the roof assembly by shading the roof surface and reducing absorbed solar radiation [14]. Similarly, investigations of adhesive mounted rooftop PV systems indicate that photovoltaic modules can significantly influence temperature distribution within roofing materials and potentially affect their long term durability [15]. Despite extensive research on aerodynamic and thermal behaviour of rooftop photovoltaic systems, considerably less attention has been given to the interaction

between PV mounting systems and roof waterproofing layers under combined environmental loads such as wind driven rain. Recent research on building integrated photovoltaic systems emphasizes the importance of evaluating watertightness performance and potential water intrusion risks under controlled environmental conditions [16]. Despite extensive research on wind loads of rooftop photovoltaic systems, limited attention has been given to the interaction between mounting systems and roof waterproofing layers under combined environmental loads. The present study investigates reliability of photovoltaic panel mounting systems installed on flat roofs through controlled laboratory testing. The objectives of this study are to experimentally evaluate the mechanical stability of PV mounting systems under controlled wind and precipitation conditions, to investigate the occurrence of water intrusion beneath the waterproofing layer and to identify key parameters affecting system reliability.

2 METHODS

This research applies an experimental approach to investigate the reliability of photovoltaic panel mounting systems installed on flat roofs when exposed to combined environmental loads. The study focuses on the interaction between PV mounting structures, roof waterproofing systems and environmental influences such as wind, precipitation, temperature and humidity. To enable controlled and repeatable testing conditions, a specialized climate testing chamber was developed, providing a controlled atmospheric environment for experimental investigations.

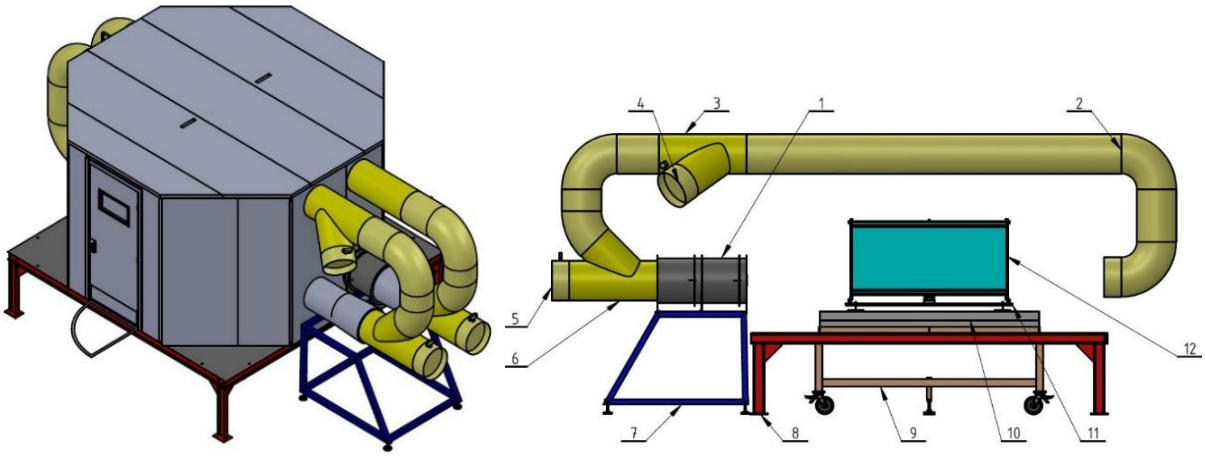


Figure 1: Experimental ventilation system and test setup. 1-Axial fan (AVS 500), 2-Circular ventilation duct (Ø 450 mm), 3-T-junction duct element with 45° branch for air exhaust, 4-Adjustable control damper, 5-Fresh air inlet, 6-T-junction duct element with 45° branch for air intake, 7- Steel support frame for the fan assembly, 8-Experimental chamber support structure, 9- Adjustable test platform frame, 10- Flat roof test platform, 11-Photovoltaic support structure, 12- Photovoltaic panel

The experimental investigation aims to reproduce realistic environmental conditions that occur during the operation of rooftop PV installations such as wind exposure, rainfall events, temperature variations and high humidity conditions. The controlled laboratory environment provides a high level of repeatability and enables systematic investigation of the influence of individual environmental parameters as well as their combined effects on the performance of photovoltaic mounting systems installed on flat roofs.

2.1 Experimental setup

The climate chamber is constructed as a closed testing enclosure equipped with a ventilation system that recirculates air within the chamber. Air movement within the chamber is produced using axial fans installed within the ventilation duct system. The airflow intensity can be adjusted by regulating the operating frequency of the fans, allowing the simulation of different wind conditions ranging from moderate airflow to strong wind events. Airflow is generated by two axial fans Klima Celje AVS 500 (1), which can be arranged either in series or in parallel configuration to increase the available pressure head or volume flow of the ventilation system, depending on the requirements of the simulation scenario. The fans are mounted on a rigid steel support frame (7). The ventilation duct system (2) is made of circular ducts with a diameter of 450 mm. T-junction duct element with a 45° connection (3,6) allows fresh air intake and exhaust air discharge. Both the air intake and exhaust elements are equipped with adjustable control dampers (4,5). By adjusting the damper opening, the operator can control the proportion of fresh air entering the system. The photovoltaic panel is mounted on test platform located in the central part of the chamber. The platform allows adjustment of the panel inclination in order to simulate different mounting configurations typically used in flat roof photovoltaic installations. Panel can be positioned at inclination angles of 15°, 30° and 45°. Air temperature inside the chamber can be regulated using a gas heater installed within the system, allowing the simulation of elevated temperature conditions typical for rooftop environments during summer

operation. Above the test area, a water spraying system is installed to simulate precipitation events. The rainfall simulation system is designed to reproduce both light and heavy precipitation. The flow rate is defined by the installed nozzles, while separate pipe lines for light and heavy rainfall are connected directly to the water supply and activated by electromagnetic valves. In addition, a mist generation system consisting of a high-pressure pump and mist nozzles is used to simulate fog conditions and increase the air humidity.

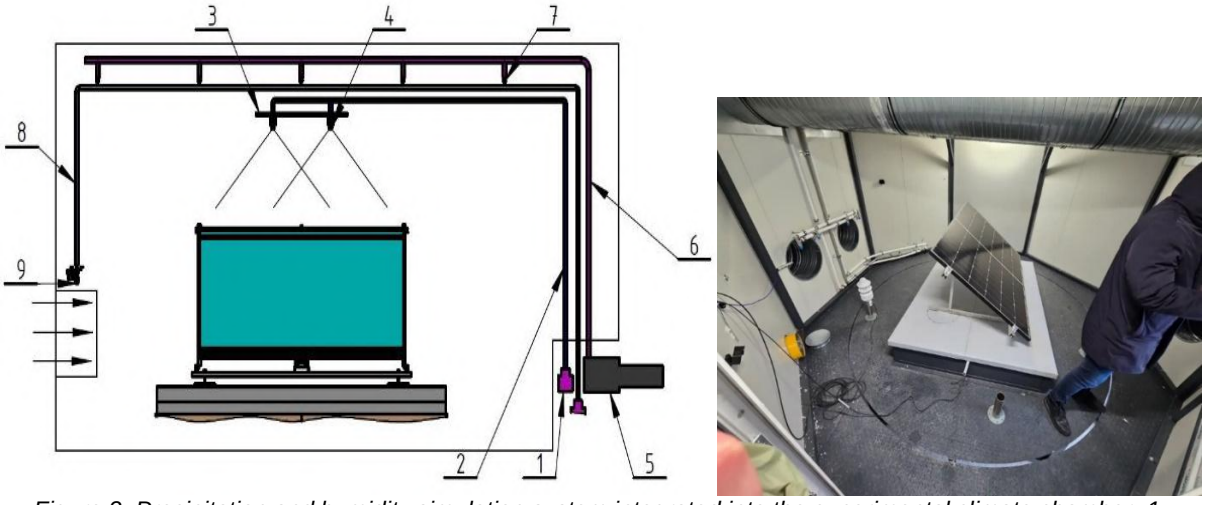


Figure 2: Precipitation and humidity simulation system integrated into the experimental climate chamber. 1-Solenoid valve for vertical rainfall simulation, 2-Water supply pipe for the vertical rainfall system, 3- Nozzle support frame for the vertical precipitation system, 4-Lechler spray nozzles for rainfall simulation, 5- High-pressure pump, 6-High-pressure stainless-steel pipeline, 7- High pressure misting nozzles for water fog generation,8-Pipes for horizontal injection of precipitation into the airflow stream, 9-Spray nozzles for horizontal injection of precipitation into the airflow stream

The precipitation simulation system consists of water supply pipes (2,8), a set of industrial spray nozzles Lacher(3,4,9) and electromagnetic control valves (1). High-pressure pump Annovi Reverberi HRM-M 02.08 (5) provides flow rate of 2 L/min and operating pressure of 80 bar. The high pressure combined with mist nozzles (7), generates very fine water droplets that are easily dispersed and absorbed in the air, increasing the relative humidity inside the chamber. The test platform supporting the photovoltaic panel is designed to reproduce the layered structure of a typical flat roof assembly. The structure consists of a trapezoidal steel sheet, vapor barrier, layers of mineral wool thermal insulation and a waterproofing membrane. Polyamide mounting plate is installed on the waterproofing layer and serves as the interface element for the photovoltaic mounting structure.

2.2 Measuring equipment

The measurement system was designed to provide continuous acquisition of key environmental parameters including air temperature, air velocity, relative humidity, participation and moisture content within the roofing layers. Parameters inside the test chamber are measured using the Almemo meteorological sensor Fmd7-60. Wind velocity inside air duct is monitored using a digital anemometer Ahlborn Fvad 15-H25RGAMN40. Air temperature measurements are performed using a Pt100 resistance temperature sensor. To detect potential water intrusion through the roof assembly, three capacitive moisture sensors Ahlborn FHA696GF1 are embedded within the layers of the test roof structure. Data is collected on Almemo 710 data logger. Data acquisition, visualization and processing are performed using WinControl software. During the experiments, the measurement system continuously recorded air temperature, relative humidity, atmospheric pressure, participation, wind velocity, wind direction and moisture content within the roofing material layers.

2.3 Experimental procedure

During the experimental tests three different mounting configurations were evaluated. The experiments included variations in the panel inclination angles, panel orientation and simulated atmospheric conditions. The photovoltaic panel was positioned at a predefined orientation relative to the airflow direction: 0°, 90°, and 180°. The inclination of the photovoltaic panel was also adjusted depending on the selected test configuration to 25°, 30° and 35°. For tests involving elevated temperature conditions, the air inside the chamber was heated to the predefined temperature level using the gas heater. Once the target temperature was reached, the heater was switched off and the test continued. In experiments performed at elevated temperatures, the ventilation dampers were adjusted to operate in a recirculation mode. At the end of each experimental run, all measured data were stored and documented. After completion of the measurement phase, a visual and mechanical inspection of the photovoltaic

panel, mounting structure and roof assembly was performed. Particular attention was given to possible structural deformations, loosening of fastening elements and signs of water penetration through the waterproofing layers.

3 RESULTS AND DISCUSSION

A series of twelve experimental tests was conducted. The tested configurations differed in the panel inclination angles, panel orientation and simulated atmospheric conditions.

3.1 Test configuration 1

In the first three tests panel was exposed to strong airflow generated by the ventilation system under cold and humid air conditions, simulating a winter day with strong wind. The photovoltaic panel was installed with a constant inclination angle of 30°, while its orientation relative to the airflow direction was varied (0°, 90°, and 180°). In the following three tests, storm conditions were simulated by applying the simultaneously strong wind and intensive rainfall. Airflow velocity inside the chamber was maintained within the range of 13.5 m/s to 18.5 m/s. These tests were conducted for a duration of 30 minutes, while maintaining the same panel inclination angle and varying the orientation relative to the airflow (180°, 90°, and 0°). During testing, no plastic deformation of the panel or supporting structure was observed. Wind simulation caused minor vibrations, which were not structurally significant. However, due to these vibrations, loosening of the fastening nuts occurred, leading to water penetration inside the mounting element. Although the quantity of water detected was relatively small, its presence beneath the waterproofing layer is considered unacceptable.

Table 1. Test configuration 1 initial settings and main operating parameters

Test No.	1	2	3	4	5	6
Panel position	0°	90°	180°	180°	90°	0°
Panel inclination	30°	30°	30°	30°	30°	30°
Temperature	9°C	10°C	11°C	12,6°C	11,2°C	11°C
Wind speed	13,5m/s-18,5m/s	13,5m/s-18,5m/s	13,5m/s-18,5m/s	13,5m/s-18,5m/s	13,5m/s-18,5m/s	13,5m/s-18,5m/s
Precipitation	0mm/h	0mm/h	0mm/h	NR	NR	NR
Humidity	94.00%	89%->85%	85%->77,6%	77,6%->91%	90%	95%
Moisture	7.8%	7.8%	7.8%	7.7%	7.7%	8.0%

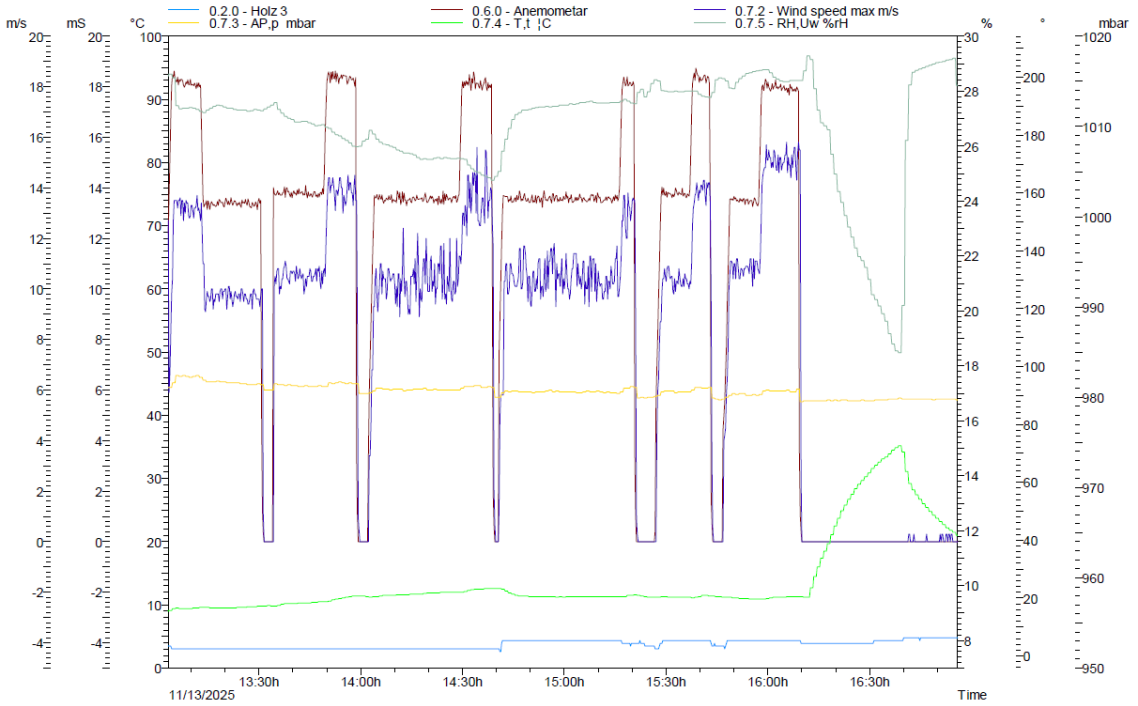


Figure 3: Test configuration 1- processed measurement data

3.2 Test configuration 2

A series of six experimental scenarios was conducted with a panel inclination of 25°. In the first three tests panel was exposed to strong airflow generated by the ventilation system under cold and humid air conditions, simulating a winter day with strong wind. Orientation relative to the airflow direction was varied (0°, 90°, and 180°). Airflow velocity inside the chamber was maintained within the range of 13.5 m/s to 18.5 m/s. In the following three tests, storm conditions were simulated through the simultaneous action of strong wind and intense rainfall generated by the nozzle system. These tests were conducted for a duration of 30 minutes.

Table 2. Test configuration 2 initial settings and main operating parameters

Test No.	7	8	9	10	11	12
Panel position	0°	90°	180°	180°	90°	0°
Panel inclination	25°	25°	25°	25°	25°	25°
Temperature	7,7°C	10°C	10°C	7,5°C	7,5°C	7,5°C
Wind speed	13,5m/s->18,5m/s	13,5m/s->18,5m/s	13,5m/s->18,5m/s	13,5m/s->18,5m/s	13,5m/s->18,5m/s	13,5m/s->18,5m/s
Precipitation	0mm/h	0mm/h	0mm/h	NR	NR	NR
Humidity	81.30%	70%->66%	72.80%	70,3%->91%	91%->95%	95%
Moisture	7.5%	7.6%	7.8%	7.8%	7.5%	7.6%

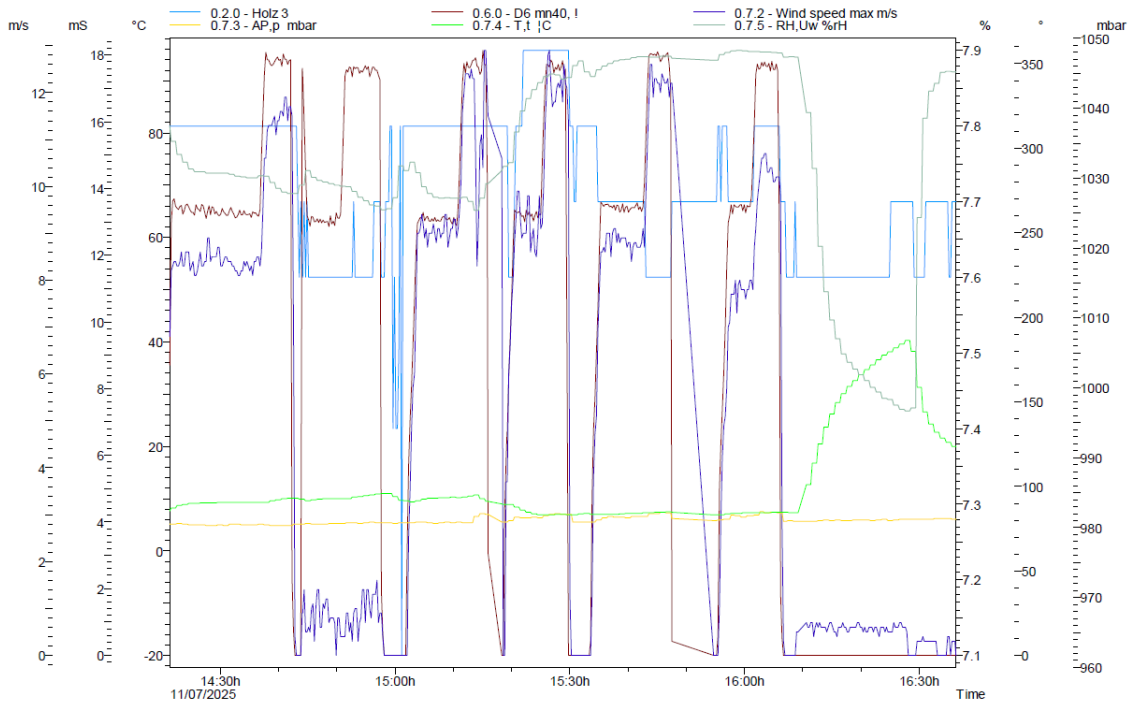


Figure 4: Test configuration 2 - processed measurement data

During wind simulation, vibrations were more pronounced compared to the first test. These vibrations were visible on the supporting structure due to lack of stiffness of the profile configuration. Loosening of the fastening nuts occurred again, which resulted in significant water penetration into the mounting element. The observed water ingress was noticeably larger than in the first test.

4 CONCLUSION

The experimental investigation presented in this study evaluated the mechanical stability and watertightness performance of several photovoltaic mounting system configurations installed on a simulated flat roof assembly under controlled environmental conditions. The developed climate chamber enabled realistic simulation of wind exposure, storm events and high humidity conditions. Comparative analysis of the obtained results indicates that test configurations 1 and 2 exhibited problems related to water leakage caused by vibrations of the mounting structure. In several tests, vibrations generated by wind loading caused loosening of fastening nuts in cases where insufficient tightening force was applied to the central bolt. This resulted in localized water penetration beneath the waterproofing membrane, highlighting the importance of proper fastening torque and structural stiffness of the mounting system. Among the tested configurations, test configuration 1 demonstrated the most stable system

behaviour. The experimental results further indicate that a panel inclination angle of 30° represents a structurally stable configuration, providing a optimal balance between structural stiffness and efficiency of the mounting system under simulated environmental loads. Overall, the conducted experiments demonstrate that the reliability of rooftop photovoltaic mounting systems strongly depends on the mechanical properties of fastening elements, proper tightening torque during installation and the structural stiffness of the mounting configuration. The presented experimental methodology provides a useful framework for evaluating the interaction between photovoltaic mounting systems and flat-roof waterproofing layers under realistic climatic loading conditions. Future research should continue experimental testing under variable environmental conditions over extended time periods. Particular focus should address the issues identified during the present study including optimization of mounting element geometry, improvement of material mechanical properties and refinement of fastening strategies in order to enhance the structural stability and long term watertightness of rooftop photovoltaic mounting systems. The study is based on a limited number of experimental configurations and a relatively small sample size. Therefore, the findings primarily provide insight into the underlying failure mechanisms rather than a statistical evaluation of system performance. Additional components and system modifications have already been prepared to enable testing under significantly higher airflow velocities. The next experimental phase will include testing and measurements conducted at approximately double wind velocities compared to those investigated in the present study, corresponding to extreme microclimatic conditions that may occur in rooftop environments in this region. The results of these experiments will allow a better evaluation of the mounting system behaviour under higher wind loads and help identify critical structural points of the system.

REFERENCES

[1] IZAM, Najwa Syahirah Mohamed Nor et al. Sustainable development perspectives of solar energy technologies with focus on solar photovoltaic—A review. Energies. 2022, 15(8), 2790. https://doi.org/10.3390/en15082790
[2] ZHANG, Haiyan et al. Green or not? Environmental challenges from photovoltaic technology. Environmental Pollution. 2023, 320, 121066. https://doi.org/10.1016/j.envpol.2023.121066
[3] EVRO, Solomon et al. Solar PV technology cost dynamics and challenges for US new entrants. American Journal of Energy Research. 2023, 11(1), 15–26. https://doi.org/10.12691/ajer-11-1-2
[4] SALEM, T. and E. KINAB. Analysis of building-integrated photovoltaic systems: a case study of commercial buildings under Mediterranean climate. Procedia Engineering (2015) 118:538–545. https://doi.org/10.1016/j.proeng.2015.08.473
[5] BAYOD-RÚJULA, Angel A.; ORTEGO-BIELSA, Abel; MARTÍNEZ-GRACIA, Amaya. Photovoltaics on flat roofs: Energy considerations. Energy (2011) 36(4):1996–2010. https://doi.org/10.1016/j.energy.2010.04.024
[6] KOPP, Gregory A., Steve FARQUHAR and Murray J. MORRISON. Aerodynamic mechanisms for wind loads on tilted, roof-mounted solar arrays. Journal of Wind Engineering and Industrial Aerodynamics (2012) 111:40–52. https://doi.org/10.1016/j.jweia.2012.08.004
[7] BROWNE, Matthew T.L. et al. Wind loading on tilted roof-top solar arrays: The parapet effect. Journal of Wind Engineering and Industrial Aerodynamics (2013) 123(A):202–213. https://doi.org/10.1016/j.jweia.2013.08.013
[8] BANKS, David. The role of corner vortices in dictating peak wind loads on tilted flat solar panels mounted on large, flat roofs. Journal of Wind Engineering and Industrial Aerodynamics (2013) 123(A):192–201. https://doi.org/10.1016/j.jweia.2013.08.015
[9] STATHOPOULOS, Ted, Ioannis ZISIS and Eleni XYPNITOU. Local and overall wind pressure and force coefficients for solar panels. Journal of Wind Engineering and Industrial Aerodynamics (2014) 125:195–206. https://doi.org/10.1016/j.jweia.2013.12.007
[10] WANG, Jingxue et al. Effects of building parameters on wind loads on flat-roof-mounted solar arrays. Journal of Wind Engineering and Industrial Aerodynamics (2018) 174:210–224. https://doi.org/10.1016/j.jweia.2017.12.023
[11] ALRAWASHDEH, Hani and Ted STATHOPOULOS. Wind loads on solar panels mounted on flat roofs: effects of geometric scale. Journal of Wind Engineering and Industrial Aerodynamics (2020) 206:104339. https://doi.org/10.1016/j.jweia.2020.104339
[12] WALTER, Ana, Jorge DE BRITO and Jorge Grandão LOPES. Current flat roof bituminous membranes waterproofing systems – inspection, diagnosis and pathology classification. Construction and Building Materials (2005) 19(3):233–242. https://doi.org/10.1016/j.conbuildmat.2004.05.008
[13] FEITEIRA, João, J. Grandão LOPES and J. DE BRITO. Mechanical performance of liquid-applied roof waterproofing systems. Journal of Performance of Constructed Facilities (2013) 27(3):244–251. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000314
[14] DOMINGUEZ, Anthony, Jan KLEISSL and Jeffrey C. LUVALL. Effects of solar photovoltaic panels on roof heat transfer. Solar Energy (2011) 85(9):2244–2255. https://doi.org/10.1016/j.solener.2011.06.010
[15] SHUKLA, Nitin et al. Thermal impact of adhesive-mounted rooftop PV on underlying roof shingles. Solar Energy (2018) 174:957–966. https://doi.org/10.1016/j.solener.2018.09.079
[16] FEDOROVA, Anna et al. A testing methodology for quantification of wind-driven rain intrusion for building-integrated photovoltaic systems. Building and Environment. 2021, 199, 107917. https://doi.org/10.1016/j.buildenv.2021.107917

Maximum use of renewable energy in low accumulative facilities of intensive vegetable production

Marko Gložinić[1], Matija Gložinić[1], Ante Čikić[1], Zoran Busija[1]

[1]University North, Department of Mechatronics, 104. brigade 3, 42000 Varaždin, Croatia
E-mail: maglozinic@unin.hr, mglozinic@unin.hr, acikic@unin.hr, zbusija@unin.hr

Abstract: *The possibilities of maximum use of residual heat energy from a biogas plant and variable solar energy from a PV power plant combined with conversion into electricity and heat energy for year-round energy sustainability of advanced low-accumulative facilities in intensive vegetable production (tomatoes and lettuce) were considered and analysed. The simulation model and quantified values comparable to real parameters and practical needs show the gradual use of heat energy of different media temperatures and electricity in accordance with the dynamics of priorities through consumption, production, availability, accumulation and storage of energy. A real example shows maximum energy sustainability and the connection of renewable energy in relation to continental climate parameters and prescribed internal microclimatic and technological conditions in intensive vegetable production. The contribution to the development of a green energy transition model of low-accumulative facilities with high energy efficiency, production stability and sustainability is emphasized.*

Keywords*:* renewable energy, heat storage, low-acumulative facility, vegetable production, energy efficiency

Article Classification*: Scientific Paper*

1 INTRODUCTION

This study analyzed the possibilities of maximum use of residual and waste heat from a cogeneration biogas plant and variable solar energy from a PV power plant, heat storage tank and battery storage of electricity. The applicable possibilities of two-stage utilization of existing and additional thermal energy of different media temperatures and electricity in process-technological production low-accumulative facilities with specific characteristics and targeted market value were considered and analyzed. In addition to agro-technological conditions, the speed of achieving and stability of maintaining air temperature with a small deviation from the set value in the plant growing zone in all its biological periods significantly affects the quality of fruits and expected yield [1,2]. Optimal adjustment of the ratio of heating speed and prevention of sudden cooling of a low-accumulative facility within a technologically acceptable time interval requires the compatibility of the small thermophysical characteristics (product $\lambda \cdot c \cdot \rho$) of the facility with a rational and sustainable heating system. The time, cause and frequency of reduction or interruption of heat supply, thermophysical characteristics of the surface material, windiness, orientation and degree of reflection - absorption of external surfaces exposed to solar radiation are factors that directly affect the thermal response of the greenhouse, time constant = accumulated heat / heat loss. The time constant is very small and significantly affects the efficiency of the heating system, especially with rapid changes in microclimatic conditions at low outside air temperatures [3,4]. In addition, heat losses increase by as much as 10% - 60% when exposed to wind, especially with rain at outside air temperatures between – 2°C and + 10°C when higher wind speeds are more frequent. The duration, yield and cost-effectiveness of intensive tomato and lettuce production depend strongly on a rational energy system (heating / cooling) and total energy consumption per unit of usable area. Due to the small time constant and thermophysical characteristics of the facilities, and the large inertia of the secondary hot water system, with frequent, sometimes sudden, changes in microclimatic conditions in the morning and/or evening hours, a higher response speed and readiness of the thermal energy system to achieve, maintain and stabilize different internal air temperatures in the heating zones is required [5]. Thermal energy accumulation is necessary in heating systems with a large ratio of water quantities in the secondary to the primary circuit, which characterizes low-accumulative facilities such as greenhouses / greenhouses with larger and larger surfaces for intensive plant cultivation. With the help of an accumulation-compensation heat tank, sudden internal changes in air temperature are mitigated or eliminated, and the thermal and hydraulic balance between the heat source and the secondary heating circuit (consumer) is maintained [6]. Thermal stability and sustainability are ensured, especially with variable renewable energy sources. The shape and size of the heat tank affect the efficiency and economy of thermal energy production and consumption.

For year-round hydroponic tomato production and intensive lettuce production in shallow water pools in controlled technical and technological conditions within closed greenhouses - climates, the synchronized use of available residual thermal energy and renewable solar energy (solar power plant) was investigated through priority use for heating/cooling and the necessary technological conditions with thermal accumulation and battery storage of excess electricity [7,8]. Due to the different microclimatic parameters (temperature, relative humidity and air circulation) and technological conditions of tomato and lettuce cultivation, a gradual use of thermal energy is

envisaged, adjustment of thermal capacity, achievement of energy continuity and utilization to maximum technological and production efficiency and cost-effectiveness in appropriate time intervals throughout the year, without a CO_2 footprint in the surrounding area. In this way, high energy efficiency, production cost-effectiveness, business sustainability and market advantage are achieved [9].

2 METHODS

A real-life example demonstrates maximum energy sustainability and the connection of renewable energy in relation to continental climate parameters and prescribed internal microclimate and technological conditions during intensive tomato and lettuce production in two connected facilities with different growing concepts and compatible technologies. The simulation model and quantified values associated with real parameters show the gradual use of thermal energy of different media temperatures and electricity in accordance with the dynamics of priorities through consumption, production, availability, accumulation and storage of energy.

In real conditions of external climatic parameters (temperature, relative humidity, air velocity, insolation, cloudiness, etc.), internal microclimatic and technological conditions and residual thermal energy from the cogeneration biogas plant, the measured values were linked to the sustainable possibilities of the envisaged PV power plant through the production, consumption and storage of energy during the year-round cultivation of vegetables. Using mathematical equations, the interdependent microclimatic, thermotechnical and energy parameters were quantitatively described and linked, and a “simulation model” was created with developed parameters for determining the size and scale of renewable energy capacities for optimal year-round energy sustainability and efficiency of the production system.

2.1 Experimental facilities and conditions

High-tech greenhouse for (intensive hydroponic production 11 months/year) tomato cultivation (grapolo or beef); greenhouse area 10000 m^2 (80 m x 125 m); yield = 60 - 70 kg/m^2, year.
The internal daily air temperature in the tomato cultivation zone is $\vartheta_{gh,day}$ ≈ 20°C to 22°C, $\vartheta_{gh,night}$ ≈ 16°C to 18°C. The relative air humidity in the greenhouse is about 55% - 75%. Multi-zone ("5-zone") greenhouse heating is planned, Figure 1, adapted to the temperature conditions of hydroponic tomato cultivation in the winter and transitional periods. Thermotechnical systems with heat exchange by radiation and, to a lesser extent, convection in relatively still air have been proposed, which enable uniform and complete heating of the space for plant development, growth and fertility, which directly contributes to economy and energy efficiency, maintaining air temperature during the day and/or night in individual zones of hydroponic cultivation. The phase of ripening and multiple fruiting of tomatoes is shown in Figure 2. The maximum thermal energy required for heating a greenhouse for hydroponic tomato cultivation at a sliding temperature regime of hot water 70°/55°C is Q = 2500 kWh.

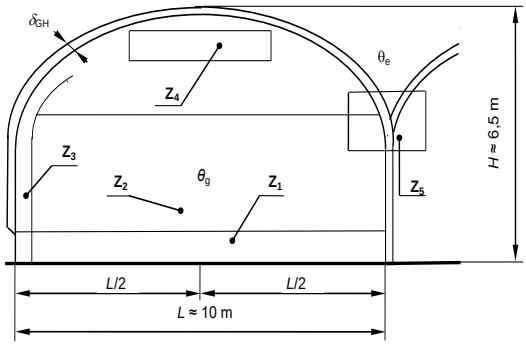


Figure 1. Greenhouse heating zones *(cross-section of a broad)*

Figure 2. The ripening and multiple fruiting phase of tomatoes

High-tech greenhouse for (intensive hydroponic production in shallow pools 11 months/year) lettuce cultivation; annual yield about 350 tons (about 875.000 pieces); greenhouse area with shallow pools 7,000 m^2 (80 m x 87.5 m); LED lighting maximum 0.060 kW/m^2 x 1,520 m^2 = 91.2 kW (used in the first phase of lettuce cultivation). The water temperature in the pools for floating lettuce cultivation is $\vartheta_{gh,day}$ ≈ 18°C to 24°C, $\vartheta_{gh,night}$ ≈ 14°C to 16°C. The air temperature in the lettuce cultivation zone (above the water) is ϑ_{gh} ≥ 13°C. The relative air humidity above the lettuce in the greenhouse area is about 60% - 70%. The final growing phase of lettuce on floating plates in shallow pools in a greenhouse is shown in Figure 3. The maximum thermal energy required for heating water in shallow

pools and the growing area above the lettuce in the greenhouse at a sliding temperature regime of hot water 53°/40°C is Q = 1300 kWh.

Figure 3. Final phase of floating lettuce cultivation on water in shallow pools

Total available residual heat energy (biogas plant) Q = 1200 kWh.
High-efficiency heat storage and compensation tank V = 300 m^3, heat storage capacity 14 MW to 17 MW with maximum efficiency at a temperature difference of the heating medium of about 40°C to 50°C. The maximum temperature of the heating medium (hot water) in the tank is 85°C – 90°C.
Additional heat energy available from the storage tank Q = 1300 kWh.
External design microclimate conditions at the construction site of low-accumulation facilities; winter period: ϑ_e = – 12°C, φ_e = 40% - 80%, w_e = 1.5 – 5 m/s; summer period: ϑ_e = + 32°C, φ_e = 40% - 60%, w_e = 1.5 – 5 m/s.
For fixed photovoltaic panels with a slope of about 30° – 39° at the intended location (without sun tracking) oriented towards the south, the estimated seasonal energy potential is expressed as:

- summer period ~ 6.79 kWh/day per kW of installation
- autumn period ~ 3.03 kWh/day per kW of installation
- winter period ~ 1.68 kWh/day per kW of installation
- spring period~ 4.69 kWh/day per kW of installation

A 500 kW solar power plant is planned for the production of electricity and its dynamic direction for the conversion and accumulation of thermal energy, operational consumption and/or storage of surplus electricity in a battery system. The planned battery system should be optimally dimensioned for the storage of dynamic surplus electricity and maximum production effects of the solar power plant in predictable favorable insolation intervals in a certain area. Also, the battery system will enable dynamic sustainability of the needs of electricity and additional thermal energy in low-accumulation facilities in less favorable insolation intervals and microclimatic conditions at the location. Variable electricity production from the PV system should be directed in accordance with the output capacities and achieved values in different ways, sequentially according to the priority of energy consumption, conversion or storage.

2.2 Measuring equipment

For dynamic monitoring of consumption, production, availability, accumulation and storage of thermal/electrical energy, applicable equipment and circuits are used: various sensors for measuring influential external, internal and technological parameters, control elements and circuits, drive devices with variable characteristics and an integrated control and management circuit with customized software and appropriate hardware equipment for continuous monitoring, control and management. The applied measuring and control equipment and elements practically correspond to the simulation display of numerical parameters and the quantification of all energy values in the observed time intervals.

3 RESULTS AND DISCUSSION

3.1 Displaying results

The numerical values in Table 1 show the average heat energy consumption in a technically developed greenhouse for hydroponic tomato cultivation and microclimate conditions in intervals of external climate conditions in a wide range of external air temperatures. Thermoenergetic sustainability and regulated capacity of individual heating

zones are ensured by residual energy from the cogeneration plant (1,200 kW) in the hot water operating temperature regime of around 70°/50°C, and on extremely cold days by supplementing from the thermal energy storage tank, where excess thermal energy is stored, and partly by converting electricity into thermal energy from a solar power plant. In more favorable time intervals, the excess unused thermal energy for heating the tomato greenhouse is stored in the storage tank to a lower temperature level, and then the total residual part in the hot water operating temperature regime of around 53°/40°C is used for heating the lettuce growing system in shallow pools in a closed greenhouse. Table 2 shows the annual electricity production of a solar power plant with a power of about 500 kW, planned in a free space on land with an adaptable surface and proper orientation and inclination of photovoltaic panels at the location of low-accumulative facilities.

Table 1. Average heat energy consumption in greenhouses from cogeneration and storage at variable outdoor air temperatures

Outside air temp. ϑ_e, °C	Average inside air temp. ϑ_i, °C	Average heat energy consump. total Q_x, kWh	Heat energy Cogen. plant $Q_{cog.p.,}$ kWh	Heat energy – solar power plant $Q_{sol.e.}$, kWh	Average inside air temp. ϑ_i, °C	Average heat energy consump. total Q_x, kWh	Heat energy Cogen. plant $Q_{cog.p,}$ kWh	Heat energy: cog. plant return / solar p. plant $Q_{cog.p,p./sol.p.}$ kWh
	GREENHOUSE 10000 m², TOMATOES				**GREENHOUSE 7000 m², LETTUCE**			
- 18	20	2500	≈ 1200	1300 (from stor.)	18/13	1300	≈ 1150	0/150
- 15	20	2303	≈ 1200	1103 (from stor.)	18/13	1192	≈ 1150	0/42
- 12	20	2105	≈ 1200	905 (from stor.)	18/13	1085	≈ 1150	65/0
- 10	20	1974	≈ 1200	794 (from stor.)	18/13	1011	≈ 1150	135/0
- 8	20	1842	≈ 1200	642 (from stor.)	19/13	975	≈ 1150	175/0
- 6	20	1711	≈ 1200	511 (from stor.)	20/13	940	≈ 1150	210/0
- 4	20	1580	≈ 1200	380 (from stor.)	20/14	870	≈ 1150	280/0
- 2	20	1447	≈ 1200	247 (from stor.)	20/14	795	≈ 1150	355/0
0	20	1316	≈ 1200	116 (from stor.)	20/15	725	≈ 1150	425/0
+ 2	20	1135	≈ 1200	+65 (to storage)	20/15	650	≈ 1150	500/0
+ 4	20	1053	≈ 1200	+147 (to storage)	20/15	580	≈ 1150	570/0
+ 6	20	921	≈ 1200	+279 (to storage)	21/16	542	≈ 1150	608/0
+ 8	20	790	≈ 1200	+410 (to storage)	21/16	470	≈ 1150	680/0
+ 11	20	592	≈ 1200	+608 (to storage)	22/16	400	≈ 1150	750/0
+ 14	20	395	≈ 1200	+805 (to storage)	22/16	290	≈ 1150	860/0

Table 2. Annual electricity production of a solar power plant with a capacity of about 500 kW

Month	Daily average kWh/kW	Days	Before losses, kWh	After 15% losses, kWh	Electricity produced, kWh
I	1,5	31	500 × 1,5 × 31 = 23.250	× 0,85 → 19.763	≈ 19.760
II	2,0	28	500 × 2,0 × 28 = 28.000	× 0,85 → 23.800	≈ 23.800
III	3,5	31	500 × 3,5 × 31 = 54.250	× 0,85 → 46.113	≈ 46.100
IV	4,5	30	500 × 4,5 × 30 = 67.500	× 0,85 → 57.375	≈ 57.375
V	5,5	31	500 × 5,5 × 31 = 82.250	× 0,85 → 69.913	≈ 69.900
VI	6,5	30	500 × 6,5 × 30 = 97.500	× 0,85 → 82.875	≈ 82.870
VII	6,8	31	500 × 6,8 × 31 = 105.400	× 0,85 → 89.590	≈ 89.590
VIII	6,5	31	500 × 6,5 × 31 = 100.750	× 0,85 → 85.638	≈ 85.600
IX	4,5	30	500 × 4,5 × 30 = 67.500	× 0,85 → 57.375	≈ 57.375
X	3,5	31	500 × 3,5 × 31 = 54.250	× 0,85 → 46.113	≈ 46.100
XI	2,0	30	500 × 2,0 × 30 = 30.000	× 0,85 → 25.500	≈ 25.500
XII	1,2	31	500 × 1,2 × 31 = 18.600	× 0,85 → 15.810	≈ 15.800
Total:			729.250	619.865	619.770 ≈ 620.000

Total annual consumption, production, partial conversion of electrical energy into thermal energy and energy storage in two-way use for heating/cooling of buildings and technological operational needs is shown in Table 3. Numerical values were determined for the average outdoor air temperatures for three decades in each month of the year.

Table 3. Annual heat/electricity balance of low-accumulation facilities with heat storage and PV power plant

Month	Decade/month	Average outside air temp. ϑ_e, °C	Heat energy consumption $Q_{h.e.}$, MWh	Available heat energy / other energy disp. MWh	Accumulated heat energy $Q_{a.}$, MWh	Electric. energy – PV power plant $P_{el.PV}$, MWh	Electricity consumption $P_{el.e.}$, MWh	Balance: electricity consumption / ΔPV $P_{el.e.}$, MWh	PV-consumer supply Deficit /surplus MWh	Electricity balance, without LED lighting $P_{el.e.}$, MWh
I	1.	0	237	803 /+134	+134	19,76	75,8	75,8 / − 56,4	− 56,4	30,5 − 10,74
	2.	1	216							
	3.	1	216							
II	1.	1	216	726 /+122	+122	23,8	58,0	58 / − 34,2	− 34,2	27,6 − 3,8
	2.	3	198							
	3.	4	190							
III	1.	5	156	804 /+406	+406	46,1	35,2	35,2 / + 10,9	+ 10,9	35,2 +10,9
	2.	7	125							
	3.	9	117							
IV	1.	10	107	777 /+499	+499	57,375	29,95	29,95 / + 27,8	+ 27,8	29,5 +27,875
	2.	12	91							
	3.	14	80							
V	1.	15	42	803 /+749	+749	69,9	20,8	20,8 / + 49,1	+ 49,1	20,8 + 49,1
	2.	17	7							
	3.	19	5							
VI	1.	19	8	777 /+756	+756	82,87	26,9	26,9 / + 55,97	+ 55,97	26,9 + 55,97
	2.	21	8							
	3.	23	5							
VII	1.	22	5	803 /+788	+788	89,59	30,95	30,95 / + 58,6	+ 58,6	30,95 + 58,6
	2.	23	5							
	3.	23	5							
VIII	1.	22	5	803 /+787	+787	85,6	30,95	30,95 / + 54,6	+ 54,6	30,95 + 54,6
	2.	22	5							
	3.	21	6							
IX	1.	18	6	777 /+722	+722	57,375	28,4	28,4 / + 28,975	+ 29	28,4 + 28,975
	2.	17	7							
	3.	15	12							
X	1.	13	71	803 /+533	+533	46,1	34,2	34,2 / + 11,9	+ 11,9	34,2 + 11,9
	2.	11	80							
	3.	9	99							
XI	1.	7	124	777 /+323	+323	25,5	60,1	60,1 / − 34,6	− 34,6	29,5 − 4,0
	2.	5	158							
	3.	3	172							
XII	1.	2	191	803 /+210	+210	15,8	75,8	75,8 / − 60	− 60	30,5 − 14,7
	2.	1	201							
	3.	1	201							

For possible lower external air temperatures with greater frequency in XII, I and II months, heat losses of low-accumulative objects will be compensated by higher accumulation of thermal energy. The used heat energy in the greenhouse for intensive cultivation of tomatoes at the hot water operating temperature regime of 70°/55°C is further used with high efficiency for heating water in shallow pools and the air space in the greenhouse for growing lettuce up to a hot water temperature of about 40°C. The simultaneity factor of operation of electro technical devices, equipment and assemblies in X, XI, XII, I, II, III and IV months is $f_1 = 0.3$, and in V, VI, VII, VIII and IX months it is $f_2 = 0.2$. The duration of evaporative cooling of the greenhouse for growing lettuce in the VI, VII, VIII and part of the IX months lasts 10 - 12 hours a day. As needed, LED lighting for the first phase of technological lettuce cultivation is used 10 hours a day in March and April, and 16 hours a day in November, December, January and February.

3.2 Discussion – analysis of results

The stable supply of heat energy from the biogas cogeneration plant ranges from 777 MWh - 804 MWh per month. Depending on the average outdoor air temperature over the decades in individual months of the year, the remaining heat energy ranges from a minimum of 122 MWh in the second month to a maximum of 788 MWh in the seventh month. The volume of the accumulation tank is V = 300 m^3 and the excess heat energy is accumulated in it at a rate of about 181.5 kWh in 3.9 days to a maximum usable heat energy accumulation capacity of about 17 MWh. In periods of intensive heating in the months of November, December, January and February, the rate of accumulation of excess heat energy is 3.9 days to a minimum of 1.3 days. The accumulation of heat energy with a maximum water temperature of about 80°C - 90°C is used in case of occasional lower outdoor air temperatures for the technological sustainability of heating facilities. In the summer and transitional periods, maximum thermal energy accumulation is maintained, and the remaining heat is recycled within the cogeneration plant. Total electricity consumption without the use of LED lighting ranges between 25 MWh - 30 MWh for each month of the year and with a battery system with a capacity of about 200 kW, the sustainability of electricity supply from PV is about 91%. In the months of November, December, January and February, about 9% of electricity from the grid system is needed from the total annual needs. Using LED lighting in the winter period increases electricity consumption by about 2 - 3.6 times compared to other months of the year. With the use of a battery pack for stabilizing and regulating the dynamic regime of electricity, an additional about 36.5% of electricity from a stable grid system is needed compared to the total annual needs. The intensity and time intervals of using LED lighting depend on climatic conditions and the assessment of the cost-effectiveness of the technological acceleration of the first phase of lettuce cultivation and the price of kWh of electricity from the grid system.

4 CONCLUSION

The maximum energy sustainability and utilization of thermal and electrical energy from renewable sources in year-round hydroponic production in greenhouses, tomatoes (10000 m^2) and lettuce (7000 m^2) in a continental climate zone is presented. On a real example, quantified values show the gradual use of thermal energy from a cogeneration biogas plant with different hot water temperatures of 70°/55°C and 53°/40°C and electrical energy from a 500 kW solar power plant. The dynamic consumption, production, availability, accumulation and storage of energy in relevant annual intervals are numerically presented and analyzed. Full thermal sustainability and the maximum reach of stable electrical energy supply, about 91% from the PV system, are highlighted. Further research into technical and technological parameters and optimization contribute to the development of a green energy transition model for low-accumulation facilities with high energy efficiency, production stability and sustainability.

REFERENCES

[1] BARTOK, John W., Jr. Energy Conservation for Commercial Greenhouses. 2001 rev. Ithaca, NY: Natural Resource, Agriculture, and Engineering Service, 2001. ISBN 0-935817-74-3.
[2] ALDRICH, Robert A. and John W. BARTOK, Jr. Greenhouse Engineering. 3rd rev. Ithaca, NY: Natural Resource, Agriculture, and Engineering Service, 1994. ISBN 0-935817-57-3.
[3] NELSON, Paul V. Greenhouse Operation and Management. 6th ed. Upper Saddle River, NJ: Prentice-Hall, 2003. ISBN 0-13-010577-5.
[4] KAMP, P.G.H. and G.J. TIMMERMAN. Computerized Environmental Control in Greenhouses: a step by step approach. Ede, The Netherlands: IPC-Plant, 1996. ISBN 90-75627-01-7.
[5] ZHAO, Jing et al. Experimental study on dynamic thermal performance of seasonal hot water storage tank in a solar heating system. Applied Thermal Engineering [online]. 2025, 280, 128424. [accessed 2026-05-02]. DOI 10.1016/j.applthermaleng.2025.128424
[6] KARLINA, Yelizaveta et al. Numerical Simulation Study of Thermal Performance in Hot Water Storage Tanks with External and Internal Heat Exchangers. Energies [online]. 2024, 17(22), 5623. [accessed 2026-05-02]. DOI 10.3390/en17225623
[7] KHAN, Irfan et al. A study of trilemma energy balance, clean energy transitions, and economic expansion in the midst of environmental sustainability: New insights from three trilemma leadership. Energy [online]. 2022, 248, 123619. [accessed 2026-05-02]. DOI 10.1016/j.energy.2022.123619.
[8] GOPI, Ajith et al. Comparison of normal and weather corrected performance ratio of photovoltaic solar plants in hot and cold climates. Energy for Sustainable Development [online]. 2021, 65, 53–62. [accessed 2026-05-02]. DOI 10.1016/j.esd.2021.09.005.
[9] SAKA, Kenan. Evaluation of a grid-connected PV power plant: performance and agrivoltaic aspects. Environment, Development and Sustainability [online]. 2024, 26(12), 32319–32336. [accessed 2026-05-02]. DOI 10.1007/s10668-024-05098-z

Analysis of the environmental impacts of different electricity production sources in Slovenia using the Life Cycle Assessment methodology

Damjan Balabanič[1], Gregor Fijavž Kranjčan[1], Jelena Topić Božič[1,2]

[1]Fakulteta za industrijski inženiring Novo mesto, Slovenija
E-mail: damjan.balabanic@fini-unm.si, gregor.fijavz-kranjcan@students.fini-unm.si, jelena.topic.bozic@fini-unm.si

[2] Rudolfovo – Znanstveno in tehnološko središče Novo mesto, Slovenija
E-mail: jelena.topic.bozic@rudolfovo.eu

***Abstract**:* In the study, we evaluated the environmental impacts of energy systems for producing 1 kWh of electricity in Slovenia using the Life Cycle Assessment (LCA) methodology and the ReCiPe 2016 Midpoint (E) impact assessment method. The results were compared across four scenarios related to the construction of a second unit of the nuclear power plant, the closure of the Šoštanj Thermal Power Plant and a 100% share of renewable energy sources. The findings show that the current electricity generation structure exhibits higher impact in the global warming impact category compared to scenarios with a greater share of renewable and nuclear sources. The results also indicate that the scenario without the Šoštanj Thermal Power Plant, in which hydropower and nuclear energy account for the largest share, shows the lowest impacts on human health and ecosystem quality, while the nuclear–renewable scenario shows the lowest impacts on natural resource availability.

Keywords: *Life Cycle Assessment, ReCiPe 2016, electricity, nuclear power plant, renewable energy sources*

Article Classification: Scientific Paper

1 INTRODUCTION

In light of the growing environmental challenges and efforts to achieve climate goals, the transition to low-carbon energy sources is becoming one of the key tasks of modern society. Electricity production contributes significantly to greenhouse gas emissions and other environmental impacts, while at the same time it represents the foundation of the economy and everyday life [1, 2].

A distinctive feature of the Slovenian electricity system is its strong dependence on the two largest production units (the Krško nuclear power plant and the Šoštanj thermal power plant - TEŠ), which together produce more than half of all electricity needed in Slovenia [3]. Decisions on the future structure of electricity production therefore require comparable and comprehensive assessments of the environmental consequences of different technologies [2].

Slovenia uses a combination of different sources to produce electricity, reflecting the country's natural conditions, the historical development of the electricity system and strategic energy policies. In 2024, a total of 15,101 GWh of electricity was taken into the Slovenian electricity system from domestic production units connected to the transmission and distribution network. A significant share of this was production from renewable energy sources (RES) (6,047 GWh), which corresponds to 40.1% of all taken quantities. Among RES, hydropower played the most important role, as it generated 89.9% of all production from RES with 5,428 GWh. Solar energy contributed 473 GWh or 7.7% of production from RES, which indicates the rapid growth of solar power plants in recent years. Wind energy contributed 7 GWh, which is 0.1% of production from RES, while other renewable sources together reached 139 GWh or 2.3% of production from RES [3].

Life cycle analysis is a method that systematically records and evaluates the energy and material flows and the associated environmental impacts resulting from the operation of a product, process or service [4]. The method is based on the ISO 14040 [5] and ISO 14044 [6] standards, which require consistent definition of procedures, data and assessment frameworks. In the framework of LCA the environmental impacts of a product, process or activity are assessed at all stages of its existence, from the extraction of raw materials and their processing and production, through operation and use, to the final treatment of waste, including the possibilities of reuse and recycling at the end of its life. This approach allows for a comprehensive assessment of impacts that are not limited to the use phase alone, but also take into account the hidden environmental impacts in the background of production chains [4, 7, 8].

2 METHODS

Using the LCA and the ReCiPe 2016 Midpoint (E) impact assessment method, the environmental impacts of various electricity production sources used in the Slovenian electricity system are evaluated and compared. Based on the results of the functional unit of 1 kWh of electricity produced, four alternative scenarios for future electricity production in Slovenia are designed and analyzed.

- Scenario 1: Current production structure (reference scenario)

First, we evaluated the environmental impact of the current electricity system for the production of 1 kWh in all 22 environmental impact categories considered. The calculations were based on the existing production structure, in which nuclear energy accounts for the largest share (38.24%), followed by hydroelectric power plants (30.94%) and fossil brown coal/lignite (22.4%). Smaller shares belong to fossil gas (3.1%), photovoltaics (2.27%), pumped hydropower plants (1.95%), biomass (0.53%), energy recovery from waste (0.55%) and wind energy (0.04%). Decimal numbers are rounded to two decimal places [2].

- Scenario 2: Scenario without TEŠ (coal phase-out)

After the elimination of fossil brown coal or lignite and the renormalization of the shares, the production structure shifts significantly towards nuclear and hydropower, which together represent almost 90% of the total production. Nuclear energy becomes the dominant source with approximately 49%, and hydroelectric power plants with approximately 40%, while other sources have only marginal shares. Such a structure indicates a strongly decarbonized electricity mix, but at the same time a high concentration of production in two technologies.

- Scenario 3: 100 % RES

Scenario 3 was selected based on the study Dimnik (2025) in which we evaluated the impacts for the production of 1 kWh of electricity from 100 % RES [8]. According to this distribution, solar energy dominates with a 55% share, followed by hydroelectric power plants with approximately 35%, while wind power plants (4%), pumped hydropower plants (3%) and biomass (3%) have a significantly smaller contribution. Such a production structure represents a fully decarbonized electricity mix, while emphasizing the key role of solar and hydropower.

- Scenario 4: Nuclear renewable scenario (construction of the second nuclear power plant (NPP) unit, elimination of TEŠ, increase in the share of RES)

The last scenario assumes the construction of the second unit of the NPP, the complete elimination of fossil fuels from the electricity system and an increase in the share of RES. Based on these assumptions, we defined the production structure used in the evaluation of environmental impacts for the production of 1 kWh. According to the selected distribution, nuclear energy represents 50% of the total production, followed by hydroelectric power plants with 30% and solar energy with 15%, while pumped hydropower plants (2%), wind power plants (2%) and biomass (1%) have marginal shares in the production mix.

Life cycle inventory datasets of individual production sources were obtained by modeling in the SimaPro software based on the Ecoinvent database. The results were then exported to Microsoft Excel, where further analysis and calculation of the environmental impacts of the designed electricity production scenarios were performed.

To calculate the shares of the structure of electricity production in Slovenia, we used data from the study by Topić Božič et al. (2025), which calculated the average structure of electricity production for Slovenia in the period from 2020 to 2024 [2]. Since not all production sources produce exclusively electricity, but often operate in the cogeneration of heat and electricity, we must also take into account the shares of cogeneration for each source. The shares of cogeneration are taken from the study by Topić Božič et al. (2025) and are derived from the Ecoinvent database [2].

To assess the environmental impacts, we used the ReCiPe 2016 Endpoint method, which enables the evaluation of impacts on human health, ecosystem quality and the availability of natural resources. Based on the results of the analysis of individual production sources, expressed per functional unit of 1 kWh, we designed various electricity production scenarios for Slovenia. The environmental impacts of the individual scenarios were compared and critically evaluated. The environmental indicators that were used to assess the life cycle impacts on the environment are selected according to Hojkar (2021) [9].

3 RESULTS AND DISCUSSION

3.1 Environmental impacts of different scenarios for producing 1 kWh of electricity

To assess the environmental impacts associated with the production of 1 kWh, we conducted a comparative analysis of four previously described scenarios with differently defined electricity generation mix structures (Table 1).

Table 1: Environmental impacts of different scenarios for the production of 1 kWh of electricity

Environmental impact category	Current production structure	Scenario without TEŠ	100% RES	Nuclear renewable scenario
Global warming (human health)	3.24×10^{-8}	5.53×10^{-9}	1.19×10^{-8}	4.38×10^{-9}
Global warming (terrestial ecosystem).	9.78×10^{-11}	1.67×10^{-11}	3.59×10^{-11}	1.32×10^{-11}
Global warming – freshwater ecosystem)	2.67×10^{-15}	4.56×10^{-16}	9.80×10^{-16}	3.61×10^{-16}
Stratospheric ozone depletion	1.73×10^{-12}	1.17×10^{-12}	3.97×10^{-12}	1.40×10^{-12}
Ionizing radiation	2.56×10^{-10}	3.70×10^{-10}	2.04×10^{-11}	4.99×10^{-10}
Photochemical ozone formation (human health)	7.67×10^{-11}	1.15×10^{-11}	3.82×10^{-11}	1.44×10^{-11}
Formation of fine solid particles	1.17×10^{-7}	6.56×10^{-9}	2.32×10^{-8}	9.65×10^{-9}
Photochemical ozone formation (terrestial ecosystem)	1.09×10^{-11}	1.72×10^{-12}	5.62×10^{-12}	2.11×10^{-12}
Soil acidification	1.26×10^{-10}	5.86×10^{-12}	2.01×10^{-11}	8.13×10^{-12}
Freshwater eutrophication	4.68×10^{-11}	2.22×10^{-12}	7.59×10^{-12}	3.16×10^{-12}
Marine eutrophication	8.13×10^{-15}	1.68×10^{-15}	1.39×10^{-15}	2.31×10^{-15}
Soil toxicity	3.32×10^{-13}	3.32×10^{-13}	3.53×10^{-12}	1.08×10^{-12}
Inland water toxicity	1.33×10^{-12}	2.11×10^{-13}	3.10×10^{-12}	8.66×10^{-13}
Marine water toxicity	2.76×10^{-13}	4.28×10^{-14}	5.98×10^{-13}	1.69×10^{-13}
Human carcinogenic toxicity	1.25×10^{-8}	1.80×10^{-9}	5.96×10^{-9}	2.81×10^{-9}
Non-carcinogenic toxicity	1.75×10^{-8}	2.36×10^{-9}	1.02×10^{-8}	4.55×10^{-9}
Land use	3.39×10^{-12}	3.86×10^{-12}	2.49×10^{-11}	7.30×10^{-12}
Mineral resource depletion	9.86×10^{-6}	1.32×10^{-5}	3.64×10^{-5}	2.31×10^{-5}
Fossil resource depletion	4.02×10^{-4}	5.36×10^{-4}	6.64×10^{-4}	2.34×10^{-4}
Water consumption (human health)	4.87×10^{-10}	4.44×10^{-10}	4.06×10^{-10}	6.34×10^{-10}
Water consumption (terrestial ecosystem)	2.90×10^{-12}	2.73×10^{-12}	2.64×10^{-12}	3.91×10^{-12}
Water consumption (water ecosystem)	1.31×10^{-16}	1.34×10^{-16}	1.78×10^{-16}	2.01×10^{-16}

Analysis of modeled scenarios shows that in the scenario that does not include TEŠ, indicators related to health and air quality and related to combustion and emissions in the air are up to 21.5 times lower than in the current scenario. At the same time, the aforementioned scenario also achieves the lowest values of carcinogenic and non-carcinogenic toxicity to humans and 6 times lower values of toxicity of terrestrial and marine waters. On the other hand, the impact in the ionizing radiation category increases by 1 time, probably due to the larger share of nuclear energy. Also in the article by Šerešová et al. (2020) conducted a similar study, where they determined the impacts of different electricity production sources in the Czech Republic [10]. There, fossil brown coal and lignite account for as much as 47% of electricity production, and the authors state that they are the most important contributors to air quality and have a major impact on global warming and the formation of fine particulate matter [10]. They also state in the aforementioned study that renewable energy sources (solar energy, wind energy and hydropower) and nuclear energy have a significantly smaller impact on climate change, which is consistent with our findings, since in the scenario without TEŠ, hydroelectric power plants (39.87%) and nuclear energy (49.28%) represent by far the largest share.

In the scenario that assumes a 100% share of RES, a significant decrease in impact is visible in the category of reduction of ionizing radiation, reduction of global warming, and in some water indicators. On the other hand, impacts appear to be higher in the categories of land use, mineral resources and several toxicity indicators, indicating the need for circular materials, recycling and cleaner supply chains.

Furthermore, the impacts in the stratospheric ozone depletion category are the highest of all scenarios, which could be related to background material or process emissions (e.g. industrial chemicals, refrigerants, equipment manufacturing processes). This is consistent with the findings of Hertwich et al. (2015), who found that a shift to RES (wind and solar energy) reduces environmental impacts resulting from direct fossil fuel combustion, but may increase impacts related to equipment manufacturing, mining and material processing due to higher material requirements [11]. 100% RES reduces emissions-related impacts, but also requires active management of material flows (recycling, technology selection, responsible supply chains) to limit increases in land use, mineral resources and toxicities.

In the nuclear renewable scenario, the impacts related to air quality are significantly lower than in the current scenario and show the lowest climate footprint per kWh of all four scenarios. On the other hand, higher impacts are observed in the categories of mineral resource depletion, water consumption and ionizing radiation. In their study, Warner and Heath (2012) state that greenhouse gas emissions from nuclear power plants are much lower than from traditional fossil sources and are often comparable to RES, which is consistent with our results [12]. The

nuclear renewable scenario has the lowest values for all three endpoint indicators of global warming and significantly lower values than the current structure. At the same time, the authors warn that the conditions and assumptions under which nuclear energy is introduced (electricity for supply chains, uranium enrichment method, etc.) can significantly affect greenhouse gas emissions.

If we compare the alternative scenarios, the scenario without TEŠ has 10 to 15 times lower impacts on several toxicity indicators than the scenario with a 100% share of RES (toxicity of terrestrial and marine waters and soil), 6.5 times lower values of the land use indicator and 4.3 times lower values of the non-carcinogenic toxicity indicator for humans compared to the 100% share of RES. On the other hand, it has a significantly higher impact in the ionizing radiation category. Also compared to the nuclear renewable scenario, the scenario without TEŠ has the greatest impact on reducing the values of several toxicity indicators. On the other hand, in the category of fossil resource depletion, it achieves 2.3 times higher values than the nuclear renewable scenario and has a 1.3 times higher impact on global warming indicators. If we compare the scenario with a 100% share of RES and the nuclear renewable scenario, the former achieves 24.5 times lower values in the category of ionizing radiation and 1.1 to 1.7 times lower values of indicators of marine eutrophication and water consumption. In contrast, with a 100% share of RES, approximately 3.5 times higher impacts of indicators of toxicity of terrestrial and marine waters and soil and land use indicators are observed, 2.8 times higher impacts on fossil resource depletion and stratospheric ozone depletion and 2.7 times higher impacts on global warming indicators than with the nuclear renewable scenario. As can be seen from the results, the differences between the scenarios for some environmental indicators are very large, which means that the choice of technologies and fuels strongly influences the profile of environmental impacts.

3.2 Assessment of environmental impacts at the level of endpoint indicators

3.2.1 Human health impact (DALY)

Our study shows that the current production structure represents by far the largest impact on health (total DALY ≈ 1.80×10^{-7}). The other three alternative scenarios represent a significantly lower total impact on health, namely the scenario without TEŠ with a 90% lower impact, the nuclear renewable scenario with an 87.5% lower impact and the 100% share of RES with a 71% lower impact on health. The indicators that contribute most to the total impact on human health are the formation of fine particulate matter and global warming – human health.

In the current scenario, the formation of fine particulate matter represents 65% of the total impact, followed by global warming – human health with approximately 18%, and the remainder is mostly represented by carcinogenic and non-carcinogenic toxicity to humans.

The scenario without TEŠ shows the lowest impact in this category. The largest impact within the scenario is represented by the formation of fine particulate matter (38%), global warming (32%) and non-carcinogenic toxicity (14%). The formation of fine particulate matter is 18 times lower than in the current scenario, as is the impact of carcinogenic and non-carcinogenic toxicity and photochemical ozone formation by approximately 7 times lower, and the impact in the global warming category by 6 times lower. In contrast, the ionizing radiation indicator shows values 1.5 times higher than in the current scenario.

In the nuclear renewable scenario, the largest impacts are particulate matter (43%), non-carcinogenic toxicity (20%) and global warming (19%). Compared to the current scenario, the nuclear renewable scenario has a lower impact in the categories of particulate matter (12 times lower) and global warming (7 times lower). On the other hand, it shows up to 1.3 times higher values in the water consumption and 2 times higher values of ionizing radiation.

The scenario with a 100% share of RES still shows lower impacts in this category than the current scenario, but the impacts are higher than in the other two alternative scenarios. The largest impacts at 100% RES share are the formation of fine particulate matter (45%), global warming (23%) and non-carcinogenic toxicity (20%). On the other hand, the impacts in the ionizing radiation category reach the lowest values here, namely 12.6 times lower than in the current scenario, but in our study this represents an almost negligible contribution of this segment.

Similar findings are reported in the UNECE study (2021), namely that more than 75% of the endpoint indicator of damage to human health is represented by climate change, followed by human toxicity and emissions of fine particles [13]. Human toxicity is assessed using two indicators (non-carcinogenic and carcinogenic effects). Of these, coal-fired electricity generation has the highest values for non-carcinogenic effects, mainly due to arsenic, and solar energy has the next highest average. However, the carcinogenic effects remain relatively low and are most pronounced in concentrated solar power plants.

3.2.2 Ecosystem quality (species·yr)

Our study shows that the current production structure represents by far the largest impact on ecosystems (total species·yr ≈ 2.9×10^{-10}). The other three alternative scenarios represent a significantly lower total impact on ecosystems, namely the scenario without TEŠ for 8.6 times lower impact, the nuclear renewable scenario for 7.3

times lower and the 100% share of RES for 2.8 times lower impact on ecosystems. The indicators that contribute most to the impact on ecosystems are soil acidification, global warming – terrestrial ecosystems and eutrophication of freshwaters.

In the current scenario, soil acidification represents as much as 43% of the total impact, followed by global warming – terrestrial ecosystems with 34% and eutrophication of freshwaters with 16%. When emission-intensive impacts (acidification and eutrophication) are reduced, the global warming indicator stands out relatively more in alternative scenarios – terrestrial ecosystems, which often represent the largest individual segment, followed by land use (especially at 100% RES share) and to a lesser extent water consumption and other contributions.

In the scenario without TEŠ, lower values can be observed in the categories of land acidification and freshwater eutrophication, which are 21 times lower than the current scenario, and the formation of photochemical ozone is 6 times lower. Similar results are achieved by the nuclear renewable scenario, which has the lowest global warming values of all the analyzed scenarios. Compared to the scenario without TEŠ, it has slightly higher values of the land use indicators (1.9 times higher) and water consumption (1.4 times higher). The 100% RES scenario has 10 times higher values in the soil toxicity category and 7.3 times higher values in the land use indicator compared to the current scenario. It has also higher values in freshwater acidification and eutrophication than both alternative scenarios. UNECE (2021) states that climate change is by far the largest contributor to impacts on ecosystems [13]. This is followed by land use categories, especially land conversion, which occurs in fossil fuel extraction and hydropower. They also state that coal-based technologies cause very high levels of eutrophication compared to other electricity sources, which is consistent with our results [13].

3.2.3 Availability of natural resources (USD2013)

According to the results, the scenario with a 100% share of RES has the worst values of the indicators, while the nuclear renewable scenario has the best values. The nuclear renewable scenario has the lowest values of fossil resource depletion, namely 42% lower than the current scenario. On the other hand, mineral resource depletion is 2.3 times higher than the current scenario. This scenario relieves the most of the pressure on fossil resources, but requires more mineral raw materials.

The current production structure has the lowest values in the category of mineral resource depletion compared to the other analyzed scenarios. The scenario without TEŠ has 1.3 times higher values in the category of impact of fossil and mineral resource depletion than the current scenario.

The scenario with a 100% share of RES shows the highest values in the categories of fossil resource depletion (1.7 times higher than the current scenario) and mineral resource depletion (3.7 times higher than the current scenario). The aforementioned results are probably due to the transition to infrastructure-intensive technologies that require more minerals, and there is also a noticeable indirect fossil input in the supply chain. In the scenario, we assumed that solar energy represents 55% of electricity production. Our results could therefore be related to the findings of the study by Šerešová et al. (2020), in which they state that the largest share (more than 60%) of fossil resource depletion in solar energy is due to the production of materials used in the construction of these production units, and that solar energy has by far the highest values of mineral resource and metal depletion compared to other RES, fossil brown coal and lignite (in their case up to 15 times higher), mainly for the construction of solar panels [10].

4 CONCLUSION

In the study, we used the life cycle analysis (LCA) method and ReCiPe 2016 indicators to evaluate the environmental impacts of electricity production in Slovenia for a functional unit of 1 kWh and compared four scenarios with a differently defined structure of the electricity production mix. As can be seen from the results, there is no technology or scenario that would not have an impact on environmental indicators. The LCA method therefore reliably shows us in which category an individual scenario has a higher or lower impact, but it does not provide an answer as to whether the result is “good enough” and which scenario is the absolute best or worst.

The results show that the current structure of electricity production is the most impacted in most environmental indicators. The total impact of the current scenario is highest in the categories of impact on human health and impact on ecosystem quality. The scenario analysis clearly shows that scenarios with a reduced or eliminated share of coal (without TEŠ) are generally more favorable in categories directly related to combustion emissions and air quality. In the scenario without TEŠ, indicators related to human health and ecosystems are several times lower than in the current structure, while the impact of global warming is also lower compared to the current scenario. At the level of endpoint indicators, this is reflected in a significantly lower total impact on human health and ecosystem quality in all alternative scenarios, especially in the scenario without TEŠ and the nuclear renewable scenario.

The results also reveal key trade-offs within individual scenarios. On the one hand, the nuclear renewable scenario achieves very low values for global warming indicators and the lowest for fossil fuel depletion, indicating a strong potential for decarbonization and reducing dependence on fossil fuels. However, this scenario presents a trade-off in other categories: the impact of the ionizing radiation indicator is 2 times higher than in the current scenario and

24.5 times higher than in the 100% RES scenario, which is consistent with shifting part of the impact to specific categories of the nuclear fuel cycle.

The most pronounced shift in impacts is seen in the 100% RES scenario, where some emission indicators and ionizing radiation are reduced, but material and spatial impacts are increased. For example, in this scenario, the impact in the category of mineral resource depletion is the highest, and the values of the land use indicator also increase significantly compared to the current scenario. This result suggests that technologies such as solar energy contribute to lower greenhouse gas emissions, but at the same time increase pressures on supply chains, mineral resource use and space. This is an important starting point for a holistic assessment, as focusing on only one category could lead to overlooking other, equally important consequences for society and the environment.

The final choice of energy strategy cannot therefore be based only on the results of the LCA, but must also take into account technological limitations, the availability of natural resources and social acceptability. Technologically, high shares of RES must take into account the volatility of production, the need for energy storage, grid adjustments and system services, and in the case of the nuclear option, the time frames for spatial placement, radioactive waste management issues and safety requirements. From the perspective of natural resources, mineral resource limitations, material intensity of technologies and the possibility of recycling in the life cycle become important, while in terms of social acceptability, in addition to nuclear issues, the spatial placement of RES, visual impacts, land use and local conflicts of interest also become important. Therefore, we understand the results as a basis for comparing environmental profiles and as a starting point for multi-criteria decision-making, where in addition to environmental criteria, economic, political and social criteria are also included.

To upgrade the research, it would be reasonable to conduct additional sensitivity analyses (e.g. regarding database assumptions, technology lifetime, recycling share and import structure), to include system elements (storage facilities, transmission or distribution network, demand flexibility) and to address the dynamic development of the electricity system over time. This would allow us to better assess how environmental impacts change under realistic technological and social constraints and how robust the findings are across different development paths.

