SCIENTIFIC EDITOR

Ryszard Szczebiot

INNOVATION TRENDS 2026

Scientific Editor

Ryszard Szczebiot

INNOVATION TRENDS 2026

University of Lomza

Łomża 2026

Scientific Editor Ryszard Szczebiot

Reviewers Ryszard Szczebiot
Pedro Vera Serna
Michal Zábovský
Romuald Kotowski
Ladislav Várkoly

Łomża 2026

ISBN 978-83-60571-89-7

CONTENT

AI APPLICATIONS IN X RAY DIFFRACTION CHARACTERIZATION

Pedro VERA-SERNA[1]

Universidad Politécnica de Tecámac, División de Ingenierías, Centro de Ingeniería Avanzada, Tecámac, Estado de México, Mexico[1]

pedrovera.upt@gmail.com[1]

ABSTRACT: This paper presents the results and variations obtained using artificial intelligence in the process of phase characterization by X-ray diffraction, as well as the results of the characterization of the synthesized material. The novelty is the approximation possible using AI, were compared the results using three different Artificial Intelligences. The synthesis was carried out via mechanochemical methods using the precursor materials Fe_2O_3 and Cr_2O_3, followed by 7 hours of milling. For traditional characterization, a Bruker diffractometer was used, along with EVA software and the ICDD PDF2+ database. The AI information was compared with traditional method find variations, but in the generation of profile was satisfactory.

Keywords: AI in characterization of materials, X Ray Diffraction, iron chromium oxide, cromite, generation PDF data with AI

INTRODUCTION

In recent decades, the use of the internet and information technologies has been making the work easier. One of the international trends with the highest demand and development is artificial intelligence, as can be seen in the results presented in DITRENDIA, which show a projected growth from 96 in 2021 to 1,847 in 2030 [1, 2]. In the case of research projects, this often offers certain benefits in the area of materials characterization, as noted in the works of Meric Eren Cakar and Lizichen Chen et al. [3, 4]. The provision of information by AI, along with information analysis and solution generation, has made it possible to explore these options as solutions [3, 4].

Apparently, thousands of researchers and professionals are increasingly generating technological development and research projects, which could be perceived as a greater challenge for presenting new advances, given the high number of publications on various topics year after year. However, social trends, the emergence of new technologies such as AI, the lack of resources or their limitations, and new trends in process changes allow for the creation of new research opportunities [5, 6].

In general terms, characterization using X-Ray Diffraction requires a sample of the material; one must also consider the type of radiation applied, the detector used, and the software that allows the data to be saved. Additionally, a comparison database containing reference patterns for the materials being analyzed and software for analyzing the matches is necessary and preferably the identification of possible elements from the periodic table present in the sample; this allows the match to be correlated and the phases present to be identified. All of the above entails a cost that not all higher education institutions can afford or sustain.

This paper presents the approaches that can be achieved using artificial intelligence to aid the research process in material characterization via X-Ray Diffraction [7]. With the appropriate options, these approaches provide accurate results and facilitate decision-making. By comparing the results offered by Claude, Chat GPT, and Gamma applications [8, 9, 10], based on a traditionally studied case that utilized the ICDD database, an X-ray diffractometer, and EVA software, a mechanochemical process is carried out using Fe_2O_3 and Cr_2O_3,

1. METHODS ANS MATERIALS

To explore the opportunities offered by artificial intelligence, we have investigated in Spanish language its use in phase identification based on results obtained from an X-ray diffractometer. The diffraction profile is analyzed with the aid of artificial intelligence, utilizing applications such as Gamma, ChatGPT, and Claude; each provides its own solution, which is then compared with the identification performed using specialized equipment and software in the traditional manner. This leads to the identification of opportunities and the realization that current limitations.

The physical study begins with high-purity precursor materials from Sigma-Aldrich, which are Fe_2O_3 and Cr_2O_3, to identify the synthesis products after 7 hours of grinding in a high-energy mill. For characterization, a Bruker D8 Advance diffractometer with copper radiation and a LINXEYE XE detector was used, along with XRD software for reading the instrument data and EVA software for phase identification.

2. RESULTS

This study is based on the need to identify the current possibilities offered by artificial intelligence in the use of web-based information. It is based on the premise that the analysis of artificial intelligence and published information may lead to an