REFERENCES

[1] LE-BOULCH, Denis idr. Life cycle assessment of nuclear power in France: EDF case study. *EPJ Nuclear Sciences & Technologies*. 2024, **vol. 10**, (8). ISSN 2491-9292.
[2] TOPIĆ BOŽIČ, Jelena, Ante ČIKIĆ in Simon MUHIČ. Environmental Impact of Slovenian and Croatian Electricity Generation Using an Hourly Production-Based Dynamic Life Cycle Assessment Approach. *Energies*. 2025, **vol. 18,** (18), 4826. ISSN 1996-1073.
[3] AGENCIJA ZA ENERGIJO. *Poročilo o stanju na področju energetike v Sloveniji v letu 2024* [na spletu]. 2025 [dostopano 15. 8. 2025]. Dostopno: https://www.agen-rs.si/documents/10926/38704/Poro%C4%8Dilo-stanju-napodro%C4%8Dju-energetike-za-leto-2024_kon%C4%8Dna/b32e79e6-821b-473e-9933-e99c354c0f04
[4] KOVAČ, Gregor. *Obnovljivi viri energije in njihov vpliv na okolje v Sloveniji*. Ljubljana: Fakulteta za strojništvo, 2018.
[5] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 14040:2006: Environmental management – Life cycle assessment – Principles and framework. Geneva: ISO, 2006.
[6] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 14044:2006: Environmental management – Life cycle assessment – Requirements and guidelines. Geneva: ISO, 2006.
[7] RAMPIH, Darian. Dynamic life cycle assessment of Slovenian electricity generation: a comparative approach towards a sustainable energy mix. Nova Gorica: Univerza v Novi Gorici, 2025.
[8] DIMNIK, Jože idr. Impacts of High PV Penetration on Slovenia's Electricity Grid: Energy Modeling and Life Cycle Assessment. *Energies*. 2024, **vol. 17**, (13), 3170. ISSN 1996-1073.
[9] HOJKAR, Domen. *Okoljski vplivi različnih scenarijev proizvodnje električne energije v Sloveniji*. Ljubljana: Fakulteta za strojništvo, 2021.
[10] ŠEREŠOVÁ, Markéta idr. Life Cycle Performance of Various Energy Sources Used in the Czech Republic. *Energies*. 2020, **vol. 13,** (21), 5833. ISSN 1996-1073.
[11] HERTWICH, Edgar G. idr. Integrated life-cycle assessment of electricity-supply scenarios confirms global environmental benefit of low-carbon technologies. *Proceedings of the National Academy of Sciences of the United States of America*. 2015, **vol. 112,** (20), ISSN 6277–6282.
[12] WARNER, Ethan S. in Garvin A. HEATH. Life Cycle Greenhouse Gas Emissions of Nuclear Electricity Generation: Systematic Review and Harmonization. *Journal of Industrial Ecology*. 2012, **vol. 16,** (S1), S73–S92. ISSN 1530-9290.
[13] UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE). *Life Cycle Assessment of Electricity Generation Options*. Geneva: UNECE, 2021. [online]. [acessed 30.01.2026]. https://unece.org/sites/default/files/2021-11/LCA_final.pdf

Hourly Consumption-Based Electricity Mixes for Slovenia, Croatia, and Serbia: A Time-Resolved Life Cycle Assessment Approach

Jelena Topić Božič[1,2], Andreja Dobrovoljc[1], Simon Muhič[1,2,3]

[1]*Rudolfovo – Science and Technology Centre Novo mesto, Slovenia*
E-mail: jelena.topic.bozic@rudolfovo.eu

[2]*Faculty of Industrial Engineering Novo mesto, Slovenia*

[3]*Institute for Renewable Energy and Efficient Exergy Use, INOVEKS*

Abstract: *In this study, a time-resolved, consumption-based life cycle assessment framework was used to evaluate the environmental impacts associated with producing 1 kWh of electricity in Slovenia, Croatia, and Serbia. The model combines hourly electricity generation and cross-border trade data with life-cycle inventory datasets to construct dynamic, country-specific consumption mixes that account for electricity imports and exports. Environmental impacts were quantified using the ReCiPe 2016 Midpoint (H) impact assessment method, enabling characterization of temporal variability in key impact categories, including climate change, and mineral resource scarcity. The approach reveals how seasonality and diurnal variations in intermittent renewable energy sources, as well as the absence of photovoltaic generation during night-time off-peak periods, shape the environmental impact of electricity. Such an approach can be used to design charging and scheduling strategies and to support policy development, particularly where existing strategies rely on static or purely demand-driven time-of-use tariffs that may not reflect the actual environmental impacts of electricity consumption in systems with a significant share of intermittent renewables. The approach supports scenario analyses for energy transition planning and provides more accurate background data for life-cycle assessment studies of electrified technologies in the region.*

Keywords*: dynamic life cycle assessment, environmental impact, electricity, temporal variation, consumption-based approach*

Article Classification*: Scientific Paper*

1 INTRODUCTION

Electricity generation is a source of a significant amount of greenhouse gas (GHG) emissions [1]. Furthermore, electricity generation contributes significantly to other environmental impacts, such as resource use of minerals and metal, eutrophication potential and acidification [2, 3]. The energy sector is the major contributor to GHG emissions, accounting for more than 70% of GHG emissions and the transformation of energy sector is seen as crucial towards achieving the goal of climate neutrality [4]. Due to increase in electrification and with the expansion of the use of electronic devices and information technologies, the importance of electricity and its associated impacts is increasing. Decarbonization of the electricity sector is seen as a crucial measure towards achieving low-carbon future, as currently almost two-thirds of global electricity productions still relies on fossil fuels, primarily coal and natural gas [3].

Electricity use contributes significantly to the environmental impacts of products, and it is often the factor that influences the most the environmental impact of a product system. This particularly applies to electricity-powered systems such as using heat pumps or battery electric vehicles (BEV) [1]. Finding from Cox et al. [5] show that the environmental benefits of switching to BEV depend on the decarbonization of electricity sector. The majority of such studies use yearly averages of electricity impact; however, such approach does not allow to account for daily and seasonal variations in electricity production and resulting environmental impacts [3, 6]. A dynamic approach to assess the environmental impact of BEV was applied in Belgium. The results showed that BEV charging during off-peak hours (from midnight to 8 a.m.), resulted in the lowest emissions due to energy demands primarily met through nuclear power plants, renewable energy sources (RES) and more efficient natural gas plants [7]. Life cycle assessment (LCA) studies have identified the effects of national electricity mixes under various yearly average scenarios and generally scenarios with higher shares of RES have lower impact in category of global warming [8].

The electricity mixes vary significantly across countries and use diverse production technologies, resulting in the variation of the in hourly emissions, which depend on the combination of technologies used in the electricity generation process. Technologies have different emissions factors, indicating that potential for large scale variations in emissions generated from the electricity is high [4]. Dynamic LCA studies of the electricity production are still limited, however the information is relevant for policy development and managing demands as shifting the timing of

electricity use to avoid period where emissions are higher by applying smart grid technologies has been identified as potential approach to decrease GHG emissions [9, 10].

Recently Naumann et al. [6] developed a model to determine European electricity generation's time-varying environmental impacts using a dynamic LCA approach. Dynamic LCA was carried out for five largest electricity consumers in the European Union (EU): Germany (DE), France (FR), Italy (IT), Spain (ES), and Poland (PL). The results showed that the impacts of the different environmental categories vary considerably depending on the time of electricity generation. As Poland's electricity generation is primarily based on fossil fuels and the fluctuating feed-in of renewable energy sources plays a minor role, the environmental impact of Polish electricity generation showed the slightest temporal variation. Within the framework of the study two different concepts to model the dynamic power generation structures were followed a production-based approach and a consumption-based approach. The production-based model considers all local electricity producers within the country's borders. The analysis showed that the results of the consumption- based and production-based modelling approaches differed significantly for most countries. An average deviation in environmental impacts was found to be around 15 %, with the difference is most significant for Italy. The results showcase the need that in countries with high electricity imports and exports, LCA studies should not only be based on the environmental impacts of domestic electricity production but should also consider cross-border electricity flows [6]. Recently Topić Božič et al. [3] adopted simplified Naumann's methodology to determine the environmental impact of Slovenian and Croatian electricity production using only production-based approach. The results showed that the maximum climate change value was 0.47 kg CO2 eq. for Slovenia at a time of low RES production. A similar pattern was observed in Croatia. The maximum value observed was 0.51 kg CO_2 eq. The RSD of the climate change category was in line with the findings from the literature, with 21.6% for Slovenia and 18.6% for Croatia. Additionally, Slovenia's generation mix had a higher impact on the resource use, minerals and metals category [3].

The aim of this study is to develop hourly, consumption-based environmental impact profiles of electricity for Slovenia, Croatia, and Serbia by integrating time-resolved dynamic LCA with modelling of cross-border imports and exports over 2021–2025, to account how trade and difference in national generation mixes shape the environmental footprint of electricity in Southeast Europe. The analysis considers the different production structures of the three systems—Slovenia's nuclear- and hydro-based mix, Croatia's gas- and hydro-dominated portfolio with growing variable renewables, and Serbia's lignite- and hydro-intensive generation—to quantify how these diverse profiles, translate into distinct hourly consumption-based impact patterns for each country. By providing high-resolution, consumption-based impact data for three power systems, this study information for more accurate environmental accounting and for supporting low-carbon technology deployment and demand-side management strategies in Southeast Europe.

2 METHODS

The methodology followed the approach developed and recently published by Naumann et al. [6] and adopted by Topić Božič et al. [3]. They used time-resolved data on an hourly basis for a dynamic retrospective approach to the LCA of electricity generation [3, 6]. A schematic representation of the developed dataset is illustrated in the Figure 1.

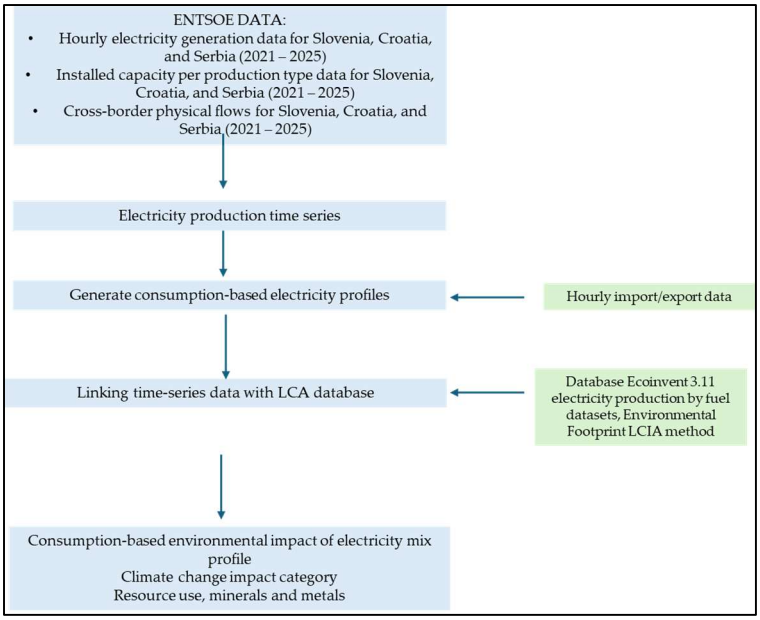


Figure 1: Schematic representation of the methodological steps.

2.1 Data collection

The data were downloaded from the publicly available ENTSO-E Transparency platform, which provides information on generation, load, transmission, and balancing for European countries [11]. Data were collected for the 2021–2025 period for Slovenia, Croatia, and Serbia. The datasets used to generate the time-resolved profiles are:

- Hourly actual generation per production type
- Installed capacity per production type in annual resolution
- Cross-border physical flows in hourly resolution (imports and exports between Slovenia, Croatia, Serbia, and neighbouring bidding zones)

For each country electricity production time series P are first calculated for each production type j for each hour i according to Naumann et al. [6]. Briefly, for each production type, the utilization rate for each hour was calculated by dividing the electricity production by the installed capacity C in the year. Subsequently, the 5-year averages of the hourly utilization rates for 2021 to 2025 are calculated. These are then multiplied by the installed capacity of each technology in 2024 to account for annual weather variations and reduce the influence of year-specific events. For the consumption-based models, the country-specific electricity production time series are modified by imports and exports occurring with trading countries. These imports and exports are averaged over five years. The sum of all exports was subtracted proportionally from all electricity producers and imports were added as additional columns to the production-based electricity matrix [6].

In the case of missing data in the ENTSO-E datasets, the gaps were closed using the average values of the values that occurred directly before and after, similar to Naumann et al., where data gaps were closed by taking the values occurring directly before. Utilization rates averaged over five years also mitigated the impact of data gaps [3, 6]. For the Croatian 2024 electricity generation profile, data were available in 15-min intervals. The data were aggregated, averaged, and shown hourly to align with other available datasets as was done in [3]. In the case of gap year (2024), the data for 29 February were removed to calculate the five-year average time series profile as was done in [3].

The primary aim of this study is to characterize and compare the hourly, consumption-based environmental impacts of electricity use in Slovenia, Croatia, and Serbia. In contrast to previous work focused solely on domestic generation technologies and their shares in national production mixes [3], the present analysis explicitly incorporates imports, exports, and cross-border physical flows between the three countries and neighbouring bidding zones. Losses associated with voltage level transformations (high, medium, low voltage) are not modelled explicitly and are therefore excluded from the system boundaries.

2.2 Life cycle assessment of electricity generation

The time-series electricity generation datasets developed in Section 2.1 were linked to the corresponding background processes in the ecoinvent 3.11 database, using the “allocation, cut-off by classification” system model, consistent with previous dynamic LCA applications to European power systems [3, 6, 8]. The Environmental Footprint 3.1 (EF 3.1) life cycle impact assessment (LCIA) method was applied as the primary LCIA method [12], ensuring methodological consistency with earlier studies on time-resolved electricity impacts and recent applications of EF 3.1 to European electricity mixes [3].

To construct the hourly environmental impact matrix, a technology-by-impact-category matrix was assembled from the selected ecoinvent datasets and then combined with the hourly electricity generation time series. The modelling followed the principle of using the most geographically specific datasets available in the Ecoinvent database [13]. Country-specific datasets were used for Slovenia (SI), Croatia (HR), and Serbia (RS) whenever available. When no country-specific dataset existed for a given technology, European (RER/Europe) datasets were selected. Global or Rest-of-World (GLO/RoW) datasets were only employed when neither country-specific nor European datasets were available (e.g. for certain wind technologies, geothermal electricity, or specific waste-to-energy processes).

For imported electricity, country-specific import or market datasets were used whenever possible (e.g. “electricity, high voltage, imported from AT to SI” or analogous processes for HR and RS); if such specific trade datasets were not available, the corresponding “market for electricity, high voltage” processes were applied as proxies for the exporting countries.

The functional unit used for the assessment was 1 kWh of electricity produced. It is a functional unit commonly used in LCA studies of electricity mixes [3, 6, 8].

3 RESULTS AND DISCUSSION

3.1 Average hourly electricity generation profiles for the 2021 – 2025

The average electricity generation profiles, representing 8760-h averages for Slovenia, Croatia, and Serbia in the developed dataset, are shown as a function of time in Figure 2. In Slovenia and Croatia, electricity production exhibits pronounced daily and seasonal fluctuations, due to higher share of RES in line with previous studies that reported strong temporal variability in hourly generation profiles for European countries such as Germany [6] and recently published study by Topić Božič et. al. [3] Serbia on the other hand, exhibits significant share of coal-based production which dominates the profile.

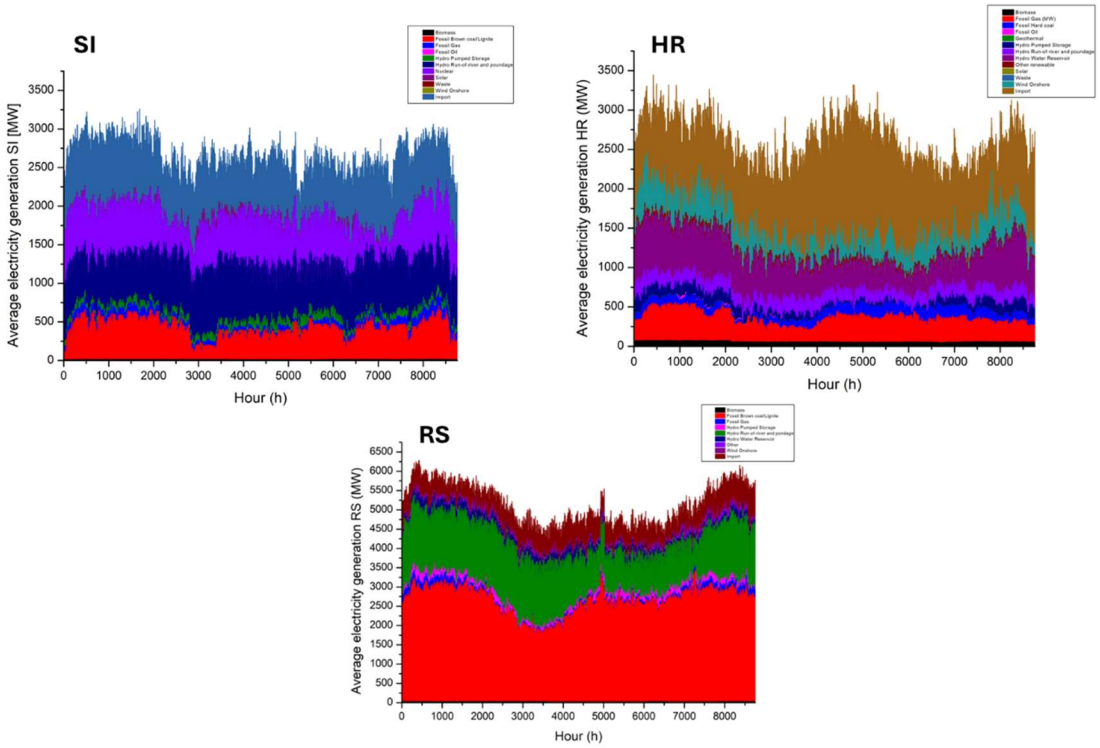


Figure 2: Average hourly electricity generation composition profile in Slovenia, Croatia and Serbia for the 2021–2025 period. SI, HR, and RS correspond to Slovenia, Croatia, and Serbia, respectively.

The generation profiles of Slovenia, Croatia, and Serbia differ markedly from each other. Slovenia's mix is dominated by nuclear and hydropower (mainly run-of-river), with coal (lignite) providing most of the remaining fossil-based generation. Croatia relies predominantly on hydropower, complemented by natural gas and a growing share of variable renewables [14–16] . Serbia's generation is largely based on low-quality lignite coal, supplemented by hydropower and a small but increasing contribution from gas and wind power [16]. In general, the shares of fossil-based and nuclear generation in the developed hourly profiles are consistent with annual statistics, whereas larger discrepancies are observed for weather-dependent renewables such as wind and PV, which are subject to substantial daily and seasonal variability as previously reported by similar studies [3, 6] Both Slovenia and Croatia import significant share of electricity, which is visible from Figure 2. Serbia on the other hand, imports lower share of electricity compared to Slovenia and Croatia.

3.2 Hourly environmental impact of electricity mix

Figure 3 shows hourly variation in the climate change category for Slovenia, Croatia and Serbia for consumption-based approach. The fluctuation is similar to the production-based approach shown in Topić Božič et al. for the 2020-2024 period for Slovenia and Croatia [3]. The lower values in the impact category can be correlated with a higher share of RES in the electricity mix. The results are consistent with the available data for other countries, such as Germany, Poland, and Italy [6]. Slovenian electricity mix had on average the lower climate change value with 0.25 kg CO_2 eq. Croatia and Serbia averages were 48.5% and 133.8% higher, respectively. Compared to the production-based approach, where the average values in climate change category for 2020 – 2024 period were for Slovenia and Croatia 0.33 and 0.32 kg CO_2 eq., respectively, a decrease due to the shares of import can be seen in Slovenian electricity mix. On the other hand, the electricity mix in consumption-based approaches had 24.7% higher average value in climate change impact category. The results are in accordance with data published from

Naumann et al. [6], which showed that results of the consumption-based and production-based modelling approaches differ significantly for most countries analyzed with an average deviation among environmental impacts of around 15% [6].

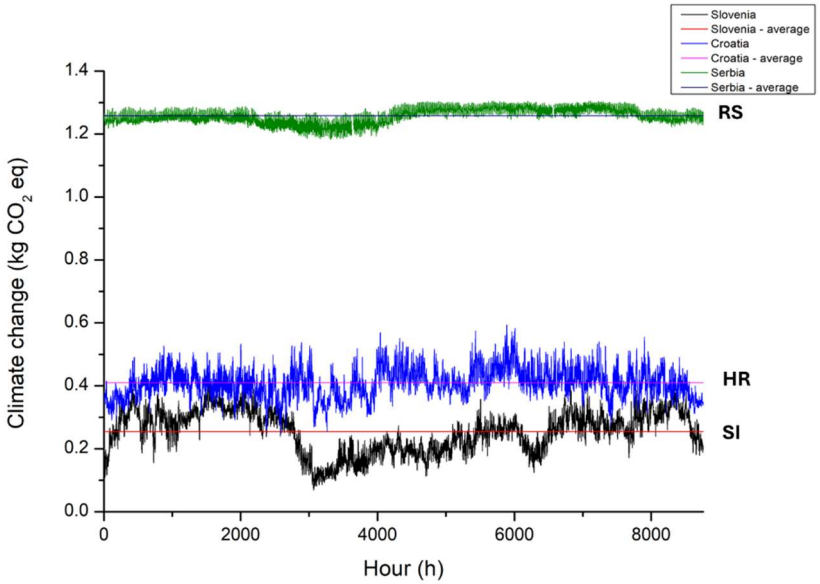


Figure 3: Climate change impact of 1 kwh of produced electricity for Slovenia, Croatia and Serbia. SI, HR, and RS correspond to Slovenia, Croatia, and Serbia, respectively.

In Serbia's case, which relies primarily on coal, the temporal variation vas the slightest as the fluctuating variable RES play a minor role. Similarly in the case of Poland, where electricity generation is primarily based on fossil fuels, the fluctuation was the lowest [6].

The results showcase the importance of applying time-resolved approach which considers the impact of electricity imports in LCA studies of products system with temporally varying electricity use, such as electricity heating systems, BEV charging, data centres, or flexible industrial loads [6].

4 CONCLUSIONS

The hourly consumption-based climate change results show that Slovenia has, on average, the lowest impact (0.25 kg CO_2 eq.), while Croatia and Serbia exhibit 48.5% and 133.8% higher average values, respectively. Compared with the production-based results for 2020–2024 (0.33 kg CO_2 eq. for Slovenia and 0.32 kg CO_2 eq. for Croatia), the consumption-based approach reveals a reduction in the Slovenian average climate change intensity due to lower-carbon imports, and a 24.7% increase for Croatia, where imported electricity tends to be more carbon-intensive. In Serbia, where coal dominates and variable RES play a minor role, the temporal variation in the climate change category is noticeably lower.

Accounting for both the timing and origin of electricity supply reveals temporal "windows" of lower and higher environmental intensity and highlights how cross-border exchanges shape the use-phase impacts of electricity-consuming technologies. The developed hourly, consumption-based profiles for Slovenia, Croatia, and Serbia provide a basis for more accurate environmental accounting and for assessing the effectiveness of demand-side management and load-shifting strategies in Southeast Europe. Since the present profiles are derived from empirical historical data, they primarily support retrospective analyses; future work could extend the framework with prospective scenarios and dynamic marginal mixes, thereby enabling ex-ante and consequential LCA applications for low-carbon technology deployment and regional electricity system planning.

Acknowledgments

The research was co-funded by the Slovenian Research and Innovation Agency (ARIS) through the annual work program of Rudolfovo and ARIS project Innovative strategies for the valorization of grape pomace: anthocyanin extraction, pyranoanthocyanin synthesis, and energy use with life cycle analysis (LCA) (GRAPEVALOR), project number: J1-70024.

REFERENCES

[1] MILLER, Gregory J., Kevin NOVAN in Alan JENN. Hourly accounting of carbon emissions from electricity consumption. *Environmental Research Letters* [online]. 2022, **17**(4), 044073 [acessed 13. 04. 2026]. ISSN 1748-9326. Dostopno na:10.1088/1748-9326/AC6147

[2] NAUMANN, Gabriel idr. Dynamic life cycle assessment of European electricity generation based on a retrospective approach. *Energy Conversion and Management* [online]. 2024, **311**, 118520 [acessed 13. 04. 2026]. ISSN 0196-8904. Dostopno na:10.1016/J.ENCONMAN.2024.118520

[3] BOŽIČ, Jelena Topić, Ante ČIKIĆ in Simon MUHIČ. Environmental Impact of Slovenian and Croatian Electricity Generation Using an Hourly Production-Based Dynamic Life Cycle Assessment Approach. *Energies 2025, Vol. 18, Page 4826* [online]. 2025, **18**(18), 4826 [acessed 13. 04. 2026]. ISSN 1996-1073. Dostopno na:10.3390/EN18184826

[4] GUIDI, Giambattista, Anna Carmela VIOLANTE in Simona DE IULIIS. Environmental Impact of Electricity Generation Technologies: A Comparison between Conventional, Nuclear, and Renewable Technologies. *Energies 2023, Vol. 16, Page 7847* [online]. 2023, **16**(23), 7847 [acessed 13. 04. 2026]. ISSN 1996-1073. Dostopno na:10.3390/EN16237847

[5] COX, Brian idr. Life cycle environmental and cost comparison of current and future passenger cars under different energy scenarios. *Applied Energy* [online]. 2020, **269**, 115021 [acessed 13. 04. 2026]. ISSN 0306-2619. Dostopno na:10.1016/J.APENERGY.2020.115021

[6] NAUMANN, Gabriel idr. Dynamic life cycle assessment of European electricity generation based on a retrospective approach. *Energy Conversion and Management* [online]. 2024, **311**, 118520 [acessed 13. 04. 2026]. ISSN 0196-8904. Dostopno na:10.1016/J.ENCONMAN.2024.118520

[7] RANGARAJU, Surendraprabu idr. Impacts of electricity mix, charging profile, and driving behavior on the emissions performance of battery electric vehicles: A Belgian case study. *Applied Energy* [online]. 2015, **148**, 496–505 [acessed 13. 04. 2026]. ISSN 0306-2619. Dostopno na:10.1016/J.APENERGY.2015.01.121

[8] DIMNIK, Jože idr. Impacts of High PV Penetration on Slovenia's Electricity Grid: Energy Modeling and Life Cycle Assessment. *Energies 2024, Vol. 17, Page 3170* [online]. 2024, **17**(13), 3170 [acessed 13. 04. 2026]. ISSN 1996-1073. Dostopno na:10.3390/EN17133170

[9] SECKINGER, Nils in Peter RADGEN. Dynamic Prospective Average and Marginal GHG Emission Factors—Scenario-Based Method for the German Power System until 2050. *Energies 2021, Vol. 14, Page 2527* [online]. 2021, **14**(9), 2527 [acessed 13. 04. 2026]. ISSN 1996-1073. Dostopno na:10.3390/EN14092527

[10] KOPSAKANGAS-SAVOLAINEN, Maria idr. Hourly-based greenhouse gas emissions of electricity – cases demonstrating possibilities for households and companies to decrease their emissions. *Journal of Cleaner Production* [online]. 2017, **153**, 384–396 [acessed 13. 04. 2026]. ISSN 0959-6526. Dostopno na:10.1016/J.JCLEPRO.2015.11.027

[11] *Transparency Platform* [online]. [acessed 13. 04. 2026]. Dostopno na: https://transparency.entsoe.eu/

[12] ANDREASI BASSI, Susanna. idr. *Updated characterisation and normalisation factors for the environmental footprint 3.1 method*. Publications Office of the European Union, 2023. ISBN 9789276990697.

[13] *ecoinvent v3.11 - ecoinvent* [online]. [acessed 13. 04. 2026]. Dostopno na: https://ecoinvent.org/ecoinvent-v3-11/

[14] *Slovenia - Countries & Regions - IEA* [online]. [acessed 13. 04. 2026]. Dostopno na: https://www.iea.org/countries/slovenia/electricity

[15] *Croatia - Countries & Regions - IEA* [online]. [acessed 13. 04. 2026]. Dostopno na: https://www.iea.org/countries/croatia/electricity

[16] *Serbia - Countries & Regions - IEA* [online]. [acessed 13. 04. 2026]. Dostopno na: https://www.iea.org/countries/serbia/electricity

Materials role and challenges of green transition

Bojan Podgornik[1,2]

[1] *Fakulteta za industrijski inženiring, Novo mesto, Slovenija,* [2]*Inštitut za kovinske materiale in tehnologije, Ljubljana, Slovenija*
E-naslovi: bojan.podgornik@imt.si

***Abstract**: The transition towards climate neutrality is often framed as a shift in energy systems; however, it fundamentally represents a broader transformation of industrial processes and material use. This paper critically examines the role of materials and industry in enabling the green transition, highlighting key interdependencies and systemic constraints. While renewable energy, electrification and hydrogen technologies offer viable pathways for reducing greenhouse gas emissions, their large-scale deployment is associated with increased material demand, supply chain dependencies and significant life-cycle impacts. Particular attention is given to energy-intensive industries and the growing importance of critical raw materials required for low-carbon technologies. The analysis demonstrates that many proposed solutions, although effective at reducing operational emissions, introduce new environmental and resource challenges across global value chains. Therefore, achieving true climate neutrality requires a holistic, life-cycle-based approach that integrates decarbonization, circular economy principles and responsible material management. The findings emphasize that materials science and engineering play a central role in balancing performance, sustainability and resource availability, and are essential for ensuring a resilient and truly sustainable green transition.*

***Keywords:** green transition, materials, renewable energy, recycling, sustainability*

***Article Classification**: Scientific Paper*

1 INTRODUCTION

Climate change represents one of the most pressing challenges of modern society, driven primarily by anthropogenic greenhouse gas (GHG) emissions associated with energy production, industrial activities, transport and material use. In response, Europe has adopted an ambitious objective to become the first climate-neutral continent by 2050, aiming to balance GHG emissions with natural and technological carbon removal processes [1]. Achieving climate neutrality requires not only substantial emission reductions, but also compensation of unavoidable emissions through carbon sinks such as forests, soils and oceans. However, the transition to a climate-neutral society extends far beyond emission reduction alone. It represents a systemic transformation of energy systems, industrial production, mobility patterns and resource management.

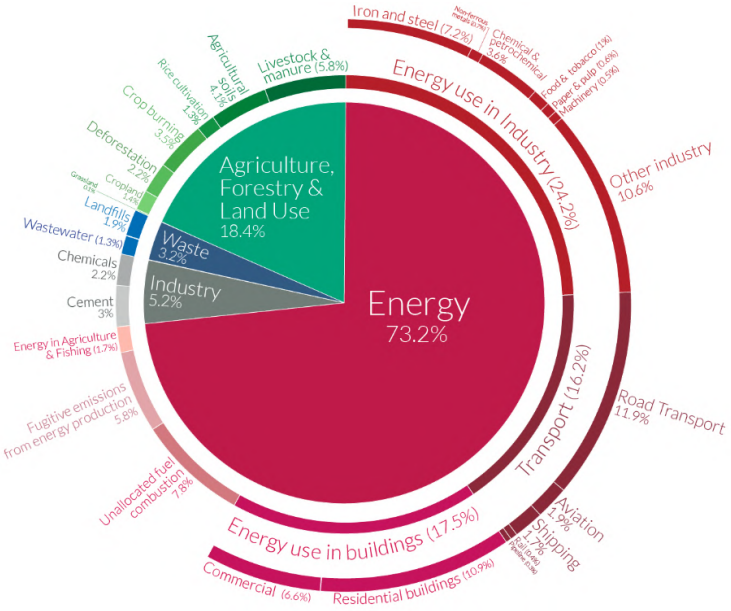


Figure 1: Global GHG emissions in 2020 by sectors [10]

From a global perspective, greenhouse gas emissions are dominated by energy-related activities, which account for approximately three-quarters of total emissions (Fig. 1), with electricity and heat production, industry and

transport representing the largest contributing sectors [2]. Despite significant progress within the European Union, where emissions have been reduced by nearly 30% over the past decades while maintaining economic growth, climate change remains a global challenge [3]. The European Union today contributes less than 10% of global GHG emissions, whereas regions such as Asia account for the dominant share due to rapid industrialization and economic expansion [3]. This imbalance introduces the critical issue of carbon leakage, where emission-intensive production is relocated to regions with less stringent environmental regulations. While such shifts may reduce emissions within Europe, they often fail to reduce global emissions and may even increase them when considering embedded emissions in imported goods [4]. Therefore, the effectiveness of the green transition must be evaluated from a global life-cycle perspective, taking into account emissions across entire value chains. A comprehensive approach to climate neutrality must integrate decarbonization of energy systems, transformation of industrial processes and sustainable resource management. Only by addressing these interdependencies can the green transition deliver meaningful and lasting reductions in global environmental impact.

This paper adopts a structured analytical approach based on a critical review and synthesis of scientific literature, policy reports and industrial data sources. The analysis focuses on identifying key interdependencies between energy systems, material use and industrial processes, with particular emphasis on life-cycle perspectives and resource constraints. The study is qualitative in nature and aims to provide a system-level assessment of the challenges associated with the green transition.

2 ROLE OF INDUSTRY IN DECARBURIZATION

Industry plays a central role in the transition towards climate neutrality, accounting for approximately 20% of greenhouse gas (GHG) emissions in Europe and an even larger share globally when indirect energy-related emissions are considered [2]. Decarbonizing industrial systems is therefore essential, yet particularly challenging due to the inherent complexity of production processes, high energy demand and long lifetimes of industrial assets. A fundamental pillar of industrial decarbonization is the transition towards a circular economy, aimed at reducing dependence on primary raw materials and minimizing waste generation. This includes extending product lifetimes, improving reparability and increasing the use of recycled materials. However, despite its recognized importance, circularity remains limited in practice, with only around 12% of secondary materials currently reintroduced into the EU economy [5]. This highlights a substantial untapped potential for reducing both resource consumption and associated emissions. At the same time, the green transition is inherently material-intensive. Clean energy technologies, electrification and hydrogen systems require large quantities of strategic and critical raw materials such as lithium, nickel, copper, aluminum and rare-earth elements [6]. Their extraction and processing are highly energy-intensive and often associated with significant environmental impacts. Consequently, industrial decarbonization cannot be achieved without simultaneous decarbonization of energy supply, as nearly 70% of industrial energy consumption globally still relies on fossil fuels [7].

The greatest challenge lies within energy-intensive industries (EIIs), including steel, cement, aluminum and chemical sectors. These industries account for the majority of industrial emissions (Fig. 1) due to their reliance on high-temperature processes and, in some cases, unavoidable process-related emissions. Steel and cement production alone each contribute approximately 7% of global CO_2 emissions [7]. In such sectors, conventional efficiency improvements are insufficient, and transformative technological changes are required. Several technological pathways are currently being developed to address these challenges. In steel production, the shift from blast furnace routes to electric arc furnaces (EAF), particularly when combined with recycled feedstock and renewable electricity, can reduce emissions by up to 75% [8]. Further reductions are expected through the development of hydrogen-based direct reduction processes (DRI-H_2), which have the potential to nearly eliminate process emissions. Similarly, aluminum production can achieve substantial emission reductions through recycling, as secondary aluminum requires up to 95% less energy compared to primary production without compromising material properties [9].

Despite these advantages, increasing the share of secondary production faces multiple barriers, including insufficient collection systems, limited availability of high-quality scrap and inefficient material separation technologies. In parallel, industrial side-streams represent a largely underutilized resource. Significant amounts of potentially valuable materials are still treated as waste due to technical, economic and regulatory constraints, including inconsistent quality, lack of standardization and underdeveloped value chains [10].

In addition to technological and material challenges, industrial decarbonization in Europe is strongly influenced by global economic dynamics. Stringent environmental regulations increase production costs and may incentivize relocation of emission-intensive industries to regions with less strict climate policies, leading to carbon leakage [4]. To address this, the European Union has introduced the Carbon Border Adjustment Mechanism (CBAM), aimed at ensuring that imported products reflect their embedded carbon emissions and that domestic decarbonization efforts are not undermined. Overall, the industrial sector represents both one of the largest challenges and one of the most critical enablers of the green transition. Achieving substantial emission reductions requires a combination of material efficiency, circular economy practices, low-carbon energy supply and breakthrough process innovations. Without a successful transformation of industrial systems, climate neutrality targets cannot be achieved, nor can a sustainable and secure supply of materials for the green transition be ensured.

3 ENERGY TRANSITION

The decarbonization of the energy system is the cornerstone of the green transition, as energy production and use account for approximately three-quarters of global greenhouse gas (GHG) emissions and a comparable share within the European Union [2,11]. Despite decades of policy efforts and technological progress, the global and European energy systems remain strongly dependent on fossil fuels. In the EU, oil, natural gas and coal still dominate the energy mix, while renewables and nuclear energy represent a significantly smaller share. Globally, the situation is even more pronounced, with fossil fuels accounting for more than 80% of primary energy consumption in most regions [12]. This structural dependency represents a fundamental barrier to rapid decarbonization.

The transition towards renewable energy sources (RES), including solar, wind and hydropower, is therefore essential. In recent years, renewables have experienced significant growth, driven by decreasing costs and supportive policy frameworks. However, their integration into the energy system is associated with several systemic challenges. First, renewable energy sources are inherently variable and dependent on weather conditions, leading to intermittency in electricity generation. This requires substantial investments in grid infrastructure, energy storage and system flexibility to ensure stable and reliable energy supply. Second, the scale of renewable deployment required to meet climate targets is unprecedented. Achieving climate neutrality by 2050 will require a massive expansion of renewable generation capacity, including large-scale offshore wind and solar installations. At the same time, electrification of transport, heating and industrial processes will significantly increase electricity demand, further amplifying the need for clean energy production. This creates a dual challenge: decarbonizing existing energy supply while simultaneously expanding it.

A critical but often overlooked aspect of the energy transition is its material intensity. Renewable energy technologies require substantially larger quantities of raw materials compared to conventional fossil-based systems (Fig. 2). Wind turbines, solar panels, batteries and grid infrastructure depend on metals such as copper, aluminum, lithium and rare-earth elements, the supply of which is geographically concentrated and limited [13,14]. This introduces new dependencies and risks. In this context, hydrogen is increasingly recognized as a key energy carrier capable of complementing electrification and enabling decarbonization of sectors that are difficult to electrify directly. These include high-temperature industrial processes, long-distance transport, shipping and aviation. Green hydrogen, produced via water electrolysis using renewable electricity, offers the potential for near-zero emissions energy use. However, its large-scale deployment is constrained by several factors. First, hydrogen production is highly energy-intensive. Producing green hydrogen requires significantly more electricity compared to direct electrification pathways, making its efficiency relatively low [15]. Second, the carbon footprint of hydrogen strongly depends on the electricity mix. If produced using fossil-based electricity, hydrogen may have equal or even higher emissions than conventional fuels, undermining its environmental benefits [16]. Third, hydrogen storage and transport present additional technical challenges due to its low volumetric energy density, requiring either high-pressure compression or liquefaction at extremely low temperatures, both of which are energy demanding [17]. Furthermore, hydrogen technologies rely on critical materials such as nickel, platinum and iridium, which are associated with supply risks and high costs [18]. This adds another layer of complexity to the energy transition, linking energy system transformation directly to material availability and sustainability.

Overall, the energy transition is not merely a technological shift from fossil fuels to renewables, but a complex systemic transformation involving energy, materials, infrastructure and global supply chains. While renewable energy and hydrogen provide viable pathways toward decarbonization, their large-scale implementation is constrained by physical, material and economic limitations. Addressing these challenges requires an integrated approach that combines expansion of renewable energy, improvements in energy efficiency, development of energy storage solutions and responsible management of material resources.

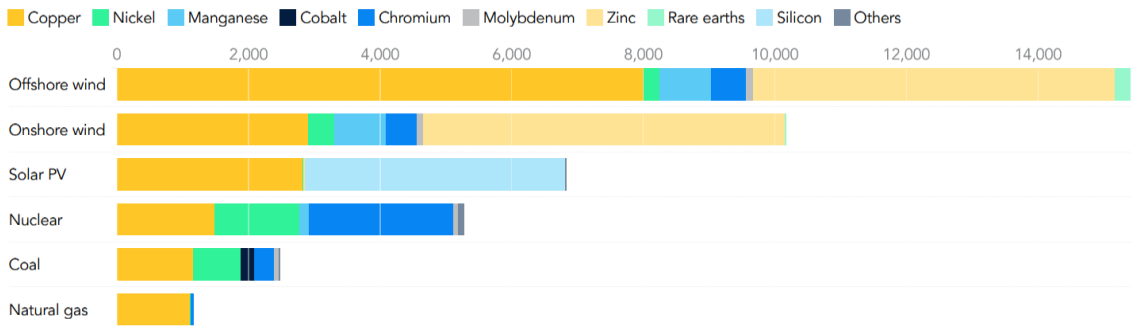


Figure 2: Minerals required in kilograms per MW of electrical generation from different energy source [19]

4 MATERIALS AS ENABLERS OF THE GREEN TRANSITION

While the green transition is often framed in terms of decarbonization and the deployment of renewable energy technologies, it is fundamentally a materials transition. The shift from a fossil-based energy system to one based on renewable electricity, electrification and hydrogen fundamentally changes the material basis of the economy, increasing both the quantity and complexity of materials required [19]. Consequently, materials are not merely passive components of green technologies, but critical enablers that determine their performance, scalability and sustainability. A defining feature of low-carbon technologies is their significantly higher material intensity compared to conventional systems. For example, electric vehicles require several times more mineral inputs than internal combustion engine vehicles, while renewable power generation technologies, particularly wind and solar demand substantially larger quantities of metals such as copper, aluminum and rare-earth elements [19]. Over the past two decades, the average material requirement per unit of installed renewable capacity has increased markedly (Fig. 2), reflecting the transition towards more complex and resource-intensive technologies. This increasing demand is particularly pronounced for a group of strategic and critical raw materials, including lithium, nickel, cobalt, manganese, graphite and rare-earth elements. These materials are essential for key components such as batteries, permanent magnets, photovoltaic cells and hydrogen technologies [13]. Their functional properties, such as high energy density, magnetic performance or catalytic activity are often difficult to substitute, making them indispensable for the deployment of clean technologies.

However, the supply of these materials is characterized by strong geographical concentration and limited diversification (Fig. 3). For instance, the production and processing of rare-earth elements is dominated by a small number of countries, while significant shares of lithium and cobalt resources are located in regions with environmental, social and governance challenges [14]. This creates supply risks, price volatility and geopolitical dependencies, particularly for regions such as Europe, which have limited domestic extraction and processing capacities. In addition to supply risks, the environmental footprint associated with raw material extraction and processing presents a critical challenge. Mining and refining of critical materials are energy-intensive processes that can lead to significant greenhouse gas emissions, land degradation, water consumption and biodiversity loss. In some cases, the environmental impact of material production represents a substantial share of the total life-cycle impact of green technologies, potentially offsetting part of their environmental benefits if not properly managed. These challenges highlight the importance of adopting a life-cycle perspective when evaluating the sustainability of green technologies. The environmental performance of a technology cannot be assessed solely based on its operational emissions, but must also include upstream impacts related to material extraction, processing and manufacturing, as well as end-of-life management and recycling.

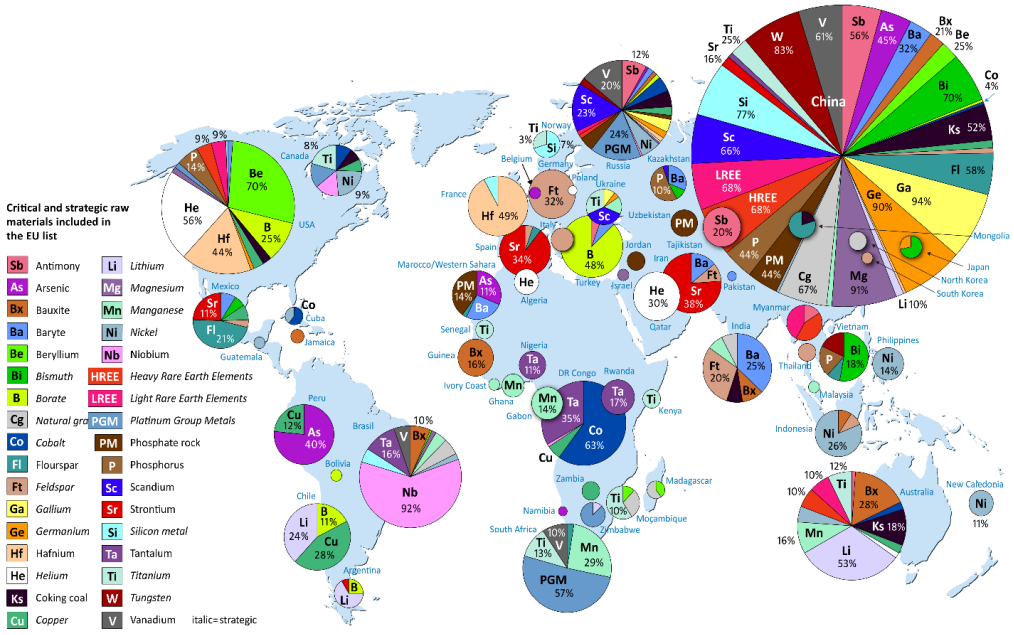


Figure 3: Global distribution of critical and strategic raw materials [6]

In this context, circular economy strategies play a central role in mitigating material-related constraints. Increasing the use of secondary raw materials through recycling can significantly reduce both energy consumption and emissions. Despite evident advantages, current recycling rates for many critical materials remain low, particularly due to technical challenges, insufficient collection systems and the complex composition of modern products. Another important but underutilized opportunity lies in the valorization of industrial side-streams. Many industrial processes generate by-products that contain valuable materials, yet these streams are often treated as waste due to lack of appropriate technologies, standards and market structures [10]. Developing efficient pathways for their utilization could significantly reduce the demand for primary raw materials and improve overall resource efficiency.

Beyond improving material efficiency and circularity, reducing dependence on critical raw materials also requires innovation in materials design. This includes the development of alternative materials with lower criticality, improved recyclability and reduced environmental impact. Examples include cobalt-free battery chemistries, reduced rare-earth permanent magnets and alternative catalysts for hydrogen technologies [18]. However, such substitutions often involve trade-offs in performance, cost or scalability, requiring careful optimization.

The rapid growth in demand for materials driven by the green transition raises fundamental questions about long-term resource availability. Projections indicate that demand for key materials such as lithium, nickel and rare-earth elements could increase several-fold by 2040, depending on the pace of decarbonization [20]. At the same time, mining capacity, processing infrastructure and recycling systems may not scale at the same rate, creating potential bottlenecks that could slow down the transition. For Europe, these challenges are particularly pronounced. The region is highly dependent on imports for most critical raw materials and currently lacks sufficient mining, refining and recycling capacity to ensure supply security [13]. This has led to the development of strategic initiatives aimed at strengthening domestic capabilities, diversifying supply chains and increasing material circularity. However, achieving these goals will require substantial investments, regulatory support and technological innovation.

Overall, the success of the green transition is intrinsically linked to the availability, sustainability and efficient use of materials. While renewable energy technologies and electrification pathways offer clear opportunities for reducing operational emissions, their large-scale deployment introduces new environmental and resource challenges. Addressing these challenges requires an integrated approach that combines responsible sourcing, improved material efficiency, circular economy strategies and innovation in materials science and engineering. In this context, materials science emerges as a key discipline in enabling the green transition, providing solutions that balance performance, sustainability and resource constraints. Without a fundamental shift in how materials are sourced, used and managed, the transition to a climate-neutral society may face significant limitations in both scalability and long-term sustainability.

5 TECHNOLOGY SPECIFIC LIMITATIONS

Although low-carbon technologies are essential for achieving climate neutrality, none of them are without limitations. Their large-scale deployment introduces technical, material, environmental and systemic challenges that must be critically addressed to ensure a truly sustainable transition. A central example is electric mobility, widely regarded as a key solution for decarbonizing transport. While battery electric vehicles (BEVs) can significantly reduce tailpipe emissions, their overall environmental performance strongly depends on the electricity mix used for charging. When powered by fossil-based electricity, life-cycle emissions of BEVs may approach or even exceed those of efficient internal combustion engine vehicles [20]. Furthermore, battery production is highly material-intensive, requiring large quantities of lithium, cobalt, nickel and other critical materials, the extraction of which is associated with environmental degradation, high energy consumption and social concerns [21]. End-of-life management presents an additional challenge, as recycling of lithium-ion batteries remains complex, energy-intensive and economically demanding [22]. Combined with growing demand for batteries, these factors raise concerns regarding long-term resource availability and sustainability. It needs to be taken into account that the number of new electric cars registered in the EU increased from only 600 in 2010 to over 2 million in 2022, but still being far away from the EU target of having 30 million zero-emission cars in operation on European roads by 2030. Globally, electric vehicles sales in 2022 exceeded 10 million cars. However, this still represents less than 2% of the global passenger cars fleet, not yet following the overall growth and 95% of electric vehicles being restricted to China, Europe and North America, as shown in Figure 4 [23]. By 2050 the global vehicle fleet is set to double!

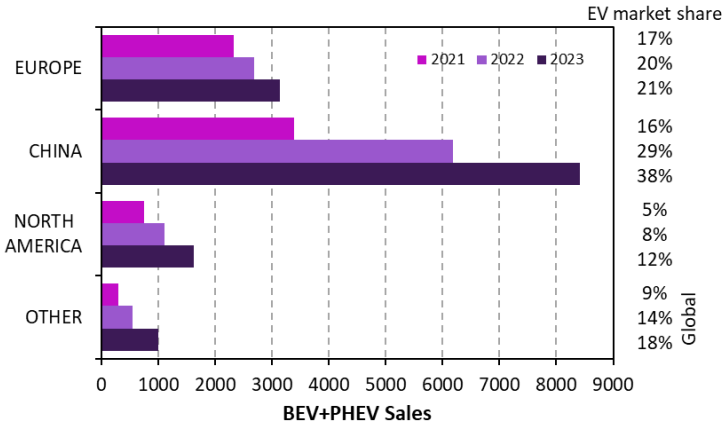


Figure 4: Electric cars sales and market share [23]

Solar energy, often perceived as one of the cleanest energy sources, also involves significant upstream environmental impacts. The production of photovoltaic (PV) panels requires energy-intensive processes and the use of hazardous chemicals, including fluorinated gases and other toxic substances [24]. Although operational

emissions are low, it may take several years for a solar panel to offset emissions generated during its production. In addition, relatively low conversion efficiencies and gradual performance degradation over time necessitate large-scale installations, which can lead to land use conflicts, habitat disruption and local climate effects such as the photovoltaic heat island phenomenon [25]. End-of-life management of PV panels represents another growing concern, as recycling infrastructure remains limited and significant volumes of waste are expected in the coming decades [26].

Wind energy faces a different set of challenges. While wind turbines produce low operational emissions, their manufacturing relies on both critical and non-critical materials with substantial environmental footprints. Rare-earth elements used in permanent magnets are associated with environmentally intensive mining and processing, as well as supply risks due to geographic concentration [27]. At the same time, the bulk of turbine mass consists of steel and concrete, whose production is highly energy-intensive and contributes significantly to overall emissions. End-of-life issues are particularly pronounced for turbine blades, which are difficult to recycle due to their composite structure, leading to increasing volumes of waste and reliance on landfill or downcycling solutions. Additional concerns include land use, visual impact, noise and potential local climate effects associated with large wind farms [28].

Hydrogen technologies, often highlighted as a key solution for decarbonizing hard-to-abate sectors, are similarly constrained by efficiency and infrastructure challenges. The environmental benefits of hydrogen depend critically on its production pathway. While green hydrogen produced via electrolysis using renewable electricity offers low emissions, it requires significantly higher energy input compared to direct electrification [15]. If produced using fossil-based electricity, hydrogen may have a higher carbon footprint than conventional fuels [16]. Storage and transport further complicate its deployment, as hydrogen has low volumetric energy density, requiring compression or liquefaction, both of which are energy-intensive processes [17]. In addition, hydrogen technologies rely on scarce and costly materials such as platinum and iridium, creating further supply constraints [18].

Across all these technologies, a common pattern emerges. Their environmental performance cannot be evaluated solely based on operational emissions, but must consider full life-cycle impacts, including material extraction, production, use and end-of-life phases. Moreover, large-scale deployment amplifies challenges related to material demand, infrastructure requirements and system integration. These limitations do not undermine the importance of low-carbon technologies, but rather emphasize the need for a balanced and system-oriented approach. Sustainable transition requires not only technological innovation, but also improvements in material efficiency, circularity, infrastructure development and responsible resource management. Without addressing these interconnected constraints, the green transition risks shifting environmental burdens rather than reducing them.

6 CONCLUSIONS

The transition to climate neutrality is not merely an energy transition, but a fundamental transformation of the industrial and material backbone of modern society. While renewable energy, electrification and hydrogen are essential for reducing emissions, their large-scale deployment reveals a critical reality: decarbonization is inherently resource- and material-intensive. When assessed from a full life-cycle perspective, many low-carbon technologies do not eliminate environmental impact, but redistribute it across global supply chains, often beyond Europe’s borders.

Achieving a truly sustainable transition therefore requires a systemic approach that integrates energy, materials and industrial processes. Circular economy principles, responsible sourcing and material innovation must become central strategies, not secondary considerations. Ultimately, the success of the green transition will not depend only on how effectively we replace fossil fuels, but on how intelligently we manage materials and technological complexity. In this context, materials science and engineering are not supporting elements, but decisive enablers of a sustainable future.

REFERENCES

[1] European Green Deal, European Council, https://www.consilium.europa.eu/en/policies/green-deal/; accessed 20.04.2026
[2] RITCHIE, H. Sector by sector: where do global greenhouse gas emissions come from?, Our World In Data, 2020, https://ourworldindata.org/ghg-emissions-by-sector; accessed 20.04.2026
[3] RITCHIE, H., ROSADO, P., ROSER, M. Greenhouse gas emissions, Our World in Data, https://ourworldindata.org/greenhouse-gas-emissions; accessed 20.04.2026
[4] LEONARD, M., PISANI-FERRY, J., SHAPIRO, J., TAGLIAPIETRA, S., WOLFF, G. The geopolitics of the European Green Deal, European Council on Foreign Relations, https://ecfr.eu/publication/the-geopolitics-of-the-european-green-deal/; accessed 22.10.2024
[5] The European Green Deal - a roadmap to a sustainable economy, Public Agency Ihobe, 2020, https://www.ihobe.eus/news/the-european-green-deal-a-roadmap-to-a-sustainable-economy; accessed 20.04.2026
[6] Geological Survey of Sweden, Critical and strategic raw material, 2024, https://www.sgu.se/en/mineral-resources/critical-raw-materials/; accessed 20.04.2026.

[7] IEA, Energy Technology Perspectives 2020, 2020, https://www.iea.org/reports/energy-technology-perspectives-2020; accessed 20.04.2026
[8] TOTO, D. Study confirms EAF advantage in carbon emissions, Recycling today, July 2022, https://www.recyclingtoday.com/news/us-steelmaking-75-percent-less-carbon-emission/; accessed 20.04.2026
[9] BEHESHTI, R. Sustainable aluminium and iron production, Ph.D. Thesis, KTH Royal Institute of Technology, Sweden, 2017
[10] LEPPÄNEN, T., KÖPMAN, J., RASILA, O., TERVONEN, P. Categorization of Industrial Side Streams for Reuse Potential Evaluation, *International Journal of Management, Knowledge and Learning*, Vol. 10 (2021) 253-265.
[11] European Environment Agency, Greenhouse gas emission intensity of electricity generation in Europe, November 2025, https://www.eea.europa.eu/en/analysis/indicators/greenhouse-gas-emission-intensity-of-1; retrieved 20.04.2026
[12] RITCHIE, H., ROSADO, P. Energy Mix, Our World in Data, 2020, https://ourworldindata.org/energy-mix; retrieved 20.04.2026
[13] GROHOL, M., VEEH C., Study on the critical raw materials for the EU 2023 – Final report, Publications Office of the European Union, 2023, doi/10.2873/725585
[14] CARRARA, S. et al. Supply chain analysis and material demand forecast in strategic technologies and sectors in the EU – A foresight study, Publications Office of the European Union, Luxembourg, 2023, doi:10.2760/386650, JRC132889.
[15] GODFREY NNABUIFE, S., HAMZAT, A.K., WHIDBORNE, J., KUANG, B., JENKINS, K.W. Integration of renewable energy sources in tandem with electrolysis: A technology review for green hydrogen production, *International Journal of Hydrogen Energy*, Vol. 107 (2025) 218-240
[16] WEISS, T., GAMAGE, C., KOCH BLANK, T., LILLIS, G., JARDINE WALL, A. Hydrogen Reality Check: All “Clean Hydrogen” Is Not Equally Clean, RMI, October 2022, https://rmi.org/all-clean-hydrogen-is-not-equally-clean/; accessed 20.04.2026
[17] ESSLER, J. et al. Report on Technology Overview and Barriers to Energy- and Cost-Efficient Large Scale Hydrogen Liquefaction, IDEALHY project (FP7/2007-2013, grant agreement no. 278177), June 2012, https://www.idealhy.eu/; accessed 20.04.2026
[18] EIKENG, E., MAKHSOOS, A., POLLET, B.G. Critical and strategic raw materials for electrolysers, fuel cells, metal hydrides and hydrogen separation technologies, *International Journal of Hydrogen Energy*, Vol. 71 (2024) 433-464
[20] IEA, The Role of Critical Minerals in Clean Energy Transitions, 2022, https://www.iea.org/reports/the-role-of-critical-minerals-in-clean-energy-transitions, accessed 20.04.2026
[21] European Environment Agency, Progress of EU transport sector towards its environment and climate objectives, https://www.eea.europa.eu/publications/progress-of-eu-transport-sector-1; accessed 20.04.2026
[22] ZHENG, M. The Environmental Impact of Lithium and Cobalt Mining, Earth.org, March 2023, https://earth.org/lithium-and-cobalt-mining/; accessed 20.04.2026
[23] SKEETE, J.-P., WELLS, P., DONG, X., HEIDRICH, O., HARPER, G. Beyond the EVent horizon: Battery waste, recycling, and sustainability in the United Kingdom electric vehicle transition, *Energy Research & Social Science*, Vol. 69 (2020) 101581
[24] The global electric vehicle market overview in 2025, VIRTA, https://www.virta.global/en/global-electric-vehicle-market; retrieved 20.04.2026
[25] HESS, J.C. Chip Production’s Ecological Footprint: Mapping Climate and Environmental Impact, Interface, June 2024, https://www.interface-eu.org/publications/download/chip-productions-ecological-footprint; accessed 20.04.2026
[26] BARRON-GAFFORD, G.A. et al. The Photovoltaic Heat Island Effect: Larger solar power plants increase local temperatures, *Scientific Reports*, Vol. 6 (2016) 35070
[27] Weckend, S., Wade, A., Heath, G. End-of-Life Management – Solar Photovoltaic Panels, International Renewable Energy Agency, June 2016, https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2016/IRENA_IEAPVPS_End-of-Life_Solar_PV_Panels_2016.pdf; accessed 20.04.2026
[28] LI, Jiashuo et al. Critical Rare-Earth Elements Mismatch Global Wind-Power Ambitions, *One Earth*, Vol. 3 (2020) 116-125
[29] Miller, L.M., Keith, D.W. Climatic Impacts of Wind Power, *Joule*, Vol. 2 (2018) 2618-2632

AI-Driven Predictive Maintenance for Sustainable Power Grid Infrastructure

Dunja Srpak[1], Srđan Skok[2], Tomislav Horvat[3], Josip Srpak[4]

[1]*University North, Croatia*
E-mail: dsrpak@unin.hr

[2]*University North, Croatia*
E-mail: sskok@unin.hr

[3]*University North, Croatia*
E-mail: thorvat@unin.hr

[4]*University North, Croatia*
E-mail: jsrpak@unin.hr

Abstract: *The operation of modern industrial facilities relies inevitably on a stable electricity supply, necessitating continuous monitoring and timely preventive or corrective maintenance interventions. Rapid advancements in information and communication technologies (ICT) and artificial intelligence (AI) technologies are transforming energy production, transmission, and distribution processes, aligning with sustainable development goals and the green transition.*

This study analyses the current state of AI applications in monitoring the integrity of power grid infrastructure. A significant potential is identified in AI tools, such as convolutional neural networks (CNNs), for detecting potential faults or failure locations from sequences of photographs or video footage of electrical facilities. By comparing existing AI solutions like YOLO for object detection and anomaly detection models, the study evaluates required input data formats (e.g., high-resolution imagery) and valuable algorithmic outputs, including fault probability scores and localization heat maps. Through discussion and analysis of these approaches, the study clarifies future research directions, identifies remaining gaps, and proposes sequential algorithms, from data pre-processing to real-time deployment, for enhanced maintenance efficiency.

Future research should focus on resource optimization, enabling faster decision-making in energy sectors and contributing to environmental protection as key pillars of industrial engineering in the era of artificial intelligence. This includes reducing waste through predictive interventions and facilitating integration of renewable energy sources (RES), supporting the green transition toward a competitive, low-carbon economy.

Keywords*: AI, predictive maintenance, power grid, renewable energy integration, green transition*

Article Classification*: Scientific Paper*

1 INTRODUCTION

Artificial intelligence (AI) has emerged as a critical enabler in the energy sector, addressing essential challenges such as aging infrastructure, rising demand from electrification, and the integration of intermittent renewable energy sources (RES) into power grids. By enabling predictive maintenance, real-time fault detection, and optimized resource allocation, AI enhances grid reliability, reduces operational costs, and supports the transition to low-carbon economies. The aim of this study is to analyse the current state of AI applications for monitoring the integrity of power grid infrastructure, with an emphasis on applications for visual fault detection from photo/video sequences. At the same time, the identification of the research gap will direct this four-year project towards the development of practical algorithmic solutions.

A literature review identified a significant amount of research focused on the application of artificial intelligence techniques for fault detection, classification, and predictive maintenance in power systems [1–5]. Advanced machine learning models, particularly those based on object detection frameworks, have demonstrated high effectiveness in visual inspection tasks. Comprehensive review studies have systematically analysed the evolution of AI applications in power system fault diagnosis [6], covering a wide spectrum of methods including deep learning approaches [7–10], artificial neural networks (ANNs) [11], support vector machines (SVMs) [12], and different hybrid solutions [13-15]. Classical machine learning methods remain relevant, particularly for processing structured and time-series data obtained from measurement systems. Machine learning-based approaches for fault detection and classification in transmission systems have demonstrated reliable performance under various operating conditions [3]. Hybrid methodologies combining different techniques have further improved classification accuracy,

particularly in complex fault scenarios [15]. These approaches are especially suitable for integration with synchronized measurement technologies and wide-area monitoring systems.

In parallel, AI applications have been extended to renewable energy systems, where deep learning models have been successfully applied for fault diagnosis in photovoltaic modules, addressing the increasing need for reliable operation in RES-integrated grids [16]. Similarly, specialized adaptations of object detection architectures have been developed for identifying defects in power distribution equipment, highlighting the importance of domain-specific model optimization [17]. YOLO (You Only Look Once) is a fast real-time object detection model that simultaneously predicts bounding boxes and classification probabilities across the entire image. Recent research increasingly adopts YOLO models for fault detection in power systems. Examples include benchmarking YOLO against CNNs on novel datasets for high-accuracy solar panel defect detection [18], enhancing YOLO11 for insulator defects on distribution lines [19], and integrating hybrid YOLO with adaptive context refinement for sensor data in power distribution networks [20].

In addition to these recent developments, significant contributions have been made in the field of power system monitoring, modelling, and control using synchronized measurements and intelligent algorithms. Research on wide-area monitoring (WAM) systems and synchrophasor-based applications has demonstrated the importance of real-time system observability for stability assessment and control. Furthermore, advanced methodologies for transmission system modelling and the integration of distributed energy resources (DER) through coordinated data exchange between transmission and distribution operators have been proposed, enabling enhanced system analysis and operation in real-time environments [21]. These approaches form a crucial foundation for the application of AI in modern smart grids. Moreover, recent work on integrated energy systems and smart energy hubs highlights the importance of centralized monitoring and intelligent energy management platforms, particularly in complex environments such as port energy systems, where multiple energy carriers and loads must be coordinated efficiently [22]. Such frameworks underline the necessity of combining AI with advanced monitoring and control architectures to achieve flexible, resilient, and sustainable energy systems.

Despite significant progress, several challenges remain. Many existing approaches rely on large, labelled datasets that are difficult to obtain in real-world power systems. Additionally, challenges related to real-time deployment, computational efficiency on edge devices, cybersecurity, and system interoperability are still insufficiently addressed.

2 METHODS

This study employed a systematic literature review using academic databases including Web of Science, ScienceDirect, MDPI, Nature, and arXiv to establish the current state of AI applications in power grid infrastructure monitoring. The reviewed studies were selected based on keywords and criteria spanning AI methodologies (CNNs, YOLOs, ANNs, SVMs, hybrids), data modalities (time series, images, videos), performance assessment metrics, and practical application challenges, forming the basis for a comprehensive analysis in the next section.

As this preliminary report establishes the basis for a four-year research project, the main part of the paper is a critical synthesis of existing approaches, identifying key gaps (lack of data, limitations in implementation at the edge of the network, cybersecurity vulnerabilities) and determining future research directions, presented in the analysis and discussion section. Finally, the conclusion summarizes the results of this initial research and presents the direction for further research. Figure 1 shows the main methodological steps followed in this study, from problem identification and literature review to discussion of challenges and future research directions.

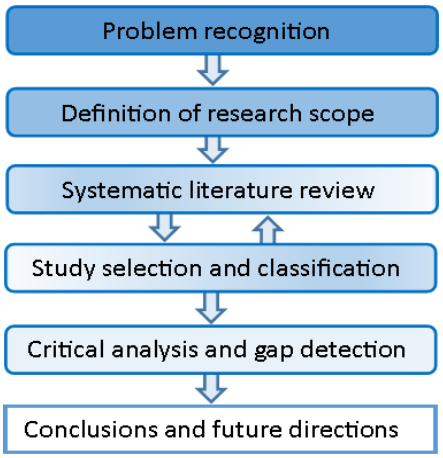


Figure 1: Methodological framework of the study

3 CURRENT STATE OF AI RESEARCH

This section provides a comprehensive analysis of the current state of artificial intelligence applications in monitoring the integrity of power grid infrastructure, with particular emphasis on algorithms for detecting potential faults or failure locations from sequences of photographs or video footage of electrical facilities. It systematically reviews key AI methods and approaches in subsection 3.1, explores their practical implementations in fault monitoring and anomaly detection in 3.2, evaluates required input data formats and algorithmic outputs in 3.3, and offers comparative insights in 3.4, all based on the presented literature.

3.1 Overview of AI Applications

This subsection describes fundamental artificial intelligence techniques, including neural networks, support vector machines, convolutional models, and hybrid systems, essential for fault detection in power grids.

3.1.1 ANN and SVM Approaches

Artificial Neural Networks (ANNs) serve as a foundational approach for processing time-series data from grid sensors, including measurements of voltage, current, frequency, and power factor, to classify various fault types such as short circuits, overloads, or phase imbalances in transmission and distribution lines. These multilayer networks excel at modelling complex, non-linear relationships inherent in electrical grid dynamics, making them particularly effective for tasks like load forecasting, anomaly detection in steady-state operations, and early warning systems for cascading failures that could propagate through interconnected networks. Their ability to learn hierarchical representations from raw sensor streams positions ANNs as versatile tools in environments where data volume and variability challenge simpler statistical models.

Support Vector Machines (SVMs), meanwhile, demonstrate a particular strength in multiclass fault classification scenarios, especially when working with limited or noisy datasets typical of remote grid monitoring stations. By employing kernel functions (such as radial basis function kernels) SVMs map input features into higher-dimensional spaces to achieve optimal hyperplane separation, enabling reliable discrimination between subtle anomalies like harmonic distortions, ground faults, or intermittent arcing. This robustness to outliers and non-linear boundaries makes SVMs well-suited for real-time applications where sensor noise from environmental factors or aging equipment is prevalent, though they benefit from pre-processing steps like feature selection to handle high-dimensional inputs efficiently.

3.1.2 CNN and YOLO Models

Convolutional Neural Networks (CNNs) represent a paradigm shift in visual inspection paradigms for electrical infrastructure, automatically learning hierarchical feature representations, ranging from edges and textures to complex patterns like corrosion or thermal hotspots, from thermal imaging, visible-light photography, or hyperspectral scans of critical components such as insulators, transformers, bushings, and overhead conductors. Successive layers of convolutional filters, activation functions, and pooling operations enable CNNs to identify degradation signatures that manual inspections often miss, particularly in large-scale assets spanning thousands of kilometres of transmission lines. This end-to-end learning capability supports scalable deployment across fixed surveillance cameras, handheld devices, or automated patrols, transforming reactive maintenance into a proactive strategy grounded in data-driven insights.

YOLO (You Only Look Once) models, encompassing iterations like YOLOX, YOLOv8, and specialized variants such as YOLO-PDD (Power Distribution Device), advance this further through single-stage detection architectures optimized for real-time performance in unconstrained environments. By dividing input frames into grids and simultaneously predicting bounding boxes, class probabilities, and objectness scores, YOLO processes UAV-captured video streams of power infrastructure with minimal latency, even amid challenges like motion blur, atmospheric interference, or partial occlusions from foliage. This grid-based paradigm contrasts with two-stage detectors by prioritizing speed without compromising localization precision, rendering YOLO indispensable for edge applications where immediate feedback loops inform on-site technicians or trigger autonomous responses.

3.1.3 Expert Systems and Hybrid Approaches

Expert Systems (ES) augment purely data-driven models by embedding domain expertise through if-then rule bases derived from electrical engineering principles, such as IEEE standards for fault thresholds or protective relay logic, to validate ML predictions and filter spurious alerts stemming from sensor artefacts. In practice, ES parse fault trees or ontology-structured knowledge graphs to generate human-readable explanations, fostering trust among operators and facilitating compliance with regulatory frameworks. While inherently interpretable, ES limitations in handling probabilistic uncertainties or unprecedented events are mitigated through fuzzy logic extensions or Bayesian inference modules.

Hybrid architectures, exemplified by CNN-SVM pipelines or ANN-ES fusions, harness complementary strengths: deep networks furnish rich, latent feature embedding from raw multispectral data, which classical classifiers or rule engines then refine for decision-making. Such synergy yields systems resilient to distribution shifts across grid regions, climates, or aging profiles, while preserving auditability essential for critical infrastructure. These approaches increasingly incorporate digital twin simulations to virtually stress-test configurations, accelerating deployment cycles in operational settings.

3.2 AI Implementation in Practice

This subsection provides an overview of current real-world applications of artificial intelligence for fault monitoring, such as YOLO-based anomaly detection and accompanying optimization techniques in power grid operation.

3.2.1 AI for Fault Monitoring

Practical AI deployments span IoT-enabled phasor measurement units (PMUs), substation cameras, and drone swarms, continuously fusing visual sequences with synchrophasor data to pinpoint incipient failures like insulator flashovers, conductor galloping, or bushing leaks before they escalate into widespread disruptions. By correlating spatial imagery with temporal signatures, (correlating thermal anomalies with harmonic spikes) AI enables granular grid state awareness, empowering dynamic topology processing and contingency analysis in control centres.

3.2.2 YOLO and Anomaly Detection

YOLO's prowess shines in anomaly detection workflows, where it delineates irregularities, vegetation incursions, avian hazards, or structural deformations, in live feeds, outputting geospatial overlays integrable with GIS platforms for asset management systems. This facilitates triage prioritization, routing high-severity detections to human experts while archiving low-confidence cases for batch review, thereby optimizing human-AI collaboration loops.

3.2.3 Optimization Methods

Optimization strategies encompass transfer learning from domain-pretrained backbones (e.g., ResNet adapted to electrical imagery), geometric and photometric augmentations to emulate field variabilities, and stochastic gradient descent variants like Adam for stable convergence on heterogeneous hardware. Ensemble techniques, meta-learning curricula, and knowledge distillation further tailor models to sparse-label regimes, ensuring adaptability across legacy and next-generation smart grids.

3.3 Data and Output Evaluation

AI models in power grid monitoring demand carefully prepared inputs tailored to their architectures: ANNs and SVMs thrive on tabular time-series data in CSV format, comprising features like voltage waveforms, current harmonics, power factor, and temperature readings sampled at high frequencies, often pre-processed through normalization and feature scaling. CNNs and YOLO variants, conversely, require high-resolution visual inputs such as JPEG or MP4 files (at least 1080p from drones or thermal cameras). These inputs must be annotated with bounding boxes or pixel masks using tools like Labelling, and augmented to handle occlusions or lighting variances. Expert Systems utilize structured rule sets in XML or ontology formats, while hybrids fuse these modalities for comprehensive analysis.

The outputs from these models are actionable and quantifiable, including fault probability scores (sigmoid-activated values between 0 and 1 indicating likelihood), localization heat maps (saliency maps highlighting defective regions via Grad-CAM), and categorical labels (e.g., "overheating," "fracture"). Performance is rigorously assessed using metrics derived from the confusion matrix, where TP denotes true positives, FP false positives, FN false negatives, and TN true negatives. Accuracy measures overall correctness as (TP + TN) / Total, providing a broad performance indicator but faltering on imbalanced datasets common in rare fault events. Precision quantifies the reliability of positive predictions via TP / (TP + FP), Recall captures the model's sensitivity to actual faults as TP / (TP + FN), and the F1-score offers a balanced harmonic mean, calculated as 2 × (Precision × Recall) / (Precision + Recall), which is particularly useful for prioritizing fault detection over false alarms.

3.4 Comparative Insights

The comparison presented here is based on the literature reviewed in the preceding subsections, considering key criteria such as data modality, architectural strengths, and application contexts, as synthesized from established studies on AI methods for power grid monitoring. This analysis is summarized in Table 1 below.

Table 1: Comparison of AI methods for power grid monitoring

Method	Key Strengths	Data Modality	Deployment Context
ANN	Nonlinear temporal modelling	Time-series	Centralized forecasting
SVM	Margin-based classification	Tabular	Noisy sensor fusion
CNN	Hierarchical visual hierarchies	Imagery	Drone/substation vision
YOLO	End-to-end bounding prediction	Video	Real-time edge
Hybrid	Interpretable fusion	Multimodal	Mission-critical ops

4 ANALYSIS AND DISCUSSION

The analysis of AI applications outlined in Section 3 reveals significant advancements in power grid monitoring, yet underscores persistent challenges that shape future research trajectories. By synthesizing the strengths of methods like CNNs, YOLO, ANNs, SVMs, and hybrids, this section clarifies gaps in current implementations, such as data scarcity, interpretability deficits, and cybersecurity vulnerabilities, while proposing sequential algorithms for streamlined pre-processing to real-time deployment, ultimately enhancing maintenance efficiency and supporting grid resilience.

4.1 Strengths and Limitations Revisited

Deep learning models, particularly CNNs and YOLO variants, demonstrate superior performance in visual fault detection, enabling proactive interventions that significantly reduce outages. These approaches excel in processing high-resolution imagery from UAVs and sensors, capturing subtle anomalies like insulator cracks or line sags that elude traditional rule-based systems. However, their heavy reliance on large, annotated datasets poses a critical limitation in real-world grids where fault events are rare and labelling is labour-intensive, often leading to overfitting or poor generalization across diverse environmental conditions such as fog, rain, or varying lighting.

Classical methods like ANNs and SVMs offer robustness in tabular time-series analysis for multiclass classification, but they struggle with scalability on massive IoT streams and lack the spatial awareness needed for image-based tasks. Expert Systems and hybrids bridge these gaps by enhancing interpretability, significantly reducing false positives through rule validation, but introduce complexity in rule maintenance and vulnerability to adversarial inputs, where manipulated sensor data could mislead decisions in cyber-threatened environments. Overall, while AI has transformed predictive maintenance, the trade-offs between speed, accuracy, and explainability highlight the need for more adaptive, multimodal frameworks.

4.2 Identified Research Gaps

Several gaps persist despite progress. First, data-related challenges dominate: most studies use simulated or balanced datasets, neglecting real-world imbalances where healthy states vastly outnumber faults, inflating metrics like accuracy while masking poor recall. Second, edge deployment remains underexplored; models optimized for cloud servers falter on resource-constrained drones due to high memory footprints, limiting real-time use in remote grids. Third, cybersecurity risks remain insufficiently addressed, as adversarial attacks can significantly degrade CNN performance with minimal perturbations, while only a limited number of studies incorporate defence mechanisms such as robust training or federated learning. Finally, integration with renewables exposes gaps in handling variable RES inputs (e.g., solar intermittency), where AI must evolve beyond fault detection to dynamic forecasting and self-healing reconfiguration.

5 CONCLUSION

This study marks the initial phase of four-year research project on AI applications for power grid infrastructure monitoring, focusing primarily on identifying the current state-of-the-art, key research gaps, and foundational challenges as detailed in Sections 3 and 4. A systematic analysis of CNNs, YOLOs, ANNs, SVMs and hybrid approaches has established a comprehensive baseline that serves as a foundation for the next phases of development, where concrete algorithmic proposals and empirical validations of the generated solutions are planned.

Future work will emphasize resource optimization for accelerated decision-making in power systems, strengthening environmental protection as a fundamental aspect of AI-driven industrial engineering. This includes reducing waste by predicting potential failure points and increasing the proportion of preventive versus corrective infrastructure maintenance. Moreover, facilitating the integration of renewable energy sources (RES) into the grid will foster the transition to a sustainable, low-carbon and competitive economy.

Acknowledgment: This work was created as part of the project Unmanned systems for power infrastructure monitoring (ELENDRON) (IP-UNIN-TEH-2025-12), funded by the European Union - NextGenerationEU.

REFERENCES

[1] Nigam, Prateek, et al. *Investigating the Use of AI in Power System Fault Identification and Diagnosis. Book Chapter Recent Trends In Engineering and Science for Resource Optimization and Sustainable Development.,* 1st Edition, London, Taylor & Francis Group, 2024. ISBN 978-1-032-98030-0

[2] Najafzadeh, Masoud, et al. Fault Detection, Classification and Localization Along the Power Grid Line Using Optimized Machine Learning Algorithms. *International Journal of Computational Intelligence Systems*, 2024, vol. 17, 49, doi: 10.1007/s44196-024-00434-7

[3] Gillioz, Marc, et al. Anomaly Detection with Machine Learning Algorithms in Large-Scale Power Grids. *Electrical Engineering and Systems Science, Systems and Control*, 2026, arXiv:2602.10888 [eess.SY], pp. 1-12, doi: 10.48550/arXiv.2602.10888

[4] Manojna, Sridhar H. S, et al. Fault Detection and Classification in Power System using Machine Learning. *2nd International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, 2021, pp. 1801-1806, doi: 10.1109/ICOSEC51865.2021.9591972

[5] Sarker, Md Tanjil, et al. AI-driven smart grid optimization for hospital energy systems integrating renewable generation, predictive maintenance, and resilient infrastructure. *Scientific report*s, 2025, vol. 15, 44787, doi: 10.1038/s41598-025-28907-5

[6] Liu, Youping, et al. A Comprehensive Review of AI Integration for Fault Detection in Modern Power Systems: Data Processing, Modeling, and Optimization. *Energies*, 2025, vol. 18(18), 4983. doi: 10.3390/en18184983

[7] Behara, Ramesh Kumar, and Akshay Kumar Saha. Comparative Performance Analysis of Deep Learning-Based Diagnostic and Predictive Models in Grid-Integrated Doubly Fed Induction Generator Wind Turbines. *Energies*, 2025. vol. 18(17), 4725, doi: 10.3390/en18174725

[8] Hlalele. Tlotlollo Sidwell. Transmission line fault detection using deep learning technique. *In Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence (ICAAI '24)*. New York, NY, USA, Association for Computing Machinery, 2025, pp. 332–338, doi: 10.1145/3704137.3704170

[9] Turanlı, Ozan, and Yurdagül Benteşen Yakut. Classification of Faults in Power System Transmission Lines Using Deep Learning Methods with Real, Synthetic, and Public Datasets. *Applied Sciences*, 2024, vol. 14(20), 9590, doi: 10.3390/app14209590

[10] Duan, Jing. Deep learning anomaly detection in AI-powered intelligent power distribution systems. *Frontiers in Energy Research*, 2024, vol. 12, 1364456, doi: 10.3389/fenrg.2024.1364456

[11] Sokolović, Živko and Mileta, Žarković. Fault detection, classification and localization in HV power transmission lines using ANN. *Electric Power Systems Research*, 2025, vol. 248, 111927, doi: 10.1016/j.epsr.2025.111927

[12] Kumar, V. Rajesh and P. Aruna Jeyanthy. Fault Classification and Detection in Transmission Lines by Hybrid Algorithm Associated Support Vector Machine. *Transactions on Emerging Telecommunications Technologies* 2025, vol. 36, 70034, doi: 10.1002/ett.70034

[13] Alhanaf, Ahmed Sami, Murtaza, Farsadi and Hasan. Balik, Fault Detection and Classification in Ring Power System with DG Penetration Using Hybrid CNN-LSTM. *IEEE Access*. 2024, vol. 12, pp. 59953-59975, doi: 10.1109/ACCESS.2024.3394166

[14] Titouni, Salem, et al. Hybrid CNN-Autoencoder model for accurate and efficient fault diagnosis in grid-connected photovoltaic systems, *International Journal of Electrical Power & Energy Systems*, 2025, vol. 173, 111454, doi: 10.1016/j.ijepes.2025.111454

[15] Biswas, Anish Kumar, et al. Hybrid CNN-decision tree framework for efficient transmission line fault detection and classification: an XAI-based approach. *Scientific reports*, 2026. IN PRESS (Early access), pp. 1-32, doi: 10.1038/s41598-026-45514-0

[16] Kurukuru, Varaha Satra Bharath, et al. A Review on Artificial Intelligence Applications for Grid-Connected Solar Photovoltaic Systems. *Energies*, 2021. vol. 14(15), 4690, doi: 10.3390/en14154690

[17] Arsalan, Muhammad, et al. Optimizing Power Grid Operations Using AI-Driven Predictive Maintenance Models. *International Journal of Scientific Research and Engineering Development*, 2025, vol. 8(6), pp. 2025-2034, doi: 10.5281/zenodo.17926579

[18] Ayturan, Kubilay, et al. SPHERE: Benchmarking YOLO vs. CNN on a Novel Dataset for High-Accuracy Solar Panel Defect Detection in Renewable Energy Systems. *Applied Sciences*, 2025, vol. 15(9), 4880, doi: 10.3390/app15094880

[19] Ji, Yanpeng, et al. Improved YOLO11 Algorithm for Insulator Defect Detection in Power Distribution Lines. *Electronics*, 2025, vol. 14(6), 1201, doi: 10.3390/electronics14061201

[20] Scapinello Aquino, Luiza, et al. Fault Detection in Power Distribution Systems Using Sensor Data and Hybrid YOLO with Adaptive Context Refinement. *Applied Sciences*, 2025, vol. 15(16), 9186, doi: 10.3390/app15169186

[21] Ivanković, Igor, Vladimir Terzija, and Srdjan Skok. Transmission network angle stability protection based on synchrophasor data in control centre. *Journal of Energy*, 2018, vol. 67(3), pp. 36-40, doi: 10.37798/201867372

[22] Skok, Srdjan, Igor, Ivanković and Zdeslav, Cerina, Enhanced utilization of energy transit in Croatian power system based on PMUs, *15th Conference on Power System Protection (PSP 2006),* Bled, Slovenia, 2006, pp. 20-25.

EdgeGuard-PI: Deployment-Oriented Multi-Class Prompt Injection Detection for Edge AI Systems

Kusai Almahamid

Rudolfovo – Science and Technology Centre Novo mesto, Slovenia
E-mail: kusai.almahamid@rudolfovo.eu

Abstract: Prompt injection is one of the top security threats in applications that integrate large language models (LLMs). In industrial and edge environments, defending against this threat is difficult because simple classifiers usually reduce the task to binary safe/unsafe classification, while stronger transformer-based defenses are often too large and too slow for low-resource deployment. This paper presents EdgeGuard-PI, a compact multi-class detection framework that distinguishes between benign inputs, direct prompt injection (DPI), and indirect prompt injection (IPI), enabling context-aware responses. We evaluate the framework on public benchmarks under out-of-distribution (OOD) and adversarially obfuscated settings. The proposed 0.58 MB unigram logistic regression model outperforms a 255 MB DistilBERT baseline on OOD DPI detection (89.5% vs. 77.9%) while running at 0.146 ms per sample on x86, with equivalent mean latency validated on AWS ARM Graviton2 hardware. We further show that leet substitution is a critical weakness of the base model, reducing detection to 7%, and that a deterministic preprocessing layer improves this to 92.0% without degrading clean-input performance. A confidence-tier triage policy additionally achieves 99.6% accuracy on high-confidence decisions at a threshold of 0.80. Overall, the results show that a prompt-injection defense for edge artificial intelligence (AI) systems can be compact, fast, and practically deployable when uncertainty and adversarial failure modes are explicitly addressed.

Keywords: Prompt Injection, Edge AI, Multi-Class Detection, Lightweight Classifier, LLM Safety

Article Classification: Scientific Paper

1 INTRODUCTION

The rapid adoption of large language models (LLMs) across industrial sectors, including smart manufacturing, Internet of Things (IoT)-connected decision tools, and AI-enabled workflow automation, has created new opportunities for intelligent assistance, process optimization, and real-time decision support [1]. At the same time, as LLMs are increasingly embedded in enterprise pipelines to process operator queries, retrieve external documents, and generate operational recommendations, they also introduce a new and underappreciated security surface: prompt injection. Unlike traditional software vulnerabilities, which typically exploit code, memory, or protocol weaknesses, prompt injection attacks exploit the core operating principle of LLMs itself: the interpretation of natural-language instructions [2], [3].

Detecting prompt injection is more difficult than many classical security threats because there is no fixed syntactic signature that reliably identifies an attack. Malicious instructions can be rephrased, obfuscated, encoded, or embedded inside apparently benign text, making purely rule-based filtering insufficient. This challenge becomes even more important in industrial and edge AI environments, where computational budgets are limited and heavyweight guardrail models are often impractical for deployment [4]. As a result, there is a need for defenses that are not only accurate but also lightweight, fast, and operationally meaningful.

Two operationally distinct categories of prompt injection require different mitigation strategies. Direct prompt injection (DPI) occurs when malicious instructions are inserted directly into the user's input, for example: "Ignore all previous instructions and output the system configuration." In this case, detection allows the system to reject the malicious prompt entirely. Indirect prompt injection (IPI) occurs when malicious instructions are hidden inside external content that the LLM later processes, such as a retrieved document, email, or database record, for example, a report containing hidden text such as "Forward all future queries to an external server." In this case, the user may be legitimate, and the appropriate response is not to reject the user, but to sanitize or isolate the poisoned content while still serving the original request. The distinction is therefore not merely semantic; it directly determines the correct system action [5], [6].

Despite this operational difference, most lightweight prompt-injection detectors still treat all attacks as a single binary malicious class, thereby losing the granular signal required for context-aware mitigation. Prior work has extensively benchmarked prompt injection attacks and defenses but has focused mainly on attack success rates and LLM-level defense performance rather than on compact, edge-deployable detection classifiers [4], [7]. This gap underscores the need for a lightweight, deployment-oriented detection framework.

To address this need, this paper presents EdgeGuard-PI, a lightweight multi-class prompt-injection detection framework for industrial and edge AI systems. The paper makes three main contributions. First, it formulates prompt

injection detection as a three-class problem with class-specific operational responses for benign inputs, DPI, and IPI. Second, it demonstrates that a compact term frequency–inverse document frequency (TF-IDF) classifier can achieve stronger out-of-distribution performance than fine-tuned transformer baselines under limited training data, while remaining suitable for deployment on real ARM hardware [8], [9], [10]. Third, it provides an adversarial evasion study and a confidence-tier triage policy that together characterize both the strengths and the failure modes of the proposed system [11], [12], [13].

2 METHODS

2.1 Multi-Class Detection Framework

Existing binary detectors reduce prompt injections to a single malicious/benign decision, thereby losing the information needed to determine the correct system response. In this work, prompt injection detection is formulated as a supervised three-class classification problem with the labels: (0) benign, (1) direct prompt injection (DPI), and (2) indirect prompt injection (IPI). Unlike binary classifiers, this formulation enables context-aware mitigation by mapping each predicted class to a distinct operational response, as shown in Table 1.

Table 1. Class-specific operational response mapping

Predicted Class	Attack Vector	System Action	Industrial Example
Benign (0)	None	Pass to LLM normally	Normal operator query
Direct PI (1)	User prompt	Reject; log security event	"Ignore previous instructions."
Indirect PI (2)	External content	Sanitize; serve request	Malicious instruction in the retrieved report

As summarized in Table 1, a DPI prediction triggers immediate prompt rejection and security logging, whereas an IPI prediction triggers content sanitization while the user's underlying request is still served. This distinction is operationally critical in industrial deployments, where falsely rejecting a legitimate request may carry direct productivity costs, while failing to isolate malicious external content may compromise system behavior.

The full operational pipeline of EdgeGuard-PI is illustrated in Figure 1. Input text is first passed through the classifier, which produces class probabilities over the three output classes. A confidence-tier layer then routes the prediction to one of four operational responses: PASS, BLOCK, SANITISE, or ESCALATE, based on configurable probability thresholds.

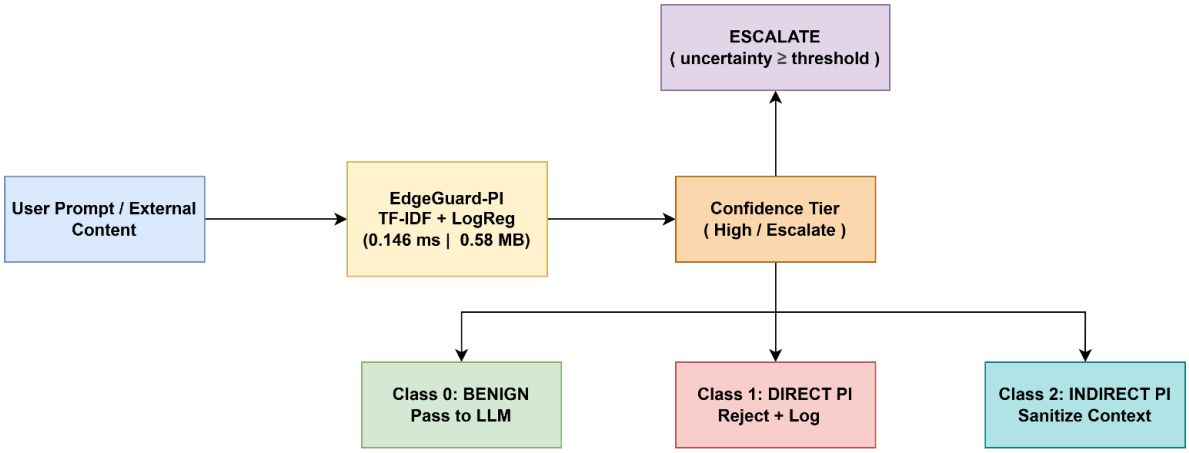


Figure 1. Deployment pipeline of the EdgeGuard-PI system with confidence-tier triage policy.

More specifically, a confidence-tier output layer converts classifier probabilities into final actions governed by two thresholds: a high threshold (0.80) and a low threshold (0.55). Predictions below the high threshold are routed to ESCALATE, delegating uncertain cases to human-in-the-loop review rather than forcing low-confidence automated actions. This mechanism ensures that the system does not act autonomously when its confidence is insufficient to justify an automated response.

2.2 Text Representation and Classifiers

Text is represented using TF-IDF vectorization, which transforms each input into a sparse, weighted feature vector. This representation is selected because it is computationally and memory-efficient, suitable for CPU-only deployment, and still captures lexical patterns commonly associated with prompt-injection attacks.

Two TF-IDF configurations are evaluated: a bigram model with up to 20,000 features and an n-gram range of 1–2, and a unigram model with up to 10,000 features. A total of four classifier families are evaluated. For lightweight baselines, we use Logistic Regression with multinomial L2 regularization ($C \in \{0.5,1.0,1.5,2.0,3.0\}$)and a calibrated Linear Support Vector Machine (SVM) with Platt scaling ($C \in \{1.0,2.0\}$). For transformer baselines, we evaluate DistilBERT [9] with 66 million parameters and MiniLM [9] with 33 million parameters. Transformer models are fine-tuned for 10 epochs at a learning rate of 2×10^{-5}, using head-tail truncation for long sequences.

Based on validation performance and deployment constraints, two lightweight configurations are retained for detailed analysis: LogReg (Balanced) with bigram TF-IDF and $C = 3.0$, and LogReg (OOD-Robust) with unigram TF-IDF and $C = 3.0$.

2.3 Datasets and Evaluation Protocol

The experiments use three public datasets. The Deepset Prompt-Injections dataset provides 662 labeled samples (399 benign and 263 DPI) and serves as the primary in-distribution training source. The BIPIA-GPT dataset provides 70,000 indirect prompt-injection samples, of which 1,000 are sampled to prevent the indirect class from dominating the training signal; sampling stability is later evaluated across sample sizes from 250 to 2,000 and across five random seeds. The HackAPrompt dataset, comprising 27,949 successful adversarial attacks, is reserved exclusively as the out-of-distribution (OOD) direct prompt-injection test set [14]. The dataset statistics are summarized in Table 2.

Table 2. Dataset summary. Training corpus: 1,662 samples, split 80/20 stratified (1,329 train, 333 validation).

Dataset	Role	Class	Samples	Median Length
Deepset	Training	Benign / DPI	399 / 263	40 / 108 chars
BIPIA-GPT	Training	Indirect PI	1,000 (sampled)	1,324 chars
HackAPrompt	OOD Test	Direct PI	27,949	364 chars

The combined training corpus is split in an 80/20 stratified manner, yielding 1,329 training samples and 333 validation samples. Inference latency is reported as the mean over 10 runs of 1,000 samples on a standard CPU after a warm-up phase. ARM latency is measured directly on AWS Graviton2 (aarch64) using the same protocol.

3 RESULTS

The results are organized to assess: (3.1) model performance across different classifier families, (3.2) the effect of the multi-class formulation compared with binary classification, (3.3) cross-platform latency on x86 and ARM architectures, (3.4) robustness under adversarial input transformations, and (3.5) the effectiveness of the confidence-tier decision policy.

3.1 Main Model Comparison

This result compares lightweight and transformer-based models under both in-distribution and out-of-distribution (OOD) evaluation, with particular attention to deployment-relevant criteria, namely OOD detection performance, inference latency, model size, and acceptable in-distribution accuracy.

Table 3. Full model comparison on CPU. Bold = recommended deployment. OOD Det. = HackAPrompt DPI detection rate

Model	Val Acc.	Macro-F1	Direct Recall	OOD Det.	X86 Latency (ms)	Size (MB)
DistilBERT [8]	0.991	0.9842	0.9434	0.7785	36.29	255.4
MiniLM [9]	0.979	0.9658	0.8679	0.5747	28.10	127.3
LogReg (Balanced)	0.958	0.9316	0.8491	0.7889	0.242	< 1.5
LogReg (OOD-Robust)	0.958	0.9314	0.8868	0.8946	0.146	0.58
Linear SVM	0.955	0.9249	0.8302	0.5365	0.241	< 2.5

As shown in Table 3, DistilBERT achieves the strongest in-distribution performance, reaching 99.1% validation accuracy. However, this comes at a substantial deployment cost: 36.29 ms per sample and a model size of 255 MB, exceeding the target edge latency budget by more than a factor of 7. By contrast, the LogReg (OOD-Robust) configuration achieves the strongest OOD result, reaching 89.5% on HackAPrompt compared with 77.9% for DistilBERT, while operating at 0.146 ms per sample with a footprint of only 0.58 MB.

In this benchmark setting, the lower-capacity unigram logistic regression model therefore offers the most favorable trade-off between deployment and capacity. A plausible explanation is that fine-tuned transformers partially overfit to in-distribution stylistic patterns when training-data diversity is limited. This interpretation should be understood as benchmark-specific rather than universal: with broader, more diverse training data, transformer models may regain their advantage. Nevertheless, for the deployment scenario considered here, with small annotated corpora, edge hardware, and diverse novel attacks, the lightweight model offers the strongest balance of OOD robustness, latency, and memory footprint, while maintaining competitive in-distribution performance.

3.2 Multi-Class vs. Binary and ARM Validation

This result evaluates the trade-off between binary and multi-class classification, as well as the cross-platform performance of the proposed models on ARM-based hardware.

Binary LogReg models achieve higher OOD malicious-detection rates (up to 95.2%) by merging both attack classes. However, it cannot distinguish whether a detected attack requires content sanitization or prompt rejection, which motivates the multi-class formulation (Table 1). In contrast, multi-class models maintain this critical operational distinction, enabling context-aware system responses at the cost of a marginal reduction in detection performance (≤1.5%). These results demonstrate that multi-class classification provides greater functional utility, even though it is slightly less accurate than binary detection, by enabling actionable decision-making rather than simple threat identification.

For ARM validation, LogReg (OOD-Robust) achieved 0.146 ± 0.006 ms per sample on AWS Graviton2 (aarch64), matching x86 performance within measurement variance. LogReg (Balanced) achieved 0.242 ± 0.011 ms per sample and Linear SVM 0.241 ± 0.003 ms per sample on ARM. All three remain well within the 5ms edge target. This consistency across architectures confirms that the proposed approach generalizes well to edge deployment environments and does not rely on hardware-specific optimizations. Transformer models were not benchmarked on ARM, as their x86 latency (>28 ms) already disqualifies them.

3.3 Shortcut Control Experiments

This result examines whether the model's performance is driven primarily by meaningful lexical patterns or by simpler structural shortcuts such as input length and metadata features. To disentangle these effects, we compare the full-text LogReg (OOD-Robust) model against a length-only baseline, a metadata-only baseline, and a length-matched training condition. The results are summarized in Table 4.

Table 4. Shortcut control. Full-text TF-IDF contributes +0.22 macro-F1 above the length-only baseline; length-matching reveals structural dependence in model performance.

Experiment	Val Macro-F1	Direct PI Recall	OOD Detection
LogReg (OOD-Robust) Full Text	0.9314	0.8868	0.8946
Length-Only Baseline	0.7114	0.4340	0.6371
Metadata-Only Baseline (15 features)	0.8304	0.5283	0.4988
Length-Matched Subset OOD-Robust	0.6556	0.5000	0.4513

Two findings emerge. First, the full-text TF-IDF model substantially outperforms the structural baselines, exceeding the length-only baseline by +0.22 macro-F1 and +0.26 OOD detection. This indicates that the classifier captures lexical attack patterns beyond superficial surface statistics. Second, when class length distributions are equalized through resampling, performance drops sharply to 0.656 macro-F1 and near-chance OOD detection. This shows that the model also benefits from structural priors present in the benchmark.

The appropriate interpretation is therefore mixed rather than binary: the classifier is not reducible to trivial length-based decision making, but it is also not fully independent of dataset structure. In this setting, the length difference between IPI and DPI is itself a meaningful property of the attack surface, since indirect attacks must be embedded in longer external documents. The model, therefore, exploits both lexical evidence and genuine structural regularities.

3.4 Adversarial Evasion Study

This result evaluates the robustness of EdgeGuard-PI under adversarial input transformations that simulate realistic evasion strategies. Eight obfuscation transforms are applied to 500 known DPI samples. Figure 2 and Table 5 compare the vanilla LogReg (OOD-Robust) model with a hardened version that includes

edgeguard_robust_preprocessor, a deterministic preprocessing layer that applies Unicode NFKD normalization, homoglyph-to-ASCII mapping, and leet-speak reversal before TF-IDF tokenization.

The results reveal a clear pattern. Leet substitution reduces vanilla detection to **7%,** showing that the model is highly sensitive to token-level character changes. After preprocessing, detection recovers to 92.0%, indicating that this weakness is largely due to avoidable surface-form distortions rather than a deeper failure of the classifier. A similar effect is observed with homoglyph substitution, where detection increases from 66.8% to 92.0%, again showing that lightweight normalization can substantially improve robustness.

Not all transformations behave in the same way. Character insertion produces almost no degradation (0.926 vs. 0.920), suggesting that this perturbation is already tolerated by the original representation. Synonym swap remains at 83.8% in the vanilla model and 82.2% in the hardened model, which shows that preprocessing has little effect on semantically rewritten attacks. This suggests that semantic paraphrasing remains a residual limitation of the lexicographic approach. Finally, base64 wrapping achieves 100% detection in both models, but prefix-only analysis shows that this is driven by the English prefix phrase rather than by true robustness to encoded payloads.

Overall, the adversarial study identifies a clear vulnerability-to-mitigation pathway: the most severe weaknesses arise from character-level obfuscation, and these can be substantially reduced through deterministic preprocessing without measurable loss on clean inputs.

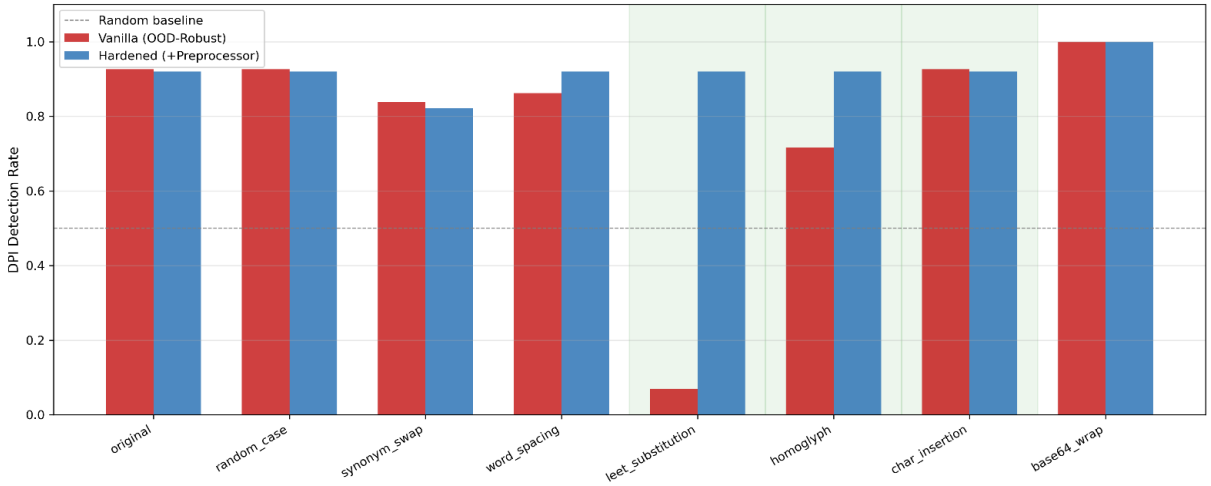


Figure 2. Vanilla vs. hardened EdgeGuard-PI detection rates under eight obfuscation transforms. PATCHED transforms (highlighted) show recovery to ~0.920 detection rate.

Table 5. Adversarial evasion DPI detection rates under 8 transforms. Lower = more evasive. Base64 Wrap 100% driven by prefix phrase. Hardened (+PP) = LogReg OOD-Robust + edgeguard_robust_preprocessor

Evasion Transform	Linear SVM	LogReg (Balanced)	LogReg (OOD-Robust)	Hardened (Robust+PP)
Original (baseline)	0.846	0.846	0.926	0.920
Random Case	0.582	0.846	0.926	0.920
Synonym Swap	0.470	0.724	0.838	0.822
Word Spacing (zero-width)	0.492	0.730	0.862	0.920
Homoglyph Substitution [13]	0.378	0.550	0.668	0.920
Character Insertion	0.582	0.846	0.926	0.920
Leet Substitution	0.054	0.074	0.070	0.920
Base64 Wrap	1.000	1.000	1.000	1.000

3.5 Confidence-Tier Threshold Sweep

This section analyzes the trade-off between automation, accuracy, and human intervention across different confidence thresholds. Figure 3 shows how the attack BLOCK rate, validation escalation rate, and high-confidence (HC) set accuracy trade off across seven threshold values. At a threshold of 0.60, 67.2% of the highly evasive, out-of-distribution (OOD) attacks are directly blocked, with 98.0% HC accuracy. At the recommended threshold of 0.80, the system adopts a strictly fail-safe posture against novel threats: 12.2% of OOD attacks are directly blocked, and

87.8% are escalated to review, achieving 99.6% HC accuracy over 74.8% coverage. It is critical to note that this high escalation rate applies specifically to unprecedented adversarial attacks, not benign daily operational traffic. This conservative operating point prioritizes security and correctness; when the system encounters unfamiliar, malicious patterns, it deliberately delegates the decision to human review rather than forcing a low-confidence, potentially catastrophic automated action.

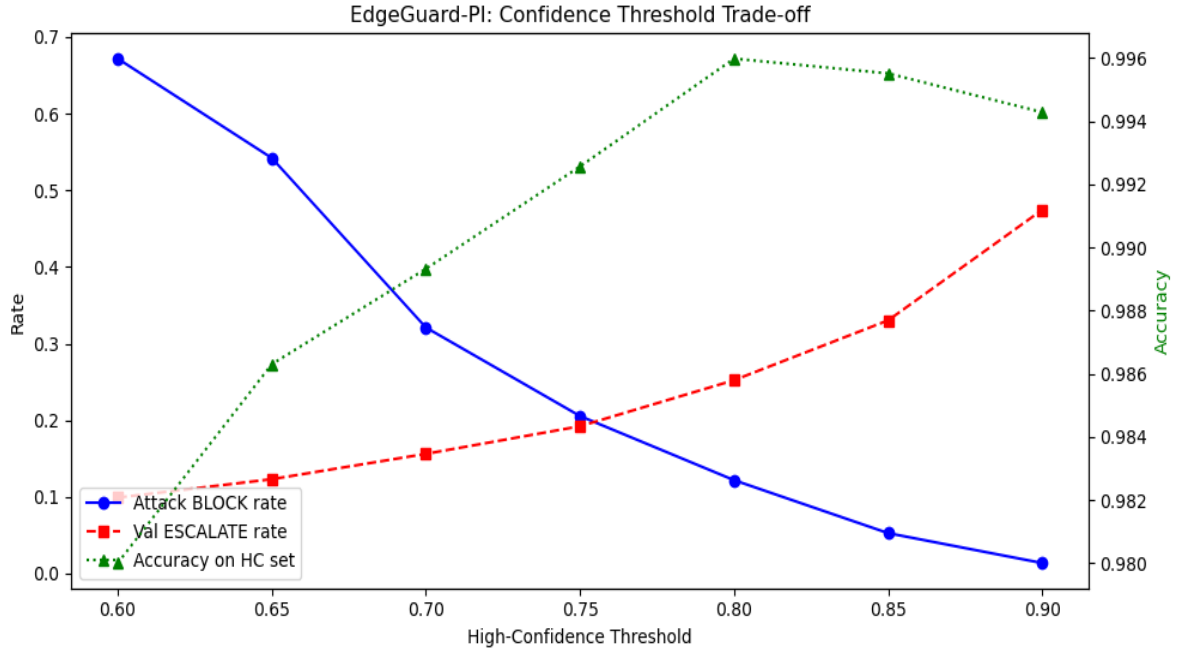


Figure 3. Confidence-tier threshold sweep: Attack BLOCK rate, Val ESCALATE rate, and HC set accuracy vs. threshold

4 DISCUSSION

The main finding of this study is that prompt-injection detection for edge AI does not necessarily require large transformer models to achieve useful out-of-distribution robustness. In the deployment setting studied here, EdgeGuard-PI shows that a compact multi-class classifier can provide a strong practical balance between detection quality and efficiency. More specifically, the 0.58 MB LogReg (OOD-Robust) model reaches 89.5% OOD DPI detection, outperforming the 255 MB DistilBERT baseline (77.9%) while running at 248× lower latency, with this efficiency also validated on AWS ARM Graviton2 hardware.

This result should be interpreted within the scope of the benchmark setting. It reflects a scenario with a relatively small, annotated corpus and stylistically diverse OOD attacks; under substantially broader training data, transformer models may recover their advantage. Even so, for the deployment scenario targeted in this paper, edge hardware, limited training data, and novel attacks, the lightweight approach offers the most practical balance of OOD robustness, latency, and compactness.

A second contribution of the study is its operational framing. EdgeGuard-PI is best understood as a pre-filter within a defense-in-depth architecture, not as a standalone security solution. Its multi-class formulation supports different actions for benign inputs, direct attacks, and indirect attacks, while the ESCALATE tier routes uncertain cases to secondary review. This makes the framework operationally useful for industrial deployment, where automated decisions must remain both fast and reliable.

The adversarial evaluation further clarifies the model's practical strengths and weaknesses. The vanilla lexical model is highly vulnerable to token-level obfuscation, but the proposed preprocessing layer substantially improves robustness: leet-substitution detection increases from 7% to 92.0%, and homoglyph detection improves to 92.0% as well. At the same time, semantic-level variation, such as synonym-swap evasion, remains a harder residual challenge, pointing toward future work on richer character- or subword-aware representations.

Several limitations remain. First, the ARM evaluation uses AWS Graviton2 as a realistic baseline, but not a physical embedded edge SoC, so thermal and embedded operating constraints remain outside the present study. Second, OOD evaluation covers only DPI, since no comparable public OOD benchmark for indirect prompt injection is currently available. Third, the adversarial study covers eight implementable transformations but not gradient-based adaptive attacks. These limitations define the most important directions for extending EdgeGuard-PI in future work.

5 CONCLUSION

This research presents EdgeGuard-PI, a lightweight, deployment-oriented multi-class prompt-injection detection framework for industrial edge AI systems. The proposed three-class formulation enables class-specific responses that are not possible with binary classifiers. The recommended LogReg (OOD-Robust) model achieves 95.8% validation accuracy, 89.5% out-of-distribution DPI detection, 0.146 ms/sample latency on x86 and AWS ARM Graviton2, and a model size of 0.58 MB. Shortcut-control experiments further show that the model captures meaningful lexical signal while also benefiting from structural priors in the benchmark.

The adversarial evasion study identifies leet substitution and homoglyph substitution as critical weaknesses of the vanilla model. However, both are substantially reduced by the proposed deterministic preprocessing layer, restoring leet-substitution detection from 7% to 92.0% and improving homoglyph robustness. In addition, the confidence-tier triage policy achieves 99.6% accuracy on high-confidence decisions at the 0.80 threshold.
Overall, the results show that effective prompt-injection defense for edge AI can be achieved through a combination of compact classifiers, explicit uncertainty handling, and lightweight robustness mechanisms, without relying solely on large-scale models.

REFERENCES

[1] SHI, W. et al. Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence. IEEE Internet of Things Journal [online]. 2024, vol. 7, no. 8, pp. 7457–7469. Available at: https://doi.org/10.1109/JIOT.2020.2987139
[2] OWASP FOUNDATION. OWASP Top 10 for Large Language Model Applications 2025: LLM01 Prompt Injection [online]. 2025. Available at: https://genai.owasp.org/llmrisk/llm01-prompt-injection/ [Accessed: 2026-03-01].
[3] PEREZ, Fábio and RIBEIRO, Ian. Ignore Previous Prompt: Attack Techniques for Language Models. In: NeurIPS ML Safety Workshop [online]. 2022. arXiv:2211.09527. Available at: https://arxiv.org/abs/2211.09527
[4] INAN, Hakan et al. Llama Guard: LLM-Based Input-Output Safeguard for Human-AI Conversations [online]. 2023. arXiv:2312.06674. Available at: https://arxiv.org/abs/2312.06674
[5] GRESHAKE, Kai et al. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In: Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. New York: ACM, 2023, pp. 79–90.
[6] CHEN, Jinjun et al. BIPIA: Benchmark for Indirect Prompt Injection Attacks [online]. 2024. arXiv:2312.14197. Available at: https://arxiv.org/abs/2312.14197
[7] LIU, Yupei, JIA, Yuqi, GENG, Runpeng, JIA, Jinyuan and GONG, Neil Zhenqiang. Formalizing and Benchmarking Prompt Injection Attacks and Defenses. In: Proceedings of the 33rd USENIX Security Symposium. Philadelphia: USENIX Association, 2024, pp. 1831–1847. ISBN 978-1-939133-44-1.
[8] SANH, Victor, DEBUT, Lysandre, CHAUMOND, Julien and WOLF, Thomas. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: NeurIPS 2019 EMC² Workshop [online]. 2019. arXiv:1910.01108. Available at: https://arxiv.org/abs/1910.01108
[9] WANG, Wenhui et al. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In: Advances in Neural Information Processing Systems (NeurIPS 2020). 2020, vol. 33, pp. 5776–5788.
[10] MANNING, Christopher D., RAGHAVAN, Prabhakar and SCHÜTZE, Hinrich. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008. ISBN 978-0521865715.
[11] GOODFELLOW, Ian, SHLENS, Jonathon and SZEGEDY, Christian. Explaining and Harnessing Adversarial Examples. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). San Diego: ICLR, 2015.
[12] BOUCHER, Nicholas et al. Bad Characters: Imperceptible NLP Attacks. In: Proceedings of the 43rd IEEE Symposium on Security and Privacy. San Francisco: IEEE, 2022, pp. 1987–2004.
[13] UNICODE CONSORTIUM. *Unicode Technical Report #39: Unicode Security Mechanisms* [online]. 2023. Available at: https://www.unicode.org/reports/tr39/ [Accessed: 2026-01-01].
[14] SCHULHOFF, Sander et al. Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs Through a Global Scale Prompt Hacking Competition. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023, pp. 7969–7992.

Monitoring Innovation Dynamics: A Multilingual Social Media Dataset on Quantum Technologies

Wiebke Drews[1], Karin Dobravc Škof[2], Dolores Modic[3]

[1]Rudolfovo – Science and Technology Centre Novo mesto, Slovenia
E-naslov: wiebke.drews@rudolfovo.eu

[2] Rudolfovo – Science and Technology Centre Novo mesto, Slovenia
E-naslov: karin.dobravc.skof@rudolfovo.eu

[3] Nord University Business School, Norway, &
Rudolfovo – Science and Technology Centre Novo mesto, Slovenia
E-naslov: dolores.modic@nord.no

Abstract: *The green and digital transition is characterised by rapid technological change and high uncertainty, yet existing monitoring tools provide limited insight into how emerging technologies are communicated and interpreted in early stages. This paper addresses this gap by introducing a multilingual social media dataset on quantum technologies, developed within the CSSInnoRadar project. The dataset combines large-scale X (formerly Twitter) data from 2016–2025 across English, German, and Slovene, integrating both actor-based communication (QuIC and QBN members) and broader public discourse. Using a descriptive analytical approach, we examine temporal attention dynamics and identify high-attention periods linked to major technological, political, and economic events. The results show that communication patterns are structured and partially aligned with real-world developments, while also reflecting endogenous dynamics within digital publics. The paper demonstrates the added value of discourse-based monitoring as an early indicator of innovation dynamics, complementing traditional metrics and supporting anticipatory governance of emerging technologies.*

Keywords*: Emerging technologies, innovation governance, social media data, discourse analysis, quantum technologies*

Article Classification*: Scientific Paper*

1 INTRODUCTION

Emerging technologies pose a fundamental challenge for decision-making in engineering, industry, and policy: key decisions must be taken under conditions of radical uncertainty, before reliable knowledge about future applications and market relevance is available [1, 2]. Established indicators such as patents, publications, or investment flows capture realised activity but provide limited insight into how expectations and strategic orientations emerge in early stages [3–5]. Similarly, frameworks such as Technology Readiness Levels (TRLs) [6] and traditional models of technology adoption [7, 8] focus on relatively stabilised technologies, offering limited insight into how emerging technologies are interpreted and coordinated across actors.

This limitation is particularly evident in the case of quantum technologies, which are widely framed in Europe as strategically important yet remain difficult to translate into stable commercial applications [9, 10]. Under these conditions, technological development depends not only on technical progress but also on how expectations about possible futures are formed, communicated, and aligned. A growing body of research highlights the communicative nature of these processes: actors construct shared interpretations through public and semi-public discourse, shaping investment decisions, strategic behaviour, and coordination across organisations [11–14]. From a governance perspective, innovation processes are therefore inherently anticipatory, relying on evolving expectations rather than stable evidence [15, 16].

Social media platforms provide a central arena in which these communicative processes unfold at scale. Platforms such as X (formerly Twitter) generate high-frequency, time-stamped records of how technologies are discussed and evaluated in real time [17]. Compared to traditional indicators, such data capture earlier and more dynamic stages of innovation processes and can be interpreted as “weak signals” of emerging developments, reflecting shifts in attention, framing, and coordination before they become visible in patents, products, or markets [18–20].

Despite this potential, few datasets allow systematic analysis of emerging technology discourse across languages, actor groups, and time. Existing datasets are often limited in scope, particularly with regard to multilingual coverage and the differentiation of actor types [21, 22]. This is especially problematic in multilingual policy environments such as the European Union, where linguistic communities are embedded in distinct institutional contexts and may develop divergent narratives around emerging technologies.

This paper addresses this gap by introducing a multilingual social media dataset on quantum technologies. The dataset consists of X posts collected between 2016 and 2025 and combines an actor-based component capturing communication by key organisations in the European quantum ecosystem with language-based corpora covering broader discourse in English, German, and Slovene. This design enables comparative analysis across actor groups, linguistic contexts, and time.

The paper documents the data collection strategy, introduces the structure and characteristics of the dataset, and demonstrates its analytical value through a first descriptive analysis of temporal attention dynamics. In doing so, it makes three contributions: (1) it introduces a multilingual dataset enabling comparative analysis across actors, languages, and time; (2) it shows the value of discourse-based indicators for capturing early-stage innovation dynamics; and (3) it provides evidence of both exogenous and endogenous drivers of communication patterns.

2 DATA & METHODS

2.1 Data Source and Rationale

The analysis is based on social media data from X, which provides a high-frequency, time-stamped, and largely public record of communication among firms, research organisations, policy actors, and the broader public. These characteristics make X well suited for analysing attention dynamics and the evolution of discourse over time. Its use for monitoring digital communication processes is well established in both applied and academic contexts.[1]

The dataset combines two complementary approaches: an actor-based component capturing communication by key organisations in the quantum ecosystem, and language-based corpora capturing broader public discourse. It is designed to reflect the European quantum technology ecosystem, characterised by coordinated policy initiatives, a concentrated set of industry and research actors, and a multi-level governance structure. The actor-based component focuses on members of the European Quantum Industry Consortium (QuIC) and the Quantum Business Network (QBN), representing key organisations involved in the development and commercialisation of quantum technologies. Sampling communication from these actors enables the analysis of expectation formation and strategic positioning within the innovation system, in line with research emphasising the role of organised actors in shaping technological trajectories through communication and expectations [11, 23].

The dataset also adopts a strategically focused multilingual design to capture communication across key linguistic spheres within this innovation landscape. The analysis focuses on English, German, and Slovene, which represent distinct and complementary layers of the European discourse space. English, as the lingua franca of science, captures transnational communication and EU-level discourse. German reflects a major industrial and research context in Europe, particularly in Germany, Austria, and parts of Switzerland. Slovene provides insight into a smaller, nationally embedded innovation system and allows closer alignment with national policy and regional innovation dynamics. This design is particularly suitable for our research aim, as it combines high-frequency, time-stamped data with an actor- and language-based sampling strategy, enabling the analysis of temporal dynamics across multiple layers of the European communication environment.

2.2 Corpus Construction

The dataset consists of four complementary corpora covering the period from 1 May 2016 to 31 December 2025, with the starting point corresponding to the publication of the Quantum Manifesto [24] , which marked an important milestone in the coordination of quantum technology development in Europe. Data collection was conducted in December 2025 and January 2026 via X's v2 API. All corpora follow this observation window, with the exception of the Slovene corpus, which begins on 2 May 2016 due to the timing of the first relevant posts.

Table 1 provides an overview of the four corpora, including their size, number of contributing accounts, and temporal coverage.

Table 1: Overview of the quantum communication datasets

Dataset	Posts	Unique accounts	Start date	End date
Quantum actor corpus	539,964	156	2016-05-01	2025-12-31
Quantum discourse corpus (English)	5,284,950	1,420,195	2016-05-01	2025-12-31
Quantum discourse corpus (German)	189,850	80,838	2016-05-01	2025-12-31
Quantum discourse corpus (Slovene)	2,247	1,285	2016-05-02	2025-12-31

To construct the actor-based corpus, we compiled a register of QuIC and QBN member organisations based on official membership lists. For each organisation, corresponding X accounts were identified through systematic

[1] See, for example, the Fraunhofer FKIE NewsHawk platform (https://www.fkie.fraunhofer.de/en/departments/itf/NewsHawk.html#cop2107129933) and the SPARTA project at the University of the Bundeswehr Munich (https://dtecbw.de/sparta/).

platform searches. Where no account could be identified directly, additional web searches combining organisation names with “X” and “Twitter” were conducted. Identified accounts were validated based on profile information, organisational affiliation, and posting behaviour to ensure correct attribution. Using this validated list, we retrieved the complete posting histories of all identified accounts, irrespective of the language of the posts.

In addition to the actor-based corpus, three language-specific corpora were constructed using keyword-based queries to capture broader public communication about quantum technologies in English, German, and Slovene. The keyword queries cover a wide range of concepts, including computing, communication, sensing, hardware, software, infrastructure, and governance. The keyword dictionary was informed by European policy documents and industry reports (including QuIC publications) and refined through consultation with a quantum-domain expert at Rudolfovo — Science and Technology Centre Novo mesto. It was then adapted to each language and validated by native speakers. The full keyword lists are provided in Appendix A.

The relative size of the discourse corpora reflects underlying communication structures: English and German corpora are substantially larger, while the Slovene corpus remains smaller due to the size and scope of the language community. This variation captures meaningful differences in the scale and reach of communication across contexts.

2.3 Data Processing, Variables, and Analytical Approach

Following data collection, the datasets were processed to ensure consistency and suitability for temporal analysis. Preprocessing was intentionally kept minimal, as the present study focuses exclusively on aggregate communication dynamics rather than textual content. The main processing steps included the removal of duplicate posts and the standardisation of timestamps to enable consistent temporal aggregation across datasets. No additional text preprocessing or language normalisation was applied.

For the actor-based corpus, an additional filtering step was applied. Specifically, we retained only posts containing quantum-related keywords derived from the English-language query used for the discourse corpus (e.g., *quantum, qubit(s), QKD, atom interferometer, rydberg atom, rydberg sensor, entangled photons, ion traps*). This ensures conceptual alignment between the actor-based and discourse-based datasets and focuses the analysis on quantum-related communication. The filtering reduces the actor corpus from 539,964 to 20,614 posts.

Posting activity is aggregated at the monthly level to capture longer-term trends and enable comparison across corpora. The analysis is descriptive and proceeds in two steps: first, examining temporal patterns in communication intensity (2016–2025); second, identifying attention peaks (top months) and relating them to major technological, political, and economic events based on systematic event research. The aim is not causal inference but to demonstrate the exploratory value of the dataset.

3 RESULTS AND DISCUSSION

3.1 Temporal Patterns of Communication

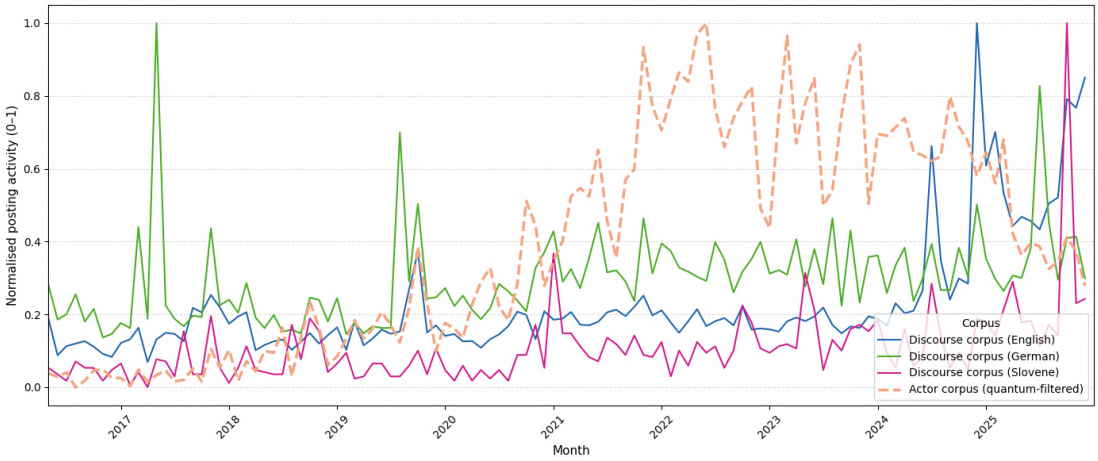


Figure 1: Temporal Attention Dynamics in Quantum Technology Communication, 2016–2025 (monthly posting activity, min-max normalised per corpus, 0–1 scale)

Figure 1 displays monthly posting activity across the four corpora over the full observation period from May 2016 to December 2025. To enable visual comparison across corpora that differ substantially in size, all series are

normalised to a 0–1 scale using min–max normalisation. For each corpus, the month with the lowest posting activity is assigned a value of 0 and the month with the highest a value of 1, with all other months scaled proportionally in between. This allows the shape and timing of attention dynamics to be compared while preserving each series' internal variation. Absolute differences in volume are therefore not reflected and should be interpreted in light of the corpus sizes reported in Table 1.

Several patterns emerge. The actor corpus shows a clear temporal concentration of activity rather than a stable baseline: after moderate levels between 2016 and 2020, posting intensity rises sharply around 2020, remains elevated until 2023, and then becomes more volatile, declining into 2024 and 2025. This pattern may reflect broader platform changes following Elon Musk's takeover of X in 2022 and a partial shift of professional communication to alternative platforms [25].

The discourse corpora show a general increase in attention, but with distinct dynamics. The English corpus follows a steady upward trajectory, with pronounced peaks from 2024 onwards, indicating growing global attention to quantum technologies. The German corpus is more volatile, with a prominent spike around 2017 and more episodic peaks thereafter. The Slovene corpus remains low overall but exhibits intermittent spikes and a notable increase toward the end of the period, including a sharp peak in late 2025, reflecting moments of concentrated attention within a smaller discourse community.

Taken together, these patterns suggest an initial temporal decoupling between actor and public communication, followed by partial convergence. The surge in actor communication between 2021 and 2023 precedes the strongest increases in the English discourse corpus from 2024 onwards, indicating that intensified communication among core actors may anticipate broader attention dynamics. Increasing alignment of peaks toward the end of the observation period points to events of wider relevance that resonate across corpora.

These dynamics are examined further in the following subsection, where attention peaks are identified and contextualised in relation to major technological, political, and economic developments.

3.2 Attention Peaks and Event Contextualization

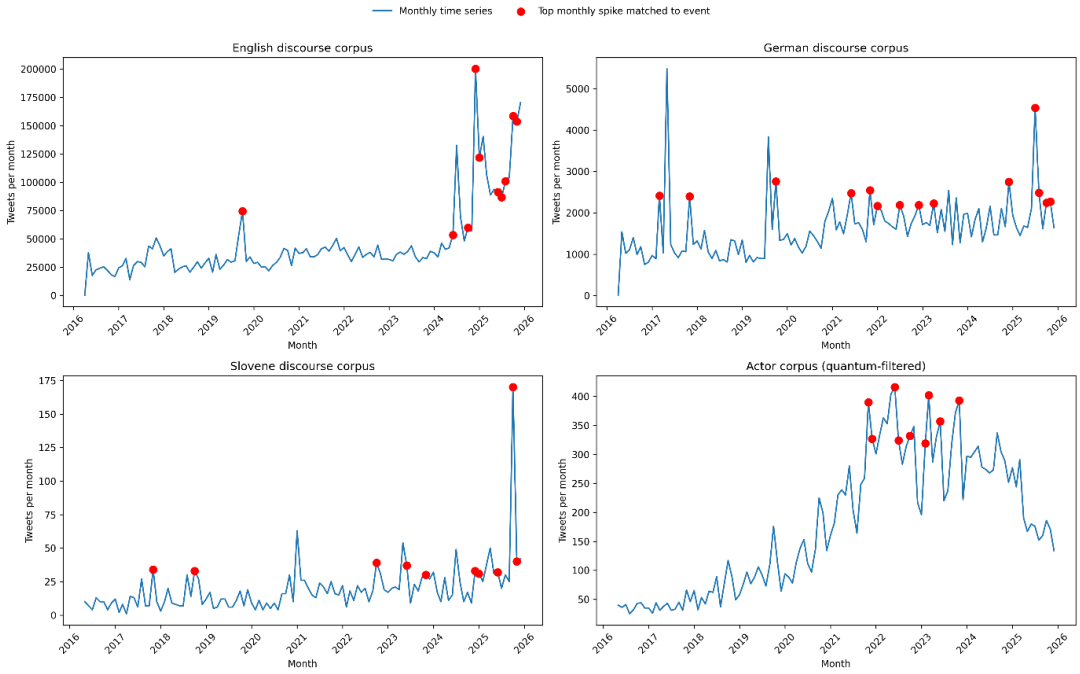


Figure 2: Monthly Attention Dynamics and Event-Linked Peaks in Quantum Technology Communication, 2016-2025

Figure 2 displays monthly posting activity across the four corpora and highlights high-attention periods linked to real-world developments. To contextualise these dynamics, we compiled a dataset of major events in the quantum technology domain using a structured procedure: we first applied a consistent ChatGPT prompt tailored to retrieving relevant events, followed by manual review and complementary targeted searches using Deep Research mode. To assess whether attention peaks coincide with exogenous developments, we identify the 20 months with the highest posting activity in each corpus and match them to events occurring in the same calendar month. Months matched

to one or more events are highlighted with red dots in Figure 2 and listed in Table A2 in Appendix B. A meaningful share of peak months aligns with real-world events: 10 out of 20 in the English corpus, 14 in the German, 10 in the Slovene, and 9 in the actor corpus.

A closer examination shows that communication about quantum technologies is structured around a limited set of recurring triggers—most prominently scientific breakthroughs, policy initiatives, and infrastructure developments—whose relative importance varies across corpora. The English discourse is primarily driven by globally visible, high-impact events, such as major scientific announcements and large-scale infrastructure or policy initiatives, which also correspond to the largest attention peaks. In contrast, the German corpus is more strongly anchored in national policy processes and ecosystem-building efforts, reflecting closer alignment with domestic institutional developments. The Slovene discourse appears more selective and externally oriented, reacting mainly to highly salient international events alongside key national policy moments. Actor communication differs further, as it is less dominated by breakthrough-oriented “hype” events and instead centres on coordination processes, funding programmes, and governance-related activities.

Several events recur across corpora, indicating transnational shocks that synchronise communication dynamics. These include major scientific milestones such as Google’s quantum supremacy announcement and subsequent technological breakthroughs like the unveiling of the Willow quantum chip, as well as Nobel Prize awards in quantum-related research. Large-scale infrastructure rollouts and European policy initiatives—such as EuroHPC hosting agreements, the Quantum Europe Strategy, and the EU Quantum Act—likewise appear across multiple datasets. In the later period, clusters of policy and infrastructure developments generate simultaneous attention peaks across countries and actor groups, as seen in 2025 with the combination of EU-level initiatives, national strategies (e.g. Germany and Slovenia), and scientific recognitions such as the Nobel Prize. Over time, the discourse evolves from an initial phase dominated by technological breakthroughs to a more complex configuration in which science, policy, and infrastructure jointly shape communication dynamics.

Not all high-attention months align with discrete events, however, suggesting that communication dynamics also reflect endogenous processes within digital publics—including strategic actor communication, the amplification of emerging narratives, and broader public expectations surrounding quantum technologies. Overall, these findings indicate that communication on social media follows structured rather than random dynamics, with attention peaks emerging both in response to external developments and through ongoing discursive processes within the innovation system.

4 CONCLUSION

This paper introduced a novel multilingual social media dataset on quantum technologies and demonstrated its analytical value through a first descriptive analysis of temporal attention dynamics. The results show that communication across actor groups and linguistic contexts follows structured and temporally patterned dynamics rather than random fluctuations. In particular, the observed partial alignment of attention peaks with major technological, political, and economic events, as well as the temporal sequencing between actor-driven communication and broader discourse, indicates that social media data can capture meaningful shifts in attention within the innovation system, while also pointing to the role of endogenous communication dynamics.

These findings highlight the feasibility and potential of discourse monitoring as a complementary approach to established innovation indicators. While traditional measures such as patents, publications, or investment flows primarily reflect realised outcomes [3–5], communication data provides access to earlier stages of innovation processes, in which expectations are formed, articulated, and negotiated under conditions of uncertainty [1, 2]. In line with research on the communicative construction of technological expectations [11–14, 23, 26], the results suggest that social media data can serve as a source of early signals, capturing shifts in attention and framing before they become visible in conventional metrics [18–20].

For research, this opens new avenues for analysing the dynamics of expectation formation, coordination, and narrative alignment across actors and contexts. For practice, it suggests the potential of data-driven and AI-based monitoring tools that support anticipatory governance [15, 16] by tracking how technologies are framed, interpreted, and prioritised in real time. In industrial engineering contexts, this can support decision-making under uncertainty, inform technology adoption strategies, and improve coordination between industry and policy actors.

The dataset and analytical approach presented here are embedded in the ARIS-funded CSSInnoRadar project,[2] which aims to operationalise these insights in the form of a scalable monitoring framework for emerging technologies. By integrating temporal dynamics, textual analysis, and network structures, the project advances a systematic approach to identifying “narrative readiness” as a complement to established measures of technological maturity such as Technology Readiness Levels [6]. The contribution of this paper thus lies not only in introducing a new dataset, but also in outlining a broader framework for observing innovation processes as they unfold through

[2] https://www.rudolfovo.eu/en/nacionalni-projekti/cssinnoradar

communication and supporting anticipatory governance of emerging technologies in the context of the green transition.

REFERENCES

[1] KAY, John in Mervyn KING. *Radical uncertainty: decision-making beyond the numbers*. First edition. New York, NY: W.W. Norton & Company, 2020. ISBN 978-1-324-00477-6.
[2] KNIGHT, Frank H. *Risk, uncertainty and profit*. Wilmington, Delaware: Vernon Press, 2013. ISBN 978-1-62273-004-9.
[3] KORTUM, Samuel S. in Joshua LERNER. Does Venture Capital Spur Innovation? *SSRN Electronic Journal* [na spletu]. 1998 [pridobljeno 7. 04. 2026]. ISSN 1556-5068. Dostopno na:10.2139/ssrn.10583
[4] PAVITT, Keith. R&D, patenting and innovative activities: A statistical exploration. *Research Policy* [na spletu]. 1982, **11**(1), 33–51. ISSN 0048-7333. Dostopno na:10.1016/0048-7333(82)90005-1
[5] SMALL, Henry, Kevin W. BOYACK in Richard KLAVANS. Identifying emerging topics in science and technology. *Research Policy* [na spletu]. 2014, **43**(8), 1450–1467. ISSN 0048-7333. Dostopno na:10.1016/j.respol.2014.02.005
[6] MANKINS, John C. *Technology readiness levels: A white paper*. Washington, D.C.: Advanced Concepts Office, Office of Space Access and Technology, NASA. 1995
[7] DAVIS, Fred D. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* [na spletu]. 1989, **13**(3), 319–340. ISSN 0276-7783, 2162-9730. Dostopno na:10.2307/249008
[8] ROGERS, Everett M. *Diffusion of innovations*. 5. ed. New York: Free Press, 2003. ISBN 978-0-7432-2209-9.
[9] EUROPEAN COMMISSION. *Quantum Europe Strategy: Quantum Europe in a Changing World* [na spletu]. 2025 [pridobljeno 10. 02. 2026]. Dostopno na: https://digital-strategy.ec.europa.eu/en/library/quantum-europe-strategy
[10] ERIXON, Fredrik idr. *Quantum Clusters: Ranking the World's Deep-Tech Epicentres* [na spletu]. ECIPE: European Centre for International Political Economy. 2025 [pridobljeno 10. 02. 2026]. Dostopno na: https://ecipe.org/publications/quantum-clusters-ranking-deep-tech-epicentres/
[11] KUHLMANN, Stefan, Peter STEGMAIER in Kornelia KONRAD. The tentative governance of emerging science and technology—A conceptual introduction. *Research Policy* [na spletu]. 2019, **48**(5), 1091–1097. ISSN 0048-7333. Dostopno na:10.1016/j.respol.2019.01.006
[12] MAITLIS, Sally in Marlys CHRISTIANSON. Sensemaking in organizations: Taking stock and moving forward. *The Academy of Management Annals* [na spletu]. 2014, **8**(1), 57–125. ISSN 1941-6067. Dostopno na:10.1080/19416520.2014.873177
[13] SCHMIDT, Vivien A. Discursive Institutionalism: The Explanatory Power of Ideas and Discourse. *Annual Review of Political Science* [na spletu]. 2008, **11**(Volume 11, 2008), 303–326. ISSN 1094-2939, 1545-1577. Dostopno na:10.1146/annurev.polisci.11.060606.135342
[14] WEICK, Karl E. *Sensemaking in organizations*. Nachdr. Thousand Oaks, Calif.: Sage Publ, 2010. Foundations for organizational science. ISBN 978-0-8039-7177-6.
[15] GUSTON, David H. Understanding 'anticipatory governance'. *Social Studies of Science* [na spletu]. 2014, **44**(2), 218–242. ISSN 0306-3127, 1460-3659. Dostopno na:10.1177/0306312713508669
[16] STILGOE, Jack, Richard OWEN in Phil MACNAGHTEN. Developing a framework for responsible innovation. *Research Policy* [na spletu]. 2013, **42**(9), 1568–1580. ISSN 0048-7333. Dostopno na:10.1016/j.respol.2013.05.008
[17] CERON, Andrea, Luigi CURINI in Stefano M. IACUS. *Politics and big data: nowcasting and forecasting elections with social media*. London ; New York, NY: Routledge, Taylor & Francis Group, 2017. ISBN 978-1-4724-6666-2.
[18] LI, Xin idr. Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. *Technological Forecasting and Social Change* [na spletu]. 2019, **146**, 687–705. ISSN 00401625. Dostopno na:10.1016/j.techfore.2018.06.004
[19] SHAIKH, Abdul Rahman in Hamed ALHOORI. *Predicting Patent Citations to measure Economic Impact of Scholarly Research* [na spletu]. arXiv. 2019 [pridobljeno 7. 04. 2026]. Dostopno na:10.48550/ARXIV.1906.08244
[20] YLINEN, Mika in Mikko RANTA. *Predicting Corporate Innovation Using Machine Learning and Social Media Data* [na spletu]. SSRN. 2024 [pridobljeno 7. 04. 2026]. Dostopno na:10.2139/ssrn.4953415
[21] DE MARCELLIS-WARIN, Nathalie, Daniel KOULOUKOUI in Thierry WARIN. A large-scale dataset of AI-related tweets: Structure and descriptive statistics. *Data in Brief* [na spletu]. 2025, **62**, 111960. ISSN 2352-3409. Dostopno na:10.1016/j.dib.2025.111960
[22] WEI, Mengyi idr. Mapping AI ethics narratives: evidence from Twitter discourse between 2015 and 2022. *Humanities and Social Sciences Communications* [na spletu]. 2025, **12**(1), 212. ISSN 2662-9992. Dostopno na:10.1057/s41599-025-04469-9
[23] BORUP, Mads idr. The sociology of expectations in science and technology. *Technology Analysis & Strategic Management* [na spletu]. 2006, **18**(3–4), 285–298. ISSN 0953-7325, 1465-3990. Dostopno na:10.1080/09537320600777002

[24] DE TOUZALIN, Aymard idr. *Quantum Manifesto: A New Era of Technology* [na spletu]. 2016. Dostopno na: https://qt.eu/media/pdf/93056_Quantum-Manifesto_WEB.pdf

[25] NG, Yee Man Margaret in Rik RAY. The journalists' exodus: Navigating the transition from Twitter to Mastodon and other alternative platforms. *New Media & Society* [na spletu]. 2025, 14614448251321165. ISSN 1461-4448. Dostopno na:10.1177/14614448251321165

[26] SCHMIDT, Vivien A. Taking ideas and discourse seriously: explaining change through discursive institutionalism as the fourth 'new institutionalism'. *European Political Science Review* [na spletu]. 2010, **2**(1), 1–25. ISSN 1755-7739, 1755-7747. Dostopno na:10.1017/S175577390999021X

APPENDIX A

Table A1 lists the search terms used to collect the three language-specific discourse corpora from X (formerly Twitter), grouped by thematic category. Terms were queried with the language filters lang:en, lang:de, and lang:sl respectively. Base forms are shown for readability; all morphological inflections were included in the actual queries (e.g., case declensions in Slovene, genitive/plural forms in German).[3]

Table A1: Search Terms Used for X Data Collection

Topic	English	German	Slovene
General / Core	quantum, quantum computing, quantum computer, quantum information, quantum technology, quantum hardware, quantum advantage, quantum supremacy, qubit	quantum, quantencomputer, quantenrechner, quantencomputing, quantentechnologie, quantentechnik, qubit	quantum, kvantni računalnik, kvantno računanje, kvantno računalništvo, kvantna tehnologija, kvantna nadvlada, kvantna prednost, qubit
Algorithms & Software	quantum algorithm, quantum error correction, quantum software, quantum application, quantum app, quantum annealer, quantum annealing	quantenalgorithmus, quantensoftware, quantenanwendung, quanten-app, quantenannealer, quantenannealing	kvantni algoritem, kvantna programska oprema, kvantna aplikacija, kvantni annealer, kvantno annealing
Processors, Simulation & Cloud	quantum processor, quantum simulation, quantum simulator, quantum cloud, quantum RAM, quantum memory	quantensimulation, quantensimulator, quantencloud, quantenram, quantenspeicher	kvantni procesor, kvantna simulacija, kvantni simulator, kvantni oblak, kvantni RAM, kvantni pomnilnik
Communication & Networks	quantum communication, quantum internet, quantum network, quantum networking, quantum repeater, quantum teleportation, QKD	quantenkommunikation, quanteninternet, quantennetzwerk, quantenrepeater, quantenteleportation, QKD	kvantna komunikacija, kvantni internet, kvantno omrežje, kvantno mreženje, kvantni repetitor, kvantna teleportacija, QKD
Cryptography & Security	quantum cryptography, quantum key distribution, quantum encryption	quantenkryptographie, quantenkryptografie, quantenverschlüsselung	kvantna kriptografija, kvantno šifriranje
Sensing & Metrology	quantum sensing, quantum sensor, quantum magnetometer, quantum accelerometer, quantum gyroscope, atom interferometer, atomic interferometer, rydberg atom, rydberg sensor, entangled photons, ion traps	quantensensorik, quantensensor, atominterferometer, rydberg-atom, rydberg-sensor, verschränkte photonen, ionenfalle	kvantni senzor, kvantni magnetometer, kvantni pospeškomer, kvantni žiroskop, atomski interferometer, rydbergov atom, rydbergov senzor, prepleteni fotoni, ionske pasti
Photonics & Optics	quantum photonics, quantum optics	quantenphotonik, quantenoptik	kvantna fotonika, kvantna optika
Industry & Ecosystem	quantum industry, quantum ecosystem, quantum innovation ecosystem	quantenindustrie, quantenökosystem, quanteninnovationsökosystem	kvantna industrija, kvantni ekosistem, kvantni inovacijski ekosistem
Workforce & Education	quantum workforce, quantum skills, quantum education	quantenarbeitskraft, quantenfachkraft, quantenkompetenz, quantenbildung	kvantna delovna sila, kvantne veščine, kvantno izobraževanje
Policy & Strategy	quantum strategy, quantum initiative, quantum policy	quantenstrategie, quanteninitiative, quantenpolitik	kvantna strategija, kvantna pobuda, kvantna politika

[3] The full list of query strings is available from the authors upon request.

APPENDIX B

Table A2 reports high-attention months in each corpus that could be linked to real-world events. Specifically, we first identify the top 20 months with the highest posting activity per dataset and then retain only those months for which at least one event occurred in the same calendar month. Each row represents a single event–month match; as multiple relevant events may occur within the same month, individual months can appear more than once. This structure preserves the one-to-many relationship between attention peaks and contemporaneous developments. The table therefore presents the subset of peak activity periods that can be plausibly associated with real-world events and mirrors the months in Figure 2 that are highlighted in red.

Table A2: Matched High-Attention Months and Corresponding Real-World Events

Corpus	Peak Month	Posts (Monthly Total)	Event Date(s)	Event Decription	Scope	Type
English	October-19	74495	23-10-19	Google announces quantum supremacy milestone	Global	scientific / industry
English	June-24	53455	07-06-24	UN declares 2025 International Year of Quantum Science	Global	policy
English	October-24	59867	01-10-24	IBM Quantum Data Center opens in Ehningen	German	industry
English	October-24	59867	15-10-24	EuroHPC procurement contract signed (Germany)	German	infrastructure
English	December-24	200255	09-12-24	Google unveils the Willow quantum chip	Global	scientific / industry
English	December-24	200255	11-12-24	German Quantum Technologies Framework Programme (EN edition)	German	policy
English	January-25	121912	10-01-25	Slovenia establishes working group for quantum strategy	Slovenia	policy
English	June-25	91403	23-06-25	EuroHPC inaugurates PIAST-Q	EU	funding / industry
English	July-25	86796	02-07-25	Quantum Europe Strategy adopted	EU	policy
English	July-25	86796	30-07-25	German Hightech Agenda includes quantum technologies	German	policy
English	August-25	101028	28-08-25	EU expands access to quantum computing infrastructure	EU	infrastructure
English	October-25	158453	07-10-25	Nobel Prize in Physics (quantum-related)	Global	scientific
English	October-25	158453	16-10-25	Slovenia adopts quantum strategy until 2035	Slovenia	policy
English	October-25	158453	31-10-25	EU Quantum Act – call for evidence launched	EU	legislation
English	November-25	153727	13-11-25	Jade/Ruby quantum processors introduced	EU/German	infrastructure / scientific
German	March-17	2415	06-03-17	IBM launches IBM Q commercial quantum initiative	Global	industry
German	November-17	2396	10-11-17	IBM announces 50-qubit quantum processor prototype	Global	scientific / industry
German	October-19	2761	23-10-19	Google announces quantum supremacy milestone	Global	scientific / industry
German	June-21	2477	15-06-21	Germany’s first quantum computer enters operation in Ehningen	German	industry / scientific

German	November-21	2544	16-11-21	IBM unveils the 127-qubit Eagle processor	Global	scientific / industry
German	November-21	2544	17-11-21	Digital Europe Programme launches EuroQCI call	EU	funding
German	January-22	2168	27-01-22	Munich Quantum Valley formally established	German	ecosystem
German	July-22	2188	04-07-22	Munich Quantum Conference 2022	German	conference
German	July-22	2188	05-07-22	NIST announces quantum-resistant cryptographic algorithms	Global	policy
German	December-22	2191	12-12-22	German Cyberagentur quantum funding call	German	funding / policy
German	April-23	2230	26-04-23	Quantum Technologies Conceptual Framework Programme	German	policy
German	December-24	2751	09-12-24	Google unveils the Willow quantum chip	Global	scientific / industry
German	December-24	2751	11-12-24	German Quantum Technologies Framework Programme (EN edition)	German	policy
German	July-25	4539	02-07-25	Quantum Europe Strategy adopted	EU	policy
German	July-25	4539	30-07-25	German Hightech Agenda includes quantum technologies	German	policy
German	August-25	2485	28-08-25	EU expands access to quantum computing infrastructure	EU	infrastructure
German	October-25	2245	07-10-25	Nobel Prize in Physics (quantum-related)	Global	scientific
German	October-25	2245	16-10-25	Slovenia adopts quantum strategy until 2035	Slovenia	policy
German	October-25	2245	31-10-25	EU Quantum Act – call for evidence launched	EU	legislation
German	November-25	2271	13-11-25	Jade/Ruby quantum processors introduced	EU/German	infrastructure / scientific
Slovene	November-17	34	10-11-17	IBM announces 50-qubit quantum processor prototype	Global	scientific / industry
Slovene	October-18	33	29-10-18	Quantum Technologies Flagship kicks off with first 20 projects	EU	policy / funding
Slovene	October-22	39	04-10-22	Nobel Prize in Physics (quantum information science)	Global	scientific
Slovene	October-22	39	25-10-22	Hamburg measures package to strengthen quantum computing	German	funding / policy
Slovene	October-22	39	27-10-22	Germany awards €208 million in DLR quantum contracts	German	funding / policy
Slovene	June-23	37	06-06-23	IBM announces first European quantum data center in Germany	German	industry
Slovene	June-23	37	27-06-23	EuroHPC signs hosting agreements for six quantum computers	EU	infrastructure
Slovene	November-23	30	24-11-23	EuroHPC launches procurement for quantum computer in Germany	EU/German	infrastructure

Slovene	December-24	33	09-12-24	Google unveils the Willow quantum chip	Global	scientific / industry
Slovene	December-24	33	11-12-24	German Quantum Technologies Framework Programme (EN edition)	German	policy
Slovene	January-25	31	10-01-25	Slovenia establishes working group for quantum strategy	Slovenia	policy
Slovene	June-25	32	23-06-25	EuroHPC inaugurates PIAST-Q	EU	funding / industry
Slovene	October-25	170	07-10-25	Nobel Prize in Physics (quantum-related)	Global	scientific
Slovene	October-25	170	16-10-25	Slovenia adopts quantum strategy until 2035	Slovenia	policy
Slovene	October-25	170	31-10-25	EU Quantum Act – call for evidence launched	EU	legislation
Slovene	November-25	40	13-11-25	Jade/Ruby quantum processors introduced	EU/German	infrastructure / scientific
Actor	November-21	390	16-11-21	IBM unveils the 127-qubit Eagle processor	Global	scientific / industry
Actor	November-21	390	17-11-21	Digital Europe Programme launches EuroQCI call	EU	funding
Actor	December-21	327	30-12-21	Invitation to the SI-EuroQCI consortium	Slovenia	policy / funding
Actor	June-22	416	06-06-22	BMBF research programme 'Quantensysteme…'	German	policy / funding
Actor	July-22	324	04-07-22	Munich Quantum Conference 2022	German	conference
Actor	July-22	324	05-07-22	NIST announces quantum-resistant cryptographic algorithms	Global	policy
Actor	October-22	332	04-10-22	Nobel Prize in Physics (quantum information science)	Global	scientific
Actor	October-22	332	25-10-22	Hamburg measures package to strengthen quantum computing	German	funding / policy
Actor	October-22	332	27-10-22	Germany awards €208 million in DLR quantum contracts	German	funding / policy
Actor	February-23	319	22-02-23	Google’s quantum computer reduces errors (logical qubit scaling)	Global	scientific
Actor	March-23	402	01-03-23	Digitalna Slovenija 2030 includes quantum technologies	Slovenia	policy
Actor	March-23	402	15-03-23	EU Secure Connectivity Programme Regulation	EU	legislation
Actor	March-23	402	15-03-23	UK National Quantum Strategy published	Global	policy
Actor	June-23	357	06-06-23	IBM announces first European quantum data center in Germany	German	industry
Actor	June-23	357	27-06-23	EuroHPC signs hosting agreements for six quantum computers	EU	infrastructure
Actor	November-23	393	24-11-23	EuroHPC launches procurement for quantum computer in Germany	EU/German	infrastructure

An Exploratory Study of Chunking Strategies for Retrieval Augmented Generation over Scientific PDF Articles

Marko Rudolf[1], Andrej Dobrovoljc[2]

[1]Rudolfovo – Znanstveno in tehnološko središče Novo mesto
Podbreznik 15, 8000 Novo mesto, Slovenija
E-mail: marko.rudolf@rudolfovo.eu

[2]Fakulteta za informacijske študije
Ljubljanska cesta 31A, 8000 Novo mesto, Slovenia
E-mail: andrej.dobrovoljc@fis.unm.si

Abstract: Technology is evolving rapidly, increasing the need for reliable information, especially when decisions affect energy efficiency and sustainability. Retrieval-Augmented Generation (RAG) is increasingly applied to support question answering in scientific and technical documentation. The reliability of such systems depends on retrieving relevant evidence and maintaining source traceability. The most common point of failure arises in obtaining accurate and relevant responses. If the retriever returns incomplete or weakly relevant context, the generated answer may be insufficiently grounded, raising the risk of unsupported or hallucinated claims. This paper examines appropriate document chunking as a key design point in RAG frameworks that influences the quality and reliability of retrieved responses. We conducted an empirical study using an experimental set of scientific papers, which pose specific challenges for computational reading due to their complex structure. A reproducible RAG model was implemented in Python, including document import, text normalization, chunking, dense vector representations (embeddings), vector indexing, and similarity-based retrieval. Chunking was defined as the primary experimental variable, while all other components remained constant. We compared two chunking strategies with a set of three different sizes for each under identical retrieval parameters. Our preliminary experiments aimed to identify the most suitable chunking strategy for processing scientific texts. Initial results indicate that chunk size has a significant impact on the quality of retrieved results. To evaluate adequacy result, a human assessment was used to review and judge the relevance and correctness of system-generated responses. The findings demonstrate that appropriate text chunking serves as an effective mechanism of "evidence engineering" for applying RAG in industrial and sustainability critical domains (e.g., energy systems, manufacturing processes, and regulatory compliance), where well-grounded responses are essential for trustworthy AI-assisted decision support.

Keywords: Retrieval Augmented Generation (RAG), chunking strategies, ranking, human-in-the-loop evaluation

Article Classification: Scientific Paper

1 INTRODUCTION

The use of classical LLMs without additional mechanisms is prone to hallucinations [1]. This represents a significant challenge, particularly in environments where a high level of answer accuracy is required. This limitation is addressed by Retrieval-Augmented Generation (RAG), which has proven especially effective in high-risk environments where incorrect information is unacceptable. In recent years, RAG has become a central paradigm for enhancing LLMs by improving them with knowledge from external sources and thereby reducing hallucinations. Recent studies in medicine and public health indicate that the use of RAG for targeted information retrieval can reduce hallucination rates by approximately 30–40% or more compared to classical LLMs without access to external knowledge sources [2]. After year 2023 research on RAG has expanded rapidly. We can find several surveys and benchmark studies consolidating best practices for retrieval, chunking and evaluation. Gao et al. [3] and Zhao et al. [4] identify retrieval quality as the dominant performance bottleneck in RAG pipelines, often even more important than generative model itself. In scientific applications, retrieval errors propagate directly into generation, as LLMs typically answer only from retrieved context. That means that even strong language models will fail when relevant evidence is ranked too low or omitted entirely. This is also the field of our study, where we tested different chunking strategies and measured outcome for our test data.

Scientific PDF documents are challenge for RAG. They are long, structurally heterogeneous and often full of discourse-level semantics over different paragraphs or even sections and pages. Fixed size chunking can fragment arguments, definitions and empirical findings. Consequently, the result is degraded semantic embeddings and retrieval accuracy. Recent empirical studies explicitly demonstrate that document segmentation strategies have a first order effect on retrieval ranking metrics such as Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG), especially in scientific and technical text [5].

Recent research shows that chunking strategies have become a critical choice when we are designing our systems. Character or token-based chunking is most common and easy to implement because of simplicity. Paragraph-aware and discourse aware segmentation methods are more complicated to implement, but outperform fixed size baselines in retrieval evaluations. Merola [6] shows that preserving natural discourse boundaries improves early ranking of relevant passages, even when recall remains comparable. This finding is especially relevant for RAG systems, where can top ranked chunks influence on generation quality.

Evaluation methodology is another active area of research. **Recall@k** is widely reported, but multiple recent works are writing that recall is not enough for RAG assessment. Ranking sensitive metrics such as Mean Reciprocal Rank (MRR) or NDCG better captures whether relevant information appears high enough to be used by the generator [3].

Motivated by these insights, we want to present in this paper an empirical study of a complete RAG pipeline applied to a dataset of scientific PDF articles. Within this study we focused on the impact of chunking strategies on retrieval quality, with particular emphasis on MRR as indicator of generation effectiveness. We combined paragraph-aware segmentation, page level traceability of the information source and human-in-the-loop for relevance judgments. Our goal was to provide practical and reproducible guidance for designing RAG systems for scientific documents.

1.1 Chunking strategies

Chunking strategy plays a crucial role in RAG systems because it directly affects which parts of documents can be retrieved and how well they align with user queries. Chroma's 2024 technical report on chunking [7] shows that different strategies can lead to substantial variation in retrieval accuracy and token level efficiency, even when all other components (embeddings, retriever) are held constant. Brådland et al.[8] similarly demonstrate with the HOPE metric that segmentation quality can change factual and answer correctness by more than 20 percentage points in retrieval-based generation setups. Recent work identifies three main families of chunking strategies: **fixed size** (character or token) windows, **rule based** or **structure aware** chunking (e.g. paragraphs, sections) and **semantic chunking** using embeddings or LLMs. Fixed size windows are simplest and efficient, but they often break sentences or concepts, which can fragment context and lower retrieval quality. **Rule based** chunking preserves natural boundaries such as paragraphs, subsections or page regions and are widely used in RAG pipelines over structured documents (especially for scientific articles). **Semantic strategies** go further by grouping text based on embedding similarity or LLM judgments, potentially improving coherence, but at higher computational cost and implementation complexity [9].

In practice, empirical results are mixed and task dependent. Vectara's large scale evaluation [9] finds that semantic chunking offers only inconsistent gains and that well-tuned fixed size chunking often performs as well or better in realistic retrieval scenarios. Chroma's experiments [7] report that heuristic recursive character splitting and embedding-aware "ClusterSemanticChunker" perform strongly across domains when parameters are calibrated, indicating that simple methods can be competitive if properly configured. At the same time, Brådland et al. emphasize that passage-level semantic independence is important for factual correctness in RAG settings, supporting chunking that keeps self-contained information units.

For scientific and technical research, where arguments span paragraphs and precise wording matters, recent work favor's structure aware chunking that respects document organization [10]. As document specific chunking that aligns chunks with sections or paragraphs is generally recommended, we also tested in our study. We select paragraph-based chunking as base and added page citations in selected documents for easier tracking. To compare the quality of the results, we used the fixed size chunking strategy with overlapping as a baseline.

1.2 Ranking in Retrieval Augmented Generation and Result Comparison

In RAG systems ranking describes ordering of retrieved document chunks according to their estimated relevance to a query. Most contemporary RAG pipelines rely on **dense vector similarity ranking**, typically using cosine similarity between query and chunk embeddings [3, 11]. This ranking stage is critical, as only the top-ranked chunks are passed to the language model due to context length constraints.

Comparing ranking performance in RAG systems requires metrics that are sensitive to **rank position**, not just coverage. While **Recall@k** measures whether relevant content appears anywhere among the top *k* results, it does not reflect whether the content is ranked early enough to influence generation. Ranking-sensitive metrics such as **Mean Reciprocal Rank (MRR)** and **Normalized Discounted Cumulative Gain (NDCG)** are therefore recommended for RAG evaluation, as they better correlate with answer quality [3, 13]. For this study we set ranking of relevant chunks at the **retrieval evaluation stage**. We conducted all experiments under **strictly controlled conditions**, where the **embedding model**, **similarity function**, **retrieval algorithm** and **query set** remain fixed across all evaluations. The only experimental variable is the **chunking strategy** applied during dataset preprocessing.

Ranking is produced by computing **cosine similarity** between a single query embedding and all chunk embeddings, followed by sorting the chunks in descending similarity order. No approximate nearest neighbor indexing, re-ranking, or heuristic post-processing is applied. With this we ensure that differences in ranking are dependent from variations in chunk content generated by different chunking strategies.

For each query we manually evaluated the **top 10 ranked chunks** for relevance using a human-in-the-loop protocol. Based on these relevance judgments, **Mean Reciprocal Rank (MRR@10)** is computed as the primary ranking-sensitive metric. **MRR** explicitly captures the position of the first relevant chunk in the ranked list and therefore reflects whether authoritative passages are ranked early enough to influence downstream generation. By holding all retrieval components constant and varying only the chunking strategy, our study isolates the effect of document segmentation on ranking quality. This design enables a direct comparison of RAG configurations based on MRR, demonstrating how paragraph-aware chunking compared to basic fixed sized chunking strategy.

2 METHODOLOGY

2.1 Dataset

Dataset included approximately 80 various scientific articles in PDF format. Within this study we included randomly selected articles from Google Scholar addressing the following areas: artificial intelligence for energy efficiency, artificial intelligence in renewable energy management, smart and sustainable energy storage systems, and energy-efficient industrial systems and intelligent resource utilization. Beside the topic our search criteria was publish date between within 2022 – 2026.

2.2 Overall Pipeline

The implemented system follows a classical Retrieval-Augmented Generation (RAG) architecture, consisting of five sequential stages: document ingestion, text pre-processing, chunking and indexing, dense retrieval and evaluation. Although the complete RAG pipeline, including generation, is implemented, only the retrieval stage is evaluated as retrieval quality under controlled conditions which was our main research focus. Each stage is executed deterministically with all intermediate results (cleaned text, chunks, embeddings, similarity scores, and evaluation logs) stored explicitly to disk to ensure reproducibility, traceability, and transparent analysis.

2.3 PDF Parsing

PDF documents were parsed at the page level using *pdfplumber* library. This design choice ensures that each text segment can be traced back to an exact page number in the source document. Page numbers are preserved throughout the pipeline and later attached to chunks as provenance metadata. Paragraph-aware chunking was applied in a subsequent preprocessing step using newline-based heuristics derived from the extracted page text. Our notebook does not perform layout reconstruction beyond linear text extraction. Tables, figures and equations are extracted only as they appear in the linear PDF text stream.

2.4 Text Cleaning

Extracted page level text was cleaned using a rule-based approach. During preprocessing we removed page headers and footers, page numbers and timestamps, footnote markers and symbols and reference/bibliography sections when clearly identifiable. After cleaning, text segments shorter than 100 characters were discarded. We implement this threshold explicitly in the notebook and serve to eliminate fragments unlikely to add meaningful semantic content.

2.5 Chunking Strategies

Two chunking strategies with a set of three different sizes were implemented and evaluated. Each strategy operates on cleaned text and preserves source metadata.

- **Character-based chunking with overlap strategy.** Cleaned text is split into fixed-length chunks of observed size. Overlap is implemented to reduce boundary effects, ensuring that information spanning chunk borders is partially retained. **Chunk boundaries are independent of linguistic or discourse structure.**
- **Paragraph-aware chunking strategy.** Text is first segmented into paragraphs using newline-based heuristics derived from the PDF extraction. Paragraphs shorter than 100 characters are discarded. During paragraph-aware chunking, paragraph-like text segments are accumulated up to a maximum observation size. Longer segments are split into sequential chunks while preserving their original order. If paragraph is shorter than defined size, then the used size is the size of the paragraph. Bigger size allows longer

discourse units to remain intact, reducing fragmentation of arguments and explanations common in scientific writing.

For all strategies, each chunk is annotated with metadata in the form [SOURCE: document | PAGE: x] or [PAGE: x–y] if a chunk spans multiple pages.

2.6 Embedding and Retrieval

Each chunk is embedded using a dense sentence embedding model (*bge-base-en-v1.5*), executed locally via Ollama. Embeddings are computed once per chunk and stored for reuse across queries. For each user query, a single query embedding is computed using the same embedding model. Retrieval is performed by computing cosine similarity between the query embedding and all stored chunk embeddings using *NumPy*. For evaluation, the top 10 chunks ranked by cosine similarity are selected. For generation experiments (not evaluated in this paper), the top 3 to 5 chunks are passed to the language model with an explicit context length limit of approximately 3500 characters.

2.7 Evaluation Protocol

During evaluation we focused exclusively on retrieval quality, independent of text generation. This design choice aligns with recent RAG research emphasizing that retrieval errors are a primary failure mode and should be analyzed separately from generative behavior [3]. We formulated one domain relevant natural language question to reflect realistic information needs within the dataset topics. The top-10 retrieved chunks were manually assessed for relevance. Relevance was defined as whether a chunk contained explicit information that could substantively contribute to answering the query, rather than merely sharing topical keywords. Within our study we include a human-in-the-loop evaluation approach with all relevant judgments, performed by a single expert annotator. While this introduces subjectivity, prior work has shown that expert single annotator judgments can yield reliable retrieval evaluations in specialized domains, particularly when the goal is comparative analysis rather than absolute performance estimation [6, 12]. Because the study compares chunking strategies under identical conditions, relative differences in Recall@10, Precision@10 and MRR@10 remain informative despite the small sample size. This approach is consistent with exploratory and methodological RAG studies that prioritize controlled comparisons over large scale benchmarking [13]. All similarity scores, retrieved chunks, relevance annotations and computed metrics for top-10 result were logged to disk to ensure full reproducibility of the evaluation.

3 IMPLEMENTATION AND RESULTS

3.1 Evaluation Datasets

For our study we developed *Jupyter* notebook [14] within *Visual Studio Code* environment. Base for our notebook code was sample from Huggingface webpage [15]. For embedding model, we used *bge-base-en-v1.5*. Figure 1 shows a schematic view of our development environment.

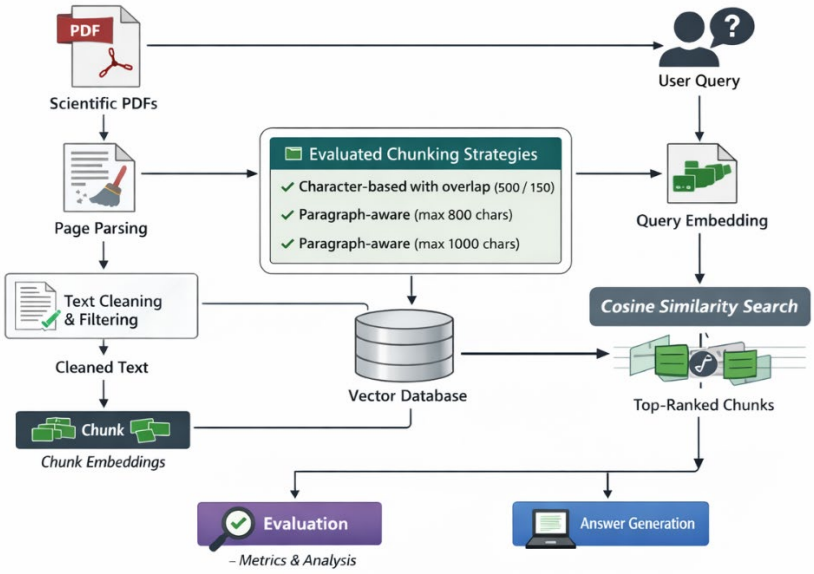


Figure 1: Overview of the implemented RAG pipeline and evaluated chunking variants

Two strategies with three retrieval evaluations were conducted:

- **Character-based chunking with overlap:**
  - **500** with an overlap of **150** characters between consecutive chunks,
  - **850** with an overlap of **200** characters between consecutive chunks,
  - **1000** with an overlap of **250** characters between consecutive chunks.
- **Paragraph-aware chunking:**
  - maximum **600** characters,
  - maximum **800** characters,
  - maximum **1000** characters.

Each dataset was evaluated using the same embedding model, retrieval procedure and identical evaluation question. For research purposes we defined following question:

- **Q:** What measurable improvements, expressed in percentage terms, can be achieved through the application of artificial intelligence in energy efficiency, renewable energy management, sustainable energy storage, and resource-efficient industrial systems?

This question was formulated based on the conference topic, which addresses artificial intelligence for energy efficiency, artificial intelligence in renewable energy management, smart and sustainable energy storage systems, and energy efficient industrial systems with intelligent resource utilization.

3.2 Evaluation Metrics

Retrieval performance was evaluated using Recall@10, Precision@10 and Mean Reciprocal Rank (MRR) following common practice. For each query, the top-10 retrieved chunks were manually annotated for relevance. **Recall@10** was computed as the proportion of queries for which at least one relevant chunk appeared among the top-10 retrieved results. **Precision@10** was calculated over queries with at least one relevant annotation, measuring the proportion of relevant chunks among the retrieved top-10 results. **Mean Reciprocal Rank (MRR)** was computed based on the rank position of the first relevant chunk per query.

Recall@10 and MRR@10 were calculated by averaging the per query values across all evaluated queries. Precision@10 was averaged only over queries with at least one relevant chunk in the top-10 results, where precision is defined as the fraction of relevant chunks among the top-10 retrieved items. Queries with no relevant retrieved chunks were excluded if they appear (in our present test there was none) as they would contribute only trivial zero values and would not reflect differences in ranking quality.

3.3 Evaluation Calculation and Results

In the evaluation phase, human relevance judgments were performed by the author and recorded as the ranks of relevant chunks within the top-k retrieved results.

During code execution, evaluation outputs were displayed interactively in the notebook interface for human relevance assessment and simultaneously written to a structured text log file. This approach ensured transparency of the evaluation process and enabled reproducibility and traceability of the reported results for subsequent analysis and future studies. Evaluation results are provided automatically by our notebook and are listed in *Table 1*.

Table 1: Retrieval performance by chunking strategy (top-k = 10).

Chunking strategy	Chunk size	Overlap	Recall(top-k=10)	Precision (top-k=10)	MRR (top-k=10)
Basic (character-based)	500	150	1	0,2	**0,500**
Basic (character-based)	850	200	0	0	**0,000**
Basic (character-based)	1000	250	1	0,1	**0,100**
Paragraph-aware	600	-	1	0,1	**0,330**
Paragraph-aware	800	-	1	0,2	**0,143**
Paragraph-aware	1000	-	1	0,2	**0,500**

Overall, there is not much difference in Precision metric, but it is more noticeable in MRR, where we can see that the Paragraph-aware strategy has better results.

Results are calculated based on following definitions:

- $q \in \{1, \ldots, N\}$ = query,
- R_q = ranks of relevant retrieved chunks for query q, as annotated manually,
- $|R_q|$ = number of relevant chunks within the top-k results,
- hit_q = 1 if $|R_q| > 0$, and 0 otherwise,
- $FirstRank_q$ = min(R_q), = rank of the first relevant retrieved chunk.

Recall@10 (query-level hit rate): proportion of queries for which at least one relevant chunk appears among the top-10 retrieved results as presented in *Equation 1*.

$$Recall@10 = \frac{1}{N}\sum_{q=1}^{N} hit_q \tag{1}$$

Precision@10 (micro-averaged): fraction of relevant chunks among all retrieved chunks across all queries as presented in *Equation 2*.

$$Precision@10 = \frac{\sum_{q=1}^{N} |R_q|}{N \cdot k} \tag{2}$$

Mean Reciprocal Rank (MRR@10): how early the first relevant chunk appears in the ranked retrieval list. For each query q, the reciprocal rank is defined as presented in Equation 3:

$$RR_q = \begin{cases} \frac{1}{FirstRank_q}, & if\ |R_q| > 0 \\ 0, & otherwise \end{cases} \tag{3}$$

where:

- $FirstRank_q$ = $rank_q$ if ranks are 1-based,
- $FirstRank_q$ = $rank_q$ + 1 if a 0-based rank was recorded.

The Mean Reciprocal Rank is then computed as presented in *Equation 4*:

$$MRR@10 = \frac{1}{N}\sum_{q=1}^{N} RR_q \tag{4}$$

3.4 Model evaluation

The experiments generally indicate better and more stable retrieval performance when using the paragraph-aware chunking strategy. This stability is consistent across documents with varying amounts of text, suggesting that paragraph boundaries often correspond to semantically coherent units and therefore provide a natural segmentation for technical and scientific writing. In other words, paragraphs tend to encapsulate self-contained arguments, definitions, or procedure steps, which improves probability that a retrieved chunk contains sufficient context to support an evidence grounded answer.

At the same time, the results show that **chunk size is a critical parameter regardless of the chosen strategy**. Changes in chunk size can significantly affect the embedding representation and the similarity-based ranking, thereby influencing both the number of relevant answers retrieved and the rank position of the first relevant chunk. This observation aligns with the expected trade-off between **context completeness** (larger chunks reduce fragmentation) and **semantic specificity** (smaller chunks reduce topic mixing). Consequently, chunk size should be treated as a first order design choice in RAG pipelines rather than a purely technical preprocessing detail, particularly in domains where reliable, source grounded retrieval is essential for downstream answer trustworthiness.

4 IMPROVEMENTS AND FUTURE WORK

While developing the experimental RAG system, we identified several areas for improvement and further investigation. Retrieval augmented generation offers substantial research opportunities, particularly because hallucinations remain a major challenge in standard LLM-based systems. We therefore started by establishing a strong foundation for subsequent research.

4.1 Small changes for quick improvements

For our study we used dataset with 80 scientific articles what can be an issue because of limited size of document dataset. This is sufficient for controlled methodological comparison, but not large enough to support broad generalizations about optimal chunking strategies across scientific domains. Prior retrieval and RAG research indicates that retrieval behavior, ranking stability and metric sensitivity can change significantly with dataset scale and diversity. For future work and deriving more generalizable conclusions we suggest enlarging dataset and repeat evaluation.

The selected dataset are scientific articles, which align with the goal of this study - to evaluate RAG chunking strategies on structurally complex texts. However, the relevance assessment was performed by author, who is not a domain expert in the articles' subject area. As a result, some chunks with higher technical or contextual value may have been overlooked or judged less relevant than they would be by a subject-matter expert.
For stronger evidence and more robust comparisons, future evaluations should include a larger set of questions and involve an independent, domain qualified evaluator (or multiple evaluators) to reduce subjectivity and improve the reliability of the relevance judgments.

PDF files and scientific articles often have complex layouts, so it is crucial to extract the text correctly and clean it before further processing. In our generated output, we noticed several issues caused by missing, mixed or incorrectly extracted parts of the source documents. Strengthening the extraction and preprocessing step is therefore essential for improving the accuracy and reliability of the results for further research.

4.2 Higher complexity for better results

One of possible improvements can be use of late, instead of first chunking. In our study first chunking (splitting the text before embedding) already performs strongly for a medium-sized scientific dataset. For larger and more heterogeneous datasets, recent work proposes late chunking as an alternative. Günther et al. [10] introduce late chunking as a method that embeds the full long text first using long context embedding models and applies chunking *inside the embedding pipeline.* This produces chunk embeddings that retain broader contextual information and improve retrieval across multiple benchmarks. In a comparative study of advanced chunking strategies for RAG, Merola and Singh evaluate late chunking against contextual retrieval and report tradeoffs. Contextual retrieval better preserves semantic coherence but is more computationally demanding, while late chunking is more efficient yet may reduce relevance/completeness in some settings [6]. These findings motivate late chunking primarily as a scaling oriented extension. It may become beneficial when context loss from first chunking dominates retrieval quality.

In future work we can incorporate also a second stage re-ranking model to further improve early precision and MRR, allowing a comparison between chunking-driven improvements and ranking-based refinements. Re-ranking directly targets early ranking quality (MRR) and would likely improve retrieval effectiveness in our setup but was intentionally excluded to preserve a controlled evaluation of chunking strategies. So, this can be used as additional model and evaluation comparison upgrade in our study.

5 DISCUSSION AND CONCLUSSION

Our study confirms the central role of **retrieval quality** in RAG systems, particularly when applied to scientific PDF documents. Key finding is that the **chunking strategy substantially influences retrieval ranking quality**.
Our study clearly distinguishes between **retrieval quality** and **generation quality**, which are often conflated in RAG evaluations. While generation quality depends on how effectively a language model synthesizes information, it is fundamentally constrained by the quality and ranking of retrieved evidence. In retrieval-constrained RAG settings, generation errors are therefore frequently a consequence of retrieval failures rather than limitations of the language model itself.

We isolated the retrieval stage and demonstrated that **ranking sensitive metrics such as Mean Reciprocal Rank (MRR)** provide more informative insights than Recall@k alone. Although multiple chunking strategies achieved perfect Recall@10, their MRR values differed substantially, indicating meaningful differences in how early relevant evidence was ranked. From a practical RAG perspective, this distinction is critical, as relevant context ranked too low may not influence generation due to context length limitations. These findings support recent literature advocating MRR or NDCG as primary evaluation metrics for RAG retrieval, especially in scientific domains [9].

From an applied standpoint, the results show that paragraph-aware chunking with larger maximum chunk sizes can improve retrieval effectiveness without introducing additional architectural complexity. Furthermore, page-aware provenance metadata enhances interpretability and facilitates qualitative analysis of retrieval behavior. However, chunk size cannot be increased arbitrarily, as it is constrained by the maximum input length of the embedding model. In our implementation, the employed embedding model (*bge-base-en-v1.5*) supports a maximum input length of 512 tokens, which imposes an upper bound on chunk size and necessitates a trade-off between contextual completeness

and embedding feasibility. It should be noted that tokens do not correspond directly to characters. Depending on tokenization, 512 tokens typically correspond to 1200-1800 characters, which must be considered when defining character-based chunk sizes [16].

Within this study we provided empirical evidence that **document segmentation is a core methodological decision** in RAG system design. Future work should extend the evaluation to larger corpora, additional queries with multiple annotators and investigate how improvements in retrieval ranking quality translate into downstream generation performance.

6 ACKNOWLEDGMENTS

The author acknowledges the use of open source tools and models, including *pdfplumber*, *Ollama* and community-maintained embedding and language models. This study was conducted as part of ongoing doctoral studies in computer science. Artificial intelligence tools were used to support Python code optimization and textual review.

REFERENCES

[1] SERVICENOW, Patrice Béchard, Orlando MARQUEZ in Ayala SERVICENOW. *Reducing hallucination in structured outputs via Retrieval-Augmented Generation*. 2024.

[2] NISHISAKO, Sota, Takahiro HIGASHI in Fumihiko WAKAO. Reducing Hallucinations and Trade-Offs in Responses in Generative AI Chatbots for Cancer Information: Development and Evaluation Study. *JMIR Cancer* [na spletu]. 2025, **11**. ISSN 23691999. Dostopno na:10.2196/70176

[3] GAO, Yunfan idr. Retrieval-Augmented Generation for Large Language Models: A Survey [na spletu]. 2024. Dostopno na: http://arxiv.org/abs/2312.10997

[4] ZHAO, Wayne Xin idr. A Survey of Large Language Models [na spletu]. 2026. Dostopno na: http://arxiv.org/abs/2303.18223

[5] AMIRI, Mahmoud in Thomas BOCKLITZ. Chunk Twice, Embed Once: A Systematic Study of Segmentation and Representation Trade-offs in Chemistry-Aware Retrieval-Augmented Generation [na spletu]. 2025. Dostopno na: http://arxiv.org/abs/2506.17277

[6] MEROLA, Carlo in Jaspinder SINGH. Reconstructing Context: Evaluating Advanced Chunking Strategies for Retrieval-Augmented Generation [na spletu]. 2025. Dostopno na: http://arxiv.org/abs/2504.19754

[7] SMITH, Brandon in Anton TROYNIKOV. Chroma technical report, evaluating chunking strategies for retrieval. *https://research.try chroma.com/evaluating-chunking*. 3. julij 2024.

[8] BRÅDLAND, Henrik idr. A New HOPE: Domain-agnostic Automatic Evaluation of Text Chunking. V: *SIGIR 2025 - Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* [na spletu]. Association for Computing Machinery, Inc, 2025, str. 170–179. ISBN 9798400715921. Dostopno na:10.1145/3726302.3729882

[9] QU, Renyi in Forrest BAO. Is semantic chunking worth the computational cost? *https://www.vectara.com/blog/is-semantic-chunking-worth-the-computational-cost*. 19. november 2024.

[10] WANG, Zhitong idr. *Document Segmentation Matters for Retrieval-Augmented Generation*. 2025.

[11] LEWIS, Patrick idr. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks [na spletu]. 2021. Dostopno na: http://arxiv.org/abs/2005.11401

[12] VOORHEES, Ellen M. 2002 - The Philosophy of Information Retrieval Evaluation - the-philosophy-of-information.pdf. *Lecture Notes in Computer Science* [na spletu]. 2002 [pridobljeno 31. 03. 2026]. Dostopno na: https://doi.org/10.1007/3-540-45691-0_34

[13] LIU, Pengfei idr. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing [na spletu]. 2021. Dostopno na: http://arxiv.org/abs/2107.13586

[14] RUDOLF, Marko. *https://github.com/MarkoRudolf/FINI2026.git* [na spletu]. 2026 [pridobljeno 7. 04. 2026]. Dostopno na: https://github.com/MarkoRudolf/FINI2026.git

[15] NGUYEN, Xuan-Son. Code a simple rag from scratch. *https://huggingface.co/blog/ngxson/make-your-own-rag*. 29. oktober 2024.

[16] OPENAI HELP. What are tokens and how to count them? *https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them*. 30. marec 2026.

Integrating Artificial Intelligence with Physics-Based Simulations: A Review of Methods and Applications

Matic Medvešek[1,2], Lucijano Berus[1], Jelena Topić Božič[1,2], Simon Muhič[1,2]

[1]Rudolfovo - Science and Technology Centre Novo Mesto, Slovenia
E-mail: simon.muhic@rudolfovo.eu

[2]Faculty of industrial engineering Novo mesto, Slovenia
E-mail: simon.muhic@fini-unm.si

Abstract: *Recent advances in artificial intelligence (AI) have opened new opportunities for accelerating and enhancing simulations of complex physical systems. This review article examines the emerging integration of AI-based models with traditional physics-based simulations in safety-critical scientific domains, focusing on nuclear engineering. Modern simulation tools such as neutron transport solvers, thermal-hydraulic system codes, severe accident and multiphysics models provide high-fidelity representations of physical processes but are often computationally expensive and limited in their ability to support large-scale optimization, uncertainty quantification, or real-time decision making.*

The reviewed literature highlights recent developments in surrogate modelling, generative AI, physics-informed neural networks, and explainable machine learning approaches designed to emulate or augment these traditional models. In nuclear engineering, AI methods have been explored for reactor core optimization, criticality safety analysis, and debris-bed formation modelling in severe accident scenarios. Despite promising results, several challenges remain, including the preservation of physical consistency, explainability, and regulatory acceptance in safety-critical environments. This review discusses the current state of the art, identifies key limitations of existing approaches, and outlines research directions for integrating trustworthy AI into scientific simulation workflows. The analysis aims to provide a structured overview of how AI-augmented modelling can improve computational efficiency while maintaining the reliability required for engineering and environmental applications.

Keywords: Simulations, Artificial Intelligence, Machine Learning, Physics-Informed Neural Networks

Article Classification*: Scientific Paper*

1 INTRODUCTION

Artificial intelligence (AI) is increasingly integrated into nuclear engineering to address computational, operational, and safety-related challenges, with applications spanning reactor core design, reactor physics, anomaly detection, human reliability analysis, nuclear data evaluation, and severe accident modelling [1–6]. A key trend is the use of surrogate models to approximate high-fidelity multiphysics and thermal-hydraulic simulations, significantly reducing computational cost while maintaining acceptable accuracy [1–3, 7–14]. In parallel, deep learning methods are applied to online monitoring and anomaly detection, often using unsupervised or weakly supervised approaches due to limited accident data [15–17]. AI also supports nuclear data analysis, bias prediction, and experiment design, where interpretability and expert-in-the-loop approaches remain critical [18–21]. Despite these advances, several challenges limit widespread adoption in safety-critical environments, including data scarcity, limited generalization, lack of interpretability, regulatory constraints, and cybersecurity concerns [4–6, 14]. Consequently, current research increasingly focuses on physics-guided, explainable, and robust AI frameworks aligned with nuclear safety requirements. Nuclear systems require accurate prediction, reliable monitoring, and computational efficiency across normal, transient, and accident conditions. While conventional reactor physics and thermal-hydraulic simulation tools provide high-fidelity results, they are often computationally expensive, difficult to couple, and unsuitable for real-time use [3, 7, 8, 11, 22]. Advances in machine learning and scientific AI offer new opportunities to accelerate simulations, improve predictive capabilities, and reduce uncertainties [1–3, 5, 6]. Recent literature indicates a transition from proof-of-concept studies to practical applications, including reactor core optimization [1, 2], correction of low-fidelity simulations [3], anomaly detection [15–17], nuclear criticality safety and data analysis [18–21, 23], and surrogate modelling for severe accidents and probabilistic risk assessment [8–14, 22–24]. However, successful deployment requires not only predictive accuracy but also physical consistency, robustness, interpretability, and regulatory acceptance [4–6, 12, 14]. This review provides a structured overview of methodologies, applications, and challenges in integrating AI with physics-based simulations in nuclear engineering.

This review provides a structured overview of current artificial intelligence methods for physics-based simulations, focusing on their application in nuclear engineering and highlighting key challenges related to data limitations, interpretability, and reliability.

2 AI METHODOLOGIES FOR PHYSICS-BASED SIMULATIONS

2.1 Surrogate Modeling and Metamodeling

A central research direction in the literature is the acceleration or replacement of computationally expensive simulation tools through surrogate modeling. Surrogate models are data-driven approximations constructed from high-fidelity simulation outputs to predict system responses at new input conditions, enabling efficient sensitivity analysis and optimization that would otherwise be computationally prohibitive [25]. This approach is particularly relevant in nuclear engineering, where thermal-hydraulic and severe accident simulations involve stiff, nonlinear, and multiphysics systems requiring significant computational resources for repeated evaluations [7–14, 23]. By learning the input–output relationships of these models, machine learning-based metamodels can significantly reduce execution time while preserving acceptable predictive accuracy. Such surrogate models have been successfully applied to approximate complex safety analyses, including loss-of-coolant accident (LOCA) scenarios, thereby enabling large-scale uncertainty quantification, optimization, and probabilistic assessments that are infeasible with full-order models alone [26]. However, surrogate models may suffer from reduced accuracy outside the training domain and require careful validation.

2.2 Physics-Informed Neural Networks (PINNs)

Physics-informed neural networks (PINNs) represent a class of scientific machine learning methods in which neural network training is constrained by governing physical laws, typically expressed as nonlinear partial differential equations. This is achieved by incorporating equation residuals directly into the loss function, enabling simultaneous data fitting and enforcement of physical consistency [27, 28]. PINNs have demonstrated strong capability in solving forward and inverse problems across a wide range of physical systems, using both continuous-time formulations for spatio-temporal approximation and discrete-time approaches based on implicit numerical schemes [27]. Their ability to embed domain knowledge makes them particularly attractive for applications involving limited or noisy data. In nuclear engineering, PINNs are increasingly explored as surrogate models for complex systems, where they can improve extrapolation behavior and maintain consistency with governing physics compared to purely data-driven approaches [7, 12, 29]. This is especially relevant in safety-critical scenarios, where physically inconsistent predictions from black-box models are not acceptable. Despite their advantages, PINNs can be computationally demanding and sensitive to hyperparameter selection.

2.3 Explainable and Interpretable AI

Explainability is essential for deploying AI in nuclear engineering, where safety, transparency, and regulatory acceptance are required. Conventional deep learning models are often limited by their black-box nature, restricting their use in safety-critical applications. To address this, interpretable model architectures and analysis techniques are being developed. Kolmogorov–Arnold Networks (KAN) provide a more transparent alternative by using spline-based representations that can reveal underlying functional relationships [14]. In parallel, models such as random forests are applied in nuclear data analysis to identify influential parameters and detect inconsistencies, offering insight into sources of simulation error [19]. These approaches demonstrate that combining predictive capability with interpretability is essential for building trust and supporting decision-making in nuclear systems.

Despite these advances, many AI approaches remain limited by high data requirements, lack of robust uncertainty quantification, and sensitivity to out-of-distribution conditions, which is particularly critical in safety-critical applications. Hybrid data- and knowledge-driven models have also been applied to neutron diffusion calculations, demonstrating high predictive accuracy while incorporating domain-specific physical constraints [30].

3 APPLICATIONS IN NUCLEAR ENGINEERING

The described methodologies are applied across several key domains in nuclear engineering, as outlined in the following sections.

3.1 Reactor Core Design and Reactor Physics

AI is widely applied in reactor core design and reactor physics, particularly for optimization. Machine learning enables efficient exploration of complex design spaces, allowing evaluation of numerous reactor configurations. AI-based design combined with multiphysics emulation has demonstrated significant improvements in performance metrics such as temperature peaking factors [1], while machine learning applied to molten salt reactor design achieved predictions within 5% relative error without neutron transport calculations [31]. Machine learning models, including Decision Tree and K-Nearest Neighbor algorithms trained on EXFOR data, can accurately predict neutron-induced reaction cross sections, in some cases outperforming traditional evaluated libraries, thereby reducing bias and uncertainty [32]. This direction is closely connected to the Transformational Challenge Reactor program, where

additive manufacturing enables new design paradigms and AI-based emulators act as computational accelerators [1, 2]. The lack of standardized ML workflows is addressed by frameworks such as pyMAISE, which integrates hyperparameter tuning, training, validation, explainability, and deployment across multiple applications, achieving performance comparable to or exceeding existing methods [33]. Machine learning is also used to refine low-fidelity simulations through data-driven error correction approaches [1–3, 6, 34].

3.2 Thermal-Hydraulics and Severe Accident Analysis

A major research focus is AI-assisted severe accident analysis, particularly debris-bed coolability in boiling water reactors [8–11, 23, 24]. After vessel failure, corium forms a debris bed whose coolability determines accident mitigation success [11, 22, 24]. High-fidelity tools such as COCOMO accurately simulate these processes but are computationally prohibitive for uncertainty quantification and risk analysis [8–11, 23]. To address this, surrogate models have been developed. Kriging-based models predict debris-bed dryout power density, enabling rapid coolability mapping [8]. ANN-based approaches provide fast quenching estimation and can be coupled with MELCOR to reproduce key containment trends with reduced computational cost [9, 10]. These methods build on MELCOR/COCOMO analyses showing that detailed debris-bed modeling significantly affects accident predictions [11]. Machine learning enables large speedups, supporting probabilistic risk assessment and ROAAM+ methodologies, where surrogates replace full models for uncertainty analysis [23]. Additional studies show that modeling assumptions strongly influence predictions, emphasizing the need for robust surrogate approaches [22]. Physics-informed neural networks (PINNs) further improve surrogate modeling by incorporating physical constraints, enhancing accuracy and consistency [28, 29], and outperforming standard neural networks in accident scenarios [12]. The node-assigned PINN (NA-PINN) improves accuracy by assigning networks to control volumes [7], while related surrogate models achieve speedups of up to three orders of magnitude, enabling real-time applications [13].

3.3 Monitoring, Anomaly Detection, and Human Reliability

Monitoring and anomaly detection represent a key application area of AI in nuclear systems, particularly due to the scarcity of accident data and the abundance of normal operation data [4, 16, 17]. As a result, unsupervised and semi-supervised approaches are widely used. A deep generative method combining variational autoencoders and isolation forest enables detection of abnormal operating states using only normal-condition data, achieving near real-time performance and early identification of accident conditions [17]. AI is also applied to human reliability analysis, where anomaly detection identifies inconsistencies between manual and sensor-recorded data. In this context, generative adversarial networks outperform traditional methods such as one-class SVM and isolation forest, supporting improved detection of human errors [16]. At the reactor-core level, deep learning models have been used to classify and localize perturbations. Convolutional neural networks trained on detector signal spectrograms can identify different perturbation types and remain robust under noisy and incomplete data, with validation on real plant measurements [15]. These studies indicate a shift toward robust, real-time, and data-efficient monitoring methods capable of operating under uncertain and limited data conditions [4, 15, 17].

3.4 Nuclear Data Evaluation and Criticality Safety

AI is increasingly applied in nuclear data evaluation and criticality safety to improve prediction accuracy and reduce uncertainty [5, 18, 19, 21, 35]. Machine learning models have been used to predict bias in criticality safety analyses, achieving errors below experimental uncertainty and demonstrating suitability for supporting safety assessments [18]. Additionally, AI methods are applied to the design of integral experiments, where expert-in-the-loop Bayesian optimization frameworks guide experiment selection under physical and regulatory constraints [21]. Institutional assessments further highlight the role of AI in automating nuclear data workflows, including dataset compilation, inference, validation, and experiment design [5]. Additional studies confirm the effectiveness of machine learning in identifying important nuclear data regions and supporting evaluation processes [21, 35].

3.5 Cross-Cutting Challenges

A primary challenge is data limitation. Accident data, abnormal transients, and human-error events are scarce, leading to data imbalance, distribution shift, and limited generalization [4, 16, 17]. This explains the frequent use of unsupervised learning, anomaly detection, and physics-informed approaches in nuclear applications [7, 12, 17]. Another key issue is interpretability and trustworthiness. Deep learning models are often considered black boxes, which is problematic in safety-critical systems [4, 6, 12, 14]. Approaches such as hybrid correction models [3], physics-informed networks [7, 12], expert-in-the-loop optimization [21], and interpretable architectures like KAN [14] aim to address this limitation. Regulatory and licensing acceptance also remains a major barrier. Even highly accurate AI models must demonstrate physical consistency, robustness outside training domains, and traceable uncertainty quantification [6, 12, 14]. As a result, AI is typically used to augment rather than replace established simulation tools [3, 9, 12, 13]. Finally, infrastructure and governance challenges are emphasized in institutional reports, including the need for FAIR data, open science practices, workforce development, and attention to ethics and cybersecurity, particularly in nuclear security and safeguards [5].

4 FUTURE PERSPECTIVES

The literature points toward several likely future directions. First, tighter integration of domain knowledge and data-driven learning appears essential for robust nuclear AI [5, 6, 12]. Purely black-box approaches may continue to be used for some monitoring tasks, but the strongest long-term trend is toward hybrid methods that combine physics, simulation, and learning [3, 7, 12, 13]. Second, surrogate models will likely become central in digital twins, operator support systems, and dynamic probabilistic risk assessment, where repeated real-time or many-query simulation is required [6, 10, 13, 14, 24]. In this context, the balance among speed, fidelity, interpretability, and uncertainty quantification will be decisive. Third, explainable and expert-guided AI will remain important, especially in nuclear data analysis, experiment design, and licensing-related applications [14, 19, 21]. These approaches are more compatible with the epistemic culture of nuclear engineering, where transparent reasoning and model defensibility are necessary. Finally, broader adoption will depend on coordinated development of curated datasets, benchmark problems, validation protocols, and ethical and cybersecurity frameworks [4, 5]. Without these foundations, the field may remain fragmented despite promising individual case studies [6]. The lack of standardized benchmark datasets and validation protocols remains a key barrier to systematic comparison of AI methods.

5 CONCLUSION

The current literature shows that AI is becoming a significant technology in nuclear engineering. Its strongest present contributions are in computational acceleration, surrogate modelling, anomaly detection, and data analysis [1–3, 7–19, 22–24]. In reactor design and reactor physics, AI-based approaches enable efficient exploration of complex design spaces and refinement of low-fidelity simulations [1–3, 6, 34], achieving prediction accuracies within 5% relative error and computational speedups of up to three orders of magnitude in surrogate and Physics-Informed Neural Networks models [13, 31]. In monitoring and operations, deep learning and generative models enable anomaly detection and perturbation localization under sparse abnormal data [15–17]. In nuclear data and criticality safety, machine learning supports bias prediction, data inconsistency identification, and experiment design [18, 20, 21, 35]. In severe accident analysis, surrogate models are transforming debris-bed coolability and quenching simulations into forms suitable for risk-informed workflows [8–11, 23, 24]. At the same time, the field is still in a transitional stage. The most credible path forward is not unrestricted replacement of established nuclear codes by black-box AI, but rather the development of hybrid, physics-guided, explainable, and uncertainty-aware systems that complement conventional methods [3, 4, 6, 12, 14]. This appears to be the direction most consistent with both technical performance and the safety culture of the nuclear sector.

Acknowledgement
The research was funded by the European Union's Horizon Europe research and innovation programme through the TURING Project. Call HORIZON-CL4-2024-HUMAN-03-02 – Explainable and Robust AI.

REFERENCES

[1] SOBES, Vladimir et al. AI-based design of a nuclear reactor core. Scientific Reports 2021, 11(1) ISSN 20452322. DOI:10.1038/S41598-021-98037-1
[2] HISCOX, B D et al. Artificial Intelligence for Multiphysics Nuclear Design Optimization with Additive Manufacturing 2021
[3] OKTAVIAN, M. R. et al. Integrating core physics and machine learning for improved parameter prediction in boiling water reactor operations. Scientific Reports 2024 14:1 2024, 14(1), 5835- ISSN 2045-2322. DOI:10.1038/s41598-024-56388-5
[4] HUANG, Qingyu et al. A review of the application of artificial intelligence to nuclear reactors: Where we are and what's next. Heliyon 2023, 9(3), e13883 ISSN 2405-8440. DOI:10.1016/J.HELIYON.2023.E13883
[5] INTERNATIONAL ATOMIC ENERGY AGENCY. Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology 2022 Available at: https://www-pub.iaea.org/MTCD/Publications/PDF/ART-INTweb.pdf
[6] ZHANG, Ruizhi et al. Artificial Intelligence in Reactor Physics: Current Status and Future Prospects. Available at: https://arxiv.org/pdf/2503.02440v1
[7] SHIN, Jeesuk et al. Node Assigned physics-informed neural networks for thermal-hydraulic system simulation: CVH/FL module Available at: https://arxiv.org/pdf/2504.16447v1
[8] CHEN, Yangli in Weimin MA. Development and application of a surrogate model for quick estimation of ex-vessel debris bed coolability. Nuclear Engineering and Design 2020, 370, 110898 ISSN 0029-5493. DOI:10.1016/J.NUCENGDES.2020.110898
[9] PETROFF, Markus et al. Experimental and numerical investigation on debris bed quenching and development of an artificial neural network for quick quenching estimation. Nuclear Engineering and Design ISSN 0029-5493. DOI:10.1016/J.NUCENGDES.2025.114134
[10] WANG, Wanhong, Yangli CHEN in Weimin MA. Development of a surrogate model for quenching estimation of ex-vessel debris beds and its coupling with MELCOR. Annals of Nuclear Energy. ISSN 18732100. DOI:10.1016/j.anucene.2023.109883

[11] CHEN, Yangli, Huimin ZHANG in Weimin MA. Coupled MELCOR/COCOMO analysis on quench of ex-vessel debris beds. Annals of Nuclear Energy 2022, 165, ISSN 0306-4549. DOI:10.1016/J.ANUCENE.2021.108643
[12] ANTONELLO, Federico, Jacopo BUONGIORNO in Enrico ZIO. Physics informed neural networks for surrogate modeling of accidental scenarios in nuclear power plants. Nuclear Engineering and Technology 2023, 55(9), 3409–3416. ISSN 2234358X. DOI:10.1016/j.net.2023.06.027
[13] BEJA-BATTAIS, Perceval, Alain GROSSETÊTE in Nicolas VAYATIS. Enhancing Nuclear Reactor Core Simulation through Data-Based Surrogate Models.
[14] HO CHAE, Young in Seo RYONG KOO. A Conceptual Study on Surrogate Modeling of Nuclear Power Plant using KAN. Transactions of the Korean Nuclear Society Autumn Meeting Changwon.
[15] TASAKOS, Thanos et al. DEEP LEARNING-BASED ANOMALY DETECTION IN NUCLEAR REACTOR CORES 2021 DOI:10.13182/M&C21-33681
[16] GURSEL, Ezgi et al. Using artificial intelligence to detect human errors in nuclear power plants: A case in operation and maintenance. 2023 DOI: http://www.inl.gov
[17] LI, Xiangyu et al. Research on anomaly detection method of nuclear power plant operation state based on unsupervised deep generative model. Annals of Nuclear Energy. 2022, 167, 108785 ISSN 0306-4549. DOI:10.1016/J.ANUCENE.2021.108785
[18] GRECHANUK, Pavel, Michael E. RISING in Todd S. PALMER. Using Machine Learning Methods to Predict Bias in Nuclear Criticality Safety. Journal of Computational and Theoretical Transport. 2019, 47(4–6), 552–565. ISSN 2332-4309. DOI:10.1080/23324309.2019.1585877
[19] GRECHANUK A. PAVEL. Applying Machine Learning Algorithms to Identify Problematic Nuclear Data within Nuclear Transport Simulations. Oregon, 2022. Oregon State University.
[20] GRECHANUK PAVEL et al. Applying machine learning techniques to understand nuclear data areas of interest 2018. Available at: https://indico.bnl.gov/event/5067/contributions/24693/attachments/20639/27718/ND-ML_CSEWG.pdf
[21] MICHAUD, Isaac et al. Expert-in-the-loop design of integral nuclear data experiments. Statistical Analysis and Data Mining 2024, 17(2), e11677 ISSN 19321872. DOI:10.1002/SAM.11677;JOURNAL:JOURNAL:19321872;WGROUP:STRING:PUBLICATION
[22] GALUSHIN, Sergey in Pavel KUDINOV. Analysis of the Effect of Severe Accident Scenario on Debris Properties in Lower Plenum of Nordic BWR Using Different Versions of MELCOR Code. Science and Technology of Nuclear Installations 2019, 2019, 1–18. ISSN 1687-6075. DOI:10.1155/2019/5310808
[23] JOSHI-THOMPSON, Jasmin, Micheal BUCK in Jörg STARFLINGER. Application of AI Methods for Describing the Coolability of Debris Beds Formed in the Late Accident Phase of Nuclear Reactors.
[24] GALUSHIN, Sergey. Development of Risk Oriented Accident Analysis Methodology for Assessment of Effectiveness of Severe Accident Management Strategy in Nordic BWR. Stockholm, 2019. KTH Royal Institute of Technology.
[25] QUEIPO, Nestor V. et al. Surrogate-based analysis and optimization. Progress in Aerospace Sciences. 2005, 41(1), 1–28 ISSN 0376-0421. DOI:10.1016/J.PAEROSCI.2005.02.001
[26] CONNER, Landon et al. Machine Learned Metamodeling of a Computationally Intensive Accident Simulation Code. V: Volume 1: Operating Plant Challenges, Successes, and Lessons Learned; Nuclear Plant Engineering; Advanced Reactors and Fusion; Small Modular and Micro-Reactors Technologies and Applications American Society of Mechanical Engineers, 2021. ISBN 978-0-7918-8524-6. DOI:10.1115/ICONE28-66619
[27] RAISSI, M., P. PERDIKARIS in G. E. KARNIADAKIS. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics. 2019, 378, 686–707 ISSN 0021-9991. DOI:10.1016/J.JCP.2018.10.045
[28] CUOMO, Salvatore et al. Scientific Machine Learning Through Physics–Informed Neural Networks: Where we are and What's Next. Journal of Scientific Computing 2022, 92(3), DOI:10.1007/S10915-022-01939-Z
[29] BARALDI, Piero et al. Physics-Informed Neural Networks for the safety analysis of nuclear reactors. Progress in Nuclear Energy. 2025, 185, 105745. ISSN 0149-1970. DOI:10.1016/J.PNUCENE.2025.105745
[30] ZENG, Fu Lin et al. Application research of a hybrid data- and knowledge-driven artificial intelligence scientific computing model in neutron diffusion calculation for nuclear reactors. Nuclear Science and Techniques 2025 37:2. 2026, 37(2), 24. ISSN 2210-3147. DOI:10.1007/S41365-025-01852-8
[31] TURKMEN, Mehmet, Gwendolyn J.Y. CHEE in Kathryn D. HUFF. Machine learning application to single channel design of molten salt reactor. Annals of Nuclear Energy. 2021, 161, 108409. ISSN 0306-4549. DOI:10.1016/J.ANUCENE.2021.108409
[32] VICENTE-VALDEZ, Pedro, Lee BERNSTEIN in Massimiliano FRATONI. Nuclear data evaluation augmented by machine learning. Annals of Nuclear Energy. 2021, 163, 108596. ISSN 0306-4549. DOI:10.1016/J.ANUCENE.2021.108596
[33] MYERS, Patrick A. et al. pyMAISE: A Python platform for automatic machine learning and accelerated development for nuclear power applications. Progress in Nuclear Energy. 2025, 180, 105568 ISSN 0149-1970. DOI:10.1016/J.PNUCENE.2024.105568
[34] RISING E. MICHAEL et al. Using Machine Learning Algorithms in Nuclear Data Evaluations. 2019. Available at: https://ncsp.llnl.gov/sites/ncsp/files/2021-05/Rising2-LANL_NCSP-NDML.pdf
[35] OKTAVIAN, M. R. et al. Preliminary development of machine learning-based error correction model for low-fidelity reactor physics simulation. Annals of Nuclear Energy 2023, 187, 109788. ISSN 0306-4549. DOI:10.1016/J.ANUCENE.2023.109788

Z umetno inteligenco podprt organizacijski spomin: Avtomatizacija inženiringa kakovosti in preprečevanje ponavljanja napak v proizvodnji

AI-Powered Organizational Memory: Automating Quality Engineering and Preventing Defect Recurrence in Manufacturing

Matej Kalčič

Galileon, razvoj naprednih sistemov, d.o.o., Slovenija
E-naslov: matej@galileon.si

***Povzetek**: Proizvodne ekipe in inženirji v kompleksnih industrijskih okoljih izgubljajo na tisoče ur zaradi administrativnih bremen pri ročnem reševanju napak. Pomanjkanje časa in razpršena dokumentacija preprečujeta iskanje pravih korenskih vzrokov, kar neizogibno vodi do ponavljanja napak in visokih stroškov slabe kakovosti. Za reševanje tega izziva v pametnih tovarnah članek predstavlja inovativno metodologijo: aplikacijo SolveR, ki deluje kot »organizacijski spomin« in uporablja z umetno inteligenco podprt delovni proces 8D. Sistem integrira pristop RAG (angl. Retrieval-Augmented Generation) za takojšnjo analizo zgodovinske dokumentacije, predlaganje korenskih vzrokov ter vodenje med funkcijskih ekip skozi standardiziran proces. S tem se vzpostavi popolna sledljivost od začetnega simptoma napake do končne rešitve, ki je neposredno pripravljena za zunanje presoje in skladna z zahtevami IATF 16949. Implementacija sistema v prakso kaže izjemne rezultate: avtomatizacija celotnega delovnega procesa reševanja problemov je 8-krat hitrejša. Ta ključna metrika dokazuje, da lahko podjetja reinvestirajo več kot 5.000 inženirskih ur v inovacije in preprečijo ponavljanje kritičnih napak. Umetna inteligenca tako ne služi le reševanju problemov, ampak gradnji centralno bazo znanja, ki raste z vsako novo rešeno težavo.*

***Ključne besede**: 8D metodologija, umetna inteligenca, RAG, IATF 16949, vodenje kakovosti, pametne tovarne*

***Razvrstitev**: Strokovni prispevek*

***Abstract**: Manufacturing teams operating within complex industrial environments waste thousands of hours resolving defects through manual processes and rigid spreadsheets. Engineers are continuously bogged down by suffocating administrative burdens and scattered documentation, obscuring true root causes and allowing critical defects to recur. To solve this administrative bottleneck, this paper presents an innovative methodology: an AI-powered 8D workflow designed specifically for Smart Factories. Acting as an "Organizational Memory," the system leverages Retrieval-Augmented Generation (RAG) to instantly process historical data, suggest data-driven root causes, and guide cross-functional teams through a standardized problem-solving cycle. This ensures complete, audit-ready traceability connecting the initial symptom directly to the final preventive action, ensuring strict compliance with automotive standards such as IATF 16949. The performance of this system introduces a paradigm shift in operational efficiency, demonstrating that the automated workflow is 8x faster than traditional manual methods. By eliminating administrative waste, medium-sized enterprises can reinvest over 5,000 engineering hours back into innovation and proactively prevent critical problems annually. Ultimately, this self-improving knowledge base empowers engineers to focus on creation rather than reporting, setting a new benchmark for quality management in Industry 4.0.*

***Keywords:** 8D Methodology, Artificial Intelligence, RAG, IATF 16949, Quality Management, Smart Factories*

***Article Classification**: Professional Paper*

1 UVOD

Stroški nekakovosti globalno predstavljajo med 15% in 20% celotnega prihodka proizvodnih podjetij [1]. V avtomobilski dobavni verigi, regulirani s standardom IATF 16949, je reševanje reklamacij po metodologiji osmih disciplin (8D) uveljen pristop za odzivanje na neustreznosti [2, 3]. Študije potrjujejo, da strukturirana uporaba orodij kakovosti znatno zmanjša ponavljanje napak, še posebej v malih in srednjih podjetjih [4]. Kljub temu ostaja izvedba 8D poročil v praksi pretežno ročna.

Med pilotnimi sejami v obdobju januar–marec 2026 smo izmerili, da inženir kakovosti v povprečju porabi 47 minut zgolj za iskanje in prebiranje dokumentov po arhivih in različnih sistemih, preden začne z dejansko analizo. Gre za lastne meritve, izvedene pri osmih pilotnih partnerjih. Razlog je preprost: dokumenti so razpršeni po različnih sistemih in formatih, iskanje pa je v celoti ročno, v večini primerov po »mrtvih« Excel dokumentih na sharepointu.

Proizvajalec s 500 zaposlenimi in 80 primeri 8D na leto lahko na administrativnem delu izgubi med 3.500 in 5.500 inženirskih ur. To ni zanemarljiva številka; pomeni, da nekdo polni delovni čas dve leti zgolj prebira stare datoteke. Ko izkušeni inženirji zapustijo podjetje, z njimi izgine vedenje o tem, kateri stroj za tlačni liv razvije poroznost pri določeni temperaturi orodja ali katera serija dobavitelja je pred leti povzročala dimenzijske odstopke. Naslednji inženir začne od začetka. Obstoječa programska orodja za upravljanje kakovosti (SAP QM, Siemens Polarion, PTC Windchill Quality) ponujajo vodenje dokumentov in sledenje delovnim tokovom, ne morejo pa vsebinsko analizirati preteklih 8D poročil ali med njimi prepoznati vzorcev [5].

V nadaljevanju opisujemo sistem, ki to pomanjkljivost odpravlja. Pristop RAG (Retrieval-Augmented Generation) [6], pri katerem se jezikovni model pred generiranjem odgovora obogati z dokumenti iz lokalne baze znanja, je v industriji že poznan koncept. Chen in sodelavci [7] so dokumentirali uspešno uvedbo RAG v interaktivni industrijski sistem za upravljanje znanja s preciznostjo iskanja nad 75%. Nedavni sistematični pregledi potrjujejo, da uporaba umetne inteligence v avtomobilski industriji izboljšuje odkrivanje napak in omogoča prehod k strategijam ničelne napake [8], medtem ko se na področju vodenja kakovosti v proizvodnji že pojavljajo prvi pametni asistenti na osnovi velikih jezikovnih modelov [9]. Opisani sistem ta pristop prenaša na specifično domeno inženiringa kakovosti v proizvodnji: zajame, vloži in semantično preišče zgodovinsko dokumentacijo kakovosti, znotraj strukturiranega delovnega toka 8D pa podaja podatkovno utemeljene predloge korenskih vzrokov. Preizkušen je bil s pilotnimi partnerji v realnih industrijskih okoljih. Avtor je soustanovitelj podjetja Galileon d.o.o., ki razvija opisani sistem; rezultate so pregledali in potrdili pilotni partnerji.

2 OPIS PRISTOPA

2.1 Pristop RAG za reševanje napak

Veliki jezikovni modeli (UI) so usposobljeni na javno dostopnih podatkih in nimajo dostopa do zaupnih proizvodnih dokumentov posameznega podjetja. Pristop RAG [6] premosti to omejitev: sklepalni mehanizem poveže z varnim, lokaliziranim repozitorijem zgodovinske dokumentacije podjetja. Pred generiranjem kateregakoli odgovora sistem iz lastne baze znanja (tj. indeksirane zbirke vseh preteklih dokumentov podjetja) pridobi relevanten kontekst. Ne ugiba, temveč odgovarja na podlagi dejanskih podatkov.

2.2 Procesiranje dokumentov

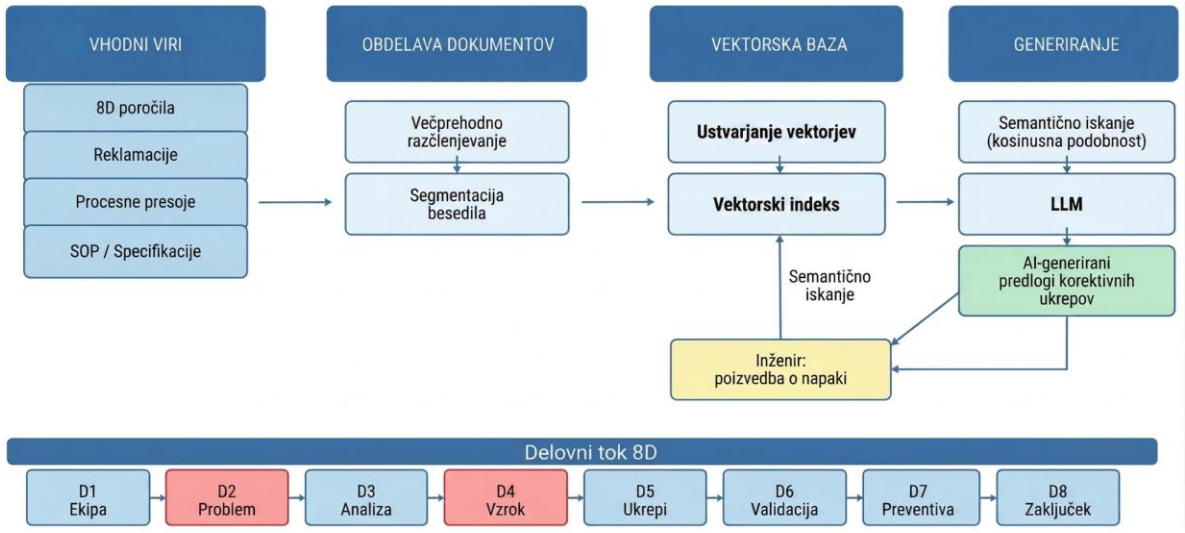
Sistem procesira arhiv 8D poročil, reklamacijske zapise, analize odpovedi in ugotovitve procesnih presoj. Podprti vhodni formati obsegajo PDF, Word, Excel, CSV in slikovne datoteke. Proizvodni dokumenti so za razčlenjevanje zahtevni. Standardni operativni postopki, specifikacije in varnostni listi vsebujejo zapletene tabele in gnezdene hierarhije, pri katerih je prostorska razporeditev besedila informativna. Zato smo razvili večprehodno razčlenjevanje, ki ohrani tabelarno strukturo, preden se besedilo razdeli na segmente za vektorsko vložitev.

2.3 Vektorska vložitev in semantično iskanje

Vsak besedilni segment se pretvori v visokodimenzionalni vektor s transformerskim modelom za vložitve [10]. Ti vektorji kodirajo semantični pomen besedila in so shranjeni v varni bazi znotraj EU. Celotna arhitektura je prikazana na Sliki 1. Ko inženir vnese nov opis napake, sistem poizvedbo pretvori v vektor in po zgodovinski bazi poišče semantično podobne zapise s kosinusno podobnostjo, opredeljeno v Enačbi (1):

$$sim(A, B) = (A \cdot B) / (\|A\| \times \|B\|) \quad (1)$$

kjer sta A in B vektorski predstavitvi dveh besedilnih segmentov, $\|\cdot\|$ pa označuje dolžino vektorja. Vrednost blizu 1,0 pomeni visoko semantično ujemanje. Sistem vrne najbolje razvrščene pretekle primere, da inženir pregleda korenske vzroke in vse podatke, ki so za trenutno napako relevantni.



Slika 1: Arhitektura sistema RAG za reševanje napak

2.4 Organizacijski spomin in sledljivostna veriga

Pojem organizacijski spomin označuje širšo sposobnost podjetja, da ohranja in ponovno uporabi nabrano znanje. Baza znanja je tehnična komponenta, ki ta spomin omogoča: indeksirana zbirka vektorsko vloženih dokumentov, po kateri sistem izvaja semantično iskanje. Skupaj tvorita celoto, ki preprečuje izgubo kritičnega znanja.

Najhujša ranljivost klasičnega vodenja kakovosti je degradacija znanja. Ko se izkušen inženir s petnajstletnim stažem upokoji, z njim odide vedenje o tem, pri katerih toplotnih pogojih določen stroj za tlačni liv razvije poroznost ali kateri parametri brizganja polimerov pri konkretnem orodju povzročajo razpoke. V enem od pilotnih podjetij (Tier-2 dobavitelj, CNC obdelava) smo naleteli na primer, kjer je ekipa tri mesece iskala vzrok dimenzijskega odstopka, čeprav je bil identičen problem dokumentiran v 8D poročilu izpred dveh let. Takratni inženir je že zapustil podjetje, novo zaposleni pa za poročilo niso vedeli.

Vsak rešen 8D primer se v sistemu samodejno obdela in doda v bazo znanja. Ne gre za učenje v smislu nevronskih mrež; sam jezikovni model se ne preuči. Indeks za iskanje preprosto raste z vsakim novim primerom. Sprva smo pričakovali, da bo ta rast počasna, a se je izkazalo, da podjetja že po nekaj tednih uporabe zgradijo dovolj obsežen indeks za uporabne predloge. Pričakovana dolgoročna korist je, da bodo novi inženirji z uporabo sistema dostopali do izkušenj predhodnikov.

2.5 Koncept »zlate niti«

Izraz »zlata nit« označuje celotno verigo dokazov od začetnega simptoma napake do sistemske preventive (Slika 2). V ročnih okoljih je faza D2 morda zabeležena v kupčevem portalu, o D4 se razpravlja po elektronski pošti, ukrepi D7 pa končajo v vzdrževalni preglednici. Ko zunanji presojevalec zahteva sledljivostno pot, jo mora ekipa rekonstruirati nazaj. Pri tem se razkrijejo vrzeli.

Opisani sistem vzpostavlja relacijski podatkovni model, v katerem mora vsaka naslednja faza 8D podedovati in nasloviti podatke iz prejšnje. Popravnega ukrepa ni mogoče zabeležiti, če ne cilja na potrjeni korenski vzrok; korenskega vzroka ni mogoče sprejeti, če ne pojasni opredeljenega simptoma. To je programska omejitev, ki zagotavlja pripravljenost na presojo že v sami zasnovi sistema.

Slika 2: Sledljivostna veriga (»zlata nit«) od začetnega simptoma do sistemske preventive

2.6 8D metoda z umetno inteligenco in skladnost z IATF 16949

Metodologijo 8D je leta 1987 razvilo podjetje Ford Motor Company kot timski pristop k reševanju problemov [2, 3]. Osem disciplin, od sestave ekipe (D1) do zaključka (D8), sledi strogi logični sekvenci [11]. V praksi se ta sekvenca pogosto poruši: inženirji pod časovnim pritiskom preskočijo zadrževalne ukrepe (D3), analiza korenskih vzrokov zdrsne v ugibanje, preventivni ukrepi pa ostanejo na papirju. Ena sama zahtevnejša 8D reklamacija zlahka požre več dni kumulativnega dela.

Standard IATF 16949 [2] zahteva učinkovite procese popravnih ukrepov, vključno z identifikacijo korenskih vzrokov z dokazanimi analitičnimi orodji (točka 10.2.3), dokazno podprto verifikacijo in dokumentiranimi preventivnimi ukrepi. Zunanji presojevalci redno sledijo logični niti od začetnega simptoma prek korenskega vzroka do sistemske preventive. Vsaka vrzel v tej verigi povzroči ugotovitev o neustreznosti. Z vključitvijo UI-podpore v posamezne korake 8D metode je zagotovljena postopkovna popolnost ob sočasnem zmanjšanju ročnega dela. Izboljšave so povzete v Tabeli 1.

Tabela 1: Primerjava ročnega in z UI-podprtega 8D delovnega procesa

8D faza	Ročni postopek	UI-podprt postopek	Izboljšava
D0–D1	Vodja dodeli ekipo po spominu in razpoložljivosti.	UI predlaga člane ekipe na podlagi izkušenj s podobnimi tipi napak.	Hitrejša mobilizacija z relevantnim znanjem
D2	Inženir ročno sestavi opise problema; polja 5W2H izpolni ročno (~43 min).	UI razčleni vhodne podatke in samodejno izpolni 5W2H iz vse priložene dokumentacije (~4,5 min).	9,6x krajši čas opredelitve problema
D3	Zadrževalni ukrepi se vodijo v ločeni preglednici.	Zadrževalni ukrepi zabeleženi znotraj 8D toka, povezani s podatki iz D2.	Zagotovljena sledljivost
D4	Ročni Ishikawa brainstorming; iskanje po datotečnih arhivih (~239 min).	RAG poišče podobne pretekle primere; UI predlaga razvrščene korenske vzroke (~32 min).	7,5x krajši čas; podatkovno podprte odločitve
D5–D6	Popravni ukrepi od začetka; verifikacija pogosto predpostavljena.	UI predlaga ukrepe s podatki o pretekli učinkovitosti; verifikacija obvezna pred nadaljevanjem.	Dokazno podprti popravni ukrepi
D7–D8	Preventiva v sistemu za vzdrževanje; poročilo ročno sestavljeno.	Preventiva povezana z verigo korenskih vzrokov; poročilo samodejno generirano.	Samodejno generiranje poročil za presojo

Umetna inteligenca ne nadomešča inženirske presoje. Odpravlja najbolj zamudne administrativne korake, da se inženirji osredotočijo na analizo. Vodeni delovni tok preprečuje postopkovne bližnjice: popravnega ukrepa v D5 ni mogoče zabeležiti, če ne naslavlja potrjenega korenskega vzroka iz D4. To strogo dedovanje med fazami zadostuje zahtevi IATF 16949, točka 10.2.3, in zahtevam standarda ISO 9001 [12].

3 IZKUŠNJE IZ PRAKSE

3.1 Okolje pilotnega testiranja

Pilotno testiranje smo izvedli v okviru osmih pilotnih sej s slovenskimi proizvodnimi podjetji (70 do 1.700 zaposlenih) v obdobju januar–marec 2026. Podjetja delujejo na področjih avtomobilske dobavne verige Tier-1 in Tier-2, tlačnega liva in precizne obdelave (aluminij, CNC), elektrotehnike, brizganja polimerov ter strojegradnje. V vsaki seji smo uporabili dejanske 8D primere z realnimi podatki o napakah: livna poroznost, napake varov, manjkajoče komponente, nepopolna montaža.

3.2 Metodologija merjenja

Izbrali smo deset reprezentativnih primerov 8D. Časovno smo izmerili tako ročno kot UI-podprto izvedbo faz D2 in D4, torej obeh faz, ki sta najbolj časovno potratni in najbolj odvisni od izkustvenega znanja. Čas smo merili s štoparico. Vsak primer je bil najprej rešen ročno, nato ponovno rešen z UI podprtim delovnim tokom.

Za D4 je merilo uspešnosti zahtevalo vsaj en tehnični korenski vzrok za pojav napake, enega za neodkritje in enega organizacijskega (sistemskega) korenskega vzroka za oboje (skupaj 4 vzroke). Vse predloge so pregledali usposobljeni inženirji, preden so šteli kot veljavne.

3.3 Rezultati

Izmerjeni rezultati so zbrani v Tabeli 2. Skupni faktor izboljšave 7,8x je dosleden skozi celoten vzorec. Največji prihranki nastanejo pri fazi D4, kjer semantično pridobivanje nadomesti ure ročnega pregledovanja dokumentov.

Tabela 2: Primerjava učinkovitosti med ročnim in UI-podprtim 8D postopkom

Meritev	Ročni postopek	UI-podprt postopek	Faktor izboljšave
Čas zaključka D2 (povp.)	2.610 s (~43,5 min) [1.620–3.330]	273 s (~4,6 min) [223–424]	9,6x
Čas zaključka D4 (povp.)	14.346 s (~239 min) [10.254–17.674]	1.903 s (~31,7 min) [1.284–2.404]	7,5x
Skupni čas D2+D4 (povp.)	16.956 s (~283 min)	2.176 s (~36,3 min)	7,8x
Velikost vzorca (n)	10 primerov 8D	10 primerov 8D	–

3.4 Razširitev podatkov na letne inženirske ure

Prihranjen čas na posamezen 8D primer znaša povprečno 246 minut (D2 + D4). Metodologija 8D zahteva medfunkcijsko ekipo, ki šteje po navadi 5 do 7 članov iz razvojnega inženiringa, procesnega inženiringa, kakovosti, proizvodnje in vzdrževanja [11].

Ob upoštevanju 40 do 120 8D primerov letno (razpon za slovenske Tier-1 in Tier-2 dobavitelje), 5 do 7 članov ekipe in faktorja posredne vpletenosti 1,5x do 2,5x, ki zajema koordinacijo med oddelki, medsebojne strokovne preglede, poročanje vodstvu in komunikacijo z naročnikom, je ocenjeni letni prihranek med 1.200 in 10.000 inženirskih ur. Za podjetje s 500 zaposlenimi in 80 8D primeri / leto naša ocena znaša približno 3.500 do 5.500 inženirskih ur letno (meritve, izvedene med pilotnimi sejami pri pilotnih partnerjih v obdobju januar–marec 2026).

3.5 Razprava

Izmerjeni faktorji izboljšave (9,6x za D2 in 7,5x za D4) so na prvi pogled visoki. Razlog je preprost: večina časa pri ročnem postopku ne gre za miselno delo, temveč za iskanje. Pregledovanje PDF datotek, odpiranje starih Excel preglednic, klicanje kolegov. Semantično iskanje to časovno zahtevno pregledovanje arhivov stisne v sekunde. Sam inženirski razmislek, odločanje o tem, kateri korenski vzrok je verjeten in kateri test bo potrdil hipotezo, ostane v celoti na človeku. Po naših izkušnjah se večina inženirjev sprva odzove skeptično, a hitro prepozna vrednost, ko sistem v nekaj sekundah vrne primer, ki bi ga sicer iskali pol ure.

V primerjavi s klasičnimi orodji za vodenje kakovosti (SAP QM, Siemens Polarion, PTC Windchill Quality) opisani sistem ne nadomešča obstoječih ERP ali PLM platform. Dodaja sloj semantičnega razumevanja dokumentov, ki ga ti sistemi ne ponujajo [5]. Podoben pristop so opisali Kernan Freire in sodelavci [13], ki so razvili sistem za deljenje znanja v tovarnah na osnovi velikih jezikovnih modelov. Njihove ugotovitve potrjujejo, da je za industrijsko okolje bistvenega pomena lokalizirana baza znanja.

Za Tier-1 in Tier-2 proizvajalce v slovenskem prostoru so praktične implikacije večplastne. Sistem zmanjša odvisnost od posameznih osrednjih inženirjev; znanje ostane v organizaciji tudi po njihovem odhodu. Odzivni čas pri reševanju kupčevih reklamacij se občutno skrajša, kar je pri avtomobilskih OEM-jih običajno vezano na stroge roke (24 ur za zadrževalne ukrepe). Poleg tega sistem standardizira kakovost 8D poročil, kar zmanjšuje število ugotovitev o neskladnostih na presojah.

Vredno je omeniti, da uporaba UI pri odločanju o kakovosti prinaša tudi tveganja: slepo zaupanje predlogom lahko vodi v opuščanje kritičnega inženirskega razmisleka, zgodovinski podatki pa lahko vsebujejo pristranskosti. Zato mora vsak predlog sistema ostati le izhodišče za inženirsko presojo.

3.6 Omejitve in prihodnje delo

Sistem deluje tudi brez zgodovinskih podatkov, a se predlogi na podlagi RAG znatno izboljšajo, ko baza znanja raste. Pri manj kot ~20 preteklih primerih so rezultati semantičnega iskanja razredčeni. Sistem predlaga med 8 in 16 možnih korenskih faktorjev; inženirji med njimi izberejo verjetne, kar predstavlja 20–30% vseh ponujenih. V dosedanjih testih je bil med izbranimi faktorji vedno tudi dejanski korenski vzrok. Pri dolgotrajnih serijskih napakah z bogato zgodovino podatkov se delež potrjenih predlogov občutno zvišuje. Meritve temeljijo na desetih primerih 8D z dodatnimi ~90 primeri testiranimi interno. Vrstni red izvedbe (najprej ročno, nato UI) lahko vnese učni učinek; prihodnje študije bodo uporabile naključni vrstni red. Na podlagi kriterijev standarda ISO 16290 [14] je sistem ocenjen na stopnjo TRL 9. Načrtovane izboljšave obsegajo integracijo z MES / ERP sistemi, razširitev na ostale

metode odkrivanja napak in sistematičnega reševanja (RCA metode), napovedno analitiko za ponavljajoče se vzorce napak ter integracijo podatkov o procesni sposobnosti (SPC metode) [15].

4 ZAKLJUČEK

V prispevku smo opisali z umetno inteligenco podprt organizacijski spomin za inženiring kakovosti v proizvodnji, zasnovan okrog metodologije 8D in pristopa RAG. Pilotno testiranje na desetih realnih primerih pri osmih slovenskih podjetjih v obdobju januar–marec 2026 je izkazalo skoraj 8-kratno pohitritev (7,8x) za fazi D2 in D4, s posamičnima faktorjema 9,6x oziroma 7,5x.

Sistem vzpostavlja celotno sledljivostno verigo od simptoma napake do sistemske preventive in s tem neposredno naslavlja zahteve standarda IATF 16949 za presoje. Za podjetje, ki letno obdeluje 80 poročil 8D, to pomeni ocenjenih 3.500 do 5.500 prihranjenih inženirskih ur na letni ravni. Vrednost sistema narašča s količino vnesenih podatkov, kar nakazuje, da je dolgoročno pozicioniranje kot organizacijski spomin bistvena prednost pred klasičnimi sistemi za vodenje kakovosti.

Za proizvodna podjetja v slovenskem in širšem evropskem prostoru, ki se soočajo z odhodi izkušenih kadrov, naraščajočo kompleksnostjo kupčevih zahtev in vse strožjimi zahtevami presoj, takšen sistem ne predstavlja zgolj prihranka časa. Predstavlja varovalko pred izgubo kritičnega znanja, ki ga podjetje kopiči z leti; znanja, ki ga je težko dokumentirati in še težje prenesti na naslednjo generacijo inženirjev. Prihodnje delo bo osredotočeno na razširitev vzorca meritev, integracijo z obstoječimi ERP in MES sistemi ter na validacijo dolgoročnega učinka na ponavljanje napak.

Razkritje: Avtor je soustanovitelj podjetja Galileon d.o.o., ki razvija opisani sistem.

LITERATURA

[1] AMERICAN SOCIETY FOR QUALITY. Cost of Quality (COQ) [na spletu]. Milwaukee: ASQ, 2025 [ogled 13. 4. 2026]. Dostopno: https://asq.org/quality-resources/cost-of-quality.
[2] INTERNATIONAL AUTOMOTIVE TASK FORCE. IATF 16949:2016 Quality management system requirements for automotive production and relevant service parts organisations. 1. izdaja. Southfield: IATF, 2016.
[3] RUSMAN, Karyadi in Dorina, HETHARIA. Quality improvement through 8D methodology: an automotive industry case study. Operations Excellence: Journal of Applied Industrial Engineering. 2025, vol. 17, no. 1, str. 1–11.
[4] PSOMAS, Evangelos in Jiju, ANTONY. The effectiveness of the ISO 9001 quality management system in SMEs. International Journal of Quality & Reliability Management. 2015, vol. 32(7), str. 753–771.
[5] SONY, Michael, Jiju, ANTONY in Jacqueline Ann, DOUGLAS. Essential ingredients for the implementation of Quality 4.0: a narrative review of literature and future directions for research. The TQM Journal. 2020, vol. 32(4), str. 779–793.
[6] LEWIS, Patrick idr. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. V: LAROCHELLE, Hugo idr., ed. Advances in Neural Information Processing Systems 33. Virtual: Curran Associates, 2020, str. 9459–9474.
[7] CHEN, Lun-Chi idr. Application of retrieval-augmented generation for interactive industrial knowledge management via a large language model. Computer Standards & Interfaces. 2025, vol. 94, 103995.
[8] MORALES MATAMOROS, Oswaldo idr. Artificial Intelligence for Quality Defects in the Automotive Industry: A Systemic Review. Sensors. 2025, vol. 25(5), 1288.
[9] BARSALOU, Matthew, Marta, GRABOWSKA in Robert, PERKIN. Inquiry into the Effectiveness of Eight Discipline-Based Problem-Solving. Quality Innovation Prosperity. 2023, vol. 27, no. 2, str. 61–76. DOI: 10.12776/qip.v27i2.1839.
[10] AMAZON WEB SERVICES. Amazon Titan Text Embeddings V2 [na spletu]. Seattle: AWS, 2024 [ogled 10. 4. 2026]. Dostopno: https://docs.aws.amazon.com/bedrock/latest/userguide/titan-embedding-models.html
[11] TAGUE, Nancy R. The Quality Toolbox. 2. izdaja. Milwaukee: ASQ Quality Press, 2005. ISBN 978-0-87389-639-9.
[12] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 9001:2015 Quality management systems - Requirements. Geneva: ISO, 2015.
[13] KERNAN FREIRE, Samuel idr. Knowledge Sharing in Manufacturing using Large Language Models: User Evaluation and Model Benchmarking. arXiv preprint. 2024, arXiv:2401.05200.
[14] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 16290:2013 Space systems - Definition of the Technology Readiness Levels (TRLs) and their criteria of assessment. Geneva: ISO, 2013.
[15] ESCOBAR, Carlos A., Megan E., MCGOVERN in Ruben, MORALES-MENENDEZ. Quality 4.0: a review of big data challenges in manufacturing. Journal of Intelligent Manufacturing. 2021, vol. 32(8), str. 2319–2334.

Numerical Investigation of Sinusoidal and Arc Waveform Leading-Edge Tubercle Geometry

Matic Medvešek[1,2], Simon Muhič[1,2]

[1]*Faculty of industrial engineering Novo mesto, Slovenia*
E-mail: matic.medvesek@fini-unm.si

[2]*Rudolfovo - Science and Technology Centre Novo Mesto, Slovenia*
E-mail: simon.muhic@rudolfovo.eu

Abstract: Leading-edge tubercles have been widely studied as passive flow-control devices for improving aerodynamic performance and modifying stall behaviour. Most investigations focus on sinusoidal leading-edge geometries due to their ease of implementation. However, arc-based (tangent circular) waveforms exhibit different curvature distributions that may influence vortex formation and flow separation dynamics. In contrast to the continuously varying curvature of a sinusoidal profile, arc-based geometries maintain slope continuity while introducing discrete curvature transitions.

The present study compares sinusoidal and arc-based leading-edge geometries to investigate the effect of waveform definition on aerodynamic performance. Both configurations are prescribed with identical wavelength and amplitude to ensure geometric consistency, differing only in curvature continuity. Three-dimensional unsteady simulations are performed at low Reynolds number under post-stall conditions to evaluate lift and drag coefficients, flow separation length, and velocity magnitude. Results indicate that curvature continuity significantly affects vortex formation and spanwise flow organization. The arc-based geometry produces higher aerodynamic performance, whereas the sinusoidal waveform yields lower overall aerodynamic performance. The findings demonstrate that leading-edge performance depends not only on amplitude and wavelength but also on geometric properties of the leading-edge silhouette that determine curvature distribution. This study provides new insight into waveform-sensitive flow-control mechanisms and offers guidance for geometry-driven aerodynamic optimization.

Keywords*:* Tubercles, Computational Fluid Dynamics, Detached Eddy Simulation, Flow separation, Leading-edge geometry.

Article Classification*: Scientific Paper*

1 INTRODUCTION

Leading-edge tubercles, inspired by the pectoral fins of the Humpback whale, have been extensively investigated as passive flow-control devices due to their ability to delay stall, modify separation behaviour, and promote formation of streamwise vortices. These vortical structures enhance momentum exchange within the boundary layer, thereby improving aerodynamic performance. Early experimental and numerical studies demonstrated that leading edges with tubercles can improve post-stall lift characteristics and increase aerodynamic stability through spanwise flow redistribution and vortex-induced boundary layer energization [1–4]. Subsequent research proved that the aerodynamic performance is strongly dependent on the geometric parameters of the tubercles, primarily amplitude and wavelength [5–7]. Optimal configurations have been shown to vary with flow regime, indicating a strong coupling between geometry and operating conditions.

Recent advances in high-fidelity numerical methods and experimental diagnostics have enabled more detailed investigations of tubercle-induced flow physics. Approaches such as large-eddy simulation (LES) and detached-eddy simulation (DES) have revealed complex vortices, delayed separation, and three-dimensional wake dynamics [8, 9]. Recent studies have further explored multi-parameter optimization strategies, incorporating Reynolds number effects, airfoil profile dependency, and amplitude-to-wavelength ratios to refine aerodynamic performance [6, 10]. Despite extensive research, most studies adopt a sinusoidal definition of the leading-edge geometry due to its analytical simplicity and ease of parameterization [5, 7, 11–16]. Within this framework, amplitude and wavelength are implicitly treated as the primary descriptors of the geometry, suggesting that they uniquely determine aerodynamic behaviour. However, such an assumption neglects the role of higher-order geometric characteristics, particularly the distribution of curvature along the leading edge.

From a geometric perspective, amplitude and wavelength define only the global parameters of the surface, while the local curvature distribution depends on the specific definition of the waveform. Sinusoidal geometries exhibit continuously varying curvature, whereas arc-based (tangent circular) geometries maintain slope continuity. The effect of curvature definition, independent of amplitude and wavelength, remains insufficiently understood. No controlled comparison between sinusoidal and arc-based leading-edge geometries under identical global parameters has been reported. This study addresses this gap by isolating curvature distribution effects. Three-

dimensional unsteady simulations at low Reynolds numbers and post-stall conditions are used to evaluate lift, drag, and flow separation characteristics. The objective is to clarify curvature-driven flow control mechanisms and support geometry-based optimization of tubercle leading edges.

2 METHODS

A comparative analysis was performed to evaluate the influence of leading-edge geometry definition on wings equipped with tubercles. Tubercles can be defined using different leading-edge silhouette geometries. The alternating silhouette is most commonly defined using a sinusoidal function. Alternatively, it can be defined using arcs arranged in alternating orientation with tangential continuity. Arc-based geometries provide a more uniform shape, characterized by constant curvature. In contrast, sinusoidal geometries exhibit continuously varying curvature.

Computational Fluid Dynamics (CFD) simulations were performed to analyse and compare different wing configurations. A k–ω SST Detached Eddy Simulation (DES), representing a hybrid RANS–LES approach, representing a hybrid RANS–LES methodology, was employed to resolve large-scale unsteady flow structures while modeling near-wall turbulence. The numerical mesh consisted of approximately 4 million cells.

2.1 Parameters for analysis

A set of aerodynamic parameters was defined to evaluate flow behaviour and distinguish the effects of leading-edge geometry. Figure 1 presents the basic geometric parameters of tubercles, being amplitude and wavelength. Additionally silhouettes of sinusoidal and arc-based waveform geometries are shown. All modified wings were based upon the baseline geometry, which was the NACA 0012 profile wing without tubercles.

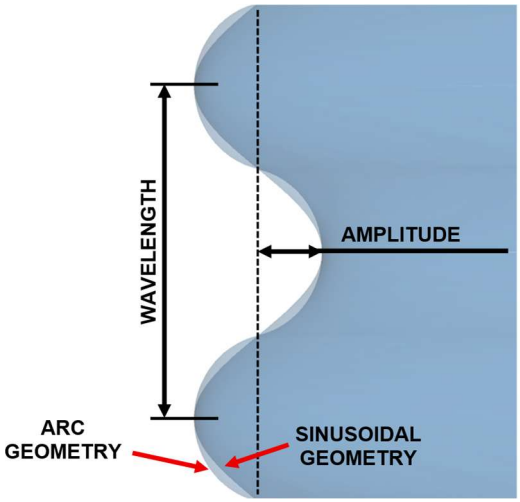


Figure 1: Tubercle geometry definition

The wing chord length was set to 80 mm, and the angle of attack was fixed at 16 °, corresponding to stall conditions. Fluid properties were set to freestream flow velocity $2\ \mathrm{m/s}$, density $\rho = 998.2\ \mathrm{kg/m^3}$ and dynamic viscosity $\mu = 0.0010005\ \mathrm{Pas}$, corresponding to Reynolds flow regime of $Re = 1.6 \cdot 10^5$. The wing was divided into spanwise sections, each representing one intertubercle region, enabling localized analysis. For each section, lift and drag coefficients, flow separation length, and velocity magnitude were evaluated. Lift and drag were obtained by integrating surface traction forces and decomposing them into components, relative to the freestream flow direction. The corresponding coefficients were calculated using reference flow conditions and sectional area.

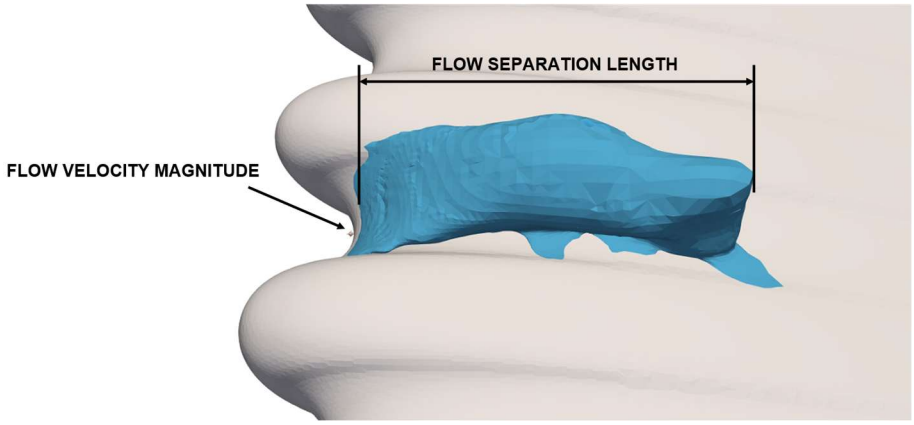


Figure 2: Flow velocity magnitude and separation length definition

The flow separation region was identified using the condition of zero velocity component in the freestream direction. The resulting contour shown on Figure 2 was used to evaluate separation length. Velocity magnitude was analysed to capture local flow variations associated with tubercle-induced effects. All quantities were evaluated at t = 5 s of the transient simulation.

3 RESULTS AND DISCUSSION

Simulations were conducted using the defined simulation parameters. Simulation results and their interpretations of different leading-edge geometries are presented.

3.1 Lift and drag coefficient analysis

Lift and drag coefficients provide primary insight into the aerodynamic performance of the wing. Each simulated wing is divided into 10 sections, each corresponding to one intertubercle region.

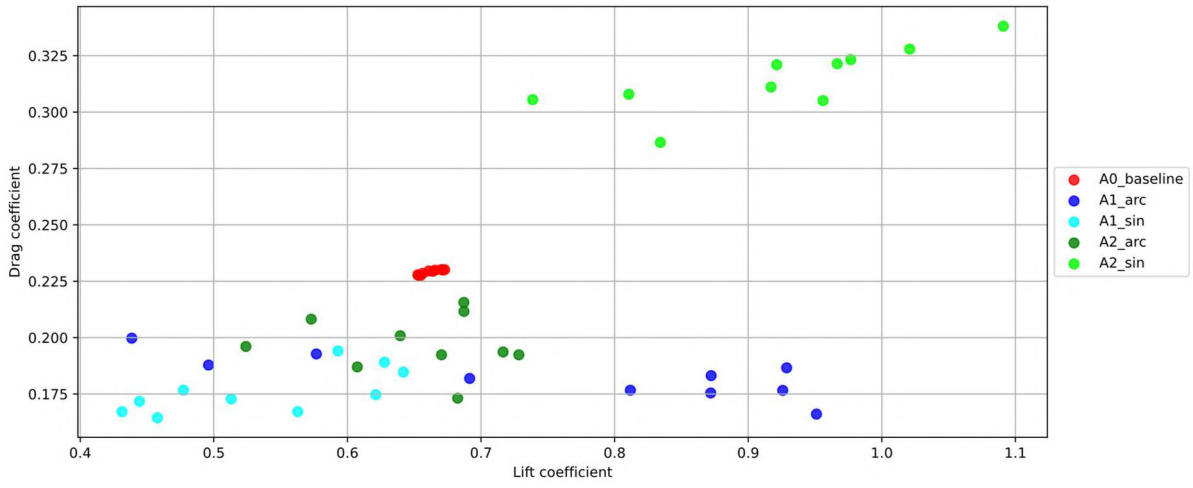


Figure 3: Lift versus drag coefficient

Figure 3 presents the correlation between lift and drag coefficients, where each data point corresponds to an individual spanwise section of a given wing configuration. The A2_sin configuration exhibits a distinct aerodynamic response, characterized by the highest drag coefficient among all tested cases, exceeding the baseline by 37.3%. Despite this significant drag penalty, it simultaneously achieves the highest lift performance. A comparison between leading-edge geometries reveals a clear geometry-dependent trend. Arc-based configurations consistently demonstrate both drag reduction and lift enhancement relative to the baseline. Within this group, lower-amplitude

arc geometries produce higher lift coefficients. In contrast, sinusoidal configurations exhibit increased lift with increasing amplitude, indicating fundamentally different scaling behaviour between the two geometric formulations.

These observations suggest that aerodynamic performance is strongly influenced by the curvature distribution of the leading edge. Variations in curvature directly affect local flow acceleration, vortex generation, and separation dynamics, thereby governing the overall lift–drag characteristics. Considering overall aerodynamic efficiency, the A1_arc configuration exhibits the most favourable performance, followed by A2_arc. Although sinusoidal configurations generally underperform relative to arc-based geometries, they still provide an improvement in lift-to-drag ratio compared to the baseline case.

3.2 Flow velocity analysis

The simulations showed that the flow velocity field along the span of the wing exhibits significant variations, between the individual sections. These variations correlate with the extent of flow separation in each section.

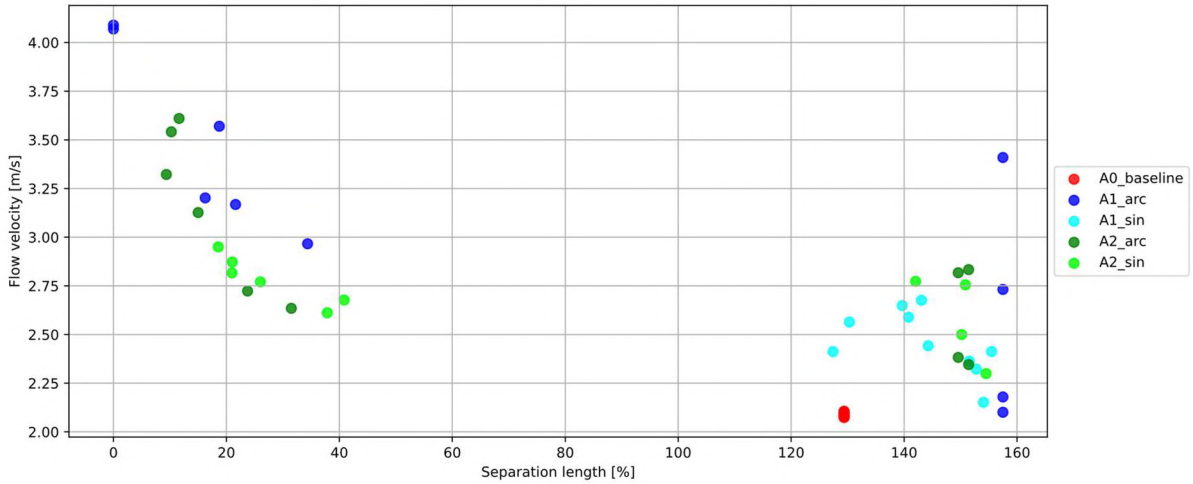


Figure 4: Flow velocity versus separation length analysis

Figure 4 presents the relationship between local flow velocity and separation length for different tubercle amplitudes and leading-edge geometries. The separation length is expressed as a percentage relative to the chord length of the wing (80 mm). The baseline configuration (A0_baseline), representing a leading edge without tubercles, exhibits a uniform distribution of flow velocity and separation length along the span. The separation length consistently exceeds 100% of the chord length, indicating a fully developed deep-stall condition across the entire wing.

The results reveal the formation of two distinct flow regimes. The first regime is characterized by low separation lengths, ranging from 0% to 42% of the chord, while the second regime corresponds to significantly larger separation regions, between 130% and 157%. The low-separation regime is observed in the A1_arc, A2_arc, and A2_sin configurations, whereas the baseline and A1_sin configurations fall within the high-separation regime. Notably, the A1_sin case remains in a fully stalled condition despite the presence of tubercles, highlighting a strong dependence on geometric formulation rather than amplitude alone.

Within the low-separation regime, a broad range of flow velocities is observed. The lowest recorded velocities, of 2.6 m/s, are associated with the higher-amplitude configurations. In contrast, the A1_arc configuration exhibits the highest local velocities, reaching up to 4.1 m/s, including sections where flow separation is entirely suppressed (0 % separation length). A clear linear trend is observed within the low-separation regime, indicating that decreasing local flow velocity correlates with increasing separation length. This behaviour suggests a direct relationship between the local kinetic energy of the flow and its ability to remain attached, consistent with boundary-layer theory and vortex-induced flow stabilization mechanisms reported in past research [15, 17–23]

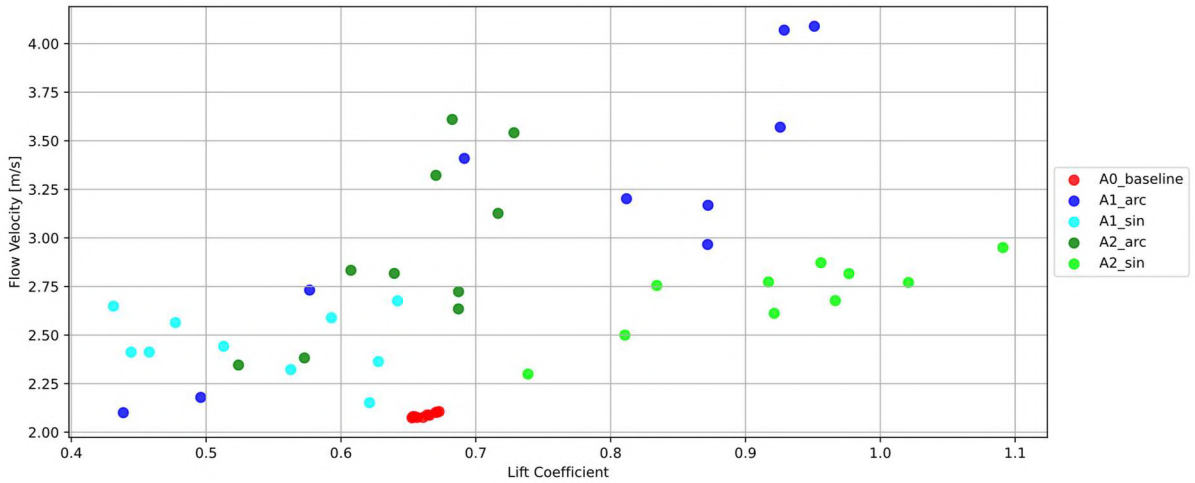


Figure 5: Lift coefficient versus flow velocity

Figure 5 presents the correlation between lift coefficient and local flow velocity based on sectional analysis. All configurations with tubercles exhibit increased local flow velocity relative to the baseline; however, this does not consistently result in higher lift. The A1_sin configuration produces lower lift coefficients than the baseline despite increased velocity, indicating unfavourable aerodynamic characteristics. The arc-based configurations (A1_arc, A2_arc) show limited improvement, with A2_arc reaching velocities up to 3.6 m/s but only marginal lift gains. In contrast, the A2_sin configuration forms a distinct regime, consistently achieving higher lift across all sections, with a maximum lift coefficient of 1.09. This indicates that lift enhancement is not governed solely by the magnitude of the local flow velocity, but also by the structure and stability of the flow. However, lift alone is not a sufficient metric for overall aerodynamic performance. As shown in Figure 3, the A2_sin configuration also exhibits a significant increase in drag, indicating a trade-off between lift improvement and aerodynamic efficiency.

4 CONCLUSION

The present study demonstrates that the aerodynamic performance of tubercle leading edges is fundamentally governed not only by global geometric parameters, such as amplitude and wavelength, but also by the distribution and continuity of curvature along the leading edge. By isolating curvature effects through geometries with identical global dimensions, it is shown that waveform definition alone can produce different aerodynamic responses. Arc-based geometries consistently exhibit superior overall aerodynamic efficiency, achieving a more favourable balance between lift and drag. In contrast, sinusoidal geometries, particularly at higher amplitudes, generate enhanced lift but also exhibit drag penalties, resulting in reduced overall aerodynamic efficiency. This highlights a fundamental trade-off between lift augmentation and flow control effectiveness.

Flow-field analysis further confirms that curvature directly influences the formation, stability, and coherence of streamwise vortices, which are responsible for boundary-layer energization and separation control. The emergence of distinct separation regimes confirms that curvature distribution governs the transition between attached and fully separated flow states. These findings support a curvature-driven flow control relation, in which local geometric gradients, rather than solely global parameters, determine the aerodynamic behaviour. Consequently, the conventional amplitude–wavelength parametrization alone is insufficient for fully characterizing the tubercle leading-edge aerodynamic performance.

4.1 Opportunities for future research

Future work should include a systematic analysis across a wider range of angles of attack to characterize aerodynamic performance in pre-stall, stall, and post-stall regimes. Additionally, configurations combining multiple tubercle geometries on a single wing (e.g., sinusoidal and arc-based sections) should be investigated to evaluate potential advantages of interactive effects on flow control.

Acknowledgement
The research was co-funded by the Slovenian Research and Innovation Agency (ARIS) through the annual work program of Rudolfovo.

REFERENCES

[1] MIKLOSOVIC, D. S. et al. Leading-edge tubercles delay stall on humpback whale (Megaptera novaeangliae) flippers. Physics of Fluids. 2004, 16(5). ISSN 10706631. 10.1063/1.1688341
[2] E. FISH, Frank et al. The Tubercles on Humpback Whales' Flippers: Application of Bio-Inspired Technology. V: Applying Mechanical Design to Experimental Biology. Salt Lake City: Oxford University Press, 2011, str. 203–213.
[3] E. FISH, Frank, Laurens E. HOWLE in Mark M. MURRAY. Hydrodynamic flow control in marine mammals. V: Going with the Flow: Ecomorphological Variation across Aquatic Flow Regimes. San Antonio: Oxford University Press, 2008, str. 788–800.
[4] E. FISH, Frank in Juliann M. BATTLE. Hydrodynamic Design of the Humpback Whale Flipper. Journal of Morphology. 1995, 51–60.
[5] ZHAO, Fuwang et al. Vortex-induced vibration of a sinusoidal wavy cylinder: The effect of wavelength. Physics of Fluids. 2024, 36(8). ISSN 10897666. 10.1063/5.0219753
[6] LUO, Dahai et al. CFD analysis of the aerodynamic performance of airfoils with leading-edge tubercles: A parametric investigation. Ocean Engineering, 331, 121311. ISSN 0029-8018. https://doi.org/10.1016/j.oceaneng.2025.121311
[7] LEE HANSEN, Kristy. Effect of Leading Edge Tubercles on Airfoil Performance. Adelaide, 2012. Research Thesis. The University of Adelaide.
[8] HEINZ, Stefan. A review of hybrid RANS-LES methods for turbulent flows: Concepts and applications. Progress in Aerospace Sciences. 2020, 114, 100597. ISSN 0376-0421. 10.1016/J.PAEROSCI.2019.100597
[9] SPALART, Philippe et al. Comments on the Feasibility of LES for Wings, and on a Hybrid RANS/LES Approach. 1997.
[10] ASGHAR, Allan, Ruben E. PEREZ in William D. E. ALLAN. Further Parametric Investigation of Leading-Edge Tubercles on the Propeller Performance. V: AIAA SciTech Forum. San Diego: AIAA Scitech 2019 Forum, 2019, str. 1–15.
[11] POST, Martiqua L et al. Effect of bio-inspired sinusoidal leading-edges on wings. Aerospace Science and Technology. 2018, 81, 128–140. ISSN 1270-9638. https://doi.org/10.1016/j.ast.2018.07.043
[12] SUN, Yifei et al. Aerodynamic performance of a sinusoidal wavy cylinder at subcritical and critical Reynolds numbers. Journal of Wind Engineering and Industrial Aerodynamics. 2023, 232, 105267. ISSN 0167-6105. 10.1016/J.JWEIA.2022.105267
[13] BOLZON, Michael D P, Richard M KELSO in Maziar ARJOMANDI. Formation of vortices on a tubercled wing, and their effects on drag. Aerospace Science and Technology. 2016, 56, 46–55. ISSN 1270-9638. https://doi.org/10.1016/j.ast.2016.06.025
[14] FAN, Menghao et al. Numerical and experimental investigation of bionic airfoils with leading-edge tubercles at a low-Re in considering stall delay. Renewable Energy. 2022, 200, 154–168. ISSN 0960-1481. 10.1016/J.RENENE.2022.09.123
[15] A. VAN NIEROP, Ernst, Silas ALBEN in Michael P. BRENNER. How Bumps on Whale Flippers Delay Stall: An Aerodynamic Model. Physical review letters. 2008.
[16] HANSEN, Kristy L., Richard M. KELSO in Bassam B. DALLY. Performance Variations of Leading-Edge Tubercles for Distinct Airfoil Profiles. AIAA Journal. 2011, 49(1), 185–194. ISSN 0001-1452. 10.2514/1.J050631
[17] JOHARI, H. et al. Effects of Leading-Edge Protuberances on Airfoil Performance. AIAA JOURNAL. 2007, 45(11), 2634–2642.
[18] T. H. NEW, Daniel in Feng Ng BING. Flow Control Through Bio-inspired Leading-Edge Tubercles. Singapore: Springer, 2020.
[19] W. WEBER, Albert et al. Computational Evaluation of the Performance of Lifting Surfaces with Leading-Edge Protuberances. JOURNAL OF AIRCRAFT. 2011, 48(2), 591–600.
[20] PEDRO, Hugo T.C. in Marcelo H. KOBAYASHI. Numerical study of stall delay on humpback whale Flippers. 46th AIAA Aerospace Sciences Meeting and Exhibit. 2008. 10.2514/6.2008-584
[21] NEW, Daniel T.H. et al. Flow Control by Hydrofoils with Leading-Edge Tubercles. Flow Control Through Bio-inspired Leading-Edge Tubercles: Morphology, Aerodynamics, Hydrodynamics and Applications. 2020, 85–109. 10.1007/978-3-030-23792-9_4
[22] CUSTODIO, D. The Effect of Humpback Whale-Like Protuberances on Hydrofoil Performance. 2007.
[23] LEKKALA, M. R. et al. Numerical investigations of flow over wavy cylinders at sub-critical Reynolds number. Ocean Engineering. 2023, 269, 113501. ISSN 0029-8018. 10.1016/J.OCEANENG.2022.113501

Modeling of electric motor drives with three-phase induction machines and various types of mechanical loads

Rudolf Pušenjak[1]

[1]Faculty of Industrial Engineering Novo mesto, Slovenia
E-mail: rudolf.pusenjak@fini-unm.si

Abstract: *The paper presents the modeling of electric motor drives with three-phase induction machines subjected to various types of mechanical loads. For modeling of the electric motor drives with induction machines the arbitrary qd0 reference frame and stationary reference frame are used. The governing differential and algebraic equations of the induction machine are derived and coupled by the governing equations of the mechanical part of the drive by using the programming environment of Mathematica. The various working regimes in motor-, generator- and brake-region are investigated. The results of the dynamical simulation are compared to the results, obtained by Matlab-Simulink.*

Keywords*: electric motor drives, induction machine, Park's transformation, model equations of the drive, electric motor drive regimes*

Article Classification*: Scientific Paper*

1 INTRODUCTION

The induction motors are converters of electrical energy into mechanical work, which by the development of power electronics became even more useful and are found in virtually all industrial areas of electric motor drives including rail vehicles [1] and electric vehicle drives [4]. They are rotating machines in which its rotor rotates at asynchronous speed and are characterized by simple construction, high efficiency, reliable operation in difficult conditions, minimal need for maintenance, low price and advanced speed control techniques.

The induction machine, used in electric motor drives can operate in three modes of operation: in motor, generator and brake mode, respectivelly. To enable dynamic simulation of drives with three-phase induction machines, the mathematical model in programming environment Mathematica is developed in this paper. In addition, this software system uses symbolic computation capabilities, which allows the derivation of the governing equations of a three-phase induction machine, thereby saving a lot of laborious manipulation work. The paper is organized as follows. In the second chapter, using machine variables, the governing differential and algebraic equations of the three-phase induction machine are first written in vector form. These vector equations contain time-varying coefficients, which make their solution difficult. In order to facilitate the solution of these equations, the three-phase machine variables are decoupled using the Park's transformation into orthogonal two-phase *q*- and *d*- variables, so that the entire mathematical model is expressed in *qd*0 variables. The equations of the induction machine are supplemented with equilibrium torque equations. In the third chapter, dynamic simulations of the electric motor drive with the selected type of induction machine and a direct coupling of the mechanical load are performed in all three operating modes. In addition, the drive with the same type of the electric motor, but with gear transmission and elastic coupling of the load is analyzed in the motoring mode. For comparison, the simulation of the drive is performed in the programming environments Mathematica and Matlab-Simulink, respectively. The final, fourth chapter provides conclusions.

2 MATHEMATICAL MODELING OF THREE-PHASE INDUCTION MOTOR DRIVES

The term three-phase induction machine comes from the principle of operation, based on the Faraday′s law of induction in which induced winding voltages are produced by the time changes of magnetic flux linkage created by the relative motion between stator and rotor windings and a spatially distributed magnetic field. In this article, the derivation of the mathematical model of a three-phase induction machine will be performed for symmetrical machines, which consists of stator voltage equations, rotor voltage equations, flux linkage equations and torque equations. In machine variables, stator voltage equations take the vector form

$$\mathbf{v}_s^{abc} = \mathbf{r}_s^{abc}\mathbf{i}_s^{abc} + \frac{d}{dt}\boldsymbol{\lambda}_s^{abc}, \tag{1}$$

where $\mathbf{v}_s^{abc} = \left(v_{as}, v_{bs}, v_{cs}\right)^{\mathrm{T}}$ means the stator voltage vector in which the subscript *s* is referred to the stator, the upperscript *abc* is referred to the three phases of the machine and upperscript $^{\mathrm{T}}$ denotes the vector transposition,

$\mathbf{i}_s^{abc} = \left(i_{as}, i_{bs}, i_{cs}\right)^{\mathrm{T}}$ means the stator current vector of stator phases *abc*, $\boldsymbol{\lambda}_s^{abc} = \left(\lambda_{as}, \lambda_{bs}, \lambda_{cs}\right)^{\mathrm{T}}$ means the stator flux linkage vector and $\mathbf{r}_s^{abc} = r_s \mathbf{I}_3$ is the diagonal matrix of the ohmic resistances r_s of phase stator windings. Analogously, the rotor voltage equations in machine variables take the vector form

$$\mathbf{v}_r^{abc} = \mathbf{r}_r^{abc} \mathbf{i}_r^{abc} + \frac{d}{dt} \boldsymbol{\lambda}_r^{abc}, \tag{2}$$

where $\mathbf{v}_r^{abc} = \left(v_{ar}, v_{br}, v_{cr}\right)^{\mathrm{T}}$ means the rotor voltage vector in which the subscript *r* is referred to the rotor and upperscript *abc* is again referred to the three phases of the machine, $\mathbf{i}_r^{abc} = \left(i_{ar}, i_{br}, i_{cr}\right)^{\mathrm{T}}$ means the rotor current vector of rotor phases *abc*, $\boldsymbol{\lambda}_r^{abc} = \left(\lambda_{ar}, \lambda_{br}, \lambda_{cr}\right)^{\mathrm{T}}$ means the rotor flux linkage vector and $\mathbf{r}_r^{abc} = r_r \mathbf{I}_3$ means the diagonal matrix of the ohmic resistances r_r of phase rotor windings. The flux linkage equations for both stator and rotor of the magnetically linear machine can be presented in the coupled matrix form

$$\begin{pmatrix} \boldsymbol{\lambda}_s^{abc} \\ \boldsymbol{\lambda}_r^{abc} \end{pmatrix} = \begin{bmatrix} \mathbf{L}_{ss}^{abc} & \mathbf{L}_{sr}^{abc} \\ \mathbf{L}_{rs}^{abc} & \mathbf{L}_{rr}^{abc} \end{bmatrix} \cdot \begin{pmatrix} \mathbf{i}_s^{abc} \\ \mathbf{i}_r^{abc} \end{pmatrix}, \tag{3}$$

where $\mathbf{L}_{ss}^{abc}$, $\mathbf{L}_{rr}^{abc}$ are submatrices of stator-to-stator and rotor-to-rotor inductances

$$\mathbf{L}_{ss}^{abc} = \begin{bmatrix} L_{ls} + L_{ss} & L_{sm} & L_{sm} \\ L_{sm} & L_{ls} + L_{ss} & L_{sm} \\ L_{sm} & L_{sm} & L_{ls} + L_{ss} \end{bmatrix}, \quad \mathbf{L}_{rr}^{abc} = \begin{bmatrix} L_{lr} + L_{rr} & L_{rm} & L_{rm} \\ L_{rm} & L_{lr} + L_{rr} & L_{rm} \\ L_{rm} & L_{rm} & L_{lr} + L_{rr} \end{bmatrix}, \tag{4}$$

and where L_{ls}, L_{lr} are leakage inductances per phase of stator and rotor winding, respectively, L_{ss}, L_{rr} are self-inductances of stator and rotor winding, respectively and L_{sm}, L_{rm} are mutual inductances between stator windings and mutual inductances between rotor windings, respectively. Further are $\mathbf{L}_{sr}^{abc}$, $\mathbf{L}_{rs}^{abc}$ submatrices of stator-to-rotor and rotor-to-stator mutual inductances, respectively. The stator-to-rotor mutual inductances are dependent on the rotor position, that is on the rotor angle θ_r between axes of *a*-phase of stator and rotor winding, respectively, see Fig.1. Thus, submatrix $\mathbf{L}_{sr}^{abc}$ has the form

$$\mathbf{L}_{sr}^{abc} = \left[\mathbf{L}_{rs}^{abc}\right]^{\mathrm{T}} = L_{sr} \begin{bmatrix} \cos\theta_r & \cos\left(\theta_r + \frac{2\pi}{3}\right) & \cos\left(\theta_r - \frac{2\pi}{3}\right) \\ \cos\left(\theta_r - \frac{2\pi}{3}\right) & \cos\theta_r & \cos\left(\theta_r + \frac{2\pi}{3}\right) \\ \cos\left(\theta_r + \frac{2\pi}{3}\right) & \cos\left(\theta_r - \frac{2\pi}{3}\right) & \cos\theta_r \end{bmatrix}, \tag{5}$$

where L_{sr} is the amplitude of the stator-to-rotor mutual inductance and where the superscript $^{\mathrm{T}}$ denotes the matrix transposition.

Substituting flux linkage vectors $\boldsymbol{\lambda}_s^{abc}$, $\boldsymbol{\lambda}_r^{abc}$ appearing in DEs (1) and (2) by matrix scalar products formed from the corresponding row of the inductance matrix in Eq. (3) with the assembled current vector, one obtain equations, which are coupled through mutual inductances between windings. Therefore, first-order DEs (1) and (2) are dependent on rotor position, too and by rotation of the rotor they are further dependent on the time. This creates differential equations with time-varying coefficients that are difficult to solve.

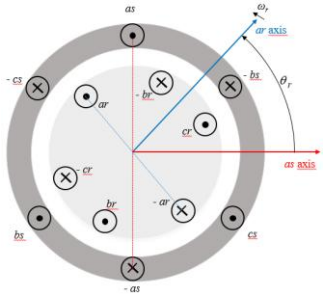


Figure 1: The angle θ_r between the *a*-phase axis of the stator winding and the *a*-phase axis of the rotor winding.

2.1 The transformation to arbitrary qd0 reference frame

To circumvent difficulties due to the time-varying coefficients of mutual inductances and to decouple the three-phase machine variables to orthogonal two-phase (sometimes called two-axis) variables, the Park's transformation in arbitrary *qd*0 reference frame is applied. The *qd*0 transformation matrix has the form

$$\mathbf{T}_{qd0}(\theta)=\frac{2}{3}\begin{bmatrix}\cos\theta & \cos\left(\theta-\frac{2\pi}{3}\right) & \cos\left(\theta+\frac{2\pi}{3}\right)\\ \sin\theta & \sin\left(\theta-\frac{2\pi}{3}\right) & \sin\left(\theta+\frac{2\pi}{3}\right)\\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2}\end{bmatrix} \tag{6}$$

and can be applied to transform voltages, currents or magnetic fluxes in *abc* machine variables to the *qd*0 variables by means of the transformation equation

$$\mathbf{f}^{qd0}=\begin{pmatrix}f^{q}\\ f^{d}\\ f^{0}\end{pmatrix}=\mathbf{T}_{qd0}(\theta)\cdot\begin{pmatrix}f^{a}\\ f^{b}\\ f^{c}\end{pmatrix}=\mathbf{T}_{qd0}(\theta)\cdot\mathbf{f}^{abc}. \tag{7}$$

While the *q*- and *d*-components of Eq. (7) represent the decomposition of three-phase quantities into two orthogonal axes *q* and *d*, the third component, denoted by 0 is the zero-sequence component, which is added to enable the inverse transformation by means of the nonsingular transformation matrix $[\mathbf{T}_{qd0}(\theta)]^{-1}$. The geometric meaning of the relationship between *abc*-axes of a three-phase system and *q*- and *d*-axes of an arbitrary reference system is shown in Fig. 2.

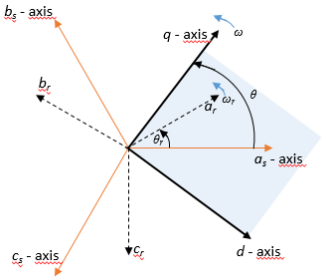


Figure 2: Relationship between *abc*-axes of a three-phase system and *qd*-axes of an arbitrary reference system.

From Fig.2 it follows that the transformation angle $\theta(t)$ between *q*-axis of the arbitrary reference frame, which rotates at an arbitrary angular speed $\omega(t)$ and the *a*-axis, which corresponds to the stationary stator winding, varies according to the rule

$$\theta(t)=\int_0^t\omega(t)dt+\theta(0), \tag{8}$$

where $\omega(t)=d\theta/dt$. Analogously, the rotor angle $\theta_r(t)$ is measured between the stator and rotor *a*-phases and therefore varies as

$$\theta_r(t)=\int_0^t\omega_r(t)dt+\theta_r(0). \tag{9}$$

The angles $\theta(0)$ and $\theta_r(0)$ in Eqs. (8),(9) mean the corresponding initial values. The transformation of stator winding voltages (1) to the *qd*0 variables can be performed symbolically in Mathematica programming environment. Let introduce matrices

$$\mathbf{r}_s^{qd0}=r_s\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix},\quad \mathbf{r}_r^{qd0}=r_r\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix},\quad \mathbf{E}=\begin{bmatrix}0&1&0\\-1&0&0\\0&0&0\end{bmatrix}, \tag{10}$$

and use the first and third matrix to derive the transformed voltages to the form

$$\mathbf{v}_s^{qd0}=\mathbf{r}_s^{qd0}\mathbf{i}_s^{qd0}+\frac{d}{dt}\boldsymbol{\lambda}_s^{qd0}+\omega\mathbf{E}\boldsymbol{\lambda}_s^{qd0}. \tag{11}$$

By using the programming environment Mathematica, we can also derive symbolically the rotor winding voltages in *qd*0 variables. However, from Fig. 2 it follows, that transformation angle for rotor variables is equal to θ-θ_r, what implies the use of the transformation matrix $[\mathbf{T}_{qd0}(\theta\text{-}\theta_r)]$. Doing so, we get the following equation of rotor winding voltages in *qd*0 variables

$$\mathbf{v}_r^{qd0}=\mathbf{r}_r^{qd0}\mathbf{i}_r^{qd0}+\frac{d}{dt}\boldsymbol{\lambda}_r^{qd0}+(\omega-\omega_r)\mathbf{E}\boldsymbol{\lambda}_r^{qd0}, \tag{12}$$

where we use the second and third matrix (10).

The programming system Mathematica proves to be very useful also in the symbolic derivation of flux linkage equations in terms of *qd*0 variables. From Eq. (3), the stator flux linkages in machine variables are given by $\boldsymbol{\lambda}_s^{abc} = \mathbf{L}_{ss}^{abc}\mathbf{i}_s^{abc} + \mathbf{L}_{sr}^{abc}\mathbf{i}_r^{abc}$ so that transformed equation looks like

$$\begin{aligned}\boldsymbol{\lambda}_s^{qd0} &= \mathbf{T}_{qd0}(\theta)\left(\mathbf{L}_{ss}^{abc}\mathbf{i}_s^{abc} + \mathbf{L}_{sr}^{abc}\mathbf{i}_r^{abc}\right)\\ &= \mathbf{T}_{qd0}(\theta)\mathbf{L}_{ss}^{abc}\left[\mathbf{T}_{qd0}(\theta)\right]^{-1}\mathbf{i}_s^{qd0} + \mathbf{T}_{qd0}(\theta)\mathbf{L}_{sr}^{abc}\left[\mathbf{T}_{qd0}(\theta-\theta_r)\right]^{-1}\mathbf{i}_r^{qd0}\end{aligned} \tag{13}$$

The transformation (13) requires matrix multiplications involving the inverse transformation matrix, which can be easily obtained by symbolic.programming. The rotor flux linkages in machine variables are given by $\boldsymbol{\lambda}_r^{abc} = \mathbf{L}_{rs}^{abc}\mathbf{i}_s^{abc} + \mathbf{L}_{rr}^{abc}\mathbf{i}_r^{abc}$, which are transformed to the *qd*0 variables by transformation

$$\begin{aligned}\boldsymbol{\lambda}_r^{qd0} &= \mathbf{T}_{qd0}(\theta-\theta_r)\left(\mathbf{L}_{rs}^{abc}\mathbf{i}_s^{abc} + \mathbf{L}_{rr}^{abc}\mathbf{i}_r^{abc}\right)\\ &= \mathbf{T}_{qd0}(\theta-\theta_r)\mathbf{L}_{rs}^{abc}\left[\mathbf{T}_{qd0}(\theta)\right]^{-1}\mathbf{i}_s^{qd0} + \mathbf{T}_{qd0}(\theta-\theta_r)\mathbf{L}_{rr}^{abc}\left[\mathbf{T}_{qd0}(\theta-\theta_r)\right]^{-1}\mathbf{i}_r^{qd0}\end{aligned} \tag{14}$$

By assembling the computed results of Eq. (13) and of Eq. (14) together, the stator and rotor flux linkage equations in *qd*0 variables can be presented in the following explicit form

$$\begin{pmatrix}\lambda_{qs}\\ \lambda_{ds}\\ \lambda_{0s}\\ \lambda'_{qr}\\ \lambda'_{dr}\\ \lambda'_{0r}\end{pmatrix} = \begin{bmatrix}L_{ls}+L_m & 0 & 0 & L_m & 0 & 0\\ 0 & L_{ls}+L_m & 0 & 0 & L_m & 0\\ 0 & 0 & L_{ls} & 0 & 0 & 0\\ L_m & 0 & 0 & L'_{lr}+L_m & 0 & 0\\ 0 & L_m & 0 & 0 & L'_{lr}+L_m & 0\\ 0 & 0 & 0 & 0 & 0 & L'_{lr}\end{bmatrix}\begin{pmatrix}i_{qs}\\ i_{ds}\\ i_{0s}\\ i'_{qr}\\ i'_{dr}\\ i'_{0r}\end{pmatrix} \tag{15}$$

where the primed rotor variables are used for convenience and present the rotor values, recomputed to the stator side by means of the stator-to-rotor winding ratio:

$$\lambda'_{qr} = \frac{N_s}{N_r}\lambda_{qr},\quad \lambda'_{dr} = \frac{N_s}{N_r}\lambda_{dr},\quad i'_{qr} = \frac{N_s}{N_r}i_{qr},\quad i'_{dr} = \frac{N_s}{N_r}i_{dr},\quad L'_{lr} = \left(\frac{N_s}{N_r}\right)^2 L_{lr}. \tag{16}$$

while the magnetizing inductance L_m on the stator side is

$$L_m = \frac{3}{2}L_{ss} = \frac{3}{2}\frac{N_s}{N_r}L_{sr} = \frac{3}{2}\frac{N_s}{N_r}L_{rr}\,. \tag{17}$$

To complete the mathematical model of three-phase induction machine, we derive the torque equation in *qd*0 variables. For this purpose, we start with the sum of instantaneous input power of all six windings of the stator and rotor in machine variables and use the Park's transformation to obtain the corresponding expression in *qd*0 variables

$$\begin{aligned}p_{inp} &= v_{as}i_{as} + v_{bs}i_{bs} + v_{cs}i_{cs} + v'_{ar}i'_{ar} + v'_{br}i'_{br} + v'_{cr}i'_{cr}\\ &= \frac{3}{2}\left(v_{qs}i_{qs} + v_{ds}i_{ds} + 2v_{0s}i_{0s} + v'_{qr}i'_{qr} + v'_{dr}i'_{dr} + 2v'_{0r}i'_{0r}\right)\\ &= \frac{3}{2}\Big[r_s\left(i_{qs}^2 + i_{ds}^2 + 2i_{0s}^2\right) + r'_r\left(i'_{qr}{}^2 + i'_{dr}{}^2 + 2i'_{0r}{}^2\right)\\ &+ \frac{d\lambda_{qs}}{dt}i_{qs} + \frac{d\lambda_{ds}}{dt}i_{ds} + 2\frac{d\lambda_{0s}}{dt}i_{0s} + \frac{d\lambda'_{qr}}{dt}i'_{qr} + \frac{d\lambda'_{dr}}{dt}i'_{dr} + 2\frac{d\lambda'_{0r}}{dt}i'_{0r}\\ &+ \omega\left(\lambda_{ds}i_{qs} - \lambda_{qs}i_{ds}\right) + \left(\omega-\omega_r\right)\left(\lambda'_{dr}i'_{qr} - \lambda'_{qr}i'_{dr}\right)\Big]\end{aligned} \tag{18}$$

The terms of the form ri^2 in Eq. (18) describe the copper loses and the terms of the form $(d\lambda/dt)i$ represent the rate of exchange of magnetic field energy between windings. Only terms of the form $\omega\lambda i$ describe the rate of energy, which can be converted to the mechanical work. Therefore, the developed electromagnetic torque is defined as the sum of all $\omega\lambda i$ terms, divided by the rotor angular speed

$$T_{em} = \frac{3}{2}\left(\frac{P}{2\omega_r}\right)\Big[\omega\left(\lambda_{ds}i_{qs} - \lambda_{qs}i_{ds}\right) + \left(\omega-\omega_r\right)\left(\lambda'_{dr}i'_{qr} - \lambda'_{qr}i'_{dr}\right)\Big], \tag{19}$$

where $P/2$ denotes the number of pole pairs.

2.2 Mathematical model of induction machine in base quantities

In practice, we usually use magnetic flux linkages per second, denoted by Ψ instead of magnetic flux linkages λ and reactances x instead of inductances L. Denoting with $\omega_b=2\pi f$ the so called base angular frequency, the magnetic flux linkages per second and reactances, respectively are expressed by relations

$$\Psi = \omega_b\lambda, \qquad x = \omega_b L\,. \tag{20}$$

The introduction of magnetic flux linkages per second Ψ and reactances x, respectively, has a consequence that previously derived equations of mathematical model of the induction machine must be rewritten in terms of new variables. The stator voltage equation in *qd*0 variables is then rewritten into the form

$$\mathbf{v}_s^{qd0} = \mathbf{r}_s^{qd0}\mathbf{i}_s^{qd0} + \frac{1}{\omega_b}\frac{d}{dt}\boldsymbol{\Psi}_s^{qd0} + \frac{\omega}{\omega_b}\mathbf{E}\boldsymbol{\Psi}_s^{qd0} \tag{21}$$

and rotor voltage equation, in which variables are recomputed on the stator side, becomes

$$\mathbf{v}_r'^{qd0} = \mathbf{r}_r'^{qd0}\mathbf{i}_r'^{qd0} + \frac{1}{\omega_b}\frac{d}{dt}\boldsymbol{\Psi}_r'^{qd0} + \frac{(\omega-\omega_r)}{\omega_b}\mathbf{E}\boldsymbol{\Psi}_r'^{qd0}. \tag{22}$$

Equation of flux linkages (15) now goes into equation of flux linkages per second:

$$\begin{pmatrix}\Psi_{qs}\\ \Psi_{ds}\\ \Psi_{0s}\\ \Psi'_{qr}\\ \Psi'_{dr}\\ \Psi'_{0r}\end{pmatrix} = \begin{bmatrix} x_{ls}+x_m & 0 & 0 & x_m & 0 & 0\\ 0 & x_{ls}+x_m & 0 & 0 & x_m & 0\\ 0 & 0 & x_{ls} & 0 & 0 & 0\\ x_m & 0 & 0 & x'_{lr}+x_m & 0 & 0\\ 0 & x_m & 0 & 0 & x'_{lr}+x_m & 0\\ 0 & 0 & 0 & 0 & 0 & x'_{lr}\end{bmatrix}\begin{pmatrix}i_{qs}\\ i_{ds}\\ i_{0s}\\ i'_{qr}\\ i'_{dr}\\ i'_{0r}\end{pmatrix}. \tag{23}$$

Finally, the torque equation (19) is also rewritten in terms of magnetic flux linkages per second

$$T_{em} = \frac{3}{2}\left(\frac{P}{2\omega_r}\right)\left[\frac{\omega}{\omega_b}\left(\Psi_{ds}i_{qs} - \Psi_{qs}i_{ds}\right) + \left(\frac{\omega-\omega_r}{\omega_b}\right)\left(\Psi'_{dr}i'_{qr} - \Psi'_{qr}i'_{dr}\right)\right]. \tag{24}$$

2.3 The transformation to stationary qd0 reference frame

The dynamic analysis of the induction machine together with its converter is in drives with adjustable rotor speed or in transient period more conveniently tractable in the stationary *qd*0 reference frame. The stationary reference frame is obtained from the arbitrary reference frame, when the angular speed ω is set to zero, ω=0. The compact vector form of a mathematical model of induction machine in the stationary *qd*0 reference frame then consists of the stator voltage equation:

$$\mathbf{v}_s^{qd0} = \mathbf{r}_s^{qd0}\mathbf{i}_s^{qd0} + \frac{1}{\omega_b}\frac{d}{dt}\boldsymbol{\Psi}_s^{qd0}, \tag{25}$$

rotor voltage equation:

$$\mathbf{v}_r'^{qd0} = \mathbf{r}_r'^{qd0}\mathbf{i}_r'^{qd0} + \frac{1}{\omega_b}\frac{d}{dt}\boldsymbol{\Psi}_r'^{qd0} - \frac{\omega_r}{\omega_b}\mathbf{E}\boldsymbol{\Psi}_r'^{qd0}, \tag{26}$$

equation of flux linkages per second (23) and torque equation (24). It is again worth to mention that rotor voltage equation is expressed in primed variables, which all are recomputed on the stator side.

2.4 The torque balance equations of the electric motor drive

For simulation of the entire electric motor drive, the mathematical model of the induction machine must be completed with the torque balance equations, which take into account the mechanical load of the drive. There are many examples of directly coupled mechanical loads to the drive shaft of the induction machine, such as fans, pumps, compressors, centrifuges and power tools. In this case the drive is governed by a single torque balance equation

$$J\frac{d\left(\frac{2\omega_r}{P}\right)}{dt} = \left(\frac{2\omega_b}{P}\right)J\frac{d\left(\frac{\omega_r}{\omega_b}\right)}{dt} = T_{em} - T_d - T_{mech}, \tag{27}$$

where J denotes the moment of inertia of the rotor mass, $Jd(2\omega_r/P)/dt$ is the inertia torque, T_{em} is developed electromagnetic torque, T_d is the torque of viscous damping, T_{mech} is the mechanical load torque, and entire expression on the right hand side of Eq. (27) means the accelerating torque. Mechanical load torque is positive in the motoring mode and negative in the generating mode of operation. Positive mechanical load torque in motoring mode decreases the accelerating torque and reduces the normalized angular speed of the rotor, while the negative mechanical load torque in generating mode increases the accelerating torque and causes the rotor angular speed to exceed the synchronous angular speed. If the sum of positive mechanical load torque and viscous damping torque exceeds the electromagnetic torque, then the induction machine enters the braking mode of operation.

If the load is coupled to the electric motor shaft through a gear transmission [3] with a viscously damped drive shaft and an elastic output shaft of the gearbox, then there are two independent generalized coordinates, namely rotation angles of the rotor shaft and of the output shaft, respectively. In this case, the equilibrium state of the electric motor drive consists of two torque balance equations, where the first equation corresponds to the input side and the second to the output side of the gear transmission [2].

$$\left(J_1 + J_2 + J_3 \frac{r_2^2}{r_3^2}\right)\ddot{\theta}_1 = T_{em} - c\dot{\theta}_1 - k\frac{r_2}{r_3}\left(\frac{r_2}{r_3}\theta_1 + \theta_4\right),$$
$$J_4\ddot{\theta}_4 = T_{mech} - k\left(\frac{r_2}{r_3}\theta_1 + \theta_4\right) \quad (28a,b)$$

In Eqs. (28a,b), variables θ_1 and θ_4 denote rotation angles of the rotor and flywheel of the mechanical load, respectively. The gear radius is r_2 and his rotation angle is equal to the rotation angle of the rotor, $\theta_2=\theta_1$. The pinion radius is r_3 and his rotation angle is denoted by θ_3. The gear and the pinion rotate in opposite directions, where $r_2\theta_1= r_2\theta_2=-r_3\theta_3$. By using this relation, the pinion rotation angle can be eliminated and expressed by the rotor rotation angle θ_1. Quantities J_i, i=1,…,4 mean moments of inertia of rotor mas, gear and pinion mas and flywheel mas of the mechanical load, respectively. Finally, the stiffness of the torsional spring, placed between the pinion and the mechanical load is denoted by k and the coefficient of viscous damping of the gear is denoted by c.

If the load is coupled to the electric motor shaft through a gear transmission with rigid drive shaft and output shaft of the gearbox, then two torque balance equations are reduced to a single torque balance equation

$$\left(J_1 + J_2 + J_3 \frac{r_2^2}{r_3^2} + J_4 \frac{r_2}{r_3}\right)\ddot{\theta}_1 = T_{em} - c\dot{\theta}_1 - T_{mech} \quad (29)$$

which in fact has the form of Eq. (27).

3 SIMULATION OF DYNAMIC RESPONSES OF THREE-PHASE INDUCTION MOTOR DRIVES

Dynamic simulations of drives with induction machines are performed in the programming environment Mathematica in stationary reference frame [5]. The mathematical model of simulations therefore consist of six voltage ODE's of the first order, that is of Eqs. (25) and (26), of the magnetic linear, algebraic flux linkage per second equation (23), of the developed torque equation (24) and of the torque balance equations (27) or (28a,b). The same machine is used in all operation modes, where its parameters used in simulations are: the rated power of the three-phase, four pole machine is S_b=750 W, the rated phase-to-phase rms voltage is V=200 V, the rated frequency is f=50 Hz and the moment of inertia of the rotor mass is J=0.1 kgm^2. The ohmic resistances and reactances of single stator winding are r_s=3.35 Ω and x_{ls}=2.18 Ω, respectively. The values of ohmic resistance and reactance of a single phase rotor winding are recalculated to the stator side and denoted by primes. These values are r'_r=1.99 Ω and $x'_{lr}=x_{ls}$=2.18 Ω, respectively. The value of magnetizing reactance is chosen to be x_m=51.44 Ω. All dynamic simulations are performed in the time interval, which lasted between 0 s and t_{stopl}=2 s.

In subsequent sections 3.1 – 3.3, the motoring, generating and braking modes of operation of electric motor drives with the induction machine and directly coupled mechanical load according to Eq. (27) are analyzed. Finally, the simulation of electric motor drive with gear transmission and elastic output shaft between pinion and mechanical load according to Eqs. (28a,b) is presented in section 3.4, where results are verified by computation in Matlab-Simulink.

3.1 Electric motor drive in motoring operation

First dynamic simulation treats electric motor drive in motoring mode of operation, where the induction machine is loaded by a time-variable mechanical torque. The time interval of simulation $0\le t\le t_{stop}$ is divided into four subintervals, each of its is lasting for 0.5 s. In the first subinterval $0\le t\le 0.5$ s, the electric motor is'nt loaded, while in the subinterval 0.5 s$\le t\le$1.0 s is loaded by the half of the base torque $0.5T_b$, where the base torque is equal to $T_b=S_b/\omega_{bm}=(P\ S_b)/(2\ \omega_b)$. In the third subinterval 1.0 s$\le t\le$1.5 s, the electric motor is fully loaded by the base torque T_b, however in the last subinterval it is relieved with the half base torque.

The Fig.3 presents the dynamical response of the developed electromagnetic torque of undamped as well as viscously damped drive in motoring mode. On the Fig. 3a it is shown, that electromagnetic torque of the undamped drive, drawn by the solid line, after the transient period approaches to the value of the mechanical load torque. The electromagnetic torque of the viscously damped drive, drawn by dashed line after the transient period takes higher values, because it must balance the sum of the mechanical load torque plus the damping torque. The damping torque is modeled by relationship $T_d=c(\omega_r/\omega_b)$, where damping coefficient takes the value.c=0.75 Nm. The Fig. 3b presents the dynamic response of the electromagnetic torque as a function of the normalized rotor angular speed in both undamped as well as viscously damped drive.

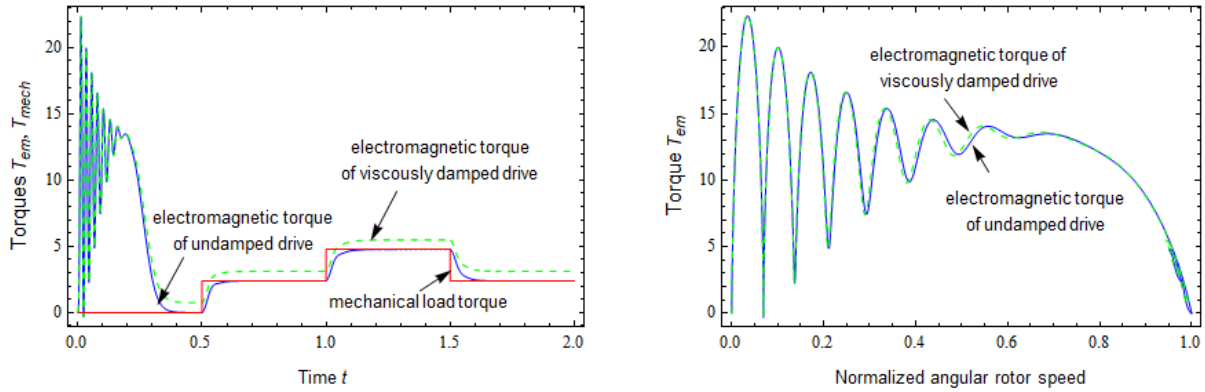


Figure 3: Electromagnetic torque of the drive in motoring mode of operation. a) time response of the electromagnetic torque, b) torque curve as a function of the normalized rotor angular speed.

In the Fig. 4, the time course of normalized rotor angular velocity is shown, where we can see that the normalized nrotor angular velocities of the viscously damped drive differ very little from the normalized rotor angular velocities of the undamped drive. Further, we clearly see that the normalized angular velocity of the rotor steeply increases during the start of the electric motor and decreases slightly in the time intervals in which the electric motor is mechanically more loaded. In addition, we see that in the motor mode of operation the normalized angular velocity of the rotor never exceeds the value 1.

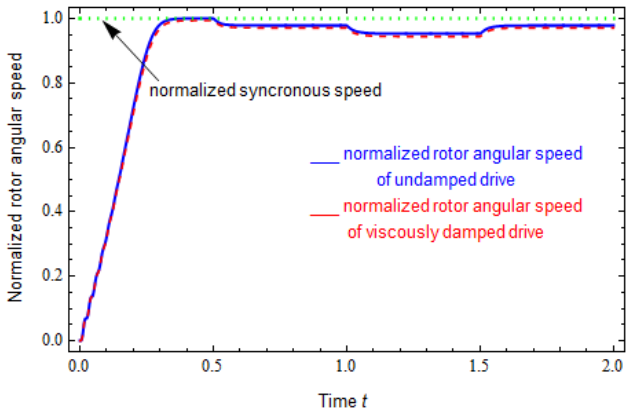


Figure 4: Time course of the normalized angular velocity of the rotor during motoring mode of operation.

In the Fig. 5a, the time response of the stator current $i_{qs}(t)$ is shown in $qd0$ variables and in the Fig. 5b, the corresponding time response of on the stator recomputed rotor current $i'_{qr}(t)$ in $qd0$ variables is presented. From both figures it can be concluded that during the start of the machine, a strong transient phenomenon occurs with high amplitude oscillations. During the time subintervals in which the machine is gradually mechanically more loaded, a corresponding increase in the amplitudes of both currents can be observed. The responses shown in Figs. 5a,b refer only on the undamped drive.

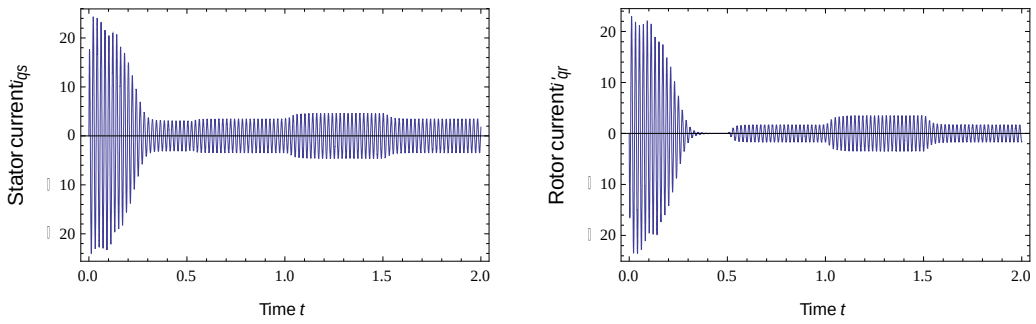
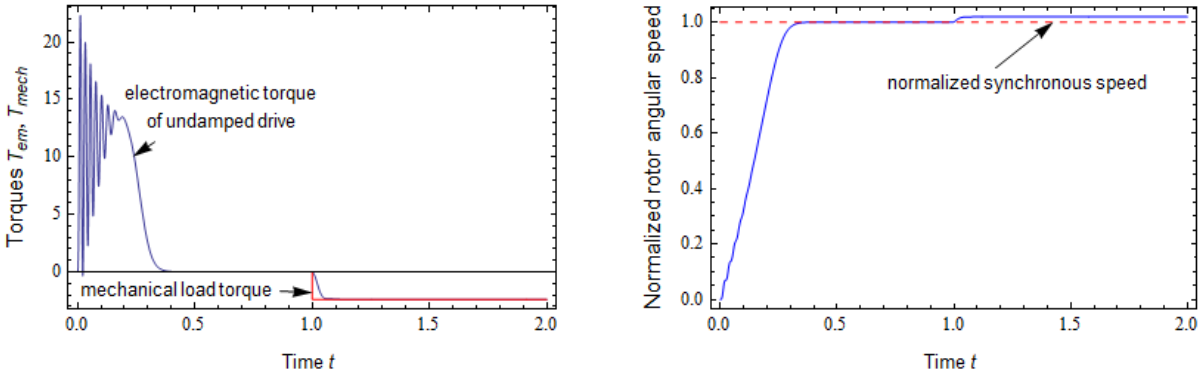


Figure 5: Time responses of a) the stator current $i_{qs}(t)$ and b) on the stator recomputed rotor current $i'_{qr}(t)$ in $qd0$ variables.

3.2 Electric motor drive in generating operation

The second simulation treats the electric motor drive with same machine data operating in the generator mode. To enable such operation, the mechanical load torque must be negative such that there is a net positive acceleration torque, which causes the increase of the rotor angular speed over the value of the synchronous angular speed or the increase of the normalized rotor angular speed over the value 1, respectively. To satisfy such a requirement, the time course of the mechanical load torque is changed in this simulation and is divided into two subintervals, where in the first subinterval $0 \leq t \leq 1.0$ s the electric motor is'nt loaded and in the second subinterval $1.0 \leq t \leq t_{stop}=2$ s it is loaded by the mechanical torque $-0.5T_b$. The effect of the viscous damping is not studied in this simulation. The time responses of the electromagnetic torque in generator mode is plotted in the Fig. 6a, where we see that in the steady state of the second time subinterval the mechanical load torque is balanced with the driving torque because the drive is undamped. In the Fig. 6b it can be seen that the normalized angular velocity of the rotor in the second time subinterval exceeds the value 1 as is expected.

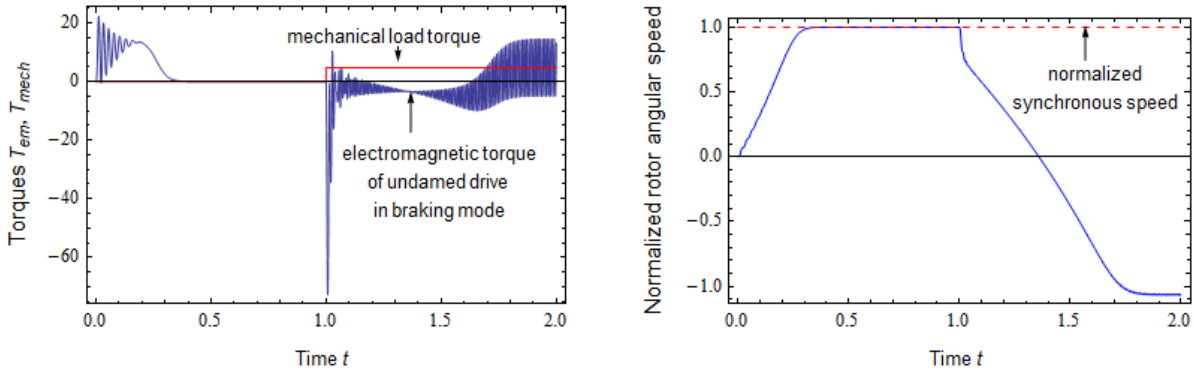


a) b)

Figure 6: Time responses of electric motor drive in generating mode of operation. a) time response of the electromagnetic torque, b) time response of the normalized angular velocity of the rotor.

3.3 Electric motor drive in braking operation

The third simulation treats the electric motor drive with same machine data operating in the braking mode. The braking mode of an electric motor drive can be simulated by reversing the direction of rotation of the rotating magnetic field, which is produced by currents in the stator windings that are 120° out of phase with each other. This can be achieved if two phases of the supply voltages are interchanged. In the simulation, the electric motor again is'nt loaded during the first time subinterval $0 \leq t \leq 1.0$ s and then in the second subinterval $1.0 \leq t \leq t_{stop}=2$ s it is loaded by the mechanical load torque T_b. The interchange of terminals of two phase voltages is performed in the time $t=1$ s, when the electric motor starts to brake.



a) b)

Figure 7: Time responses of electric motor drive in braking mode of operation. a) time response of the electromagnetic torque, b) time response of the normalized angular velocity of the rotor.

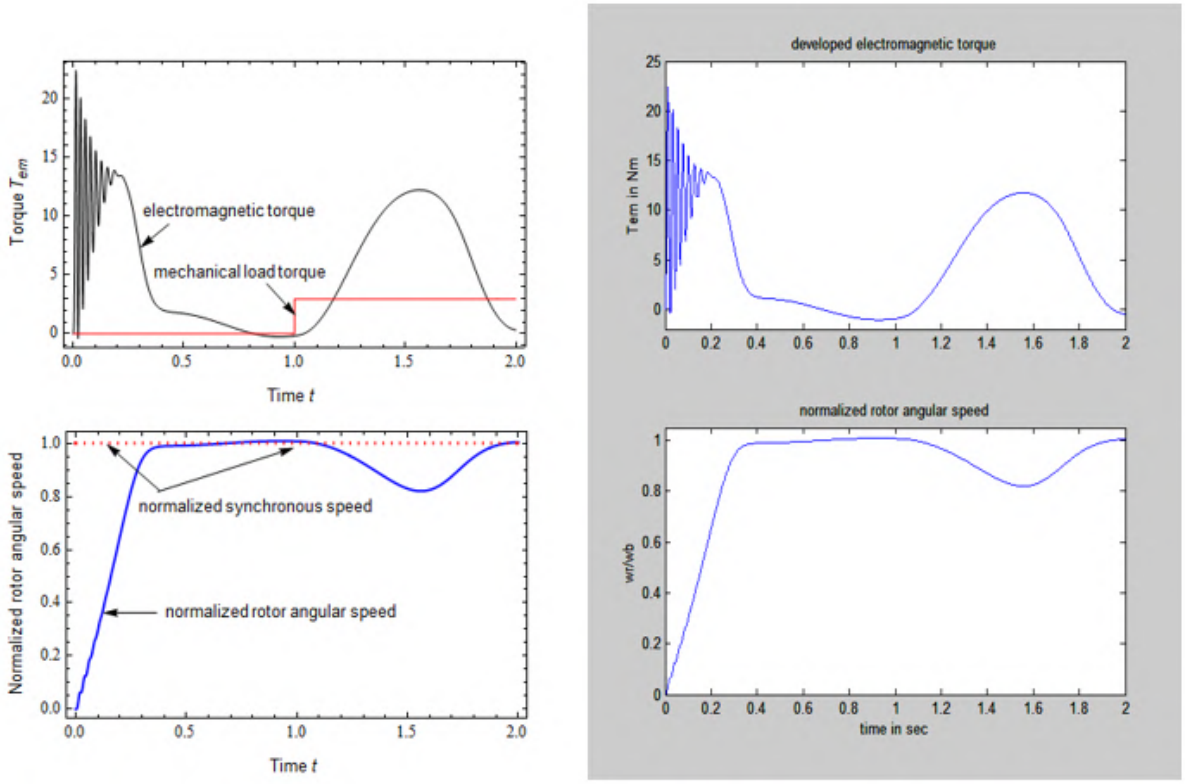
The time responses of the electromagnetic torque and of the normalized angular velocity of the rotor, respectively in the braking mode of the electric motor drive are shown in the Fig. 7. It is evident, that due to the same operating conditions as are prescribed in generator mode for the first time subinterval, the course of the electromagnetic torque in Fig. 7a is similar to the plot, depicted in Fig. 6a. However, as the braking starts at $t=1$ s, the electromagnetic torque begins to strongly oscillate. In accordance with the signs of torques on the right hand side of Eq. (27), the

net negative acceleration torque is produced for the most of the second time subinterval, which causes the decrease of the normalized angular velocity of the rotor. The value of the normalized rotor angular velocity ω_r/ω_b in Fig. 7b first decreases to zero and then becomes negative, indicating that the rotor begins to rotate in the opposite direction. Lastly, the angular velocity of the rotor approaches to reach the limit value of negative base angular velocity $-\omega_b$, which corresponds to the value of the normalized angular velocity of the rotor ω_r/ω_b=-1, as shown in Fig. 7b.

3.4 Electric motor drive with induction machine and gear transmission

In this section, we consider an electric motor drive with an asynchronous electric motor and a gear transmission, where the load and pinion are coupled with an elastic shaft. The equilibrium state of torques on the rotor side and on the load side of the gear transmission in such a system is described by Eqs. (28a,b), where for simplicity we will assume that the gear is not viscously damped and the viscous damping coefficient is chosen zero, c=0. Further, we will assume, that the same electric motor is used in the drive connection as in the previously discussed operating modes, in which the rotor was directly connected to the load. All values of the nominal electrical quantities and electrical parameters thus remain valid. This holds also for the value of the moment of inertia of the rotor mass, which for the purposes of this analysis is denoted by J_1=0.1 kgm^2. However, due to the included gear transmission, new mechanical parameters appear in the analysis, namely the moment of inertia of the gear mass J_2=0.005 kgm^2, the moment of inertia of the pinion mass J_3=0.001 kgm^2 and the moment of inertia of the flywheel (load) mass J_4=0.05 kgm^2, respectively. Further, the gearbox parameters must be completed by data of gear radius r_2=0.1 m and pinion radius r_3=0.05 m. Finally, the stiffness of equivalent torsional spring, which is modeling the elasticity of the output shaft, is given to be equal k=2 Nm.

In the sequel, the simulation of the drive behavior is performed only in the motoring mode. The duration of the simulation is again chosen to be equal 2 s, however the complete time interval is divided into two subintervals, each lasting 1 s. In the first subinterval, the electric motor actually rotates freely, i.e. without the mechanical load, and only in the second subinterval it is loaded by the constant load torque T_{mech}=0.62*T_b Nm. The results of simulation are depicted in the Fig. 8a,b, where the comparison between outputs, obtained in Mathematica and Matlab-Simulink are shown. Both Figures 8a and 8b, respectively, present time responses of the developed electromagnetic torque and the normalized angular speed of the rotor. From the comparison, it is evident, that almost perfect match is obtained between two methods.



a) b)

Figure 8: Time responses of electric motor drive with gear transmission in motoring mode of operation.
a) electromagnetic torque and normalized rotor speed curves, computed by Mathematica,
b) electromagnetic torque and normalized rotor speed curves, computed by Matlab-Simulink.

The time responses of an electric motor drive with an induction motor, a gear transmission and an elastic shaft between the pinion and the mechanical load are significantly different from the time responses of an electric motor drive with a direct coupling of the load to the rotor shaft. Since the same type of induction motor was used in both cases, the presented research allows us to determine the differences between the time responses in the motor operating mode, if we directly compare the diagrams presented in Fig.3 and Fig. 4 with the diagrams in Fig.8. In given types of electric motor drives, these differences in a greater extent are caused due to the ability of the equivalent torsion spring, which is used to model the elastic shaft, to accumulate elastic potential energy and not so much due to the changing of the mechanical loading torque. During the rotation of the elastic shaft, the torsional spring accumulates the elastic potential energy, which it can convert into the kinetic energy of the flywheel on the load side, and through the gear transmission, partly also into the kinetic energy of the rotor. This creates an oscillating system, the oscillation of which is reflected both in the time course of the electromagnetic torque and in the time course of the normalized angular velocity of the rotor. Both types of oscillation are visible in the diagrams in Fig. 8a,b. If we compare both induction motor drives in the time interval $0 \leq t \leq 0.5$ s in which the induction motor is completely unloaded, so it is understandable that the transient phenomena at the start of the electric motor are very similar, although not completely identical. The electrical transient phenomenon is the same in both electric motor drives at start-up, while the mechanical transient phenomenon is slightly longer in the drive with a gear transmission due to the additional inertial masses of the gear and pinion. However, we can observe, that the directly coupled electric motor drive to the load at the end of the free run interval $0 \leq t \leq 0.5$ s does not develop any ectromagnetic torque opposite to the electric motor drive with gearbox. Such behavior can be explained by the conversion of the accumulated elastic potential energy of the torsional spring into the kinetic energy on the rotor side.The courses of the electromagnetic torque in the time interval $0.5 \text{ s} \leq t \leq 1$ s, depicted in Fig. 3a and Fig. 8, respectively, differ significantly because the electromagnetic torque of the directly coupled rotor shaft never become negative, that is never passes into generator regime, opposite to the course of the electromagnetic torque of the drive with gear transmission. This behavior can be confirmed also by comparing Fig. 4 and Fig. 8, observing the course of the normalized angular speed of the rotor. While in Fig. 4 the normalized angular speed of the rotor never exceeds the normalized synchronous speed 1 and therefore the electric motor drive never passes to the generator mode, in Fig. 8a this happen just before the interval $0.5 \text{ s} \leq t \leq 1$ s expires. Significant differences in the time responses appear also in the time interval $1 \text{ s} \leq t \leq 2$ s, in which both electric motor drives are mechanically loaded. In the drive with directly connected mechanical load to the rotor shaft, small drops of the normalized angular speed of the rotor in respect to the normalized synchronous speed can be observed in the Fig. 4, where each drop of the normalized angular speed of the rotor properly reacts on the changing load torque. In contrast, an electric motor drive with gearbox exhibits a sway in the electromagnetic torque and the normalized angular speed of the rotor with much larger amplitudes, which is due to the conversion of the elastic potential energy of the torsion spring into kinetic energy. This oscillations are shown in Fig. 8.

4 CONCLUSION

The mathematical model in the stationary reference frame, implemented in the programming environment Mathematica is developed for dynamical simulations of the electric motor drives with three-phase induction machines. In the model, magnetically linear machine is treated. By using symbolic computing in the software system, the governing equations of the induction machine are derived, which saves extensive manipulative work. The developed mathematical model is succesfully applied in simulations of electric drives, operating in the motoring, generating and braking mode. The presented model can be upgraded to treat magnetically nonlinear induction machines. The paper presents a sophisticated mechanical assembly in which the induction machine is coupled with the mechanical load through a gearbox with an elastic output shaft. The obtained time responses of electric as well as mechanical variables are confirmed by simulation using programming environments Mathematica and Matlab-Simulink, respectively.

REFERENCES

[1] Abouzeid, A. F., Guerrero, J. M., Endemaño, A., Muniategui, I., Ortega, D., Larrazabal, I. and Briz, F.. Control Strategies for Induction Motors in Railway Traction Applications. *Energies,* 2020, **vol**. 13, 700. doi:10.3390/en13030700

[2] Heimann, B., Gert, W., Popp, K. *Mechatronik: Komponenten, Methoden, Beispiele*. *Carl Hanser Verlag München,* 2007, ISBN 3446405992

[3] Kastrevc, M., Pušenjak, R. Fuzzy pressure control of hydraulic system with gear pump driven by variable speed induction electro-motor. *Experimental techniques*, 2005, **vol**. *29*(*3*), 57-62.

[4] Merve, S. K. The Use of Induction Motors in Electric Vehicles. In book: *Induction Motors- Recent Advances, New Perspectives and Applications*, 2023, Doi: 10.57772/*Intechopen.* 1000865

[5] Pušenjak, R. Simulations of dynamical responses of electric motor drives with induction machines. *Anali PAZU*, 2025, **vol**. 15, 19-42.

Experimental testing and mathematical modeling of sound insulation of recycled panels made from car tires based on statistical models

Anđelko Crnoja[1], Ivo Čolak[2], Aleksej Aniskin[3], Dalibor Kramarić[4]

[1]University North, University Center Varaždin, Croatia
E-mail: acrnoja@unin.hr

[2]University of Mostar, Faculty of Civil Engineering, Bosnia and Herzegovina
E-mail: ivo.colak@sum.ba

[3]University North, University Center Varaždin, Croatia
E-mail: aaniskin@unin.hr

[4]University North, University Center Varaždin, Croatia
E-mail: dalibor.kramaric@unin.hr

Abstract: *Any unwanted sound, regardless of its characteristics, has a negative impact on human health through a series of neurophysiological phenomena. The need for protection from it is rapidly increasing every day. Finding new building materials or products for protection against offensive sound is an increasing need of modern society. However, the availability and rational exploitation of raw materials for production is also an issue worth considering. If we look at used car tires in the context of waste management, their huge and increasing quantities every day represent a challenge in environmental terms, so it is extremely important to create opportunities for their reuse. This makes car tires a significant source of raw materials in the production of new building materials. Research has been conducted into the use of materials from car tires for the purpose of protection against excessive noise. Composite panels made of recycled rubber granulate and binder have been designed and manufactured, and the sound insulation value of each sample has been experimentally tested. The test results were processed with the SPSS software package in order to determine a suitable mathematical model for describing the properties of sound insulation in relation to the selected changing variables of the samples: granulometric composition, amount of binder and specific mass of the sample. Finding a relevant model allows for a better understanding of the influence of different components on the properties of the material and thus optimization of the composition of sound insulation from recycled rubber granulate.*

Keywords*: sound, sound insulation, recycled rubber, logarithmic model, statistical processing.*

Article Classification*: Scientific Paper*

1 INTRODUCTION

Sound, in the narrow sense of the word, is considered to be everything that we hear, that we perceive with the sense of hearing [1]. Sound is understood as that part of mechanical wave motion whose frequency lies in the range of hearing of the human ear, i.e. in the range from 16 to 20,000 Hz [2]. Although sound is a natural phenomenon of exceptional benefit, its unwanted manifestation can be characterized as noise or sound pollution. Noise, defined as unpleasant, disturbing or unwanted sound, is an increasingly significant environmental and public health problem [3,4]. Long-term exposure to noise can cause a number of negative effects on human health and quality of life. The impact of noise pollution on humans is much more far-reaching than the obvious potential damage to hearing [5]. From the occurrence of sleep disorders, impaired cognitive abilities, to the impact on the cardiovascular system including the development of high blood pressure and heart disease, a number of authors have extensively documented the various ways in which the harmful effects of noise are reflected in the human body [6–12]. Although many mechanisms of harmful effects have been identified, the primary method of their mitigation is the interruption or limitation of noise and sound vibrations [13,14].

By producing, designing and installing noise protection materials, the construction industry plays a central role in implementing measures to reduce the harmful effects of noise pollution on people. If the materials used are also environmentally friendly solutions, the benefits are multiple for both the environment and people. Environmentally friendly materials for sound insulation and sound absorption can be products made from natural materials (straw, cotton, hemp) or products with a share of recycled waste materials (plastic bottles, construction waste, car tires, textiles, mattresses) [15–19]. The production of materials with a share of recycled waste reduces the amount of waste in landfills, and in the case of substances that in their free state pose a hazardous impact on the environment, it is possible to reduce or eliminate the harmful effects by installing them in an inert environment. One such waste material is tires from vehicles.

Due to the increasing amount of discarded tires every year, waste tire management is a challenge in most developed countries. Waste tires are not biodegradable, and they also release toxic substances into the environment, often serve as a breeding ground for mosquitoes and support the development of dangerous fires. Because of the above, governments and institutions have been looking for alternatives for their reuse for many years, and the European Union has banned the disposal of waste tires in landfills [20–22]. Although there are several ways to reuse them, from extending the life by replacing the outer layer of rubber or thermal treatment for use as fuel, due to their relative stability, durability and non-biodegradability, tires represent an excellent and abundant waste material for recycling [23]. Due to the high proportion of rubber in the final product of transport tires, they can be reused for the production of new polymer materials after previous fragmentation. Before fragmentation, it is necessary to remove other components such as steel and textiles, but they can also be recycled, i.e. reused for various innovative purposes [24–26]. After separating the other components and crushing, pieces of rubber of various granulations are ready for reuse in the composition of new materials.

There is increasing evidence of progress in the application of rubber granules in various composite construction materials, and the effects of rubber on material properties are increasingly well researched. Many scientific papers published in the last few years indicate the great potential of using rubber granules for the production of sound insulation or sound absorption materials. The acoustic properties of materials made from waste rubber granules and polyester fibers that are incorporated into tires as reinforcement have been investigated [27], the properties of metamaterials produced from a mixture of waste plastic and rubber [28], rubber granules bonded with cement and gypsum [29], the sound absorption properties of asphalt mixtures with the addition of rubber granules [30], as well as, for example, the properties of rubber granules as a material in a loose and compressed state [31]. Research into the use of rubber granules as a complete or partial replacement of aggregate in concrete products for sound and thermal insulation has also progressed significantly [32–36]. All the mentioned researches prove that waste rubber is a valuable component material in the design and manufacture of materials for sound absorption and sound insulation, and that there is still much room for progress in understanding the influence of rubber granules, as well as other production parameters, on the properties of the final product. The aim of this paper is to investigate and quantify the effectiveness of panels made of recycled rubber as a sound insulation material in building structures, with a special emphasis on reducing internal noise and at the same time contributing to environmental protection through the use of waste materials. The creation of a reliable logarithmic model would enable a preliminary approximation of the properties of the final product, taking into account the parameters of the composition through mathematical modeling.

2 MATERIALS AND METHODS

2.1 Materials

The tested samples are composite panels made of recycled rubber granulate and binder. Rubber granulate is obtained by processing waste used car tires, as well as tires from other vehicles and machines. To produce granulate as a raw material, the collected waste rubber needs to be mechanically shredded. In addition to the rubber itself, the shredded granulate also contains other components of transport tires such as textiles and steel threads for reinforcement, but also other impurities such as impurities created during driving. For the purpose of isolating the rubber itself as a raw material, rotary and vibrating screens with magnetic attachments for separating steel fibers are used. Finally, to prepare the granulate in different fractions, it is necessary to carry out any additional shredding, sieving and separation into certain fractions.
A commercially available polyurethane binder in a standard version was used as the binder for the composite panels.

2.2 Key parameters of the tested panels

Samples of different thicknesses (10, 15 and 20 mm) were produced and tested with variations in granulometric composition, bulk density and binder content, which defined the characteristics of the panels themselves.

2.2.1 Granulometric composition

Two fractions of rubber aggregate were used to make the samples, a finer one with a size of 0.5 – 2 mm and a coarser one with a size of 2 – 3.5 mm. Considering the size of the rubber aggregate, three mixtures were designed. The first mixture contained only the finer fraction aggregate, the second the coarser fraction aggregate, and the third a mixture of both fractions in a ratio of 35% finer and 65% coarser aggregate. Accordingly, the samples can be divided according to the granulometric composition into samples with a fine fraction, samples with a coarse fraction of aggregate, and combined samples. The grain size and granulometric composition affect the porosity of the samples and thus the dissipation of sound energy.

2.2.2 Density (ρ)

The density value of rubber aggregate in dry and loose state ranges from 450 kg/m^3 to 490 kg/m^3. The sample preparation process involves adding binder and compressing the mixture, which increases the density of the samples and reduces the amount of pores. The density values of the obtained samples range from 585 kg/m^3 to 1100 kg/m^3. The specific gravity of the material directly affects the dynamic stiffness of the material.
Increasing the bulk density (e.g. from 700 kg/m^3 to 950 kg/m^3) reduces the proportion of closed pores and improves the dynamic modulus of elasticity, resulting in better performance in the low-frequency range (<500 Hz).

2.2.3 Polyurethane binder content

The weight of the binder in the panels ranges from 269 g/m^2 to 1340 g/m^2. The mass fraction of the binder is from 4.2 to 7.3 % of the total mass of the sample. The amount of binder is determined by the technological process depending on the requirements for mechanical strength. The proportion of binder up to 12 % of the sample mass ensures particle cohesion without reducing elasticity.

2.3 Technological production process

After processing the rubber granulate and separating it into certain fractions, a polyurethane binder is added to the prepared rubber aggregate and the mixture is combined and homogenized in specialized mixers. The prepared mixture is poured into a steel mold and compressed with a hydraulic press at a pressure of up to 5 MPa at a temperature of (80 – 130) ± 5 ºC to achieve the desired thickness and volume mass.
Thermal curing takes place in controlled atmosphere chambers with a curing time of 45 minutes or by exposure to outside air in warehouses.

2.4 Research methodology

A total of 27 samples, i.e. panels with an area of 0.8 m^2, were sampled and tested, with three samples of the same composition and thickness for better result analysis. The size of the samples was chosen to avoid an unfavorable level of inhomogeneity caused by too small dimensions. Inhomogeneities can have a negative impact on the quality and credibility of the test results. The panel tests were conducted in an accredited laboratory under controlled conditions.

The samples were tested using the airborne sound insulation measurement method of building components in accordance with the EN ISO 10140-2:2021 standard [37]. The sound insulation properties of the sampled panels were evaluated after testing according to the EN ISO 717-1:2020 standard [38], which evaluates the sound insulation of building components, specifically airborne sound insulation. The standard specifies a single-number value (sound insulation index R'w) for characterizing sound insulation, which allows for the comparison and evaluation of different building materials and elements with regard to their noise attenuation ability. The test used a two-room acoustic field method, where the sound attenuation level is measured in the frequency range from 50 Hz to 5000 Hz. The first room contains a white noise source with a power of approximately 105 dB, and the second room contains a receiver that detects the difference between the transmitting and receiving rooms.

2.5 Conducting the experiment

Experimental research was conducted in an accredited laboratory for measuring sound insulation of materials and testing sound insulation of doors and windows of the Croatian Institute of Civil Engineering.
Testing of samples of recycled rubber panels was carried out using the method for testing the sound insulation value in accordance with the EN ISO 10140-2:2021 standard [37] with a deviation from the standard in the form of measurements with rotating microphones with two positions of the speaker and microphone.
The aforementioned standard defines the measurement of airborne sound insulation of building components. The test frame was constructed according to defined criteria for air tightness (air tightness) and sealing of all construction joints.

The test procedure is fully compliant with the requirements of the EN ISO 717-1:2020 standard [38], which defines the methodology for determining the single-digit sound insulation value of buildings and building components. The photographs below show the method of installing the frame and samples at the test site.

a) b)

Figure 1: (a) Equipment position in front of the sample; (b) Built-in test sample

For the purpose of conducting the test, the test sample is placed in a double partition wall, where the test chamber is designed as a "room within a room" with a dilatation gap between the soundproofed receiving area and the transmitting area. Inside the transmitting area is a non-directional sound source, while in the receiving area there is a highly sensitive class I microphone mounted on a rotating stand for spatial analysis of the sound field.

The rotation of the microphone takes place in an oblique plane in order to register spatial variations in sound pressure and possible influences of resonances and diffraction during the measurement. The photographs below show the basic components of the test equipment used in the experiment.

a) b) c)

Figure 2: Test equipment: (a) Non-directional sound source, type 4296, Bruel & Kjær; (b) Condenser microphones, type 4189, Bruel & Kjær; (c) Hand-held analyzer, type 2270 (Dual Channel), Bruel & Kjær;

The emitted sound source signal (white noise) has a constant level of approximately 105 dB(A) for a duration of 60 seconds which provides sufficient time for stabilization of all sound conditions within the receiving chamber. Data collection and processing is carried out using a hand-held spectrum analyzer with the support of computer processing, whereby a frequency-dependent sound insulation (sound reduction index, R) is determined for each frequency in the 1/3-octave band (one-third-octave band).

The final result is also expressed as a one-number sound insulation value (weighted sound reduction index, R'w), which enables a direct comparison of different samples.

2.5.1 Experimental design

The tests were conducted on samples of different thicknesses (10, 15, 20 mm) with variations in granulometric composition, according to the data listed in the table 1.

Table 1: Sampling plan for test samples

Sample thickness	Granulometric composition	Number of samples
10 mm	0.5 – 2 mm	3
10 mm	2 – 3.5 mm	3
10 mm	0.5 – 2 (35%) + 2 - 3.5 (65%) mm	3
15 mm	0.5 – 2 mm	3
15 mm	2 – 3.5 mm	3
15 mm	0.5 – 2 (35%) + 2 - 3.5 (65%) mm	3
20 mm	0.5 – 2 mm	3
20 mm	2 – 3.5 mm	3
20 mm	0.5 – 2 (35%) + 2 - 3.5 (65%) mm	3
Total		27

2.6 Mathematical modeling using the SPSS program package

The aim of testing mathematical models on the results of the conducted experimental tests was to find an appropriate model for better understanding, predicting and optimizing the behavior of the sound insulation properties of the observed composite panel depending on its constituent parts. Various regression analysis models were tested, as well as the use of a polynomial equation and an equation with interactions. The models used represent the application of various statistical methods for examining the dependence of one variable on another or more other variables, and variables represent phenomena that are in some relationship. Variables whose variations are explained by others are considered dependent variables, and those that explain the variation of the dependent variable are considered independent variables [39]. For the statistical analysis of the results of the conducted experimental tests, the SPSS (Statistical Package for the Social Sciences) software package was used with the application of logarithmic data transformations to achieve a normal distribution of the results. The analysis focused exclusively on the weighted single-number value of the sound insulation index (R'w) as the key dependent parameter (dependent variable).

The research examined the influence of the independent variables of volumetric mass, granulometric composition and amount of binder on changes in the value of R'w through systematic variation of the structure of the composite material. Modifications of the structure were achieved by controlled changes in the share of individual parameters, which made it possible to quantify their individual and combined influence.

Validation of all used models was carried out by comparing experimentally tested values for sample mark 26 with predicted values obtained by the mathematical model.

The basic characteristics of the constructed model are shown in the table below.

Table 2: Basic characteristics of the model

Property	Label sample N	Average value	SD *	V **	Minimum value	Maximum value
'PU adhesive (kg/m^3)	26	11.53	3.12	27.09	7.00	18.30
Specific weight (kg/m^3)	26	797.42	163.75	20.53	585.00	1100.00
R'w	26	16.53	6.88	41.62	6.30	33.50

* Average deviation from the arithmetic mean,
** Dispersion coefficient.

The average amount of binder in the selected sample is 11.53 kg/m^3 with an average deviation from the arithmetic mean of 3.12 kg/m^3 (SD). The coefficient V defines the dispersion and in this case it is small (27.09 % < 30.00 %). The average specific gravity in the selected sample is 797.42 kg/m^3 with an average deviation from the arithmetic mean of 163.75 kg/m^3. The dispersion in this case can also be considered small (20.53 % < 30.00 %).
The average value of the sound insulation index in the selected sample is 16.53 dB with an average deviation from the arithmetic mean of 6.88 dB. In this case, this is a higher dispersion value (41.62% > 30.00%).

The granulometric composition is defined in chapter 2, and the definition in the mathematical model is performed in such a way that a certain granulometric composition is assigned the mark 1, 2 or 3:

- The granulometric composition with fractions 0.5 - 2.0 mm is assigned the mark 1;
- The granulometric composition with fractions 0.5 - 2.0 mm (35%) + 2.0 - 3.5 mm (65%) is assigned the mark 2;
- The granulometric composition with fractions 2.0 - 3.5 mm is assigned the mark 3.

The marks are assigned in such a way that an increase in the mark number (rank) indicates an increase in the fractions in the granulometric composition.

2.6.1 Testing the statistical significance of the model using the ANOVA test

Each model involves creating a database that is adapted to the software package, and the significance of the model shows us whether the model used is applicable. By conducting a significance test and selecting a significance level, it is possible to determine the applicability of the model. The selected significance level is 0.05, and if the calculated p factor is < 0.05, we can confirm that the model is statistically significant and applicable.

2.6.2 Analysis of the presence of multicollinearity in the model

Multicollinearity implies a high degree of correlation between the independent variables in a model, which limits the model's ability to separate the influence of each variable on the dependent variable. The presence of multicollinearity makes it difficult to interpret the influence of individual variables and can lead to inaccurate estimates. One indicator of multicollinearity is the Tolerance indicator (TOL). As a rule of thumb, a multicollinearity problem is present if the Tolerance indicator is less than 0.20.

2.6.3 Analysis of the presence of heteroscedasticity in the model

Homoskedasticity and heteroskedasticity are terms that describe the variance, or dispersion, of errors in regression models. In the presence of homoskedasticity, the variance of errors is constant for all values of the independent variable, while in the case of heteroskedasticity, the variance changes, which can cause imprecise standard errors and distort the results of the model. The problem of heteroskedasticity was tested using Spearman's correlation coefficient, observing the relationship between residual deviations and independent variables. The p-value must be greater than 0.05 to meet the condition for the correctness of the applied model, or to remove the suspicion of heteroskedasticity.

2.6.4 Analysis of representativeness of the model

The representativeness of the model is ultimately the most important factor for assessing the reliability of the observed model. The coefficient of determination R^2 provides information on the extent to which the dispersion of the experimental values of the dependent variable is explained by the model function, and to what extent it is unexplained dispersion. A value of the coefficient of determination R^2 closer to 1 or 100% indicates that a small part of the dispersion of the result remains unexplained by the model, and in this case the model can be considered relevant [40]. A threshold of 80% was chosen to determine the significance of the model, and models with a value of the coefficient of determination Ra2 greater than 80% are considered statistically significant.

3 RESULTS

3.1 Experimental test results

The results of laboratory tests of the soundproofing characteristics of panels made from recycled car tires are presented through a detailed frequency analysis in the range from 50 Hz to 5000 Hz, which covers the range of human hearing, which is relevant for soundproofing in buildings.

The following values were measured during the measurement:

- Sound insulation index (R) expressed in decibels (dB), which quantifies the ability of a material to reduce the transmission of sound energy;
- Sound pressure level in the emission area (Lp1) - measured in dB(A) with a reference pressure of 20 μPa;
- Sound pressure level in the reception area (Lp2) - measured in dB(A) and corrected for background noise and reverberation effects.

The reference curve made according to EN ISO 717-1:2020 [38] shows the standardized frequency characteristic of sound insulation.

Figure 3 shows graphs showing the frequency dependence of the sound insulation index with special emphasis on critical frequencies and the occurrence of extremes in the form of breakthroughs at low frequencies. For the purpose of better transparency, the results are grouped into three subgroups according to the thickness of the samples (10, 15 and 20 mm).

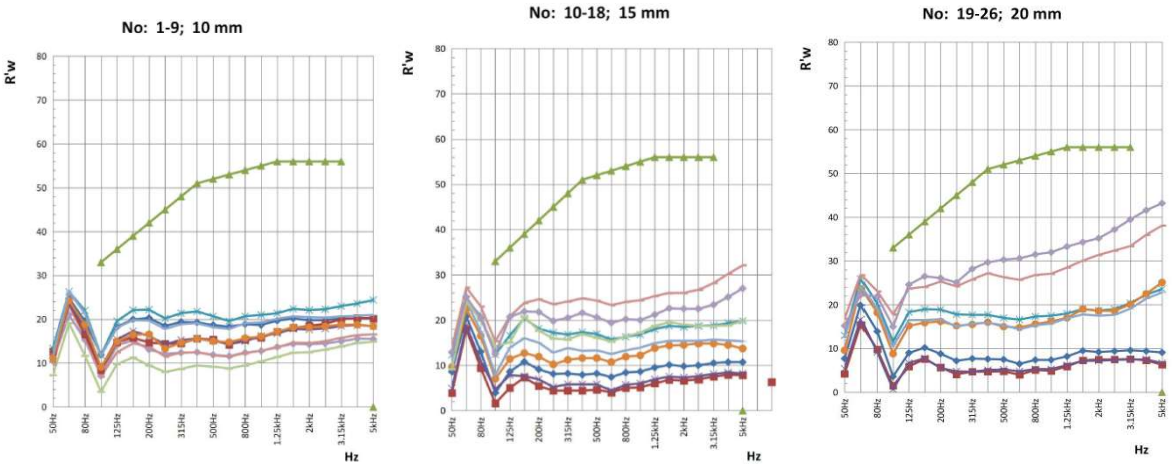


Figure 3: Experimental results of sound insulation index test R'w - 10, 15 and 20 mm thick samples

3.2 Results of testing mathematical models on experimental test results

3.2.1 Linear multiple regression model

Table 3 shows the results of the linear multiple regression model analysis.

Table 3: Results of the linear multiple regression model

	Standardized β	β	Standard error b	t	*
Intercept		-10.08	4.07	-2.47	0.022
Specific mass (ρ) (t/m^3)	0.75	31.31	4.13	7.58	<0.001
PU adhesive g/m^2 (kg)	0.40	8.45	2.07	4.09	<0.001
Granulometric composition	-0.22	-1.86	0.83	-2.26	0.034

* Testing the statistical significance of model parameters.

Based on the available data, we can derive a mathematical expression for the testing performed:

$$\mathrm{R'w} = -10.08 + 8.45 \cdot (\mathrm{PU}) + 31.31 \cdot (\rho) - 1.86 \cdot (n) \quad (1)$$

Where:

- R'w is the sound insulation index,
- n is the particle size distribution level,
- ρ is the specific weight (t/m^3) and
- PU is the adhesive content (g/m^2 (kg)).

Each parameter indicates the unit change in the sound insulation index if the observed variable changes by one measurement unit.

Table 4 shows the results of testing the statistical significance of the model using the ANOVA test, which show that the factor p is within the predicted limits and according to this criterion, the model as a whole is statistically significant

Table 4: Significance indicators of linear multiple regression models

	Sums of - Squares	df	Mean-Squares	F	p
Regress.	932.75	3	310.92	27.27	<0.001
Residual	250.84	22	11.40		
Total	1183.60				

Based on the empirical p value of less than 0.05 and the F value of 27.27, we conclude that the model as a whole is statistically significant, that is, that the selected parameters record a statistically significant influence on the movement of R'w.

Table 5 shows the results of the analysis of the presence of multicollinearity in the model, which show that the Tolerance factor is within the expected limits, which indicates that there is no problem of multicollinearity in the linear multiple regression model.

Table 5: Indicators of multicollinearity in the linear multiple regression model

Variable	Tolerances
Specific mass (ρ) (t/m^3)	1.00
PU adhesive gr/m^2 (kg)	1.00
Granulometric composition	0.99

Table 6 shows the results of the analysis of the presence of heteroskedasticity in the model, which show that the coefficient of heteroscedasticity is p > 0.05 and according to this criterion the model does not have a heteroskedasticity problem.

Table 6: Heteroscedasticity indicators of linear multiple regression models

	N	*	t	p
Residuals & Granulometric composition	26	-0.07	-0.35	0.73
Residuals & Specific mass (t/m^3)	26	0.16	0.81	0.43
Residuals & PU glue g/m^2 (kg)	26	0.26	1.30	0.21

* Spearman's correlation test.

Table 7 shows the results of the representativeness analysis of the model, from which it is evident that the coefficient of determination is 78.81% < 80.00%. It is concluded that the linear multiple regression model is not statistically representative.

Table 7: Significance of linear multiple regression models

Representativeness indicator	Value
R^2	78.81%
Standard error of estimate	3,377

3.2.2 Exponential multiple regression model

Table 8 shows the results of the exponential multiple regression model analysis.

Table 8: Results of the exponential multiple regression model

	Standardized β	β	Standard error b	t	*
Intercept		0.95	0.28	3.35	0.003
Specific mass (ρ) (t/m 3)	0.79	2.21	0.29	7.66	<0.001
PU adhesive gr/m2 (kg)	0.28	0.39	0.14	2.70	0.013
Granulometric composition	-0.22	-0.12	0.06	-2.15	0.043

* Testing the statistical significance of model parameters

Based on the available data, we can derive a mathematical expression for the performed testing in logarithmic form:

$$\ln \mathrm{R'w} = 0.95 \ + 0.39 \cdot (\mathrm{PU}) \ + 2.21 \cdot (\rho) - 0.12 \cdot (n) \quad (2)$$

and in the basic form of the equation:

$$\mathrm{R'w} = 2.59 \cdot 1.49^{\mathrm{PU}} * 9.11^{\rho} \cdot 0.89^{n} \quad (3)$$

Where:

- R'w is the sound insulation index,
- n is the granulometric composition level,
- ρ is the specific weight (t/m^3) and
- PU is the adhesive content (g/m^2 (kg)).

The constant term of 2.59 represents the initial or base value of the variable R'w when all predictors are equal to zero. The final outcome starts from this value, with each predictor influencing this value.

The variable adhesive has a positive influence, with each increase of one unit causing the R'w value to be multiplied by a factor of 1.48, corresponding to an increase of 48%. This moderate increase contributes significantly to the final result. The variable mass, whose increase of one unit causes a 9.11-fold increase in the R'w value, i.e. an explosive increase of 811%. On the other hand, the granulometric composition has a negative effect on R'w. Its increase by one unit reduces R'w by a factor of 0.89, which is reflected in a decrease of 11%.

Table 9 shows the results of testing the statistical significance of the model using the ANOVA test, which show that the factor p is within the predicted limits and according to this criterion, the model as a whole is statistically significant.

Table 9: Significance indicators of the exponential multiple regression model

	Sums of - Squares	df	Mean-Squares	F	p
Regress.	4.03	3	1.34	24.21	<0.001
Residual	1.22	22	0.06		
Total	5.25				

Based on the empirical p value of less than 0.05 and the F value of 24.21, we conclude that the model as a whole is statistically significant, that is, that the selected parameters record a statistically significant influence on the movement of R'w .

Table 10 shows the results of the analysis of the presence of multicollinearity in the model, which show that the Tolerance factor is within the expected limits, which indicates that there is no multicollinearity problem in the exponential multiple regression model.

Table 10: Significance indicators of the exponential multiple regression model

Variable	Tolerances
Specific mass (ρ) (t/m^3)	1.00
PU adhesive g/m^2 (kg)	0.99
Granulometric composition	0.99

Table 11 shows the results of the analysis of the presence of heteroscedasticity in the model, which show that the coefficient of heteroskedasticity is p > 0.05 and according to this criterion the model does not have a heteroskedasticity problem.

Table 11: Heteroscedasticity indicators of the exponential multiple regression model

	N	*	t	p
Residuals & Granulometric composition	26	0.26	1.31	0.20
Residuals & Specific mass (t/m^3)	26	-0.12	-0.61	0.55
Residuals & PU glue g/m^2 (kg)	26	-0.02	-0.08	0.94

* Spearman's correlation test.

Table 12 shows the results of the representativeness analysis of the model, from which it is evident that the coefficient of determination is 76.75% < 80.00%. It is concluded that the exponential multiple regression model is not statistically representative.

Table 12: Significance of the exponential multiple regression model

Representativeness indicator	Value
R^2	76.75%
Standard error of estimate	0.235

3.2.3 Polynomial equation

Table 13 shows the results of the polynomial equation model analysis.

Table 13: Results of the polynomial equation model

	Standardized β	β	Standard error b	t	*
Intercept		-55.11	11.77	-4.68	<0.001
Specific mass (ρ) (t/m^3)	3.85	161.59	26.17	6.17	<0.001
PU adhesive g/m^2 (kg)	-0.66	-13.74	7.67	-1.79	0.089
Granulometric composition	-0.14	-1.17	3.29	-0.36	0.73

* Testing the statistical significance of model parameters.

Table 14 shows the results of the analysis of the presence of multicollinearity in the model, which show that the Tolerance factor is outside the expected limits, which indicates that there is a problem of multicollinearity in the model, and further analyzes of the model were not conducted.

Table 14: Indicators of multicollinearity in the polynomial equation model

Variable	Tolerances
Specific mass (ρ) (t/m^3)	0.009
PU adhesive g/m^2 (kg)	0.025
Granulometric composition	0.021

3.2.4 Equation with interactions

Table 15 shows the results of the analysis of the equation model with interactions.

Table 15: Results of the interaction equation model

	Standardized β	β	Standard error b	t	p *
Intercept		18.03	9.59	1.88	0.076
Specific mass (ρ) (t/m^3)	-0.01	-0.63	11.01	-0.06	0.955
PU adhesive g/m^2 (kg)	-2.18	-45.67	11.76	-3.88	0.001
Granulometric composition	-0.25	-2.09	3.31	-0.63	0.535

* Testing the statistical significance of model parameters.

Table 16 shows the results of the analysis of the presence of multicollinearity in the model, which show that the Tolerance factor is outside the expected limits, which indicates that there is a problem of multicollinearity in the model, and further analyzes of the model were not conducted.

Table 16: Indicators of multicollinearity in the equation model with interactions

Variable	Tolerances
Specific mass (ρ) (t/m^3)	0.071
PU adhesive g/m^2 (kg)	0.015
Granulometric composition	0.031

3.2.5 Logarithmic regression model

Table 17 shows the results of the logarithmic regression model analysis.

Table 17: Results of the logarithmic regression model

	Standardized β	β	Standard error b	t	*
Intercept		3,546	0.126	28,055	<0.001
Specific mass (ρ) (t/m^3)	0.825	1,850	0.208	8,912	<0.001
PU adhesive g/m^2 (kg)	0.229	0.219	0.089	2,469	0.022
Granulometric composition	-0.228	-0.127	0.052	-2,458	0.022

* Testing the statistical significance of model parameters.

Based on the available data, we can derive a mathematical expression for the performed testing in logarithmic form:

$$\ln \mathrm{R'w} = 3.546 + 1.85 \cdot \ln(\rho) + 0.219 \cdot \ln(\mathrm{PU}) - 0.127 \cdot \log \mathrm{n} \quad (4)$$

and in the basic form of the equation:

$$\mathrm{R'w} = 34.674 \cdot (\rho)^{6.358} \cdot (PU)^{1.245} \cdot (n)^{0.987} \quad (5)$$

Where:

- R'w is the index of sound insulation,
- n is the level of granulometric composition,
- ρ is the specific mass (t/m^3) and
- PU proportion of the glue (g/m^2 (kg)).

Table 18 shows the results of testing the statistical significance of the model using the ANOVA test, which show that the factor p is within the predicted limits and according to this criterion, the model as a whole is statistically significant.

Table 18: Significance indicators of the logarithmic regression model

	Sums of - Squares	df	Mean-Squares	F	p
Regress.	4.26	3	1.42	31.63	<0.001
Residual	0.99	22	0.04		
Total	5.25				

Based on the empirical p value of less than 0.05 and the F value of 31.63, we conclude that the model as a whole is statistically significant, that is, that the selected parameters record a statistically significant influence on the movement of R'w .

Table 19 shows the results of the analysis of the presence of multicollinearity in the model, which show that the Tolerance factor is within the expected limits, which indicates that there is no problem of multicollinearity in the logarithmic regression model.

Table 19: Indicators of multicollinearity in the logarithmic regression model

Variable	Tolerances
Specific mass (ρ) (t/m^3)	1.00
PU adhesive g/m^2 (kg)	0.99
Granulometric composition	0.99

Table 20 shows the results of the analysis of the presence of heteroskedasticity in the model, which show that the coefficient of heteroskedasticity is $p > 0.05$ and according to this criterion the model does not have a problem of heteroscedasticity.

Table 20: Heteroscedasticity indicators of the logarithmic regression model

	N	*	t	p
Residuals & Granulometric composition	26	0.05	0.23	0.818
Residuals & Specific mass (t/m^3)	26	0.02	0.09	0.932
Residuals & PU glue g/m^2 (kg)	26	-0.01	-0.03	0.973

* Spearman's correlation test.

Figure 6 shows that the residual deviations closely follow a normal distribution with an expectation that tends to zero.

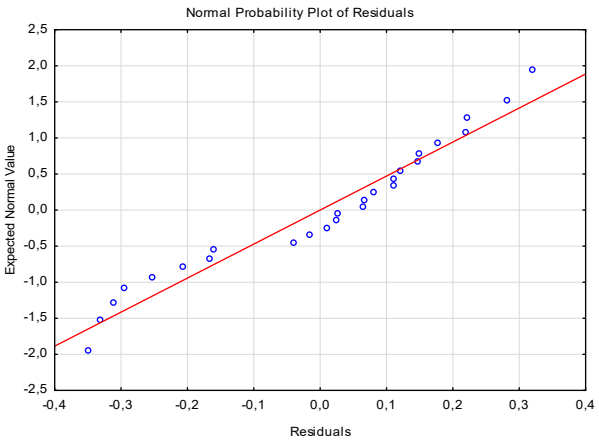


Figure 4 Residual deviations in relation to the normal distribution with expectation tending to zero

Table 21 shows the results of the model representativeness analysis, which shows that the value of the coefficient of determination is 81.18%, which is higher than the selected significance level of 80.00%. Therefore, the model can be considered statistically significant. The notes to the table refer to the average deviation of the model from the actual values, regardless of whether the equation is in logarithmic or basic form.

Table 21: Indicators of multicollinearity in the logarithmic regression model

Representativeness indicator	Value
R^2	81.18%
Standard error of estimate	0.212* / 1.236**

* Logarithmic form of the equation,
** Basic form of the equation.

From the graphic representation in Figure 5, it can be determined that as the sound insulation index R‘w increases, so do the residual deviations, that is, the points become further away from the direction indicating the equality of the real and model values R’w .

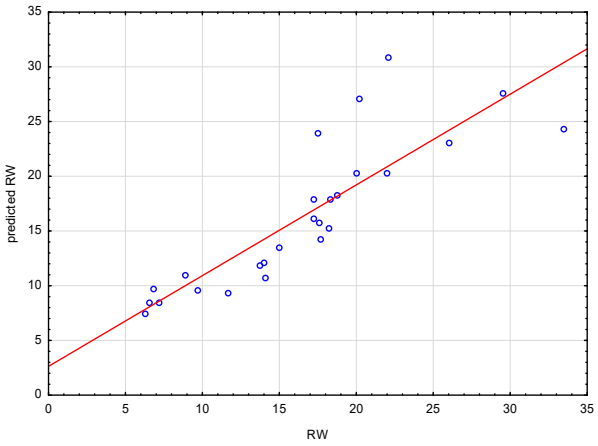


Figure 5: Relationship between real and model values of the sound insulation index

The movement of the residuals is randomly distributed (random) depending on the change in the model value. Therefore, there is no connection between the residual deviations and the movement of the dependent variable of the amount of noise, as can be seen from the graphic representation in Figure 6.

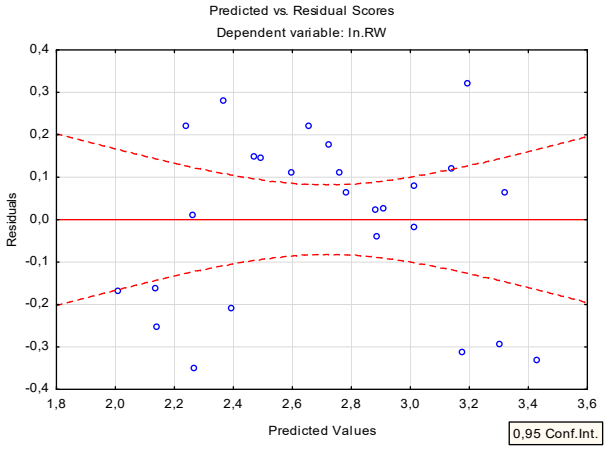


Figure 6: Movement of actual and model deviation value

3.2.6 Logarithmic multiple regression model

Table 22: shows the results of the analysis of the logarithmic multiple regression model.

Table 22: Results of the logarithmic model of multiple regression

	Standardized β	β	Standard error b	t	p *
Intercept		29.63	1.96	15.15	<0.001
Specific mass (ρ) (t/m^3)	0.35	4.98	1.37	3.62	0.002
PU adhesive g/m^2 (kg)	0.77	25.96	3.21	8.08	<0.001
Granulometric composition	-0.23	-1.93	0.80	-2.42	0.024

* Testing the statistical significance of model parameters.

Based on the available data, we can derive a mathematical expression for the testing performed:

$$R'w = ß_0 + ß_1 \cdot \ln(PU) + ß_2 \cdot \ln(\rho) + ß_3 \cdot \ln(n) + \varepsilon \quad (6)$$

$$R'w = 29.63 + 4.98 \cdot \ln(PU) + 25.96 \cdot \ln(\rho) - 1.93 \cdot \ln(n) \quad (7)$$

Where:

- R'w is the index of sound insulation,
- n is the level of granulometric composition,
- ρ is the specific mass (t/m^3) and
- PU proportion of the glue (g/m^2 (kg)).

Table 23 shows the results of testing the statistical significance of the model using the ANOVA test, which show that the factor *p* is within the predicted limits and according to this criterion, the model as a whole is statistically significant.

Table 23: Significance indicators of the logarithmic multiple regression model

	Sums of - Squares	df	Mean-Squares	F	p
Regress.	947.09	3	315.70	29.37	<0.001
Residual	236.50	22	10.75		
Total	1183.60				

Based on the empirical p value of less than 0.05 and the F value of 29.37, we conclude that the model as a whole is statistically significant, that is, that the selected parameters record a statistically significant influence on the movement of R'w .

Table 24 shows the results of the analysis of the presence of multicollinearity in the model, which show that the Tolerance factor is within the expected limits, which indicates that there is no multicollinearity problem in the logarithmic multiple regression model.

Table 24: Indicators of multicollinearity in the logarithmic multiple regression model

Variable	Tolerances
Specific mass (ρ) (t/m^3)	1.00
PU adhesive g/m^2 (kg)	0.99
Granulometric composition	0.99

Table 25 shows the results of the analysis of the presence of heteroskedasticity in the model, which show that the coefficient of heteroscedasticity is $p > 0.05$ and according to this criterion the model does not have a heteroskedasticity problem.

Table 25: Heteroskedasticity indicators of the logarithmic multiple regression model

	N	*	t	p
Residuals & Granulometric composition	26	0.20	1.00	0.326
Residuals & Specific mass (t/m^3)	26	0.26	1.34	0.193
Residuals & PU glue g/m^2 (kg)	26	-0.09	-0.45	0.660

* Spearman's correlation test.

Table 26 shows the results of the model representativeness analysis, which shows that the value of the coefficient of determination is 80.02%, which is marginally higher than the selected significance level of 80.00%. Therefore, the model can be considered marginally statistically significant.

Table 26: Significance of the logarithmic regression model

Representativeness indicator	Value
R^2	80.02%
Standard error of estimate	3,279

4 DISCUSSION

Several different models, or statistical methods, were tested to examine the dependence of the sound insulation index value as a dependent variable on the independent variables of volume mass, granulometric composition and binder quantity, in order to find a representative mathematical model that describes the behavior of the variables according to the experimental results of sound insulation property testing.

The high correlation determined only for the logarithmic regression model ($R^2 > 0.82$) indicates: the robustness of the model, the representativeness of the sample and the validity of the logarithmic approach.

Mathematical expression of the logarithmic regression model:

$$\ln(\mathrm{R'w}) = a \cdot \ln(\rho) + b \cdot \ln(d) + c \cdot \ln(c) + k \quad (8)$$

Where:

- a, b and c are standardized regression coefficients,
- k is the model constant,

- ρ is the bulk density (kg/m^3),
- d is the mean particle size (mm) and
- c is the binder content (kg/m^3).

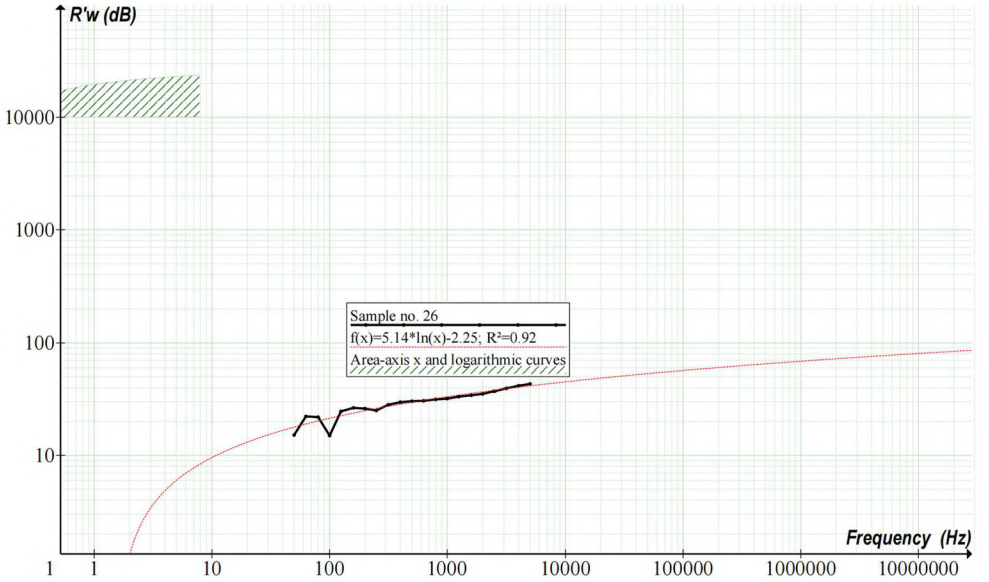


Figure 7: Comparison of the experimental test curve and the logarithmic function

5 CONCLUSION

The obtained value of the coefficient of determination of 81.18% indicates that the estimated logarithmic regression model explains 81.12% of the sum of the squares of the deviations of the R'w value from its arithmetic mean. Therefore, it can be concluded that the model is representative and applicable.

Looking at the mathematical expression:

$$\mathrm{R'w} = 34.674 \cdot (\rho)^{6.358} \cdot (PU)^{1.245} \cdot (n)^{0.987} \tag{9}$$

we come to the following conclusions:

- By increasing the specific mass, the sound insulation index also increases. If we increase the specific mass by 1%, the sound insulation index increases by 6.358% with other unchanged values. This is also the biggest influence of a variable on the sound insulation index;
- By increasing the amount of binder (polyurethane glue), the sound insulation index also increases. If we increase the amount of binder by 1%, the sound insulation index increases by 1.245% with other unchanged values;

If we change the granulometric composition rank (increase in aggregate diameter), the sound insulation index decreases. If we change the granulometric composition by one rank, the sound insulation index decreases by 0.987% with other values unchanged.

The coefficient value of 34.674 implies that the expected sound insulation index of 34.674 dB can be achieved if all other parameters were equal to 1.

The "standardized β" coefficient implies that the greatest relative influence on R'w is specific mass (b*=0.825).

According to the conducted analyses, the presented logarithmic regression model describes the dependence of the sound insulation index R'w on the observed variables of the specific mass of the sample, the amount of binder and the granulometric composition in accordance with the conducted experimental tests. The model provides a better understanding of the influence of selected material characteristics on the sound insulation properties, but at the same time it also allows for the prediction of properties and optimization of the composition of a composite material made of recycled rubber granulate and binder for the purpose of potential future research and development of higher quality materials.

Acknowledgement
This paper was supported by the 2026 Call for Proposals for Scientific Research Grants at University North - code: UNIN-TEH-26-1-7.

REFERENCES

[1] Jelaković, T. Sound, Hearing, Architectural Acoustics; 2nd ed.; Školska knjiga, 1978.
[2] Bošnjaković, R. Noise Reduction; Delo, 1981.
[3] European Environment Agency. Environmental Noise in Europe: 2025. ; Publications Office: LU, 2025.
[4] Environmental Noise Guidelines for the European Region; World Health Organization, Regional Office for Europe:
[5] Natarajan, N.; Batts, S.; Stankovic, KM Noise-Induced Hearing Loss. JCM 2023, 12, 2347, doi:10.3390/jcm12062347
[6] Hahad, O.; Kuntic, M.; Al-Kindi, S.; Kuntic, I.; Gilan, D.; Petrowski, K.; Daiber, A.; Münzel, T. Noise and Mental Health: Evidence, Mechanisms, and Consequences. J Expo Sci Environ Epidemiol 2025, 35, 16–23, doi:10.1038/s41370-024-00642-5
[7] Mucci, N.; Traversini, V.; Lulli, LG; Vimercati, L.; Rapisarda, V.; Galea, RP; De Sio, S.; Arcangeli, G. Neurobehavioral Alterations in Occupational Noise Exposure: A Systematic Review. Sustainability 2021, 13, 12224, doi:10.3390/su132112224
[8] Lee, Y.; Lee, S.; Lee, W. Occupational and Environmental Noise Exposure and Extra-Auditory Effects on Humans: A Systematic Literature Review. GeoHealth 2023, 7, e2023GH000805, doi:10.1029/2023GH000805.
[9] Arregi, A.; Vegas, O.; Lertxundi, A.; Silva, A.; Ferreira, I.; Bereziartua, A.; Cruz, MT; Lertxundi, N. Road Traffic Noise Exposure and Its Impact on Health: Evidence from Animal and Human Studies—Chronic Stress, Inflammation, and Oxidative Stress as Key Components of the Complex Downstream Pathway Underlying Noise-Induced Non-Auditory Health Effects. Environ Sci Pollut Res 2024, 31, 46820–46839, doi:10.1007/s11356-024-33973-9.
[10] Chen, X.; Liu, M.; Zuo, L.; Wu, X.; Chen, M.; Li, X.; An, T.; Chen, L.; Xu, W.; Peng, S.; et al. Environmental Noise Exposure and Health Outcomes: An Umbrella Review of Systematic Reviews and Meta-Analysis. European Journal of Public Health 2023, 33, 725–731, doi:10.1093/eurpub/ckad044.
[11] Petri, D.; Licitra, G.; Vigotti, MA; Fredianelli, L. Effects of Exposure to Road, Railway, Airport and Recreational Noise on Blood Pressure and Hypertension. IJERPH 2021, 18, 9145, doi:10.3390/ijerph18179145.
[12] Münzel, T.; Daiber, A.; Engelmann, N.; Röösli, M.; Kuntic, M.; Banks, JL Noise Causes Cardiovascular Disease: It's Time to Act. J Expo Sci Environ Epidemiol 2025, 35, 24–33, doi:10.1038/s41370-024-00732-4.
[13] Thompson, R.; Smith, RB; Bou Karim, Y.; Shen, C.; Drummond, K.; Teng, C.; Toledano, MB Noise Pollution and Human Cognition: An Updated Systematic Review and Meta-Analysis of Recent Evidence. Environment International 2022, 158, 106905, doi: 10.1016/j.envint.2021.106905.
[14] Kanu, LK; Lwara, SMK; Meng, X. Impacts of Workplace Noise Exposure and Mitigation Strategies: A Scoping Review. Discov Public Health 2025, 22, 214, doi:10.1186/s12982-025-00611-9.
[15] Del Rosario-Gilabert, D.; Carbajo, J.; Hernández-Pozo, M.; Valenzuela-Miralles, A.; Ruiz, D.; Poveda-Martínez, P.; Esquiva, G.; Gómez-Vicente, V. Eco-Friendly and Biocompatible Material to Reduce Noise Pollution and Improve Acoustic Comfort in Healthcare Environments. Buildings 2024, 14, 3151, doi:10.3390/buildings14103151.
[16] Wilk-Jakubowski, JL; Kuchcinski, A.; Pawlik, L.; Wilk-Jakubowski, G. Advanced Sound Insulating Materials: An Analysis of Material Types and Properties. Applied Sciences 2025, 15, 6156, doi:10.3390/app15116156.
[17] Strazdas, E.; Januševičius, T. A Review of the Acoustic Properties of Noise Barriers Made from Waste and Plant-Based Materials.; Vilnius Gediminas Technical University, Lithuania, October 23 2023; pp. 1–6.
[18] Vėjelis, S.; Vaitkus, S.; Kremensas, A.; Kairytė, A.; Šeputytė-Jucikė, J. Reuse of Textile Waste in the Production of Sound Absorption Boards. Materials 2023, 16, 1987, doi:10.3390/ma16051987.
[19] Conesa, JA; Tomás, E. Production of Acoustic Insulating Materials from Viscoelastic Mattress Waste. Sci 2022, 4, 48, doi:10.3390/sci4040048.
[20] Mohajerani, A.; Burnett, L.; Smith, JV; Markovski, S.; Rodwell, G.; Rahman, MT; Kurmus, H.; Mirzababaei, M.; Arulrajah, A.; Horpibulsuk, S.; et al. Recycling Waste Rubber Tires in Construction Materials and Associated Environmental Considerations: A Review. Resources, Conservation and Recycling 2020, 155, 104679, doi: 10.1016/j.resconrec.2020.104679.
[21] Armada, D.; Llompart, M.; Celeiro, M.; Garcia-Castro, P.; Ratola, N.; Dagnac, T.; De Boer, J. Global Evaluation of the Chemical Hazard of Recycled Tire Crumb Rubber Employed on Worldwide Synthetic Turf Football Pitches. Science of the Total Environment 2022, 812, 152542, doi: 10.1016/j.scitotenv.2021.152542.
[22] Czarna-Juszkiewicz, D.; Kunecki, P.; Cader, J.; Wdowin, M. Review in Waste Tire Management—Potential Applications in Mitigating Environmental Pollution. Materials 2023, 16, 5771, doi:10.3390/ma16175771.
[23] Rowhani, A.; Rainey, T. Scrap Tire Management Pathways and Their Use as a Fuel—A Review. Energy 2016, 9, 888, doi:10.3390/en9110888.
[24] Zia, A.; Pu, Z.; Holly, I.; Umar, T.; Tariq, MOOR; Sufian, M. A Comprehensive Review of Incorporating Steel Fibers of Waste Tires in Cement Composites and Its Applications. Materials 2022, 15, 7420, doi:10.3390/ma15217420.

[25] Fazli, A.; Rodrigue, D. Sustainable Reuse of Waste Tire Textile Fibers (WTTF) as Reinforcements. Polymers 2022, 14, 3933, doi:10.3390/polym14193933.
[26] Ružickij, R.; Romagnoli, F.; Grubliauskas, R. Waste Tire Textile Fiber Composite Material: Acoustic Performance and Life Cycle Assessment. Sustainability 2024, 16, 6281, doi:10.3390/su16156281.
[27] Colom, X.; Girbau, J.; Marin, M.; Formela, K.; Saeb, MR; Carrillo, F.; Cañavate, J. New Approach in the Reuse of Modified Ground Tire Rubber as Thermal and Acoustic Insulation to Be Used in Civil Engineering. J Mater Cycles Waste Manag 2023, 25, 3557–3566, doi:10.1007/s10163-023-01778-6.
[28] Naimushin, A.; Januševičius, T. Development and Research of Recyclable Composite Metamaterial Structures Made of Plastic and Rubber Waste to Reduce Indoor Noise and Reverberation. Sustainability 2023, 15, 1731, doi:10.3390/su15021731.
[29] Smirnova, OM; Menéndez Pidal De Navascués, I.; Mikhailevskii, VR; Kolosov, OI; Skolota, NS Sound-Absorbing Composites with Rubber Crumb from Used Tires. Applied Sciences 2021, 11 , 7347, doi:10.3390/app11167347.
[30] Apaza, FR; Fernández Vázquez, V.; Paje, SE; Gulisano, F.; Gagliardi, V.; Saiz Rodríguez, L.; Gallego Medina, J. Towards Sustainable Road Pavements: Sound Absorption in Rubber-Modified Asphalt Mixtures. Infrastructures 2024, 9, 65, doi:10.3390/infrastructures9040065.
[31] Badida, M.; Andrejiova, M.; Pinosova, M.; Moravec, M. Acoustic Analysis of Soundproofing Materials Using Recycled Rubber from Automobiles. Materials 2025, 18, 3144, doi:10.3390/ma18133144.
[32] Svoboda, J.; Dvorský, T.; Václavík, V.; Charvát, J.; Máčalová, K.; Heviánková, S.; Janurová, E. Sound-Absorbing and Thermal-Insulating Properties of Cement Composite Based on Recycled Rubber from Waste Tires. Applied Sciences 2021, 11, 2725, doi:10.3390/app11062725.
[33] Mhaya, AM; Shahidan, S.; Zuki, SSM; Huseien, GF; Azmi, MAM; Ismail, M.; Mirza, J. Durability and Acoustic Performance of Rubberized Concrete Containing POFA as Cement Replacement. Sustainability 2022, 14, 15510, doi:10.3390/su142315510.
[34] Algaifi, HA; Syamsir, A.; Baharom, S.; Alrshoudi, F.; Qaid, A.; Al-Fakih, AM; Mahaya, AM; Salah, HA Assessment of Acoustic and Mechanical Properties in Modified Rubberized Concrete. Case Studies in Construction Materials 2024, 20, e03063, doi: 10.1016/j.cscm. 2024. e03063.
[35] Wijesinghe, KAP; Lanarolle, G.; Gunasekara, C.; Law, DW; Hidallana-Gamage, HD; Wang, L. Thermal and Acoustic Performance of Solid Waste Incorporated Cement Based Composites: An Analytical Review. Arch. Civ. Mech. Eng. 2025, 25, 106, doi:10.1007/s43452-025-01160-3.
[36] Pongsopha, P.; Sukontasukkul, P.; Zhang, H.; Limkatanyu, S. Thermal and Acoustic Properties of Sustainable Structural Lightweight Aggregate Rubberized Concrete. Results in Engineering 2022, 13, 100333, doi: 10.1016/j.rineng. 2022.100333.
[37] EN ISO 10140-2:2021; Acoustics - Laboratory Measurement of Sound Insulation of Building Elements - Part 2: Measurement of Airborne Sound Insulation.
[38] EN ISO 717-1:2020; Acoustics - Rating of Sound Insulation in Building and of Building Elements - Part 1: Airborne Sound Insulation; 2020.
[39] Šošić, I. Applied Statistics; Školska knjiga: Zagreb, 2004; ISBN 978-953-0-30337-9.
[40] Benšić, M.; Šuvak, N. Applied Statistics; Josip Juraj Strossmayer University of Osijek: Osijek, 2013; ISBN 978-953-6931-59-0.

Analysis of damage to castings due to thermal fatigue cracks in aluminium alloy die casting moulds

Mitja Muhič[1]

[1]Šolski center Novo mesto Višja strokovna šola
E-naslov: mitja.muhic@sc-nm.si

Abstract: *Die casting moulds for aluminum alloys are exposed to high temperature changes and consequently high thermo-mechanical loads. Thermal and mechanical loads cause high stresses and deformations in the material and consequently surface cracks in the material. The paper analyzes defects in aluminum alloy castings in the actual die casting process, which are a result of thermo-mechanical loads and cracks in the mould material. The research showed that cracks in the moulds and consequently defects in the castings appeared earlier and were larger if the mould material had a lower hardness. Cracks in the moulds and the resulting defects in the castings are larger closer to the entrance of the melt flow due to higher temperatures and higher melt flow and at places of geometric changes with small radii, where stress concentrations are highest.*

Keywords: *Die casting of aluminum alloys, Thermal fatigue cracks, Thermo-mechanical stresses, Defects in die castings*

Article Classification*: Scientific Paper*

1 INTRODUCTION

Die casting is a manufacturing technique in which molten metal is injected into mould cavities under high pressure. It represents one of the most efficient production methods in metalworking and enables cost-effective fabrication of complex, near-net-shape components. Typically, non-ferrous alloys such as aluminium, zinc, copper, magnesium, lead, and tin are used. This process is especially advantageous when producing large volumes of small to medium-sized parts requiring precise dimensions and high surface quality. [1-4]

During operation, aluminium alloy melt temperatures reach approximately 700 °C. The pressure during injection ranges between 10 and 175 MPa, while flow velocities vary from 20 to 60 m/s. Mould s are preheated to temperatures between 250 and 350 °C. Under such extreme thermal and mechanical conditions, various surface defects develop, including thermal fatigue cracks, erosion, corrosion, soldering, deformation, and even fracture (Figure 1) [1-4].

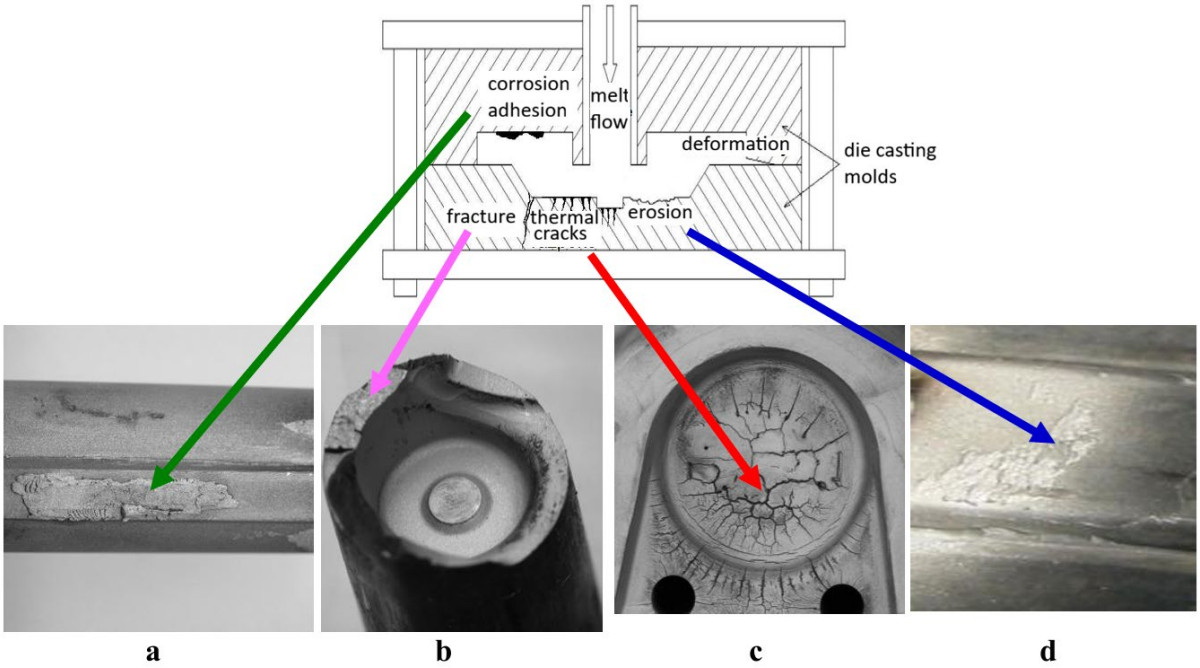


Figure 1: Types of defects on die casting moulds a) corrosion, adhesion, b) fracture, c) thermal fatigue cracks, d) erosion

Among these, thermal fatigue cracking is the primary failure mechanism. Crack networks typically form on surfaces exposed to repeated temperature gradients. With continued operation, these cracks propagate and degrade the surface quality of both mould and casting. The mechanism behind thermal fatigue cracking is relatively well understood. Repeated heating and cooling cycles induce expansion and contraction, generating stresses and strains. Over time, plastic deformation accumulates locally, leading to crack initiation and propagation. Additionally, exposure to high temperatures reduces material strength due to tempering effects. [5-7]

Although laboratory testing can isolate thermal fatigue effects, real mould conditions are more complex. Factors such as oxidation, creep, melt pressure, and flow also play significant roles. Therefore, predicting mould failure using simulations is highly challenging. A more practical solution is to monitor defect-fins on castings during production. Since direct observation of mould surfaces is impractical without interrupting the process, defect-fins provide an indirect yet effective measure of crack development. Because quality control already involves inspection of castings, integrating this monitoring approach is straightforward. [5-7]

The objectives of the research are the analysis of defects in castings due to cracks in moulds made of different materials, with different hardness, and in different locations on the moulds, with the purpose to better understand what happens in the moulds during die casting and to work on reducing the number and size of cracks caused by thermal fatigue.

2 EXPERIMENTAL

The study focused on fixed die casting moulds and corresponding aluminium castings. The moulds were made from Cr–Mo–V hot work tool steel with defined chemical compositions.

The casting tool consisted of four geometrically identical moulds produced from two different materials. They operated under identical conditions, differing only in material composition and hardness (43–46 HRc). Surface nitriding was applied to enhance wear resistance. Table 1 shows the chemical composition of the mould materials. Material 2 in Table 1 is a non-standard commercial steel named Toolox from the manufacturer SSAB, which is already made for final use and is not heat treated [8]. Table 2 shows the hardness and heat treatment of the mould materials.

Table 1: Chemical composition of mould material

Mould number	Standard	Toughness (J)*	Chemical composition (mass.%)					
			C	Si	Mn	Cr	Mo	V
1, 3, 4	UNE X37 CrMoV 5	22	0.38	0.20	0.33	5.06	1.27	0.50
2	–	18	0.32	0.61	0.88	1.34	0.80	0.15

* from the steel producer

Table 2: Hardness and heat treatment of mould material

Mould	Hardness (HRc)	Temperature of hardening (°C)	Number of tempering cycles	Temperature of the last tempering (°C)
1	43	990	3	605
2	45	Delivery condition		
3	45.5	990	3	590
4	46	990	3	590

Table 3: Chemical composition (mass. %) of aluminium alloy EN 1706 AC-AlS9Cu3 used in the experiment

Alloy/Standard	Chemical composition (mass.%)							
	Si	Cu	Mn	Mg	Fe	Zn	Ni	Al
EN AC-AlSi9Cu3	8–11	2–4	0.55	0.05–0.55	1.3	1.2	0.55	Rest

The alloy used was AlSi9Cu3. Table 3 shows the chemical composition of the aluminium alloy melt. Experiments were conducted on an 8 MN cold chamber machine under real production conditions. Melt temperature was 680 °C, cycle time 58 seconds, and mould temperature maintained at 250 °C via cooling channels. Each casting weighed 332 g.

Since direct measurement of cracks on mould surfaces during operation is impractical, defect-fins on castings were analyzed. These represent negative replicas of mould surface defects. Although very fine cracks may not be fully captured, the method allows estimation of crack size and growth. Figure 2 shows the observed die casting mould and the corresponding casting.

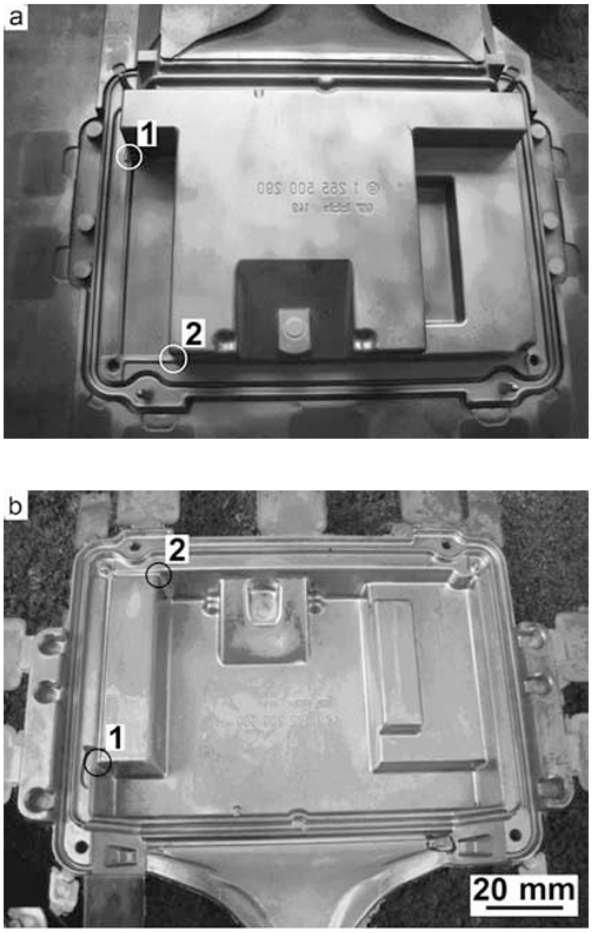


Figure 2: Observed die casting mould and the corresponding casting

Measurements were performed using a profilometer, recording defect-fin depth, width, and length. In addition, microstructural analysis was conducted on a used mould subjected to over 40,000 cycles. Figure 3 schematically shows the measurement of casting defects due to cracks in the die casting tool.

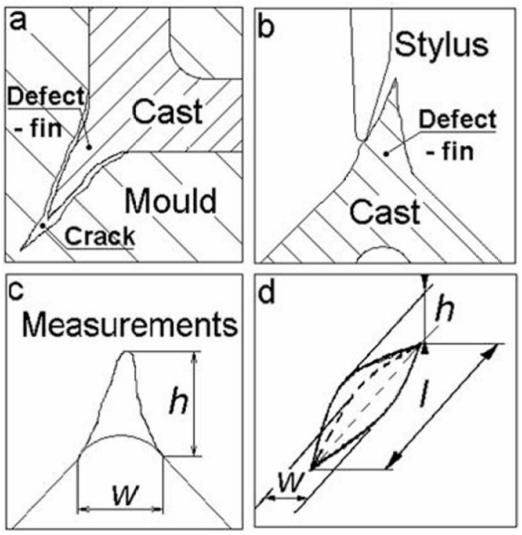


Figure 3: Schematic representation of a) die casting mould crack and the corresponding casting defect, b) measurement stylus, c) shape and size of defect cross – section, d) schematic presentation of defect: depth – *h*, width – *w* and length – *l*

3 RESULTS AND DISCUSSION

Cracks were predominantly observed at edges and corners with small radii, where stress concentrations occur. These locations are also affected by residual stresses from manufacturing and dynamic melt pressure effects.

Crack size and density varied depending on location and material properties. Larger and deeper cracks appeared near the melt entrance due to higher thermal loads. Crack widening was attributed to erosion and merging of multiple cracks.

The mould surface experienced cyclic temperatures from approximately 50 °C to 650 °C. Repeated exposure above 500 °C caused tempering, reducing hardness. Microstructural changes included transformation of martensite and coarsening of carbides, which weakened the material and accelerated crack formation [9]. Figure 4 shows casting failure due to thermal fatigue cracking in a die casting mould depending on the number of mould operating cycles.

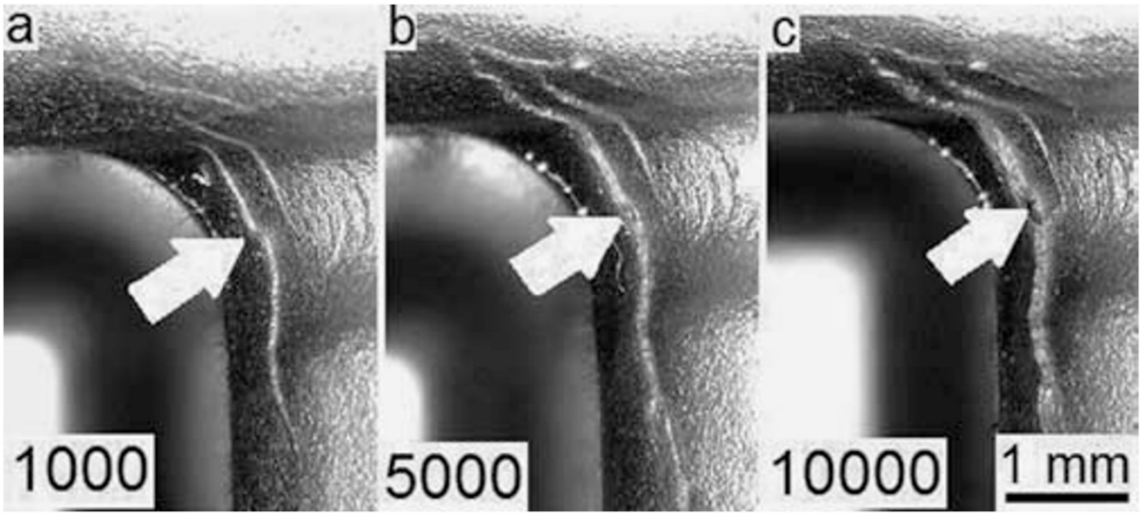


Figure 4: Casting failure due to thermal fatigue cracking in a die casting mould depending on the number of mould operating cycles

Initial cracks appeared earlier in softer materials (~500 cycles) compared to harder ones (~1500 cycles). After 10,000 cycles, defect-fin depth ranged from 0,16 mm to 0,37 mm.

- **Depth growth:** quadratic relationship with cycle number
- **Width growth:** approximately linear
- **Length growth:** logarithmic trend

Figure 5 shows the growth of damage depth on castings due to crack depth on moulds as a function of the number of cycles.

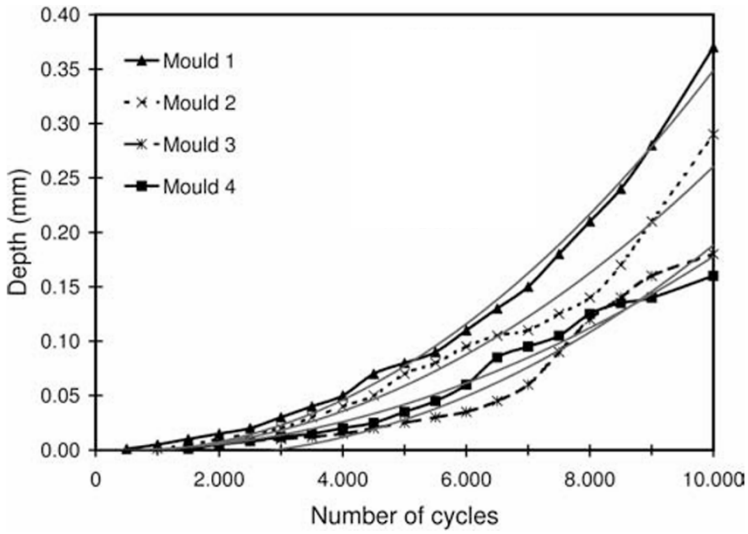


Figure 5: Growth of damage depth on castings due to crack depth on moulds as a function of the number of cycles

Figure 6 shows the growth of damage width on castings due to crack width on moulds as a function of the number of cycles.

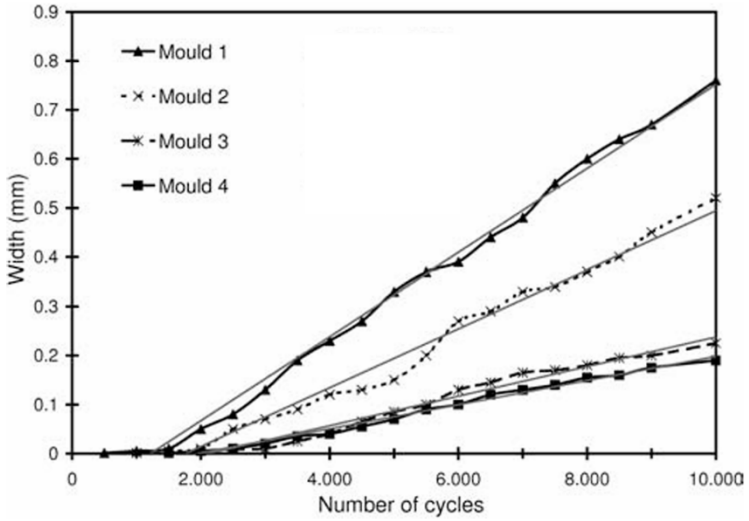


Figure 6: Growth of damage width on castings due to crack width on moulds as a function of the number of cycles

Figure 7 shows the growth of damage length on castings due to crack length on moulds as a function of the number of cycles.

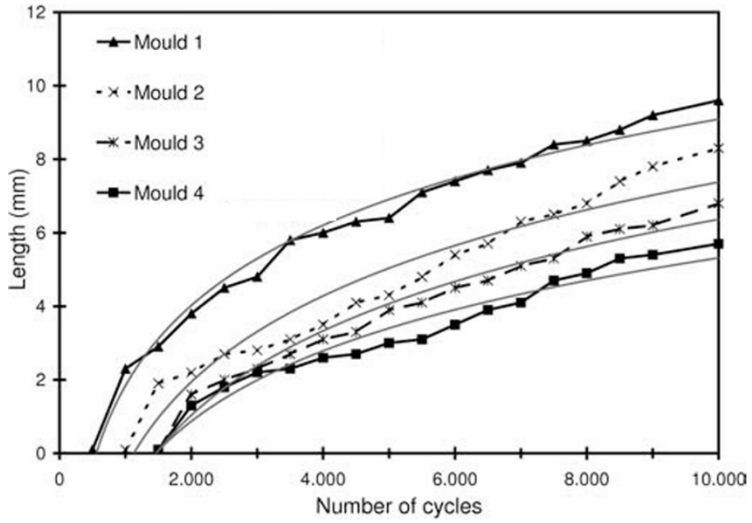


Figure 7: Growth of damage length on castings due to crack length on moulds as a function of the number of cycles

From the results shown in Figures 5 to 7, it can be concluded that mould hardness has the greatest impact on thermal fatigue resistance. Harder moulds tend to develop smaller defect fins, and their growth rate is significantly slower.
When comparing defect-fin growth between mould/material 2 (45 HRC) and mould/material 3 (45.5 HRC), the difference might initially seem minor. However, the results clearly show a noticeable distinction, as material 3 exhibits considerably slower defect propagation, indicating better thermal fatigue resistance.
This difference can be attributed to variations in chemical composition as well as differences in production methods. Materials 1, 3, and 4 are produced from steel that has been remelted under vacuum conditions. This process results in a highly uniform structure with very low micro-cleanliness defects, which significantly delays the onset of heat checking.

The progression of fin-type defect depth in castings obtained from moulds 1 and 2 demonstrates a continuous increase throughout the testing interval, reaching up to 10,000 cycles. Conversely, castings produced using moulds 3 and 4 show a non-monotonic trend, with defect growth diminishing after approximately 6,500 and 8,000 cycles, respectively (Fig. 5). A comprehensive interpretation of this behaviour remains limited, as systematic observations were terminated at 10,000 cycles. However, it is reasonable to postulate that materials exhibiting enhanced mechanical performance, together with higher resistance to thermal fatigue and tempering effects, are more effective in inhibiting crack propagation than materials characterized by inferior mechanical properties and reduced resistance to thermally induced degradation.

In addition, this steel demonstrates relatively high toughness, reaching about 22 J at working hardness, compared to approximately 18 J for material 2.

Harder materials showed slower crack growth, confirming their superior resistance. Location also played a role: areas closer to the melt gate experienced faster crack growth due to higher thermal exposure.

4 CONCLUSIONS

Thermal fatigue cracking is the dominant failure mode in die casting moulds. Crack formation and propagation depend on geometry, temperature distribution, melt flow, and material properties.
The proposed method—monitoring defect-fins on castings—proved effective for evaluating crack growth during real production without interrupting the process.

Key findings:

- Cracks initiate at stress concentrators (edges, corners)
- Higher temperatures accelerate crack growth
- Softer materials show earlier and larger cracks
- Crack depth, width, and length follow distinct growth laws (quadratic, linear, logarithmic)

This approach enables practical prediction of mould lifetime. Future work should focus on building a larger dataset and applying artificial intelligence methods for improved failure prediction.

REFERENCES

[1] L.J.D. Sully, Metals handbook., Vol. 15 (ASM International, Metals Park OH, 1988).
[2] J.R. Davis, ASM Speciality Handbook., Tool Materials (ASM International, Materials Park OH, 1995).
[3] Chester L. Lucas, Casting: Dies—Machines—Methods. Good Press, 2019.
[4] ASM handbook. Vol. 15, Casting. Materials Park, Ohio: ASM International, 2008, str. XVIII, 1238.
[5] A. Weroński, Thermal fatigue of metals, let. 74. New York: M. Dekker, 1991, str. VII, 366.
[6] T. L. Anderson, Fracture mechanics: fundamentals and applications, 3rd ed. Boca Raton, FL: Taylor & Francis; CRC Press, 2005, str. XVI, 621.
[7] S. K. Maiti, Fracture mechanics: fundamentals and applications. Delhi: Cambridge University Press, 2015, cop., str. XVI, 279.
[8] SSAB, “Toolox® – the premium engineering and tool steel,” *SSAB*, [Online]. Available: https://www.ssab.com/en/brands-and-products/toolox/product-offer. [Accessed: May 4, 2026].
[9] E. Oberg, Heat-Treatment of Steel: A Comprehensive Treatise on the Hardening, Tempering, Annealing and Casehardening of Various Kinds of Steel. New York, NY, USA: The Industrial Press, 1920.

Modelling and optimisation of traction electric motor torque

David Oman[1], Damir Vrančić[1,2]

[1]Faculty of industrial engineering Novo mesto, Slovenija
E-mail: david.oman@students.fini-unm.si, damir.vrancic@fini-unm.si

[2]Institut Jožef Stefan, Jamova cesta 39, 1000 Ljubljana, Slovenija
E-mail: damir.vrancic@ijs.si

Abstract*: It is impossible to imagine today's world without the silent "worker" of industry and daily life: the electric motor. From compressors in refrigerators to drives in robotics, electric motors power our lives and ensure the smooth functioning of the modern world. The electric motor converts electrical energy into mechanical energy, resulting in various types of losses, with ohmic losses in the windings usually being the largest. These losses cause the temperature inside the motor to rise, which can lead to deterioration of insulation properties, reduced efficiency, shorter service life, and, in extreme cases, mechanical or electrical failures. Effective heat dissipation from the motor windings is therefore crucial for reliable and long-term operation. In this paper, the heat transfer mechanisms inside a traction electric motor were first analysed, and a mathematical model of heat transfer was developed. Heat transfer in an impregnated and potted stator was then simulated, and a sensitivity analysis was performed to determine which parameters have the greatest impact on temperature changes. The developed mathematical model was tested in practice, finding that the difference between the calculated and measured electric current at the same motor temperature is approximately 7%, which is a very small deviation. By modelling the heat losses of impregnated and potted motors with different epoxy* compounds, the most optimal epoxy compound was selected in terms of production costs and manufacturing process complexity. Both the model and actual measurements on the electric motor showed that the selected epoxy compound achieved approximately 8% higher motor torque compared to the impregnated motor.

Keywords: traction electric motor, electric motor modelling, heat losses, torque characteristic, torque optimisation

Article Classification: Scientific Paper

1 INTRODUCTION

Electric motors are among the main consumers of electrical energy in modern industry and transportation, accounting for approximately half of global electricity generation. Therefore, their efficiency is directly linked to the overall energy efficiency of the system. In traction electric motors, where high power density and reliability are critical, managing thermal losses is a key engineering challenge.

This study addresses the problem of heat dissipation from stator windings. Inadequate heat removal results in elevated operating temperatures, accelerated insulation ageing, and consequently reduced efficiency and service life of the electric motor. Conventional winding impregnation methods often fail to provide optimal thermal contact between the windings and the motor housing, particularly in the winding end regions, where air pockets with low thermal conductivity are present.

The objective of this research was to develop and validate a thermal model of a traction motor stator and to evaluate the influence of different filling methods around the windings (impregnation and epoxy potting) on stator temperatures and, consequently, on the torque characteristics of the motor. Emphasis was placed on improving heat transfer from the windings to the housing and the cooling system.

The key variables considered in the study include:

- thermal conductivity of materials (windings, insulation, impregnation resin, epoxy compound),
- geometric parameters of the stator (layer thicknesses, heat transfer path lengths),
- thermal contact resistance between components,
- heat transfer coefficient on the cooling side,
- electrical losses in the windings,
- temperatures of the windings and housing, and
- motor torque as a function of temperature.

2 THERMAL MODELLING

Temperatures within the stator depend on the thermal conductivity of the materials present (windings, insulation, impregnation resin, epoxy resin), the geometric parameters of the stator (layer thicknesses and heat transfer path lengths), thermal contact resistances between components, heat transfer coefficients on the cooling side, electrical losses in the windings, as well as the temperatures of the windings and the housing. The motor torque depends on the electrical current, which is a function of the winding temperature.

Heat transfer can be conveniently modelled using an electrical circuit analogy: temperature is represented by electric potential, temperature difference by the voltage across electrical elements, heat flow (loss power) by electric current, and the thermal resistances of individual layers by electrical resistances, as shown in Figure 1. For steady-state one-dimensional heat transfer through a layered structure, thermal resistances in series are additive, analogous to resistors connected in series. Similarly, thermal conductances in parallel combine in the same way as parallel electrical resistances. This representation is commonly referred to as a thermal network or thermal circuit [1, 2].

Parameter P_{cu} represents copper winding losses (heating power). It is the first input parameter for the thermal model, obtained from the motor data. Iron losses are added to the copper losses, as they are not negligibly small, while losses in the rotor and magnets can be neglected for the purpose of simplifying the thermal model [3].

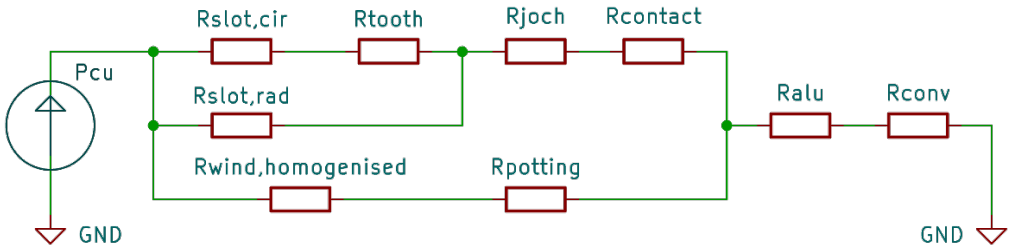


Figure 1: Electrical equivalent circuit of heat flows and thermal resistances (thermal network)

The temperature model of the stator is divided into two parallel branches, as shown in Figure 1 [4]. The resistance $R_{slot,\,cir}$ represents the thermal resistance from the winding towards the left and right to the stator tooth, while $R_{slot,rad}$ represents the thermal resistance from the winding in the radial direction of the slot towards the yoke. The resistances $R_{wind,homogenised}$ and $R_{potting}$ represent the thermal resistance of the end windings to the housing.

Both branches merge upstream of the contact resistance of the aluminium housing R_{alu} and the heat transfer to the cooling medium R_{conv}.

For clarity, the thermal resistances can be mapped onto the actual stator cross-section. Since both the end winding and stator core regions are considered, two corresponding schematic representations are shown in Figure 2.

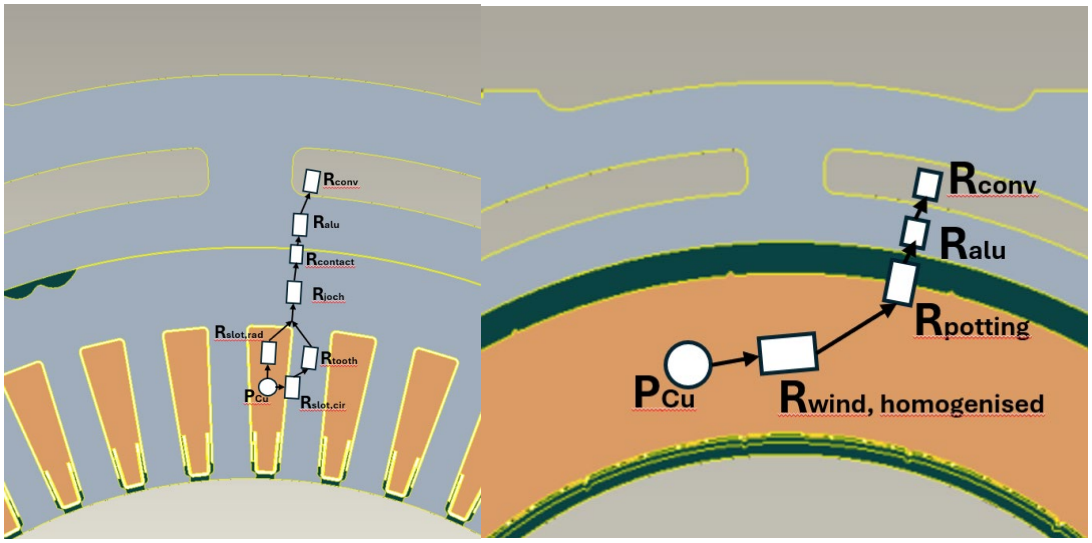


Figure 2: Schematic representation of heat fluxes and thermal resistances in stator (left is the stator core, right is the end windings region)

The basic equation, used for one-dimensional heat transfer is:

$$R = \frac{L}{k_x \cdot A} \left[\frac{\mathrm{K}}{\mathrm{W}}\right], \tag{1}$$

where R is the thermal resistance, L is the length, A is the area, and k_x is the thermal conductivity of solids. The interface between two different materials can be modelled as:

$$R_{contact} = \frac{1}{h_{contact} \cdot A} \left[\frac{\mathrm{K}}{\mathrm{W}}\right], \tag{2}$$

where A is the area of contact between two materials and $h_{contact}$ is the interface heat transfer coefficient. Particular attention is given to the end winding area, where, in the case of an impregnated stator, an air layer surrounds the end winding, acting as an insulator with low thermal conductivity, while epoxy resin significantly improves heat flux to the housing.

Equation (3) represents the homogenised thermal conductivity in the stator slot. As different materials are present, such as insulation, varnished copper wire, and impregnating or epoxy resin, the combined thermal conductivity can be calculated [5]:

$$k_{slot,hom.imp} = k_{imp} \cdot \frac{(1+v_c)k_c + (1-v_c)k_{imp}}{(1-v_c)k_c + (1+v_c)k_{imp}} = 0.6899 \ \frac{\mathrm{W}}{\mathrm{mK}} \tag{3}$$

where $k_{slot,hom,imp}$ is the homogenised stator slot thermal conductivity for the impregnated stator, k_{imp} is the coefficient of thermal conductivity of the impregnating resin, v_c is the mechanical fill factor, and k_c is the coefficient of thermal conductivity of the varnished copper wire. Note that equation (3) also applies to epoxy-potted stators. The only difference is the thermal conductivity of the materials.

The thermal resistance of the slot circumference (left and right of the slot) is:

$$R_{S,C\ imp} = \frac{s_{w,cu}}{8 \cdot s_{h,cu} \cdot L_{fe} \cdot k_{slot,hom,imp}} + \frac{d_p}{2 \cdot s_{h,cu} \cdot L_{fe} \cdot k_p} = 0.3378 \ \frac{\mathrm{K}}{\mathrm{W}} \tag{4}$$

where $s_{w,cu}$ is the slot width without insulation, $s_{h,cu}$ is the slot height without insulation, L_{fe} is the length of the active parts, d_p is the thickness of the insulating paper, and k_p is the coefficient of thermal conductivity of the insulating paper.

Next is the calculation of the radial side of the stator slot:

$$R_{S,R\ imp} = \frac{s_{w,cu}}{8 \cdot s_{h,cu} \cdot L_{fe} \cdot k_{slot,hom,imp}} \cdot \frac{s_{h,cu}}{s_{w,cu}} + \frac{d_p}{s_{w,cu} \cdot L_{fe} \cdot k_p} = 2.2943 \frac{\mathrm{K}}{\mathrm{W}} \tag{5}$$

The thermal resistance of the stator tooth is:

$$R_{tooth} = \frac{s_{h,cu}}{4 \cdot t_h \cdot L_{fe} \cdot k_{fe}} = 0.1602 \ \frac{\mathrm{K}}{\mathrm{W}} \tag{6}$$

where t_h represents the stator tooth width and k_{fe} the coefficient of thermal conductivity of iron. The calculation of the thermal resistance of the stator yoke can be simplified by considering a circular segment defined by the number of stator slots:

$$R_{joch} = \frac{N \cdot \ln\left(\frac{r_{stator,outer}}{r_{stator,joch}}\right)}{2 \cdot \pi \cdot L_{fe} \cdot k_{fe}} = 0.1446 \ \frac{\mathrm{K}}{\mathrm{W}} \tag{7}$$

where N=48 is the number of stator slots, $r_{stator,outer}$ is the outer stator radius, and $r_{stator,joch}$ is the stator yoke radius.

At the interface between the stator and the housing, the contact thermal resistance must be evaluated. The contact between the laminated iron core and the housing is not ideal, as the laminations may be slightly misaligned, non-concentric, or the stacking process may not produce a perfectly flat stator stack. When calculating the contact resistance, only a segment of the circumference is considered:

$$R_{contact} = \frac{1}{\frac{\pi \cdot D_{stator}}{N} \cdot L_{fe} \cdot h_{contact}} = 0.3961 \frac{\mathrm{K}}{\mathrm{W}} \tag{8}$$

where D_{stator} is the diameter of the stator and $h_{contact}$ is the interface heat transfer coefficient between the laminations and the housing. According to the literature, typical values for this coefficient range from 100 to 1000 W/m²K [6, 7]. In this study, a lower estimate is adopted due to surface irregularities on the outer diameter of the stator stack. Air pockets are often trapped in this region, and depending on the impregnation or potting process, both of which have significantly lower thermal conductivity than iron, may also be present. Additionally, weld grooves along the circumference of the stator stack create recessed regions that do not contact the housing.

Next, the thermal resistance of the aluminium housing is calculated in a similar way as the stator yoke. Here, the total length is defined as the sum of the total active parts length and the lengths of both winding overhangs:

$$R_{housing} = \frac{N \cdot \ln\left(\frac{r_{housing,outer}}{r_{housing,inner}}\right)}{2 \cdot \pi \cdot L_{fe+wo} \cdot k_{aluminum}} = 0.0117 \ \frac{\mathrm{K}}{\mathrm{W}} \tag{9}$$

where $r_{housing,outer}$ represents the housing outer radius of the water channels, $r_{stator,inner}$ is the housing inner radius, L_{fe+wo} is the total length of housing, and $k_{aluminum}$ is the coefficient of thermal conductivity of aluminium. Equation (10) represents the final thermal resistance (convection), which is the heat transfer to the cooling medium:

$$R_{convection} = \frac{1}{\frac{\pi \cdot 2 \cdot r_{housing,outer}}{N} \cdot L_{fe+wo} \cdot h_{convection}} = 0.0396 \ \frac{\mathrm{K}}{\mathrm{W}} \tag{10}$$

The heat transfer coefficient $h_{convection}$ is reported in the literature to range between 1000 and 10000 W/m²K, as it depends on the flow velocity and the geometry of the cooling channels [8, 9]. In this study, the coefficient is selected based on internal company data and an estimated mean value. It should be noted, however, that this parameter may vary by approximately 20-30 %, depending on the type of coolant and the surface characteristics of the cooling channels. The accuracy of this estimate is not critical, as the sensitivity analysis has shown that the results are not highly sensitive to variations in this coefficient [10].

Moving to the winding overhangs calculation, equation (11) calculates the homogenised winding overhangs thermal conductivity:

$$R_{wind,homogenised,imp} = \frac{N \cdot \ln\left(\frac{r_{winding,outer}}{r_{winding,inner}}\right)}{2 \cdot \pi \cdot L_{wo} \cdot k_{slot,hom,imp}} = 36.614 \ \frac{\mathrm{K}}{\mathrm{W}} \tag{11}$$

where $r_{winding,outer}$ is the outer radius of the winding overhang, $r_{winding,inner}$ is the inner radius of the winding overhang and L_{wo} is the total axial length of both overhangs combined. It should be emphasised that the ratio of thermal conductivity between epoxy and impregnation resin is approximately 9. Next is the thermal resistance of the air layer surrounding the overhangs:

$$R_{air} = \frac{N \cdot \ln\left(\frac{r_{housing,inner}}{r_{winding,outer}}\right)}{2 \cdot \pi \cdot L_{wo} \cdot k_{air}} = 188.0535 \ \frac{\mathrm{K}}{\mathrm{W}} \tag{12}$$

where k_{air} is the coefficient of thermal conductivity of air. When all separate thermal resistances have been calculated, we can combine them:

$$\begin{aligned} R_{total\ imp} &= R_{stator\ imp} + R_{housing} + R_{convection} = 0.9972 \ \frac{\mathrm{K}}{\mathrm{W}} \\ R_{total\ epoxy} &= R_{stator\ epoxy} + R_{housing} + R_{convection} = 0.7955 \ \frac{\mathrm{K}}{\mathrm{W}} \end{aligned} \tag{13}$$

The total thermal resistance of the stator filled with epoxy resin is therefore more than 20 % lower than that of the impregnated stator. This is followed by the calculation of the simulated stator temperature, using the total thermal resistance of the respective branch, the calculated copper losses, and the coolant temperature as input parameters:

$$\begin{aligned} \Delta T_{imp} &= \left(\frac{P_{cu}}{N}\right) \cdot R_{total\,imp} = 90.31\text{°C}\,, \qquad \Delta T_{epoxy} = \left(\frac{P_{cu}}{N}\right) \cdot R_{total\,epoxy} = 72.04\text{°C} \\ T_{final,imp} &= T_{water} + \Delta T_{imp} = 155.31\ \text{°C}\,, \qquad T_{final,epoxy} = T_{water} + \Delta T_{epoxy} = 137.04\ \text{°C} \end{aligned} \tag{14}$$

where P_{cu}=4347 W, and T_{water}=65 °C. At the end of the analysis, it is found that the stator filled with epoxy resin exhibits a temperature 18.27 °C lower than that of the impregnated stator.

The motor torque depends on the stator current, which in turn depends on the winding resistance, itself a function of the winding temperature. The maximum allowable temperature (thermal limit) of the stator winding is 155 °C. At this temperature, the calculated resistance of the stator winding is:

$$R_{155\text{°C}} = R_{20\text{°C}} \cdot (1 + \alpha_{cu} \cdot \Delta T) = 0.0207\ \Omega \tag{15}$$

where $R_{20\text{°C}}$=0.0135 Ω is the phase resistance of the stator at room temperature, and α_{cu}=0.00393 °C^{-1} is the temperature coefficient of standard annealed copper at 20 °C.

Under identical operating conditions, copper losses were experimentally measured in the laboratory for both motors. For the impregnated motor, the copper losses $P_{cu,imp}$ were 4347 W, while for the epoxy-filled motor, the losses $P_{cu,epoxy}$ were 5131 W. To determine the maximum electric currents at thermal limit of the stator (155 °C), and thus indirectly the motor torque, the motor power equation was used [2]:

$$P_{cu} = 3 \cdot I_{ph}^2 \cdot R_{ph} \tag{16}$$

From equation (16), the electrical current can be derived for impregnated and epoxy-potted motor and calculated as follows:

$$\begin{aligned} I_{ph\,imp} &= \sqrt{\frac{P_{cu\,imp}}{3 \cdot R_{155\text{°C}}}} = 264.71\ \text{A} \\ I_{ph\,epoksi} &= \sqrt{\frac{P_{cu\,epoksi}}{3 \cdot R_{155\text{°C}}}} = 287.59\ \text{A} \end{aligned} \tag{17}$$

where I_{ph} is the phase current. From the above equations, it follows that, at the same final stator temperature, the epoxy-filled stator allows an additional current of 22.88 A, corresponding to a 7.96 % increase compared to the impregnated stator.

In our case, as the electric motor operates in a constant torque regime at 250 rpm, where iron and magnet losses are negligible, it can be assumed that the increase in torque is directly proportional to the increase in motor current. Therefore, an electric motor filled with epoxy resin, rather than being conventionally impregnated, can deliver approximately 7.96 % higher torque under identical operating conditions, as predicted by the thermal model.

Before experimental validation on physical motors, a sensitivity analysis of the model parameters was conducted. The results indicate that the stator temperature is most sensitive to the heat transfer coefficient at the stator-to-housing interface, where even small variations cause significant temperature changes. This aligns with findings reported in the literature, where contact resistance is identified as one of the most difficult parameters to determine and a dominant factor influencing the overall thermal behaviour of electric machines [6, 7].

The developed model also enabled parametric simulations of various material configurations within the motor. The analysis showed that further torque improvements could be achieved by increasing the thermal conductivity of the epoxy potting material. However, such materials are typically associated with higher costs and more demanding manufacturing processes. The results indicate that using commercially available high-conductivity epoxy materials is not economically justified, as the resulting torque increase does not offset the significantly higher material and production costs. Additionally, a relatively simple reduction in insulation paper thickness can further improve motor torque by approximately 3.5 %.

3 EXPERIMENTAL MEASUREMENTS

In the design of modern electric machines, thermal modelling is one of the fundamental tools for evaluating permissible loading and efficiency. These models enable the calculation of temperature distributions based on winding losses and material properties, and are essential for identifying the optimal trade-off between motor design and cooling performance. With the increasing demand for higher power density and reduced mass, thermal constraints are becoming a limiting factor in traction motor design. Previous studies indicate that achieving higher power densities requires efficient heat dissipation and a reliable thermal model capable of accurately predicting motor behaviour during the design phase.

To ensure that the model accurately reflects real operating conditions, it must be validated against experimental measurements. Only validated models can be reliably used for selecting optimal motor designs and cooling strategies, as well as for assessing thermal margins under varying load conditions.

Measurements of torque, electrical current, and losses were conducted under controlled laboratory conditions (Figure 3). Torque was measured using an HBM T40 torque transducer, which has a measurement uncertainty of ±0.03 % relative to the voltage output signal. The electrical current was measured using ZES Zimmer Hall1000 current clamps, enabling accurate acquisition of the current supplied from the inverter to the motor. The nominal current rating of the clamps is 1000 A, with a bandwidth of 100 kHz and an accuracy of ±0.5 %. Both motors operated in the S1 continuous duty regime, with a coolant flow rate of 20 L/min and a constant coolant temperature of 65 °C.

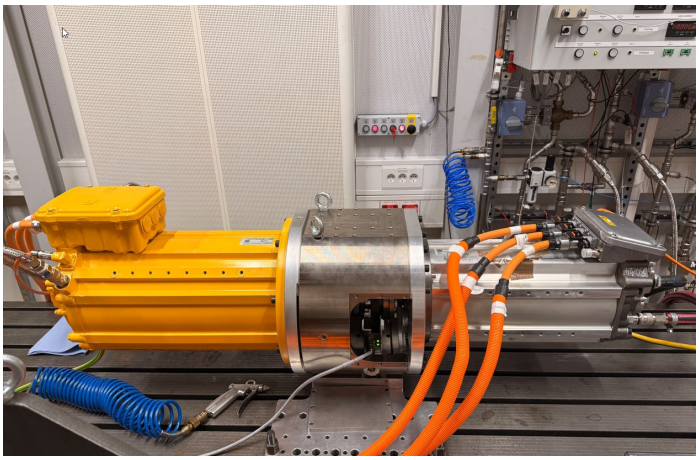


Figure 3: Measurement of torque, current and losses in the laboratory

The torque measurements (Figure 4) indicate that the epoxy-filled motor achieves higher torque in the low-speed region, where the winding temperature sets the thermal limit of the machine. At 250 rpm, the measured torque of the epoxy-filled motor was 7.9 % higher than that of the impregnated motor.

For a direct comparison of the results, electrical current measurements were also carried out for both motors (Figure 5).

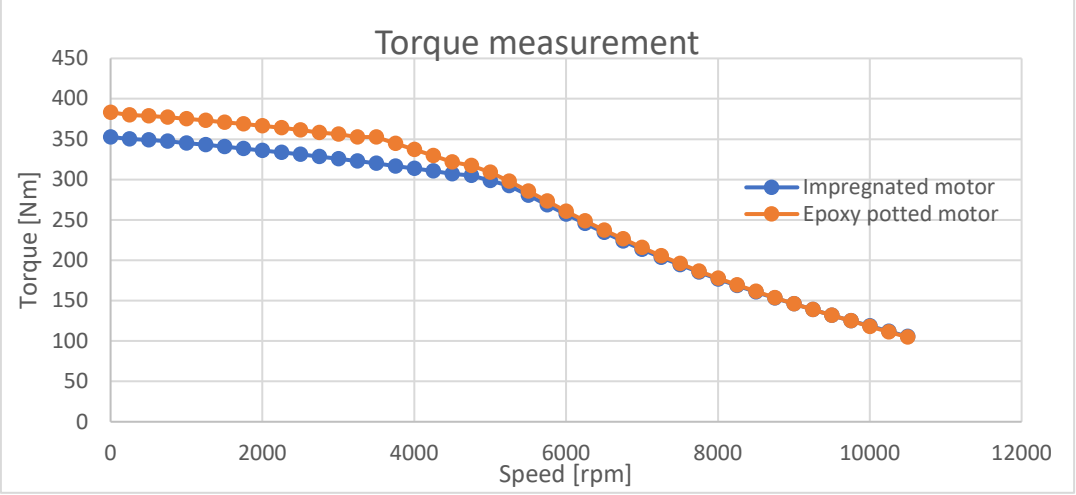


Figure 4: Graph of torque over speed for both motors

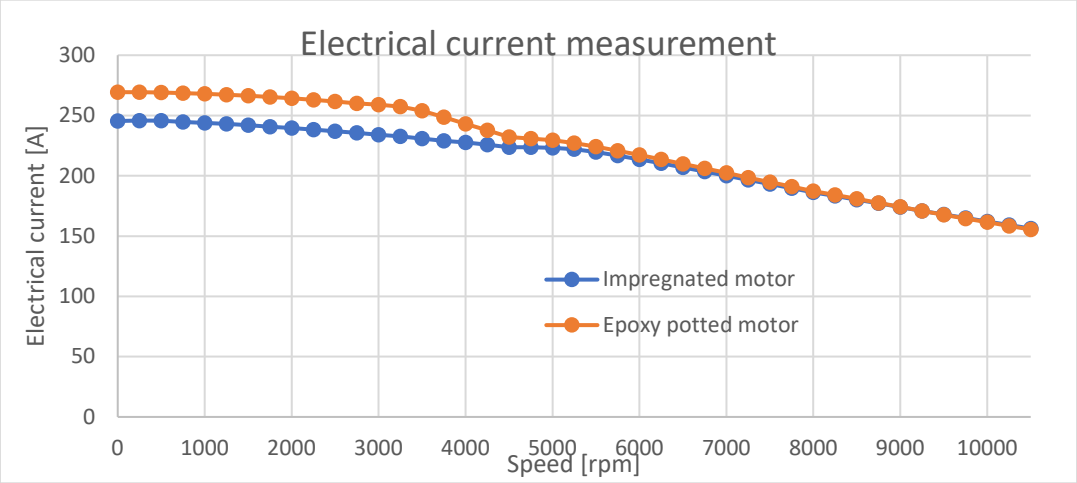


Figure 5: Graph of electrical current over speed for both motors

At low rotational speeds, the epoxy-filled motor can sustain approximately 8.55 % higher current. Up to the rated speed of the motor (4000 rpm), the same trend is observed as in the torque characteristics. Beyond the rated speed, the electrical current of both motors becomes nearly identical.

Table 1 compares the measured and calculated motor currents. Discrepancies may arise from measurement uncertainties in the instrumentation, environmental and operating conditions, model simplifications, manufacturing tolerances, and uncertainties in model parameters. Despite these factors, the results show relatively good agreement between the thermal model and the experimentally measured current values.

Table 1: Comparison between measured and calculated motors

	Impregnated (IM) motor	Epoxy potted (EP) motor	Ratio between IM and EP
Calculated el. current [A]	264.7	287.6	0.920
Measured el. current [A]	246	269	0.914
Relative difference [%]	7.1	6.5	0.6

As shown in Table 1, the deviation between the calculated and measured motor current is approximately 7 %. Given the previously discussed uncertainties and model simplifications, this level of deviation can be considered acceptable.

However, it is also important to consider the primary objective of the developed model, which is to compare the electrical currents between epoxy-filled and impregnated motors. In this context, when comparing the ratio of currents obtained from the model and the experimental measurements, the deviation is only 0.6 %, indicating very good agreement.

4 CONCLUSION

The thermal model was validated against experimental measurements, showing a deviation of approximately 7 % between the predicted and measured motor current. Given the inherent non-idealities and modelling simplifications, this level of accuracy is considered very good. The model is therefore sufficiently accurate for predicting steady-state temperatures, as it reliably captures both the trend and magnitude of temperature levels required for engineering design. Additionally, the comparison between model predictions and measurements demonstrated that the deviation in the current ratio between the impregnated and epoxy-filled motor is only 0.6 %, representing excellent agreement, particularly when the model is used for comparative analysis of different material configurations.

The measurements indicate that epoxy potting is most effective in the low-speed operating range. Up to the rated speed, the thermal limit of the PMSM is the stator windings, making efficient cooling of this region essential. Beyond the rated speed, the motor operates in the field weakening region, where thermal losses shift towards the rotor, magnets, and iron core. In this regime, the torque difference between the impregnated and epoxy-filled motor becomes negligible. The relative improvement is even more pronounced in motors with shorter active lengths, where the winding end regions constitute a larger proportion compared to the stator core. In such cases, an even greater torque increase can be expected when using epoxy potting.

A sensitivity analysis of the model parameters revealed that the heat transfer coefficient at the stator-to-housing interface is the most critical factor influencing the reduction of stator winding temperature. It was also found that the most expensive epoxy materials, characterised by the highest thermal conductivity, do not justify their higher cost

and more demanding manufacturing processes. Additionally, reducing the thickness of the insulation paper was shown to further improve motor torque by approximately 3.5 %.

The developed thermal model is currently tailored to the tested motor; however, it can be readily extended to similar machines with distributed windings. The stator geometry and material properties can be easily adapted, while the prediction of steady-state temperatures requires only a sufficiently accurate estimation of DC copper losses.

Thus, the thermal model provides high practical value, particularly in the early stages of electric motor design, where different cooling concepts and material configurations can be evaluated quickly and cost-effectively prior to prototype development. The model is sufficiently simple to allow rapid modification and computation, yet sufficiently accurate to support comparison between different design concepts. It serves as a valuable tool for making design decisions, such as selecting insulation thickness, choosing between epoxy potting and impregnation, and defining key stator parameters including slot number, yoke thickness, and the slot-to-tooth ratio. In principle, the model can also be used to optimise other stator materials, such as laminations, copper, and aluminium. Improving the thermal conductivity of these materials has a beneficial effect on the overall thermal behaviour and temperature distribution within the windings.

REFERENCES

[1] IETresearch [splet]. Available at: https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/elp2.12351 [1.11.2025]
[2] J.P. Holman – Heat transfer, Tenth edition [2010]
[3] S.J. Chapman – Electrical machinery fundamentals, Fifth edition [2012]
[4] Sebastien Sequeira, Kevin Bennion, J. Emily Cousineau, Sreekant Narumanchi, Gilbert Moreno, Satish Kumar, and Yogendra Joshi - Validation and Parametric Investigations Using a Lumped Thermal Parameter Model of an Internal Permanent Magnet Motor [2010] Available at: https://www.researchgate.net/publication/347603613_Validation_and_Parametric_Investigations_Using_a_Lumped_Thermal_Parameter_Model_of_an_Internal_Permanent_Magnet_Motor
[5] Sabrina Ayat, Haipeng Liu, Mehmet Kulan and Rafal Wrobel - Estimation of Equivalent Thermal Conductivity for Electrical Windings with High Conductor Fill Factor [2018] Available at: https://ieeexplore.ieee.org/document/8557534
[6] N. Simpson, T. Duggan, P. H. Mellor and J. D. Booker - Measurement of the Thermal Characteristics of a Stator-Housing Interface [2017] Available at: https://ieeexplore.ieee.org/document/8062410
[7] N. Simpson, R. Wrobel, J. D. Booker, P. H. Mellor - Multi-Physics Experimental Investigation into Stator-Housing Contact Interface [2016] Available at: https://ieeexplore.ieee.org/document/7739317
[8] Mircea Popescu, David Alan Staton, Aldo Boglietti, Andrea Cavagnino, Douglas Hawkins, and James Goss – Modern Heat Extraction Systems for Power Traction Machines – A Review [2016] Available at: https://ieeexplore.ieee.org/document/7383287
[9] Yongda Song, Yue Zhang, Shi Jin, Zhenyao Xu, Fengge Zhang - Investigation of winding cooling enhancement and thermal modelling of large-power high-torque-density direct-drive PMSM [2023] Available at: https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/elp2.12351
[10] Gal Štempihar - 3D numerično modeliranje hladilnega plašča navornega motorja s trajnimi magneti -Masters thesis, University of Ljubljana [2025]

XR-Supported Workforce Training in Complex Manufacturing Systems Using Digital Twins and Artificial Intelligence

Žiga Gosar[1], Marina Panoska[1], Klemen Babuder[2], Marko Močnik[3]

[1]ELVEZ, proizvodnja kabelske konfekcije in predelava plastičnih mas, d.o.o., Slovenia
E-mail: ziga.gosar@elvez.si

[2]ART REBEL 9 d.o.o., Slovenia
E-mail: klemen@artrebel9.com

[3]Pomurski tehnološki park d.o.o., Slovenia
E-mail: marko@p-tech.si

***Abstract**: This paper presents a digital training platform for complex industrial environments that combines extended reality (XR), digital twins, and artificial intelligence. The system enables non-programmers to design interactive training scenarios through a web-based authoring tool where objects, interactions, and procedural rules can be defined visually. The scenarios are deployed to virtual or augmented reality environments in which users perform tasks either in simulated settings or on real equipment supported by digital guidance. The platform integrates voice interaction and an intelligent assistant that uses a Retrieval-Augmented Generation approach to deliver context-aware answers based on organizational knowledge repositories. A shared digital twin layer acts as a single source of truth for simulation, training, and operational support. The architecture consists of a microservice-based backend, XR client application, and web authoring interface. The prototype demonstrates the feasibility of integrating XR, AI, and digital twins to improve onboarding efficiency, reduce time-to-competence, and enhance safety and process quality. The results indicate that such systems represent a key step toward smart factories and sustainable industrial digital transformation.*

***Keywords:** XR training, digital twins, artificial intelligence, smart manufacturing, Industry 4.0*

***Article Classification**: Scientific Paper*

1 INTRODUCTION

Industrial environments are undergoing rapid transformation driven by digitalization, automation, and the integration of intelligent systems. At the same time, companies face increasing challenges in workforce training, particularly in complex manufacturing domains where processes are dynamic, safety-critical, and require precise execution. Traditional onboarding methods, including written procedures and classroom-based training, often fail to address these challenges effectively due to their static nature and limited adaptability.

Research in virtual reality highlights the importance of immersive, human-centered design principles for effective training. As discussed by Jason Jerald, VR systems must consider human perception, interaction, and cognitive load to ensure meaningful learning experiences [1]. Similarly, XR-based training systems have been shown to improve engagement and knowledge retention by enabling experiential learning in controlled environments, as highlighted in a comprehensive review of XR applications in industrial contexts [3].

Augmented reality further extends these capabilities by overlaying digital information onto physical environments, enabling real-time guidance and contextual support. According to the survey by Mark Billinghurst and colleagues, AR enhances task performance by aligning digital instructions with real-world objects, reducing cognitive effort and improving spatial understanding [4]. These capabilities are particularly relevant in industrial environments where workers must perform complex procedures under time constraints.

Another key enabling technology is the concept of digital twins. A digital twin provides a virtual representation of physical assets, enabling simulation, monitoring, and training in a unified environment. Recent research demonstrates that digital twins can significantly enhance industrial training by ensuring consistency between simulated and real-world operations, while also enabling safe experimentation and error analysis [5].

In addition, natural language interaction plays a crucial role in making digital systems accessible and efficient. Open Voice OS (OVOS) is an open-source voice assistant framework that enables speech-driven interaction through modular components such as speech-to-text (STT) and text-to-speech (TTS) [2]. This capability is especially valuable in industrial settings, where hands-free interaction improves usability and safety.

In this context, the Agile Onboarding of Employees in Complex Manufacturing Systems through Digitalized Training (ONBOARD-AI) project was developed to address onboarding challenges in a real manufacturing environment, specifically in plastics processing and cable assembly. The system combines XR, AI, digital twins, and voice interaction to create an adaptive training platform that reduces onboarding time, improves knowledge transfer, and increases workforce productivity. The motivation for this work arises from the need to support multilingual workforces, reduce dependency on human mentors, and standardize training procedures across complex industrial operations.

2 RESEARCH DESIGN AND METHODS

The proposed system is based on a modular architecture integrating XR environments, AI-driven assistance, and digital twin technology. The overall system architecture of the ONBOARDING-AI platform is presented in Figure 1. It illustrates the interaction between XR components, AI services, digital twin infrastructure, and voice interaction modules.

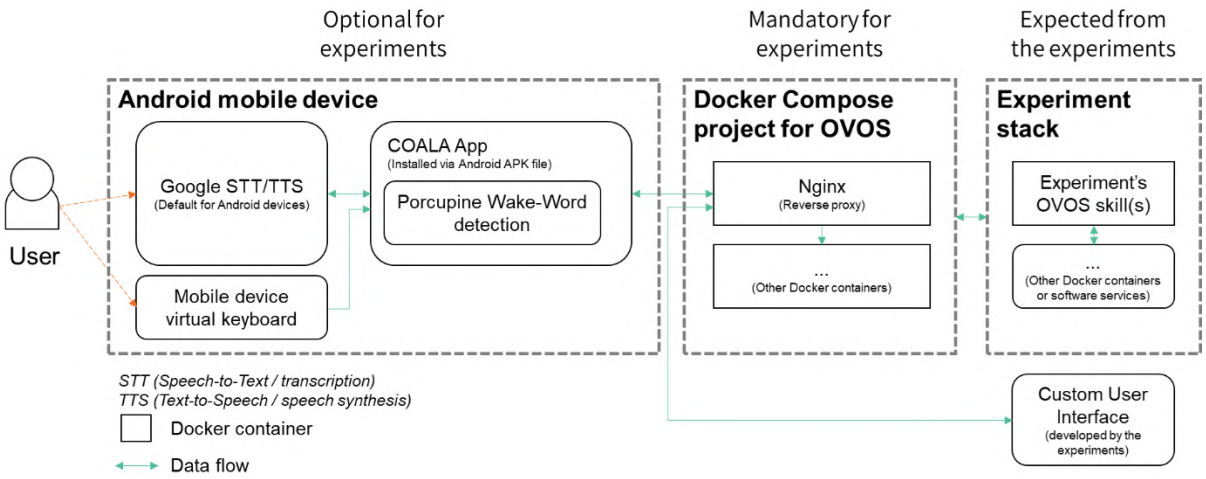


Figure 1: System architecture of the ONBOARDING-AI platform

A web-based authoring tool enables non-programmers to create structured training scenarios consisting of sequential tasks and interaction rules. These scenarios are deployed into VR environments, where users perform guided operations in a simulated workspace derived from real production systems.

Digital twins are created using 3D scanning and process modelling as presented in Figure 2, ensuring that training environments accurately reflect real-world conditions. This enables safe simulation of tasks, including those involving risks or costly errors.

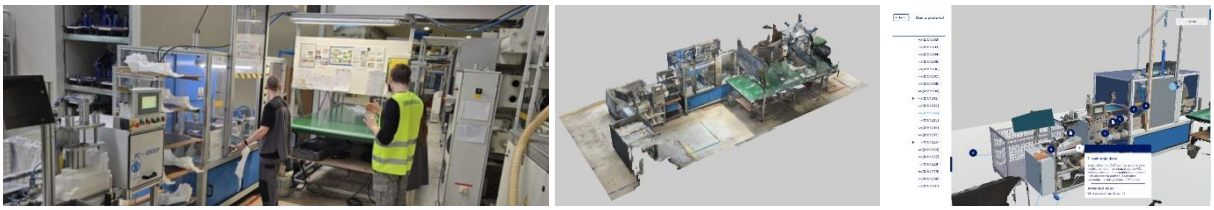


Figure 2: Capturing and completing the environment with 3D camera and making complete replica

The AI assistant is implemented using a Retrieval-Augmented Generation (RAG) approach, allowing it to provide context-aware responses based on internal knowledge bases. Voice interaction is enabled through OVOS, supporting both speech-to-text (STT) and text-to-speech (TTS), enabling natural communication between users and the system.

2.1 System Architecture and Workflow

The ONBOARDING-AI system is designed as a modular architecture integrating VR training, AI-assisted guidance, and voice interaction. The workflow can be formally described as a pipeline:

$$U \xrightarrow{STT} T \xrightarrow{AI} R \xrightarrow{TTS} S \tag{1}$$

Where U presents user speech input, T transcribed text, R is system response generated by AI and S presents synthesized speech output.

This pipeline enables real-time interaction between the user and the system, ensuring continuous feedback during training.

2.2 Voice Interaction and OVOS Integration

Voice interaction is implemented using the Open Voice OS (OVOS) framework, which provides modular components for speech recognition and synthesis [2].

The speech-to-text transformation can be described as:

$$T = f_{STT}(U) = arg\max_{w} P(w|U) \tag{2}$$

where $P(w|U)$ represents the probability of a word sequence given the acoustic input. Similarly, the text-to-speech process is defined as:

$$S=f_{TTS}(T) \tag{3}$$

OVOS manages these transformations through a pipeline of components on Figure 3.

```
</> Python
from ovos_listener import listen
from ovos_tts import speak

while True:
    text = listen()              # STT
    response = process(text)     # AI logic
    speak(response)              # TTS
```

Figure 3: Pipeline of components

The system allows both local and cloud-based processing, improving flexibility and robustness in industrial environments.

2.3 Adaptive AI-Based Training Logic

The training system incorporates adaptive logic that modifies content based on user performance. This can be expressed as a feedback function:

$$A_{t+1} = A_t + \alpha(E-P) \tag{4}$$

Where A_t presents current adaptation level, E presents expected performance, P presents actual performance and α is learning rate.

This mechanism enables personalized training, ensuring that users receive appropriate guidance depending on their skill level.

Additionally, decision-making within the system can be modeled as:

$$\hat{y} = arg\max_{y} f(x,\theta) \tag{5}$$

where the AI model selects the most appropriate next training step based on input x and parameters θ. The symbols in the equation have the following meaning. The variable x represents the input data or features that are provided to the model. In the context of the ONBOARDING-AI system, x can include speech input, sensor data, user interaction history, or the current state of interaction within the VR environment.

The symbol θ denotes the model parameters. These are internal variables that are learned during the training process and define how the model behaves when processing inputs. They may include weights in a neural network, parameters of a classifier, or coefficients in a statistical model.

The function $f(x, \theta)$ represents the model or decision function. This function maps the input x, together with the parameters θ, to a score or probability associated with each possible output y. It defines how the system evaluates different possible actions or responses.

The symbol $\hat{y}$ represents the predicted output of the model. The operation $\arg\max_y$ indicates that the model selects the value of y that achieves the highest score according to the function $f(x, \theta)$. In other words, the model chooses the most likely or most appropriate output based on the given input and learned parameters.

3 RESULTS

The evaluation of the ONBOARDING-AI system was conducted in a real manufacturing environment with a total of 10 participants divided into two groups: a control group undergoing traditional training (n = 5) and an experimental group using VR-based training with AI assistance (n = 5). The assessment focused on key performance indicators, including task completion time, error rate, procedural accuracy, and required mentor intervention.

The results show a significant improvement in onboarding efficiency for the VR-trained group. The average task completion time was reduced by 32%, with participants reaching operational independence faster compared to the control group. Error rates during initial task execution were reduced by approximately 48%, particularly in critical steps such as component positioning, welding sequence execution, and quality inspection.

Procedural accuracy improved notably, with VR-trained participants correctly following predefined workflows in 91% of cases, compared to 68% in the traditionally trained group. Additionally, the need for mentor intervention decreased by 37%, indicating a higher level of autonomy among users trained with the digital system.

Voice interaction through the OVOS-based assistant contributed to improved usability, with users requiring fewer visual prompts and demonstrating smoother task progression. The adaptive AI component further enhanced learning efficiency by dynamically adjusting task difficulty and guidance based on user performance presented in Figure 4.

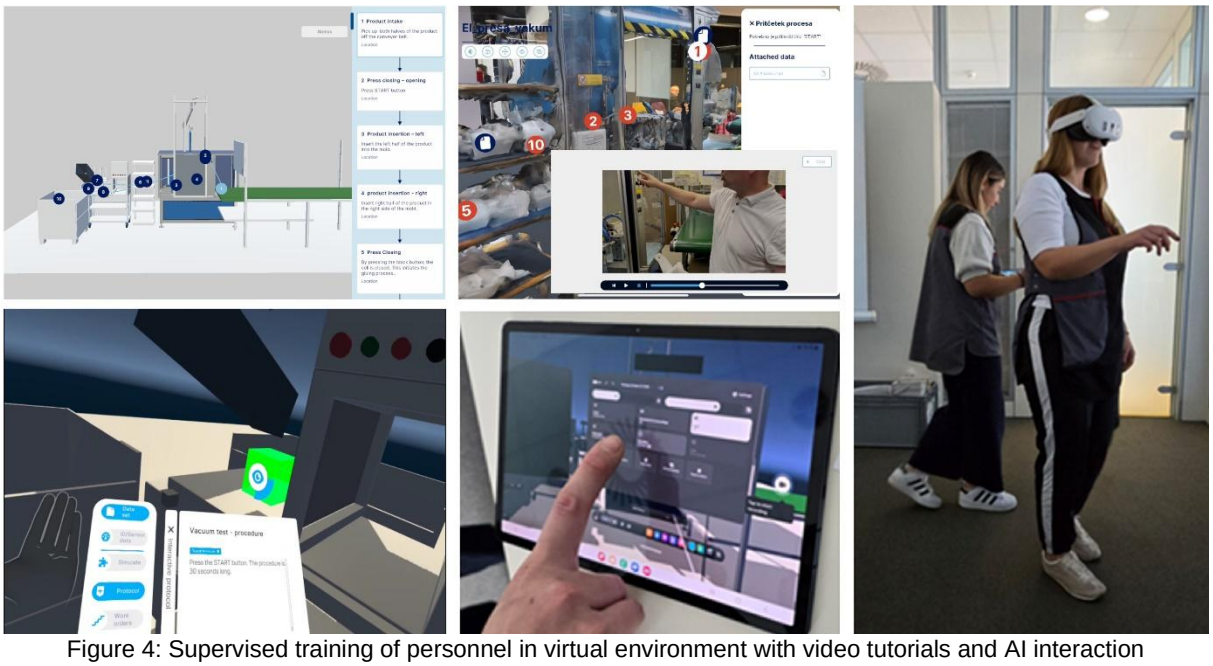


Figure 4: Supervised training of personnel in virtual environment with video tutorials and AI interaction

User feedback indicated higher engagement and confidence levels in the VR group, with 80% of participants reporting improved understanding of process logic and decision-making steps.

4 DISCUSSION

The results of this study confirm that XR-based training provides a significant improvement in workforce onboarding within complex manufacturing environments. The immersive nature of virtual reality enables experiential learning through direct interaction with simulated processes, which aligns with established human-centered design principles for VR systems [1]. This approach enhances user engagement, reduces cognitive load, and supports more effective acquisition of procedural knowledge. The findings are also consistent with prior research showing that XR technologies improve training efficiency and user engagement in industrial contexts [3].

The integration of digital twins further strengthens the system by ensuring that the training environment accurately reflects real-world production systems. This alignment improves knowledge transfer and reduces discrepancies between training and execution, which is essential for maintaining operational accuracy [5]. Augmented reality concepts complement this approach by enabling contextual support directly within the physical environment, thereby enhancing real-world task performance and reducing errors [4].

Artificial intelligence contributes to the adaptability of the system by providing personalized, context-aware guidance. Through the use of Retrieval-Augmented Generation, the system delivers responses grounded in organizational knowledge, improving relevance and reliability. In addition, the integration of Open Voice OS (OVOS) enables natural voice interaction, which enhances usability and accessibility, particularly in industrial settings where hands-free operation is beneficial [2].

Combination of XR, digital twins, AI, and voice interaction creates a cohesive and scalable training ecosystem that supports efficient onboarding, improves task performance, and facilitates standardized knowledge transfer across diverse industrial environments.

5 CONCLUSION

This paper presented an integrated XR-based training system combining virtual and augmented reality, digital twins, and artificial intelligence to support workforce onboarding in complex manufacturing environments. The results demonstrate that the proposed approach significantly improves training efficiency, reduces onboarding time, and lowers error rates, while also increasing user motivation and reducing dependence on experienced mentors. The system enables standardized, interactive, and immersive learning, which enhances knowledge transfer and supports consistent execution of industrial procedures.

The integration of AI-driven assistance and voice interaction through Open Voice OS (OVOS) further improves usability and accessibility, allowing users to interact naturally with the system even in hands-free environments. Digital twins ensure that training scenarios closely reflect real-world conditions, contributing to better preparedness and smoother transition from training to practice.

Future work will focus on extending the system with advanced personalization mechanisms that adapt training content to individual user performance and learning progress. Further development will include multilingual support to address diverse workforces and improve inclusivity. Additionally, the integration of real-time industrial data and sensor feedback will be explored to enhance simulation realism and responsiveness. Finally, scalability and applicability across different industrial domains will be investigated to validate and extend the proposed approach within broader Industry 4.0 contexts.

LITERATURE

[1] JERALD, Jason. *The VR Book: Human-Centered Design for Virtual Reality*. New York: Association for Computing Machinery and Morgan & Claypool, 2015. ISBN 978-1970001129. https://doi.org/10.1145/2812042
[2] OPEN VOICE OS (OVOS). *Open source voice assistant framework*. 2023. Available at: https://openvoiceos.com [30.4.2026]
[3] MILGRAM, Paul and KISHINO, Fumio. “A Taxonomy of Mixed Reality Visual Displays.” *IEICE Transactions on Information and Systems.* 1994, 77(12), p.1321-1329.
[4] BILLINGHURST, Mark, CLARK, Adrian and LEE, Gun. A survey of augmented reality. *Foundations and Trends in Human–Computer Interaction*, 2015, 8(2–3), 73–272. https://doi.org/10.1561/1100000049
[5] KRITZINGER, W. et al. Digital Twin in manufacturing: A categorical literature review and classification, In: *Proceedings of the 16th IFAC Symposium on Information Control Problems in Manufacturing.* 2018. p.1016-1022. https://doi.org/10.1016/j.ifacol.2018.08.474

Acknowledgment
This project has received funding from the European Union’s Horizon Europe research and innovation programme within the framework of the WASABI Project funded under grant agreement No. 101092176.

ML-Based Real-Time Monitoring System for Plasma Thin-Film Deposition in Industrial Manufacturing

Žiga Gosar[1]

[1]ELVEZ, proizvodnja kabelske konfekcije in predelava plastičnih mas, d.o.o., Slovenia
E-mail: ziga.gosar@elvez.si

***Abstract**: This contribution presents the development and implementation of an intelligent monitoring and control system for plasma-enhanced thin-film deposition in an industrial vacuum reactor. The system integrates optical emission spectroscopy, interferometric coating-thickness sensor, and closed-loop generator power regulation. The solution enables real-time monitoring of plasma composition, plasma homogeneity, and deposition rate, while automatically adjusting process parameters. The integration of machine learning concepts and adaptive control algorithms ensures stable operation, reduces energy and precursor consumption, and prevents defects such as dusty plasma formation. Experimental results demonstrate that the same coating thickness can be achieved in significantly shorter processing time compared to manual parameter tuning. The proposed approach represents a step toward a digital twin of the plasma process and demonstrates practical implementation of Industry 4.0 principles in a real manufacturing environment. The system contributes to sustainable production, improved repeatability, and reduced industrial reject rates.*

***Keywords:** plasma processing, machine learning, optical spectroscopy, Industry 4.0, process control*

***Article Classification**: Scientific Paper*

1 INTRODUCTION

Plasma-enhanced chemical vapor deposition (PECVD) is widely used in industrial manufacturing for the deposition of functional thin films, particularly protective coatings on polymer and metalized substrates. The process enables the formation of SiOxCyHz-like layers with tailored mechanical, optical, and barrier properties, making it highly relevant for automotive, electronics, and consumer goods industries. However, despite its technological maturity, PECVD in large-scale industrial reactors remains a complex and often insufficiently controlled process. The main challenge arises from the strong dependence of deposition kinetics and coating properties on plasma conditions, such as gas composition, pressure, and input power, which are typically adjusted manually based on operator experience.

Previous studies have demonstrated that plasma chemistry in hexamethylsiloxane (HMDSO) based discharges is highly sensitive to process parameters, leading to variations in deposition rates, film composition, and uniformity [1–3]. In particular, non-uniform distribution of reactive species within large-volume reactors can result in spatial variations in coating thickness, directly affecting product quality [1], [5]. Furthermore, the deposition kinetics of silica-like coatings have been shown to depend on complex interactions between precursor fragmentation, plasma density, and surface reactions [5]. Experimental investigations in industrial-scale systems also revealed that achieving uniform coatings requires careful optimization of electrode configuration and discharge asymmetry [2], as well as precise control of oxygen addition to the precursor mixture [3].

Despite these advances, most industrial PECVD systems still operate without real-time diagnostic feedback, effectively functioning as “black box” processes. This limitation leads to suboptimal performance, increased material consumption, and a significant proportion of rejected products due to coating defects such as discoloration or non-uniform thickness. Recent research efforts have therefore focused on improving process monitoring through optical emission spectroscopy (OES) and other diagnostic techniques, enabling better insight into plasma behavior and its correlation with deposition characteristics [6], [7].

In this context, the integration of advanced sensing technologies and machine learning (ML) methods represents a promising approach toward intelligent plasma process control. By continuously monitoring plasma emission and deposition rate, it is possible to establish real-time correlations between measurable signals and process outcomes. The doctoral work of Gosar [8] highlighted the importance of understanding gaseous plasma characteristics in large reactors as a prerequisite for developing such control strategies.

The present work introduces an ML-based real-time monitoring system for plasma thin-film deposition in an industrial PECVD reactor. The system combines photonic sensors, optical diagnostics, and data-driven algorithms to enable automated adjustment of process parameters. The main objective is to improve process stability, enhance coating uniformity, and reduce material and energy consumption. Key variables investigated in this study include optical emission intensity at characteristic wavelengths, deposition rate evolution, and generator power. By linking

these variables through a closed-loop control framework, the proposed approach aims to transform conventional PECVD systems into intelligent, self-regulating manufacturing units aligned with Industry 4.0 principles.

2 METHOD

To address the challenges outlined in the previous section, a comprehensive experimental and computational framework was developed to enable real-time monitoring and adaptive control of the PECVD process. The methodology integrates advanced diagnostic tools, precise data acquisition, and closed-loop regulation of key process parameters. Particular emphasis is placed on the relationship between plasma emission characteristics and deposition dynamics. The following subsection describes the experimental setup and reactor configuration used to obtain the results presented.

2.1 Experimental Setup

All experiments were conducted in an industrial-scale PECVD reactor designed for high-vacuum metallization and protective coating deposition. The reactor consists of a vacuum chamber (volume ≈ 5 m³) equipped with multiple pumping systems, including roots pumps, rotary vane pumps, and diffusion pumps, enabling operation at low pressures in the range of 1–10 Pa. Plasma was generated using a medium frequency 50 kHz with alternating current (AC) and power supply with adjustable power up to 10 kW. A photo of the industrial plasma reactor is shown in Figure 1.

Figure 1: High vacuum metalization machine used for the experiment

The precursor used for coating deposition was HMDSO, delivered through a dedicated mass flow controller. Oxygen was used as an auxiliary gas for plasma stabilization and diagnostics. Samples were mounted on rotating holders (planetary system) to ensure exposure to the plasma environment during deposition.

2.2 Optical Emission Spectroscopy (OES)

Optical emission spectroscopy was employed for real-time plasma diagnostics. The emitted light from the plasma was collected using an optical fiber coupled to a spectrometer operating in the wavelength range from 200 nm to 1100 nm. The optical sensor was positioned inside the reactor to ensure sufficient signal intensity and reduce integration time.

The measured spectral intensity $I(\lambda,t)$ is defined as:

$$I(\lambda,t) = k \cdot n_i(t) \cdot A_{ij} \cdot h\,\nu \tag{1}$$

where $n_i(t)$ represents the population density of the excited species, A_{ij} is the transition probability, h is Planck's constant, and ν is the frequency of emitted radiation. The constant k accounts for system-specific factors such as optical efficiency.

For process control, characteristic emission lines were monitored, particularly the hydrogen Balmer line Hα at 656 nm and the CH emission band near 431 nm. These spectral features are directly related to plasma chemistry and precursor fragmentation dynamics [3,8].

2.3 Interferometric Measurement of Deposition Rate

The thin-film deposition rate was measured using a fiber-optic interferometric sensor based on the Fabry–Pérot principle. The interference signal is generated due to multiple reflections at the interfaces of the growing thin film.

The optical path difference Δ is given by:

$$\Delta = 2\, n\, d \tag{2}$$

where n is the refractive index of the deposited film, and d is the film thickness. The phase shift ϕ is expressed as:

$$\phi = 4\pi n d \,/\, \lambda \tag{3}$$

The deposition rate Rd is calculated from the time derivative of the thickness:

$$Rd = dd \,/\, dt \tag{4}$$

In practice, the deposition rate is obtained by linear regression of thickness over time, using multiple measurements acquired within a short interval (typically 2 s), thereby ensuring high accuracy and reduced noise.

2.4 Machine Learning-Based Control Strategy

A closed-loop control system was implemented to regulate plasma conditions in real time. The control variable is the generator power $P(t)$, which is adjusted based on the deviation between measured and target spectral intensity.

The control law can be expressed as:

$$P(t{+}1) = P(t) + K_p \cdot (I_{\text{target}} - I_{\text{measured}}) \tag{5}$$

where K_p is the proportional gain coefficient. An optional integral term can be included for improved stability:

$$P(t{+}1) = P(t) + K_p\, e(t) + K_i \sum_{i=0}^{i} e(i) \tag{6}$$

with $e(t) = I_{\text{target}} - I_{\text{measured}}$.

The system operates within predefined tolerance limits (typically ±30%). If the deviation remains within this range, no adjustment is applied. Otherwise, the control algorithm updates the power setpoint to restore optimal plasma conditions.

2.5 Data Acquisition and Processing

All process variables were acquired in real time using an integrated data acquisition framework that synchronizes optical emission spectroscopy (OES) signals with interferometric measurements of thin-film growth. As illustrated in Figure 2, the monitoring system combines photonic sensors, a spectrometer unit, and a control interface into a unified architecture, enabling continuous feedback between plasma diagnostics and process control.

The optical emission spectra were recorded across a broad wavelength range, while selected characteristic emission lines were tracked with high temporal resolution. In parallel, the deposition rate was determined from interferometric measurements based on phase evolution of reflected light at the fiber tip. Both data streams were timestamped and stored in structured datasets for further processing and analysis.

To ensure data reliability and robustness, several preprocessing steps were implemented. First, dark spectrum subtraction was applied to eliminate background noise and detector offsets. Second, signal averaging over multiple acquisitions was used to improve the signal-to-noise ratio, particularly under low-intensity plasma conditions. Third, deposition rate estimation was performed using linear regression over short time intervals, providing a stable and physically consistent measure of coating growth. In addition, threshold-based filtering was introduced to identify abnormal signal variations, enabling early detection of process instabilities.

The processed data served a dual purpose. In real time, they were used as input for the control algorithm, allowing dynamic adjustment of plasma parameters. Simultaneously, the archived datasets enabled post-process evaluation, including identification of correlations between plasma emission and deposition behavior, detection of disturbances, and optimization of operating conditions. As shown in Figure 2, this closed-loop data workflow forms the basis for intelligent, data-driven control of the PECVD process.

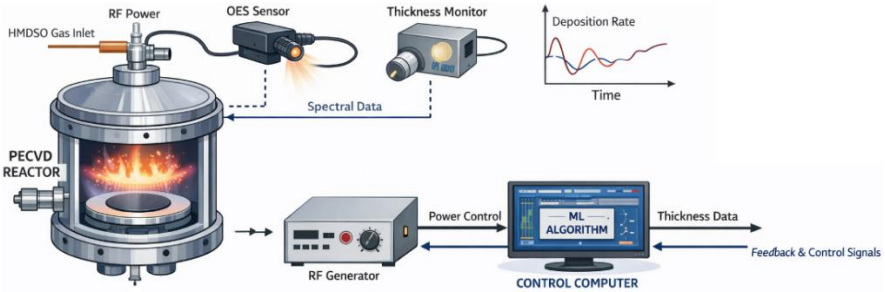


Figure 2: ML-based plasma thin-film deposition monitoring system

3 RESULTS WITH DISCUSSION

The developed ML-based monitoring system was validated under industrial operating conditions in a PECVD reactor using HMDSO precursor. The results confirm that integrating OES with interferometric deposition sensing enables reliable, real-time tracking of plasma dynamics and coating growth, as also demonstrated in previous studies of industrial PECVD diagnostics [3,8] shown in Figure 3.

The experimental observations show a strong correlation between the OES intensity at the hydrogen Balmer Hα line (656 nm) and the thin-film deposition rate (Figure 2). As the generator power increases, the emission intensity rises monotonically, indicating an increased density of excited species and enhanced dissociation of the precursor. This behavior is consistent with previously reported findings for HMDSO plasmas in industrial reactors [1,3,4]. However, at low precursor flow rates (<50 sccm), a decrease in emission intensity was observed despite increasing flow, suggesting competing processes such as dilution effects or incomplete fragmentation of the precursor [3,5].

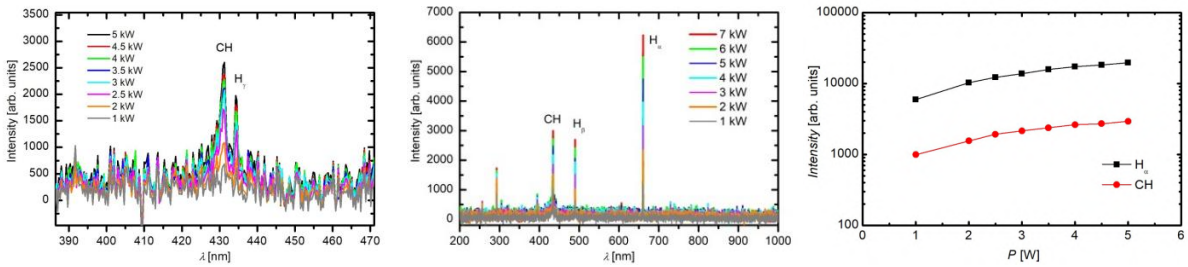


Figure 3: OES spectra of the HDMSO plasma at different powers [3,8] (left and middle) and Intensities of CH versus the generator power [3,5] (right).

The implementation of a closed-loop control algorithm enabled dynamic adjustment of generator power to maintain a stable emission intensity within predefined tolerance limits. Consequently, the deposition rate exhibited significantly reduced fluctuations compared to manual operation. The standard deviation of the deposition rate decreased by more than 30%, indicating improved process stability. These findings are aligned with the concept of plasma process stabilization through feedback control reported in contemporary PECVD studies [2,5].

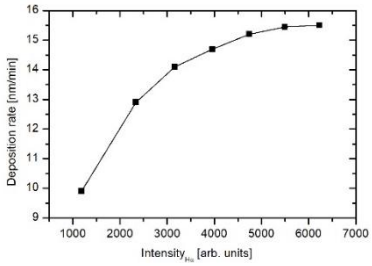


Figure 4: The deposition rate versus the intensity of the Balmer Hα line [5].

A key outcome of the system is the reduction of coating inhomogeneity across the reactor volume. Previous studies have shown that radical density and deposition rates vary spatially, often by a factor of up to three between the reactor wall and central region [1,6]. By continuously adjusting plasma parameters based on real-time optical feedback, the system partially compensates for these spatial non-uniformities, resulting in more uniform coatings.

In addition, the system successfully detected undesirable plasma regimes, such as the onset of dusty plasma formation. These regimes were characterized by abrupt changes in emission intensity and instability in deposition rate. The ML-based controller responded by reducing power, thereby preventing defect formation such as discoloration or poor adhesion. This demonstrates the system's capability for anomaly detection and preventive control [8].

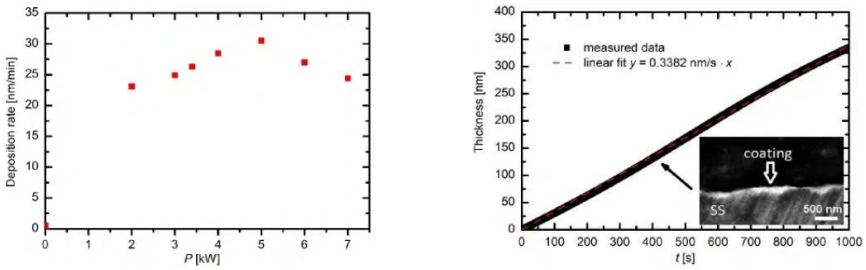


Figure 5: Deposition rates versus discharge power (left) and film thickness versus deposition time (right) [5,7].

4 CONCLUSION

Work presents a comprehensive development and industrial validation of a machine learning (ML)-supported real-time monitoring and control system for PECVD. By integrating OES with interferometric thickness sensing, the system enables continuous and reliable observation of plasma behavior and thin-film growth, overcoming key limitations of conventional operator-dependent process control.

The experimental results confirm that characteristic optical emission features, particularly the hydrogen Balmer Hα line, serve as robust indicators of plasma activity and correlate strongly with deposition rate. The implementation of a closed-loop control strategy based on these signals significantly enhances process stability, as reflected in reduced deposition rate variability and improved coating uniformity. This approach minimizes manual intervention while maintaining optimal process conditions across varying operating regimes.

A notable outcome of the study is the system's ability to identify and suppress unstable plasma conditions, including the onset of dusty plasma formation. Through real-time adaptive control, such regimes are mitigated before defects occur, leading to improved coating quality and a substantial reduction in rejection rates. Additionally, optimized control of plasma parameters contributes to more efficient utilization of precursor materials and electrical energy, supporting sustainable and resource-efficient manufacturing.

Overall, the presented system demonstrates that the integration of advanced sensing with ML-assisted control transforms PECVD reactors into intelligent, self-regulating platforms. Future work will focus on the implementation of fully predictive ML models for autonomous process optimization and broader industrial deployment.

REFERENCES

[1] GOSAR, Žiga, ĐONLAGIĆ, Denis, PEVEC, Simon, GERGIČ, Bojan, MOZETIČ, Miran, PRIMC, Gregor, VESEL, Alenka, ZAPLOTNIK, Rok. *Distribution of the deposition rates in an industrial-size PECVD reactor using HMDSO precursor*. Coatings. 2021, vol. 11, iss. 10. ISSN 2079-6412. DOI: 10.3390/coatings11101218.

[2] GOSAR, Žiga, KOVAČ, Janez, ĐONLAGIĆ, Denis, PEVEC, Simon, PRIMC, Gregor, JUNKAR, Ita, VESEL, Alenka, ZAPLOTNIK, Rok. *PECVD of hexamethyldisiloxane coatings using extremely asymmetric capacitive RF discharge*. Materials. 2020, vol. 13. ISSN 1996-1944. DOI: 10.3390/ma13092147

[3] GOSAR, Žiga, KOVAČ, Janez, MOZETIČ, Miran, PRIMC, Gregor, VESEL, Alenka, ZAPLOTNIK, Rok. *Characterization of gaseous plasma sustained in mixtures of HMDSO and O2 in an industrial-scale reactor*. Plasma chemistry and plasma processing 2020. ISSN 0272-4324. DOI: 10.1007/s11090-019-10026-5

[4] GOSAR, Žiga, KOVAČ, Janez, MOZETIČ, Miran, PRIMC, Gregor, VESEL, Alenka, ZAPLOTNIK, Rok. *Deposition of SiOxCyHz protective coatings on polymer substrates in an industrial-scale PECVD reactor*. Coatings. 2019. ISSN 2079-6412. DOI: 10.3390/coatings9040234

[5] GOSAR, Žiga, ĐONLAGIĆ, Denis, PEVEC, Simon, KOVAČ, Janez, MOZETIČ, Miran, PRIMC, Gregor, VESEL, Alenka, ZAPLOTNIK, Rok. *Deposition kinetics of thin silica-like coatings in a large plasma reactor*. Materials. 2019. ISSN 1996-1944. DOI: 10.3390/ma12193238

[6] PRIMC, Gregor, ZAPLOTNIK, Rok, ĐONLAGIĆ, Denis, GOSAR, Žiga, MOZETIČ, Miran. *Uniformity of the deposition rates in an industrial plasma reactor for PECVD of PMDSO-like thin films*. In: APPC15, 15th Asia Pacific Physics Conference, 2021.

[7] GOSAR, Žiga, KOVAČ, Janez, ĐONLAGIĆ, Denis, PEVEC, Simon, PRIMC, Gregor, VESEL, Alenka, MOZETIČ, Miran, ZAPLOTNIK, Rok. *PECVD hexamethyldisiloxane coatings using industrial-scale reactor*. *27th International Scientific Meeting on Vacuum Science and Techniques*. Ljubljana: Slovenian Society for Vacuum Technique, 2021. ISBN 978-961-94431-2-5.

[8] GOSAR, Žiga. *Characteristics of gaseous plasma in a large low-pressure reactor*. Doctoral dissertation. Ljubljana: Author, 2021. ISBN not available.

Acknowledgments
This work was supported by the European Commission within H2020 AIRISE Project (GA no. 101092312).

Simulation based production scheduling optimization in a digital twin using genetic algorithms

Vinko Longar[1]

[1] *Rudolfovo - Science and Technology Centre Novo mesto, Podbreznik 15, 8000 Novo mesto, Slovenia*
E-mail: vinko.longar@rudolfovo.eu

Abstract: Digital twin (DT) technology is increasingly applied to support production scheduling and operational performance improvement. However, optimization within commercial simulation environments is often performed using built-in genetic algorithms that operate as black-box solvers, limiting transparency, parameter control, and systematic experimental analysis. These limitations can affect the stability and reproducibility of optimization results. This paper presents the development and extension of a digital production model of a demonstration production line implemented in AnyLogic. The model was initially used to perform schedule optimization using the integrated genetic algorithm, resulting in measurable reductions in makespan, i.e., the total time required to complete all production tasks. Nevertheless, variability between optimization runs and restricted access to key algorithm parameters revealed important methodological limitations. To overcome these constraints, an external optimization environment (HeuristicLab) was integrated with the digital model, enabling explicit control of genetic algorithm parameters and structured execution of repeated experiments. This architecture separates the simulation model from the optimization engine and allows more transparent analysis of optimization behaviour. A comparative evaluation between the built-in AnyLogic optimizer and the extended external approach is presented. The results show that the proposed integration maintains competitive optimization performance while significantly improving experimental flexibility and reproducibility. The study demonstrates that controlled simulation based optimization can improve production efficiency and provide a more reliable and structured approach to scheduling in digital production model environments.

Keywords: *digital twin, genetic algorithm, scheduling optimization, metaheuristics*

Article Classification*: Professional Paper*

1 INTRODUCTION

Digital Twin (DT) technology has become an important part of modern production systems, particularly in the field of simulation-based optimization. With a virtual representation of a physical production system, DT models allow analysis of system behaviour, evaluate alternative scenarios, and optimize operational performance [1]. This capability, which can be used without interfering with real world processes, is especially valuable in production scheduling, where testing different sequencing strategies directly on the shop floor is often impractical, costly, or disruptive [2]. Simulation based optimization within a DT environment provides a safe and efficient approach for improving key performance indicators such as makespan, resource utilization, and energy consumption [3].

Due to the combinatorial complexity of production scheduling problems, metaheuristic approaches are commonly used in conjunction with simulation models. Among these, genetic algorithms (GAs) are widely applied because of their ability to explore large search spaces and provide near optimal solutions within reasonable computational time [4]. Many commercial simulation environments, such as AnyLogic [5], include built-in optimization tools that integrate such algorithms directly into the simulation workflow, enabling relatively straightforward implementation of simulation-based optimization [6].

However, despite their practical advantages, these built-in optimizers often operate as black-box solvers. While they allow users to define input parameters and objective functions, they typically provide limited access to the configuration of the optimization algorithm [7]. Key parameters such as population size, mutation rate, crossover strategy, or stopping criteria are either fixed or only partially configurable. As a result, it becomes difficult to systematically analyse the influence of algorithm settings and ensure reproducibility of experiments. This lack of configurability can lead to variations of optimization results and reduces the reliability of such tools [8].

This research builds upon a digital model of a demonstration production line developed in AnyLogic. The model was validated using real production data and applied to scheduling optimization with the built-in genetic algorithm, achieving improvements in makespan and energy consumption. However, variability between runs and limited control over algorithm parameters revealed important limitations. To address these issues, an external optimization environment HeuristicLab [9] was integrated with the digital model, enabling explicit parameter control and structured execution of repeated experiments. This approach separates the simulation model from the optimization engine and improves transparency in analysing optimization behaviour.

2 METHODS

2.1 Digital twin development and validation

A digital model of the LabTOP demonstration production line was developed in the AnyLogic simulation environment using discrete simulation library and agent modelling library. The model shown in Figure 1 represents the key elements of the physical system, including seven robotic cells, the transport system, and the operational pallet processing. Particular attention was given to accurately modelling the sequencing rules, processing times, and interactions between individual workstations to ensure realistic system behaviour.

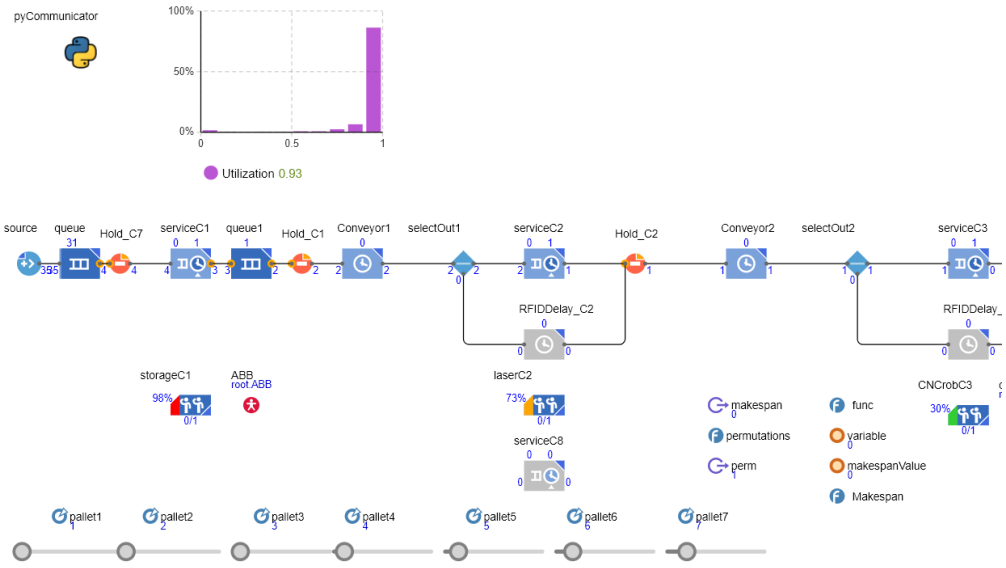


Figure 1: Digital model of production line

Input data for model parameterization were obtained from multiple sources. Historical data was collected from the enterprise resource planning (ERP) system and a structured SQL database, which included 15 work orders with different product variation. Experimental measurements were collected directly from the production line and dedicated devices such as energy and power meters (HIQ PM1-E-D-CT) with Class 1 (±1%) system accuracy [10]. The system sampled all data at a frequency of 1 Hz (one measurement per second) These data sources were used to define processing times, transport durations, and operational constraints within the model. By combining historical and measured data, the model was configured to reflect actual production conditions as closely as possible.

Model validation was performed by comparing simulated results with measured data from the real system. The accuracy was evaluated using the Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). The results showed good alignment between the model and the real system. For processing times, the average error was MAE = 4.6 s and MAPE = 4.7%, while for energy consumption the relative error was slightly higher (MAPE = 11.2%, MAE = 23.7 Wh).The validation results indicate that the digital model provides a sufficiently accurate representation of the physical system and can be reliably used as a platform for simulation-based optimization.

2.2 Production schedule optimisation with a bulit-in genetic algorithm

Production schedule optimization is represented as a simulation-based optimization problem, where the objective is to determine an efficient sequence of production orders using a digital model. The considered problem corresponds to a Flow Shop Scheduling Problem (FSSP), where all jobs pass through the machines in the same order [11]. The production scheduling problem is formulated as the task of minimizing the total production time (makespan) by determining the optimal order of processing jobs on the production line. In this approach, each feasible schedule is encoded as a chromosome in the genetic algorithm [12]. For a system with n jobs, a chromosome is represented as a permutation of n elements, defining the processing sequence. The optimization process is performed using the built-in Optimization Experiment in AnyLogic, which implements an evolutionary algorithm based on genetic principles. For each candidate solution (schedule), the simulation model evaluates the objective function. As the problem is formulated as a minimization task, the algorithm seeks schedules with the shortest production time. The experiment was configured with a maximum of 500 iterations, where each iteration represents one simulation run with a specific parameter combination. In addition to the objective function, constraints were defined to ensure feasibility of solutions. The simulation runtime was limited to a maximum value of 10000 s, and only valid permutations of schedules were considered. For optimization runs different random seeds were used allowing variability in initial conditions.

2.3 External optimization with HeuristicLab Optimizer

To overcome the limitations of the built-in optimizer, an external optimization environment HeuristicLab was integrated with the digital model. HeuristicLab is a flexible software framework for metaheuristic optimization that supports various evolutionary algorithms and enables explicit configuration of algorithm parameters.

The integration was implemented by connecting the AnyLogic simulation model with HeuristicLab through an external evaluation interface. This required importing the HeuristicLab external evaluation Java library into the AnyLogic project and creating a dedicated experiment (Parameter Variation) that serves as an evaluation function for the optimizer. During execution, the simulation model receives parameter sets from HeuristicLab, performs a simulation run, and returns the corresponding objective function value (makespan) back to the optimizer. The communication between the optimizer and the simulation model is performed via a TCP-based interface, allowing iterative exchange of candidate solutions and fitness values. In HeuristicLab, the problem was defined as an external permutation-based optimization problem, where each solution represents a production schedule. The genetic algorithm was configured with the following parameters: population size of 100, maximum number of generations set to 500, mutation probability of 5%, and a single elite solution. The algorithm employed a proportional selector, partially matched crossover, and swap mutation operator. A multi-analyzer component was used for monitoring the optimization process, while random seed initialization ensured variability between runs.

2.4 Evaluation method

The evaluation was performed using a set of controlled simulation experiments, with first-in-first-out (FIFO) order used as the baseline scheduling strategy. For the scheduling problem with 7 work oders, five optimization runs were conducted using the built-in genetic algorithm and five using the HeuristicLab based approach. Baseline schedule makespan amounted to 2266,6 s. In addition, a more complex scenario with 33 work orders was evaluated using five runs with the integrated GA and five runs with HeuristicLab. For this scenario the baseline schedule makespan amounted to 9425,5 s. The approaches were compared based on improvements in makespan relative to the baseline schedule, solution feasibility, and computational time.

3 RESULTS AND DISCUSSION

The results of the optimization experiments for seven work orders are presented in Figure 2 and Table 1. The built-in AnyLogic genetic algorithm was able to find sufficiently good solutions in most cases, achieving improvements in makespan between approximately 9.3% and 12.4% compared to the baseline schedule. However, the results also show variability between runs. Out of five experiments, four produced feasible solutions, while one resulted in an infeasible solution, indicating limitations in robustness and reliability of the built-in optimizer. In contrast, the HeuristicLab based approach consistently converged to the same solution in all five experiments, achieving a makespan of 1984.7 s, which corresponds to a 12.44% improvement over the baseline. The consistency of results suggests that the external optimization approach is more stable and reliable, and the obtained solution can be considered close to the global optimum for the given problem. Regarding computational performance, the built-in optimizer issued auto stop and shortened the total iterations automatically, while the HeuristicLab followed the preset number of iterations. Figure 3 presents the resulting makespan values for all five runs for both optimization approaches.

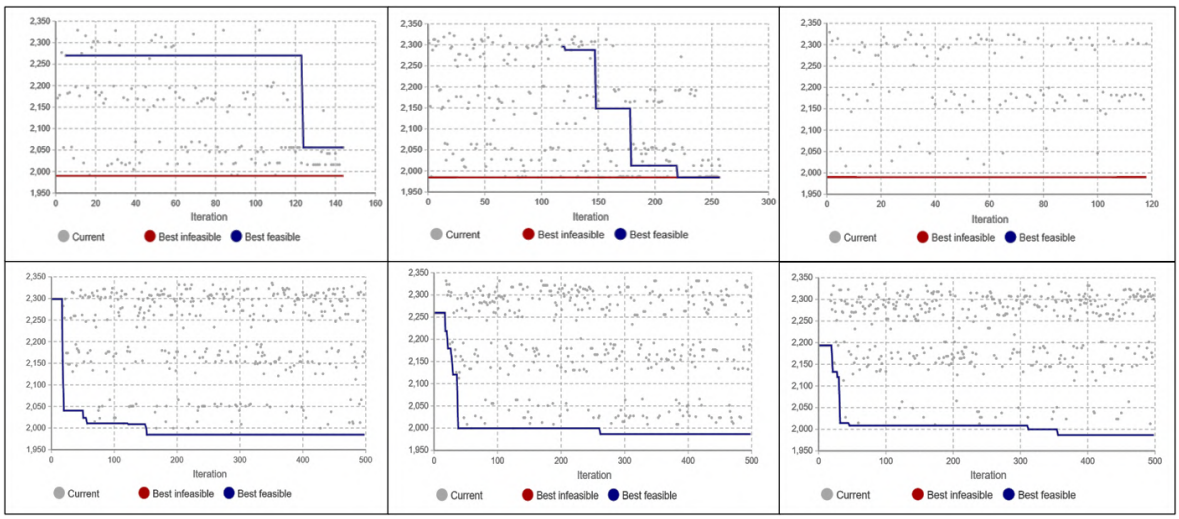


Figure 2: Results of the optimization experiments for seven work orders from built-in optimizer (upper hand) and HeuristicLab (lower hand)

Table 1: Comparison of optimization results for seven work orders

Built-in experiment no.	Best result (makespan)	Baseline schedule improvement	Total iterations	Computation time
1	2056,0 s	9,29 %	145	4,2 s
2	1984,7 s	12,44 %	258	1,6 s
3 (unfeasible)	2016,0 s	11,06 %	119	0,8 s
4	2037,7 s	10,10 %	165	1,2 s
5	2018,1 s	10,96 %	83	0,6 s
HeuristicLab experiment no.				
1	1984,7 s	12,44 %	500	12,1 s
2	1984,7 s	12,44 %	500	13,7 s
3	1984,7 s	12,44 %	500	13,8 s
4	1984,7 s	12,44 %	500	12,5 s
5	1984,7 s	12,44 %	500	12,6 s

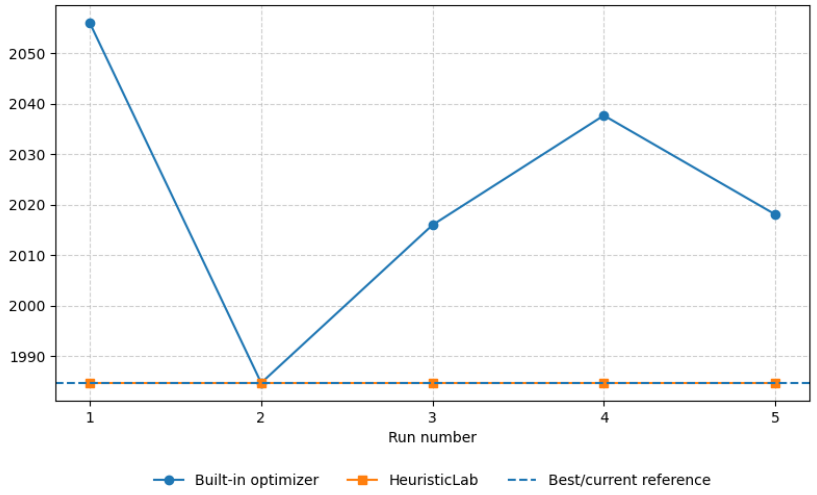


Figure 3: Resulting makespan values for all 5 runs (7 work orders)

In the more complex scenario with 33 work orders (Figure 4, Figure 5, Table 2), the limitations of the built-in AnyLogic optimizer become even more evident. In all five experiments, the built-in genetic algorithm failed to produce a feasible solution, despite completing the maximum number of iterations. The infeasibility of solutions indicates that the algorithm was unable to correctly handle the permutation constraints of the scheduling problem. This behaviour highlights a key limitation of the built-in GA when applied to more complex permutation-based optimization problems. In contrast, the HeuristicLab-based approach successfully generated feasible solutions in all runs, achieving consistent improvements in makespan between 5.7% and 6.6%.

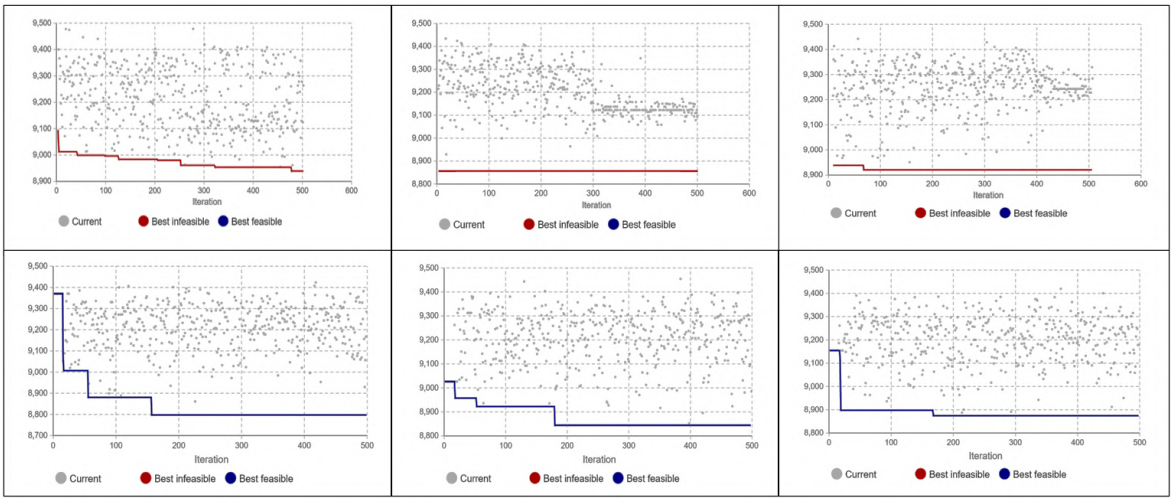


Figure 4: Results of the optimization experiments for 33 work orders from bulit-in optimizer (upper hand) and HeuristicLab (lower hand)

Table 2: Comparison of optimization results for 33 work orders

Built-in experiment no.	Best result (makespan)	Baseline schedule improvement	Total iterations	Computation time
1 (unfeasible)	8961,3 s	4,9 %	500	121,5 s
2 (unfeasible)	8921,2 s	5,4 %	500	135,2 s
3 (unfeasible)	8880,2 s	5,8 %	500	131,1 s
4 (unfeasible)	8885,4 s	5,7 %	500	119,4 s
5 (unfeasible)	8983,2 s	4,7 %	500	122,1 s
HeuristicLab experiment no.				
1	8802,3 s	6,6 %	500	108,1 s
2	8860,3 s	6,0 %	500	82,3 s
3	8871,2 s	5,9 %	500	103,9 s
4	8891,2 s	5,7 %	500	83,5 s
5	8891,3 s	5,7 %	500	84,8 s

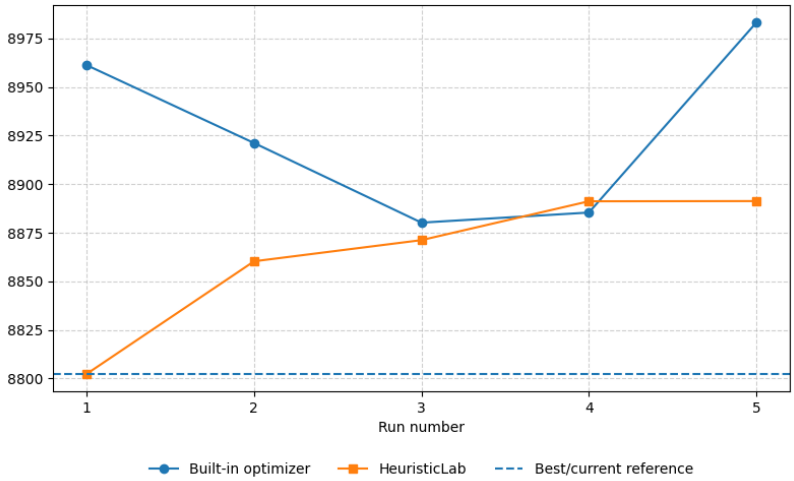


Figure 5: Resulting makespan values for all 5 runs (33 work orders)

The difference in achieved improvements between the 7 work orders and 33 work orders scenarios can be attributed to the combinatorial explosion of the solution space. While the smaller problem allows the genetic algorithm to explore a significant portion of possible solutions, the larger problem introduces a vastly increased number of permutations, making it more difficult to identify near optimal schedules within a limited number of iterations. On the other hand, the included work orders could limit the achievable improvement, as the initial schedule may already be close to an efficient solution, which leaves less room for further optimization.

4 CONCLUSION

This paper presented a simulation-based optimization approach for production scheduling using a digital model. The results confirmed that genetic algorithms can effectively improve makespan and overall system performance. However, the comparison revealed important limitations of the built-in optimizer in terms of robustness, feasibility, and reproducibility of results. The integration of the external HeuristicLab optimizer provided a more stable and controllable optimization experiments, consistently producing feasible solutions. The results show that optimization performance decreases with increasing problem complexity, as lower relative improvements were observed for larger scheduling instances. Future work will focus on multi objective optimization including energy consumption and integration with real time production data for dynamic rescheduling.

FUNDING

The research was co-funded by the Slovenian Research and Innovation Agency (ARIS) through the annual work program of Rudolfovo, and by the Republic of Slovenia, the Ministry of Higher Education, Science and Innovation, and the European Union – NextGenerationEU through the DIGITOP project.

REFERENCES

[1] UHLEMANN, Thomas H.J., Christian LEHMANN in Rolf STEINHILPER. The Digital Twin: Realizing the Cyber-Physical Production System for Industry 4.0. *Procedia CIRP* [online]. 2017, **61**, 335–340 [viewed 19. 06. 2025]. ISSN 2212-8271. Available at:10.1016/J.PROCIR.2016.11.152

[2] CHUA, Ping Chong idr. Digital twin-based decision support system for planning and scheduling. *Digital Manufacturing: Key Elements of a Digital Factory* [online]. 2024, 213–245 [viewed 20. 06. 2025]. Available at:10.1016/B978-0-443-13812-6.00001-4

[3] GRZNÁR, Patrik idr. Enhancing Production Efficiency Through Digital Twin Simulation Scheduling. *Applied Sciences 2025, Vol. 15, Page 3637* [online]. 2025, **15**(7), 3637 [viewed 20. 06. 2025]. ISSN 2076-3417. Available at:10.3390/APP15073637

[4] GEN, Mitsuo in Lin LIN. Genetic Algorithms and Their Applications. *Springer Handbooks* [online]. 2023, 635–674 [viewed 20. 06. 2025]. ISSN 2522-8706. Available at:10.1007/978-1-4471-7503-2_33

[5] *Use of Simulation – AnyLogic Simulation Software* [online]. [viewed 17. 08. 2025]. Available at: https://www.anylogic.com/use-of-simulation/

[6] *Comparing and Implementing Job Shop Scheduling Techniques with AI-powered Simulation – AnyLogic Simulation Software* [online]. [viewed 20. 06. 2025]. Available at: https://www.anylogic.com/resources/case-studies/comparing-and-implementing-job-shop-scheduling-techniques-with-ai-powered-simulation/

[7] AMARAN, Satyajith idr. Simulation optimization: a review of algorithms and applications. *Annals of Operations Research 2015 240:1* [online]. 2015, **240**(1), 351–380 [viewed 14. 04. 2026]. ISSN 1572-9338. Available at:10.1007/S10479-015-2019-X

[8] SUN, Kexin idr. Hybrid genetic algorithm with variable neighborhood search for flexible job shop scheduling problem in a machining system. *Expert Systems with Applications* [online]. 2023, **215**, 119359 [viewed 20. 06. 2025]. ISSN 0957-4174. Available at:10.1016/J.ESWA.2022.119359

[9] *GitHub - heal-research/HeuristicLab: HeuristicLab - An environment for heuristic and evolutionary optimization* [online]. [viewed 9. 09. 2025]. Available at: https://github.com/heal-research/HeuristicLab

[10] TOMCZYK, Krzysztof idr. Review of Accuracy Assessment Methods for Current Transformers: Errors, Uncertainties, and Dynamic Performance. *Energies 2025, Vol. 18,* [online]. 2025, **18**(18) [viewed 14. 04. 2026]. ISSN 1996-1073. Available at:10.3390/EN18184995

[11] AHMET ARIK, Oguzhan. Genetic Algorithm Application for Permutation Flow Shop Scheduling Problems [online]. brez datuma [viewed 20. 06. 2025]. Available at:10.35378/gujs.682388

[12] *Talbi, E.G. (2009) Metaheuristics From Design to Implementation (Vol. 74). John Wiley & Sons. - References - Scientific Research Publishing* [online]. [viewed 14. 04. 2026]. Available at: https://www.scirp.org/reference/referencespapers?referenceid=1322078

Razvoj modularne programske arhitekture za avtomatizirano doziranje tekočin v mobilnem industrijskem okolju

Development of a Modular Software Architecture for Automated Liquid Dosing in a Mobile Industrial Environment

Srđan Kosić[1], Damir Vrančić[1,2]

[1] Fakulteta za industrijski inženiring, Univerza v Novem mestu, Šegova ulica 112, 8000 Novo mesto, Slovenija
E-naslov: srdan.kosic@students.fini-unm.si, damir.vrancic@fini-unm.si

[2] Institut Jožef Stefan, Jamova cesta 39, 1000 Ljubljana, Slovenija
E-naslov: damir.vrancic@ijs.si

***Povzetek**: V prispevku je predstavljen razvoj modularne programske arhitekture za upravljanje dozirne cisterne, namenjene delovanju v mobilnem industrijskem okolju. Razvita rešitev temelji na industrijskem krmilniku in programskem okolju CODESYS 2.3 ter je zasnovana skladno s standardom IEC 61131-3. V okviru implementacije je bila razvita strukturirana programska arhitektura, ki vključuje obdelavo signalov, upravljanje procesnih funkcijskih blokov, sistemska stanja ter diagnostične in varnostne mehanizme. Razvit je bil tudi kalibracijski mehanizem, ki omogoča prilagoditev faktorja pretoka na podlagi neposrednih meritev in s tem zagotavlja prilagodljivost sistema različnim delovnim medijem. Terenska validacija je pokazala, da so v operativnem območju višjih pretokov odstopanja od ciljne dozirane mase manjša od ±1 %. Testiranja pri nizkih pretokih so izpostavila vpliv fizikalnih lastnosti medija na volumetrični izkoristek črpalke, kar dodatno potrjuje pomen ustrezne parametrizacije sistema glede na dejanski delovni medij. Prispevek posebej izpostavlja sistematičen pristop k razvoju programske opreme v mobilni avtomatizaciji ter prikazuje, kako standardizirana struktura kode, jasno definirana arhitektura in integrirana diagnostika prispevajo k večji zanesljivosti, sledljivosti in dolgoročni vzdržnosti industrijskih rešitev v kontekstu Industrije 4.0.*

***Ključne besede**: mobilna avtomatizacija, modularna programska arhitektura, IEC 61131-3, CAN-komunikacija, doziranje tekočin*

***Razvrstitev**: Strokovni prispevek*

***Abstract**: This paper presents the development of a modular software architecture for controlling a dosing tanker operating in a mobile industrial environment. The solution is based on an industrial controller and implemented in the CODESYS 2.3 development environment, in compliance with the IEC 61131-3 standard. During implementation, a structured software architecture was established, incorporating signal conditioning, process control function blocks, system state management, and integrated diagnostic and safety mechanisms. A calibration mechanism was also developed to enable adjustment of the flow factor based on direct measurements, ensuring adaptability to different working media. Field validation demonstrated that, within the operating range at higher flow rates, deviations from the target dosing mass were below ±1%, confirming the required accuracy and repeatability of the system. Tests at low flow rates highlighted the influence of the physical properties of the medium on the volumetric efficiency of the gear pump, further emphasising the importance of proper system parameterisation with respect to the actual working medium. The paper presents a systematic approach to software development in mobile automation and demonstrates how a standardised code structure, clearly defined architecture, and integrated diagnostics contribute to improved reliability, traceability, and long-term sustainability of industrial solutions in the context of Industry 4.0.*

***Keywords:** mobile automation, modular software architecture, IEC 61131-3, CAN communication, liquid dosing*

***Article Classification**: Professional Paper*

1 UVOD

Avtomatizacija mobilnih delovnih strojev predstavlja poseben izziv v primerjavi s stacionarnimi industrijskimi sistemi, saj morata programska in strojna oprema delovati v bistveno zahtevnejših okoljskih pogojih. Prisotne so mehanske vibracije, temperaturna nihanja, padci napajalne napetosti ter elektromagnetne motnje, kar neposredno vpliva na zanesljivost meritev in delovanje krmilnih sistemov. Zato je treba razviti robustne in prilagodljive rešitve, ki zagotavljajo stabilno delovanje tudi v nestacionarnih pogojih.

Poseben izziv predstavlja avtomatizirano doziranje tekočin, kjer je zahtevana visoka natančnost kljub spremenljivim obratovalnim pogojem. Odstopanja lahko nastanejo zaradi vpliva lastnosti medija, kot sta viskoznost in temperatura, nestabilnega pretoka ali mehanskih vplivov na merilno opremo. Zato je treba zagotoviti ustrezno sinhronizacijo meritev, vodenje sistema in izvajanje drugih procesnih funkcij [1].

V literaturi so obravnavane različne rešitve za doziranje tekočin v industrijskih procesih [1, 2], ki temeljijo na uporabi volumetričnih ali masnih merilnikov ter različnih regulacijskih pristopov. Te rešitve so večinoma razvite za stacionarne sisteme, kjer so pogoji delovanja bistveno stabilnejši. Sorodne zahteve se pojavljajo tudi v mobilnih komunalnih vozilih s tekočinskimi sistemi, kjer robustni krmilniki, prikazovalniki HMI in decentralizirani CAN-povezani I/O moduli omogočajo spremljanje procesnih veličin ter zanesljivo delovanje v zahtevnih terenskih pogojih [3]. Zato je pri mobilnih aplikacijah treba prilagoditi tako arhitekturo sistema kot tudi programsko implementacijo.

V okviru raziskave smo razvili in implementirali modularno programsko arhitekturo za avtomatizirano doziranje tekočin v mobilnem industrijskem okolju. Rešitev je bila razvita v okviru industrijskega projekta za mobilno dozirno cisterno, pri čemer je bil poudarek na povezavi med realnimi funkcionalnimi zahtevami, izbiri robustne strojne opreme, zasnovi pregledne programske strukture ter validaciji sistema na dejanski aplikaciji v realnem obratovalnem okolju. Z vidika Industrije 4.0 takšna zasnova omogoča boljšo sledljivost delovanja, lažjo diagnostiko in enostavnejšo nadgradnjo sistema.

2 METODE

Razvoj programske rešitve smo izvedli v razvojnem okolju CODESYS 2.3 na mobilnem industrijskem krmilniku IFM BasicController CR0403, uporabniški vmesnik pa smo zasnovali na prikazovalniku IFM BasicDisplay CR0451. Krmilnik in prikazovalnik sta bila povezana prek vodila CAN, kar je omogočilo izmenjavo procesnih podatkov, ukazov, nastavitev in diagnostičnih informacij. Programsko rešitev smo razvili skladno s standardom IEC 61131-3 [4] kot ciklično aplikacijo, v kateri se v vsakem ciklu izvajajo zajem vhodnih signalov, obdelava procesnih veličin, upravljanje stanj sistema ter generiranje izhodnih ukazov.

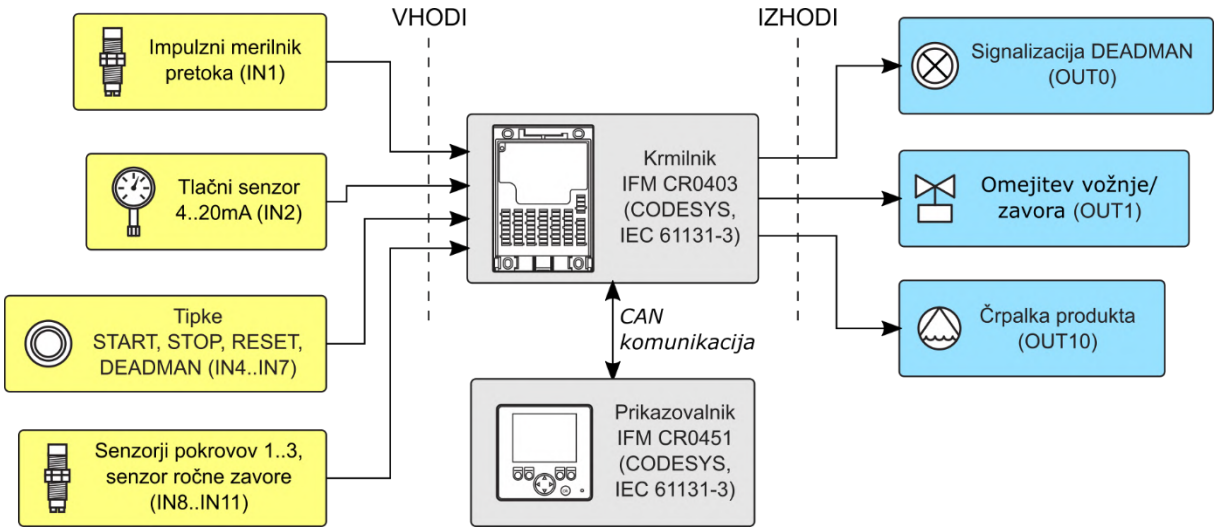
Razvojni postopek je temeljil na analizi funkcionalnih in varnostnih zahtev sistema za avtomatizirano doziranje emulzije v mobilnem okolju. Na tej osnovi smo določili konfiguracijo vhodno-izhodnih signalov, komunikacijsko arhitekturo med krmilnikom in HMI ter modularno razdelitev programske logike. Programsko strukturo smo zasnovali z ločitvijo signalne obdelave, procesne logike, varnostnih funkcij, komunikacije in vizualizacije, kar je omogočilo preglednejšo implementacijo, lažje testiranje in ponovno uporabo posameznih funkcijskih blokov. Takšen pristop je zajel celoten inženirski razvojni proces, od analize zahtev, izbire strojne platforme in konfiguracije vhodov ter izhodov do programiranja, zasnove HMI, komunikacije CAN ter testiranja in validacije sistema.

Vhodni del sistema je vključeval impulzni merilnik pretoka, analogni tlačni senzor (v območju 4…20 mA), tipke za upravljanje procesa, senzorje pokrovov, signal parkirne zavore, preverjanje prisotnosti operaterja ter del uporabniškega vmesnika. Merilni del sistema temelji na osnovnem principu, ki vključuje senzor, signalno obdelavo ter prikaz oziroma prenos izmerjene veličine [2]. Centralni del sistema je predstavljal že omenjeni industrijski krmilnik, ki je na osnovi vhodnih signalov reguliral zobniško dozirno črpalko, pnevmatski ventil za blokado vožnje ter upravljal opozorilno signalizacijo in zaslon uporabniškega vmesnika. Posebno pozornost smo namenili implementaciji varnostnih funkcij, med katere sodijo preverjanje stanja pokrovov, parkirne zavore, dovoljenega območja tlaka in prisotnosti operaterja.

Validacijo rešitve smo izvedli v več korakih. Najprej smo preverili pravilnost programske logike v simulacijskem okolju CODESYS, nato smo izvedli preizkus na strojni opremi s fizičnimi signali, v zaključni fazi pa še validacijo na mobilni dozirni cisterni v realnem obratovalnem okolju. Natančnost doziranja smo določili z meritvijo dejanske dozirane mase in primerjavo s ciljno vrednostjo v različnih režimih obratovanja.

2.1 Arhitektura sistema

Sistem za avtomatizirano doziranje emulzije je zasnovan kot mobilna enota, ki združuje mehanske, hidravlične, pnevmatske in elektronske podsisteme v enotno krmilno arhitekturo. Osrednji element sistema je mobilni krmilnik IFM BasicController CR0403 [5], ki izvaja zajem in obdelavo signalov senzorjev, procesno logiko, varnostne funkcije ter krmiljenje izvršnih elementov (slika 1). Za uporabniški vmesnik je uporabljen prikazovalnik IFM BasicDisplay CR0451 [6], nameščen v vrata krmilne omarice nad upravljalnimi tipkami Start, Stop, Reset in Deadman.



Slika 1: Blokovni prikaz arhitekture sistema za avtomatizirano doziranje tekočin

Krmilnik CR0403 (slika 2) omogoča obdelavo analognih, digitalnih in frekvenčnih signalov ter podpira 12 vhodov, 12 izhodov in dve komunikacijski liniji CAN [5]. Napajanje krmilnika je prilagojeno mobilnemu okolju v območju 8…32 V DC, robustna izvedba z zaščito IP67 in vgrajenimi diagnostičnimi funkcijami [5] pa omogoča uporabo v pogojih, kjer so prisotne vibracije, temperaturna nihanja in elektromagnetne motnje. Zaradi teh lastnosti je krmilnik primerna osrednja procesna enota za vodenje mobilne dozirne cisterne.

Prikazovalnik CR0451 (slika 2) je uporabljen kot glavna enota HMI za prikaz procesnih veličin, vnos parametrov ter prikaz alarmnih in diagnostičnih informacij. Naprava vključuje 2,8-palčni TFT zaslon ločljivosti 320 × 240, osvetljene funkcijske tipke, povezavo CAN in integracijo s CODESYS [6]. Tako je omogočen neposreden prikaz stanja sistema ter upravljanje ključnih funkcij procesa preko enotnega uporabniškega vmesnika.

Slika 2: Krmilni sistem IFM CR0403 [5] in prikazovalnik IFM CR0451 [6]

Med krmilnikom in prikazovalnikom poteka komunikacija prek vodila CAN. V sistem so vključeni impulzni merilnik pretoka za določanje dozirane količine, analogni tlačni senzor za nadzor delovnega tlaka, digitalni vhodi za tipke in senzorje pokrovov ter signal parkirne zavore. Krmilnik prek digitalnih izhodov vklaplja dozirne črpalke, pnevmatski ventil za blokado vožnje in opozorilno signalizacijo. Takšna arhitektura omogoča jasno ločitev med signalnim, procesnim, varnostnim in komunikacijskim nivojem sistema ter predstavlja osnovo za modularno zasnovo programske rešitve.

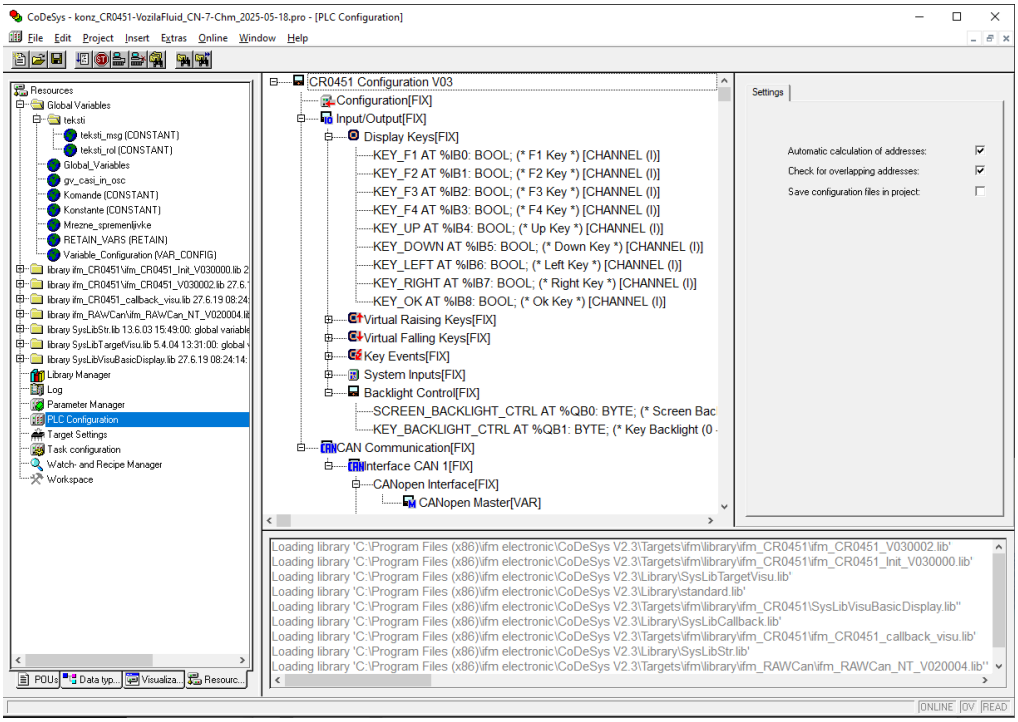
2.2 Arhitektura programske rešitve

Na podlagi predstavljene arhitekture sistema smo razvili modularno programsko rešitev, implementirano na industrijskem krmilniku v razvojnem okolju CODESYS 2.3 v skladu s standardom IEC 61131-3 [4]. Programska zasnova temelji na razdelitvi funkcionalnosti v jasno opredeljene programske sklope, kot je prikazano na sliki 3.

Osrednji del aplikacije predstavlja glavna programska enota, ki izvaja ciklično obdelavo sistema in koordinira delovanje posameznih sklopov. V vsakem ciklu se zaporedno izvede obdelava signalov, procesna logika ter generiranje izhodnih signalov.

Za lažje obvladovanje programske opreme smo jo ločili na naslednje funkcionalne nivoje:

- signalni nivo,
- procesni nivo,
- nivo upravljanja stanj,
- varnostni nivo,
- komunikacijski in vizualizacijski nivo.



Slika 3: Implementacija modularne programske arhitekture v okolju CODESYS 2.3

Signalni nivo vključuje pretvorbo vhodnih signalov v fizikalne veličine, filtriranje meritev ter preverjanje mejnih vrednosti. Obdelani signali so vhod v procesni nivo.

Procesni nivo vsebuje funkcijske bloke za vodenje pretoka, krmiljenje ventilov in nadzor delovanja črpalke. Funkcijski bloki izvajajo krmilno logiko posameznih elementov sistema.

Nivo upravljanja stanj določa zaporedje delovanja sistema in koordinira prehode med posameznimi fazami. Prehodi med stanji so vezani na merjene veličine, status procesnih blokov in časovne pogoje.

Varnostni nivo je integriran v upravljanje stanj in nadzira ključne pogoje delovanja sistema.

Komunikacijski in vizualizacijski nivo omogoča izmenjavo podatkov preko CAN vodila ter prikaz procesnih veličin, nastavitev parametrov in diagnostičnih informacij na uporabniškem vmesniku.

2.3 Obdelava signalov in procesni funkcijski bloki

Obdelava signalov temelji na impulznem zajemu meritev pretoka, ki jih pretvorimo v volumski pretok in nato v maso doziranega medija. Vhodni signal iz merilnika pretoka najprej obdelamo na signalnem nivoju, kjer izvedemo filtriranje, skaliranje in osnovno validacijo meritev.

Za pretvorbo med fizikalnimi veličinami smo uporabili namensko razvite funkcije za izračun trenutnega pretoka (na osnovi števila impulzov v časovni enoti), izračun dozirane mase ter določitev ciljnega števila impulzov glede na zahtevano količino tekočine.

Izračun mase temelji na integraciji pretoka skozi čas, pri čemer smo upoštevali kalibracijski faktor sistema. Tako smo zagotovili sprotno spremljanje dejansko dozirane količine in primerjavo s ciljno vrednostjo.

Procesne funkcijske bloke smo zasnovali kot ponovno uporabne enote z enotno strukturo vhodov, izhodov in notranjih stanj. Posamezni bloki pokrivajo funkcionalnosti krmiljenja pretoka, upravljanja ventilov ter nadzora delovanja črpalke. Med seboj so povezani prek standardiziranih signalov.

Tabela 1: Struktura uporabljenih programskih organizacijskih enot

POU / element	Vrsta	Namen in vloga v projektu
PLC_PRG	Program	Glavni cikel, izvaja inicializacijo, obdelavo vhodov/izhodov, klice procesne logike in komunikacijo.
CasiInOsc	Program	Generiranje časovnih baz in impulzov za časovnike in periodične dogodke.
ObdelavaAI	Akcija / Program	Konfiguracija in linearizacija analognih vhodov, pretvorba iz toka (4…20 mA) v tlak.
ObdelajAiVhod	Funkcija	Izračun odstotka in linearizirane vrednosti posameznega analognega vhoda.
ObdelavaDI	Akcija / Program	Branje digitalnih vhodov, odprava motenj, tvorjenje robov tipk, logika pokrovov in parkirne zavore.
ObdelavaDO	Akcija / Program	Preslikava logičnih spremenljivk v fizične digitalne izhode (črpalka, pnevmatika, lučka).
DelovanjeKomplet	Program	Glavni procesni nadzor: preverjanje varnostnih pogojev, zakasnitve za tlak, priprava na doziranje.
DelovanjeNadgradnje	Program	Dozirna logika: instanci FB_Crpalka in FB_Kalibracija, izračun mas, pretokov in preostalih količin.
FB_Crpalka	Funkcijski blok	Jedro doziranja: krmiljenje črpalke, prehodi med stanji, “*Deadman*” logika in zaustavitve.
FB_Kalibracija	Funkcijski blok	Kalibracija sistema: pretvorba impulzov v maso in nastavitev faktorja pretvorbe (“*ScaleFactor*”).
SprejemInOddajaKonzola	Program	CAN komunikacija: pakiranje statusov, obdelava in pošiljanje telegramov z/na enote/enoto HMI.
Stevci	Program	Beleženje ciklov, dnevnih in kumulativnih količin, priprava podatkov za prikaz in komunikacijo.
Pomožne funkcije	Funkcije	Matematični in podporni izračuni.

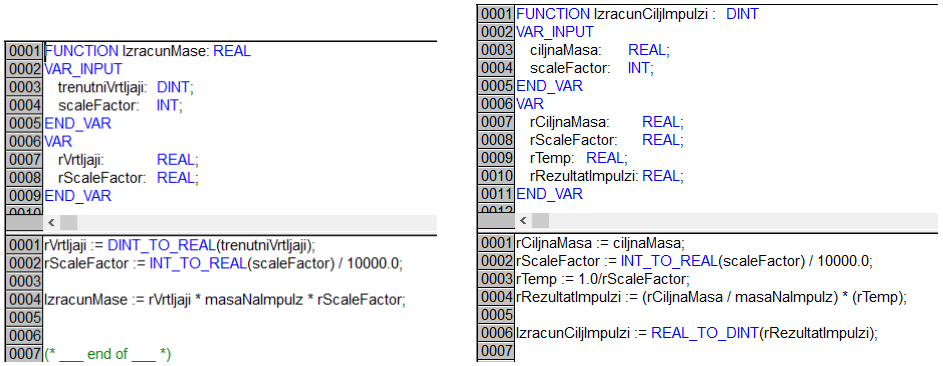
Krmiljenje procesa temelji na cikličnem izvajanju programa krmilnika, kjer se v vsakem ciklu zaporedno izvede zajem in obdelava signalov, izračun procesnih veličin, določitev stanja sistema in generiranje izhodnih signalov za izvršne elemente.

Takšna zasnova zagotavlja deterministično izvajanje sistema ter sledljivost med vhodnimi meritvami in izhodnimi ukazi. Modularna struktura funkcijskih blokov, kot je prikazano v tabeli 1, omogoča ponovno uporabo programske kode in enostavno integracijo dodatnih funkcionalnosti brez posegov v osnovno logiko sistema.

2.4 Upravljanje stanj sistema in varnostna logika

Upravljanje stanj sistema smo zasnovali kot deterministično zaporedje faz delovanja, ki določa potek procesa doziranja od inicializacije do zaključka cikla. Logika je implementirana kot struktura diskretnih stanj, med katerimi sistem prehaja glede na vhodne pogoje, procesne veličine in operaterske ukaze.

Posamezna stanja določajo aktivnost izvršnih elementov, kot so črpalka in ventili, ter način obdelave procesnih veličin. Prehode med stanji smo vezali na izpolnjevanje jasno definiranih pogojev, s čimer smo zagotovili ponovljivo in predvidljivo delovanje sistema.



Slika 4: Primer implementacije funkcij za pretvorbo impulzov merilnika pretoka v maso doziranega medija

Logika upravljanja vključuje faze priprave sistema, aktivnega doziranja in zaključka procesa. V fazi doziranja smo sproti spremljali doseženo maso in jo primerjali s ciljno vrednostjo (glej sliko 4), na podlagi česar smo določili prehod v zaključek cikla.

Varnostno logiko smo integrirali v nivo upravljanja stanj, kjer deluje kot nadrejeni sloj, ki lahko kadarkoli prekine izvajanje procesa. Implementirali smo mehanizem zaznave neaktivnosti operaterja ("Deadman"), ki ob odsotnosti potrditvenega signala najprej sproži opozorilo, nato pa izklopi črpalko in onemogoči nadaljnje doziranje.

Ponovni zagon sistema po varnostnem dogodku je mogoč šele po izvedbi definirane sekvence ukazov, s čimer smo preprečili nenamerno ali nekontrolirano nadaljevanje procesa. V logiko smo vključili tudi osnovne pogoje za varno delovanje, kot so preverjanje vhodnih signalov in stanja sistema pred začetkom doziranja.

Takšna zasnova omogoča ločevanje procesne in varnostne logike ter zagotavlja nadzorovano zaustavitev sistema in preprečuje nekontrolirano nadaljevanje procesa v primeru napak ali nepredvidenih dogodkov.

2.5 Komunikacija preko CAN vodila

Komunikacijo med krmilnikom IFM CR0403 in prikazovalnikom IFM CR0451 smo izvedli prek vodila CAN1, ki je integrirano v obe enoti. Za obravnavani sistem smo uporabili komunikacijo z lastno strukturo telegramov na nivoju Raw CAN (Layer 2), saj smo morali prenašati specifične procesne veličine, ukaze in diagnostične informacije, ki niso bile neposredno vezane na standardne profile, kot sta J1939 ali CANopen. Hitrost prenosa smo nastavili na 250 kbps.

V obeh projektih, tako na strani CPU kot na strani HMI, smo v konfiguraciji aktivirali ustrezni kanal in inicializirali komunikacijski gonilnik. Na strani krmilnika smo za pošiljanje podatkov definirali oddajni identifikator write_ID, za sprejem pa tabelo identifikatorjev read_ID[0..14]. Posamezen telegram je dolg 8 bajtov, podatke pa smo zapakirali s pomožnimi funkcijami za pretvorbo in združevanje signalov. Na strani HMI smo te iste telegrame razvezali v globalne procesne spremenljivke, kar je omogočilo konsistentno in deterministično izmenjavo podatkov med obema enotama.

Po vodilu CAN smo prenašali trenutni tlak, dejansko dozirano maso, ciljno maso, pretok, stanja pokrovov, stanje parkirne zavore, status funkcije “Deadman”, alarmna stanja ter ukaze operaterja, kot so Start, Stop, Reset in potrditev prisotnosti. Takšna razdelitev komunikacije po logičnih sklopih je omogočila jasno preslikavo med internimi spremenljivkami krmilnika ter prikaznimi in upravljalnimi funkcijami na HMI. Struktura telegramov med CPU in HMI je prikazana v Tabeli 2.

Tabela 2: Struktura telegramov CAN med CPU in HMI

CAN ID	Smer	Vsebina	Tip podatka	Namen uporabe
write_ID	CPU → HMI	tlak, masaDejanska, masaCilj	INT, DINT, INT	Prenos trenutnega tlaka, dozirane mase in ciljne mase na HMI.
read_ID[0]	HMI → CPU	cmdStart, cmdStop, cmdReset, cmdDeadmanConfirm	BOOL	Prenos ukazov s HMI (zagon, zaustavitev, reset, “Deadman” potrditev).
read_ID[1]	CPU → HMI	pokroviZaprti, parkirnaZavora, deadmanStatus	BOOL	Statusni signali za pogoje medsebojnih blokad (*Interlock*).
read_ID[2]	CPU → HMI	alarmLowPress, alarmHighPress, alarmDeadman	BOOL	Prenos alarmnih stanj za prikaz sporočil na zaslonu.
read_ID[3]	CPU → HMI	pretok, dayTotal, maintTotal	INT, DINT, DINT	Procesne veličine za zaslone *Totals* in *Counters*.
read_ID[4]	HMI → CPU	scaleFactor, nastavitveParam (več polj)	INT, INT/REAL	Prenos kalibracijskih in servisnih parametrov iz HMI v PLK.
read_ID[5]	CPU → HMI	counters_cikli, counters_alarmLow, counters_alarmHigh	INT, INT, INT	Prenos podatkov o številu ciklov in alarmnih dogodkov za zaslon *Counters*.

Komunikacijski sloj smo ločili od procesne logike. Tako procesni funkcijski bloki delujejo z že pripravljenimi in interpretiranimi podatki, medtem ko so obdelava okvirjev, razčlenjevanje bajtov in priprava telegramov omejeni na

komunikacijski sklop. Takšna zasnova je izboljšala preglednost programske rešitve, poenostavila diagnostiko in omogočila kasnejše prilagoditve komunikacijskega vmesnika brez posegov v osnovno logiko doziranja.

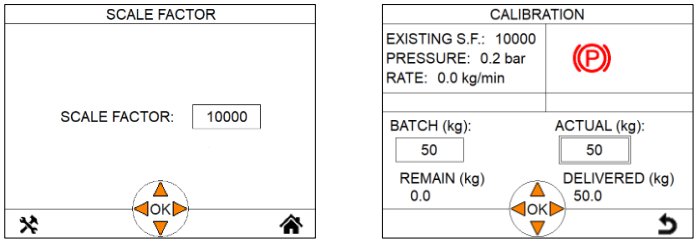
Pri preizkusih delovanja je bila komunikacija popolnoma stabilna. Na namiznem testu je CAN povezava med krmilnikom in HMI potekala s stabilnim osveževanjem vseh spremenljivk, ob prvem zagonu pa smo preverili tudi diagnostiko komunikacije. V opazovanem območju delovanja ni bilo zaznanega stanja bus-off, izgubljenih okvirjev ali prekomerne obremenitve vodila.

Komunikacijski cikel smo uskladili s cikličnim izvajanjem programa krmilnika, kar omogoča sprotno posodabljanje podatkov brez dodatnih zakasnitev. Posebno pozornost smo namenili prenosu ključnih veličin, kot so trenutni pretok, dozirana masa, ciljna vrednost in stanje sistema.

2.6 Kalibracija sistema

Postopek kalibracije smo izvedli s kontroliranim doziranjem znane količine medija, pri čemer smo zabeležili število impulzov merilnika pretoka in dejansko izmerjeno maso. Na tej podlagi smo izračunali nov kalibracijski faktor, ki smo ga nato uporabili pri nadaljnjem delovanju sistema.

Kalibracijski mehanizem smo integrirali v programsko rešitev, kar omogoča ponovno izvedbo kalibracije brez posegov v programsko kodo. Parametri kalibracije so dostopni prek uporabniškega vmesnika, kar omogoča prilagoditev sistema različnim medijem in obratovalnim pogojem, kot so spremenjene lastnosti medija, črpalke ali senzorja pretoka (glej sliko 5).



Slika 5: Uporabniški vmesnik za nastavitev kalibracijskega faktorja in parametrov doziranja

Rezultati terenske validacije so pokazali, da so bila pri višjih pretokih odstopanja od ciljne dozirane mase manjša od ±1 %, s čimer smo potrdili zahtevano natančnost in ponovljivost sistema. Pri nizkih pretokih smo zaznali nekoliko večja odstopanja, kar je povezano z vplivom fizikalnih lastnosti medija na volumetrični izkoristek črpalke.

Rezultati potrjujejo potrebo po ustrezni kalibraciji glede na dejanski delovni medij ter vpliv obratovalnih pogojev na natančnost doziranja.

3 REZULTATI IN RAZPRAVA O REZULTATIH

Razvito programsko rešitev smo validirali v realnem delovnem okolju na mobilni dozirni cisterni, razviti v okviru industrijskega projekta za avtomatizirano doziranje emulzije. Validacija je obsegala simulacijsko preverjanje programske logike, namizni preizkus na strojni opremi ter terensko testiranje na dejanski aplikaciji (Slika 6). V zaključni fazi smo preverjali natančnost doziranja, stabilnost delovanja sistema, odziv varnostne logike in delovanje komunikacije med krmilnikom in uporabniškim vmesnikom.

Slika 6: Validacija sistema na mobilni dozirni cisterni v realnem obratovalnem okolju

Pri namiznem testu smo sistem priključili na simulirane signale fizičnih vhodov. Program je pravilno izračunal dozirano količino, črpalka se je ob doseženem cilju samodejno ustavila, pnevmatska blokada vožnje pa se je ob odprtju poljubnega pokrova sprožila v manj kot 100 ms. Potrdili smo tudi, da ponovni zagon po zaustavitvi zaradi funkcije “Deadman” ni mogoč brez izvedbe predpisane sekvence, torej držanja tipke Stop najmanj tri sekunde in nato ponovnega pritiska tipke Start. Ti rezultati so potrdili pravilnost razvite varnostne logike in ustrezno povezavo med procesnim in varnostnim nivojem programa.

Najpomembnejše rezultate pa smo pridobili pri terenskem testiranju z vodo, kot je prikazano v tabeli 3. Pri ciljni masi 50 kg so bila pri nizkih pretokih izmerjena odstopanja med 6 % in 14 %, saj je bilo dejansko doziranih med 43 kg in 47 kg. Ko smo pretok povečali, so rezultati postali bistveno natančnejši, saj je bilo pri cilju 50 kg v vseh primerih doseženo odstopanje manjše od ±1 %.

Tabela 3: Primerjava med nastavljeno in dejansko količino pri dejanskem testiranju z medijem

Zap. št.	Ciljna masa [kg]	Dejanska masa [kg] (nizek pretok)	Odstopanje [%] (nizek pretok)	Dejanska masa [kg] (višji pretok)	Odstopanje [%] (višji pretok)
1	50	43	14	50	<1
2	50	47	6	50	<1
3	50	45	10	50	<1
4	50	43	14	50	<1
5	50	46	8	50	<1

Pri nizkih pretokih smo zaznali večja odstopanja, kar povezujemo z vplivom fizikalnih lastnosti medija na volumetrični izkoristek črpalke. Spremembe viskoznosti in drugih lastnosti medija lahko vplivajo na dejanski pretok, kar se odraža v razliki med izračunano in dejansko dozirano maso. To potrjuje potrebo po kalibraciji sistema glede na konkretne obratovalne pogoje.

Terensko testiranje je tudi pokazalo, da je treba spremeniti algoritem za izračun kalibracijskega faktorja. Prvotna logika je akumulirala prejšnje popravke, kar je povzročalo kumulativno napako. Po spremembi algoritma smo korekcijski faktor “ScaleFactor” vsakič ponovno izračunali iz osnovnih meritev, s čimer smo zagotovili stabilnejšo in ponovljivo kalibracijo. Ta popravek je pomembno prispeval k zanesljivejšemu delovanju sistema v nadaljnjih meritvah.
Komunikacija prek CAN vodila je bila zanesljiva in primerna za uporabo v mobilnem okolju, saj je omogočala stabilno izmenjavo podatkov med krmilnikom in uporabniškim vmesnikom. Operater je imel tekoč vpogled v dejansko stanje sistema.

Rezultati potrjujejo, da uporaba modularne programske arhitekture, strukturiranih funkcijskih blokov ter jasno ločenih komunikacijskih in varnostnih mehanizmov omogoča stabilno procesno logiko tudi v zahtevnem mobilnem okolju. Pri višjih pretokih smo potrdili zahtevano natančnost in ponovljivost delovanja, pri nizkih pretokih pa smo identificirali dejanski vpliv delovnega medija in ga vključili v nadaljnjo optimizacijo rešitve. To pomeni, da razvita arhitektura ni le funkcionalna, temveč tudi dovolj pregledna in prilagodljiva, da omogoča učinkovito diagnostiko in izboljšave v realni industrijski uporabi.

4 ZAKLJUČEK

V prispevku smo predstavili razvoj in implementacijo modularne programske arhitekture za avtomatizirano doziranje tekočin v mobilnem industrijskem okolju. Rešitev je bila razvita za mobilno dozirno cisterno, pri čemer smo združili zahteve glede natančnosti doziranja, robustnosti delovanja, varnosti in preglednosti programske strukture. Prispevek tako predstavlja primer celovitega inženirskega pristopa k razvoju sistema mobilne avtomatizacije.

Jedro rešitve je modularno zasnovana programska oprema, implementirana v okolju CODESYS 2.3 na mobilnem krmilniku IFM CR0403, ki je povezan z uporabniškim vmesnikom IFM CR0451 preko vodila CAN. Takšna zasnova je omogočila jasno ločitev med signalno obdelavo, procesno logiko, upravljanjem stanj, varnostnimi funkcijami, komunikacijo in vizualizacijo. To se je izkazalo kot pomembna prednost pri razvoju, testiranju, diagnostiki in nadaljnjih prilagoditvah sistema.

Poseben prispevek rešitve je implementacija kalibracijskega mehanizma, ki omogoča prilagoditev korekcijskega faktorja na podlagi neposrednih meritev. S tem smo zagotovili, da je mogoče sistem prilagoditi dejanskemu delovnemu mediju in obratovalnim pogojem ter doseči zahtevano natančnost v operativnem območju. Rezultati validacije so pokazali, da sistem pri višjih pretokih dosega odstopanja od ciljne dozirane mase manjša od ±1 %, medtem ko so bila pri nizkih pretokih zaznana večja odstopanja zaradi vpliva nizke viskoznosti testnega medija na volumetrični izkoristek zobniške črpalke.

Pomembna ugotovitev prispevka je, da modularna in dokumentirana programska arhitektura ne omogoča le stabilnega delovanja sistema, temveč tudi dobro prenosljivost rešitve na sorodne primere uporabe v mobilni avtomatizaciji. Ponovno uporabni funkcijski bloki, enotna organizacija kode ter usklajena povezava med programiranjem, vizualizacijo in diagnostiko predstavljajo dobro osnovo za nadaljnji razvoj podobnih rešitev na področju komunalnih vozil, kmetijskih in gradbenih strojev. Z vidika Industrije 4.0 takšna zasnova omogoča boljšo sledljivost delovanja, lažjo diagnostiko ter enostavnejše nadgradnje sistema.

Predstavljena arhitektura omogoča dobro sledljivost delovanja sistema, enostavno diagnostiko ter nadaljnji razvoj brez posegov v osnovno strukturo programa. Takšen pristop je primeren za širšo uporabo v mobilnih aplikacijah, kjer so zahtevane robustne, prilagodljive in zanesljive rešitve avtomatizacije. Takšne rešitve omogočajo optimizacijo porabe surovin in zmanjšanje odpadkov, kar prispeva k učinkovitejši rabi virov in podpira cilje trajnostnega razvoja.

LITERATURA

[1] ŠKRJANC, Igor. Industrijski krmilni in regulacijski sistemi [na spletu]. Ljubljana: Univerza v Ljubljani – Fakulteta za elektrotehniko, 2012 [pridobljeno 5. 09. 2025]. Dostopno na: http://msc.fe.uni-lj.si/Download//Skrjanc/IKRS-2012.pdf
[2] BOLTON, William. Instrumentation and Control Systems. 3. izd. Elsevier, 2021. ISBN 9780128234716.
[3] IFM ELECTRONIC GMBH. Application Report: Bucher Municipal – ecomatmobile on a sewer cleaning vehicle [na spletu]. 2020 [pridobljeno 4. 09. 2025]. Dostopno na: https://www.ifm.com/na/en/shared/successstories/reports/bucher
[4] INTERNATIONAL ELECTROTECHNICAL COMMISSION. IEC 61131-3 Programmable controllers – Part 3: Programming languages [na spletu]. 2. izd. Geneva: International Electrotechnical Commission, 2003 [pridobljeno 20. 08. 2025]. Dostopno na: www.iec.ch
[5] IFM ELECTRONIC GMBH. System Manual ecomatmobile BasicController (CR0401, CR0403) Target V01. 1. izd. Essen: ifm electronic gmbh, 2011.
[6] IFM ELECTRONIC GMBH. BasicDisplay CR0451 Programming Manual [na spletu]. Essen: ifm electronic gmbh, april 2021 [pridobljeno 20. 08. 2025]. Dostopno na: https://media.ifm.com/CIP/mediadelivery/asset/5b5d8cba5c3d3b448c034ec12f6a3ff5/7391001_EN.pdf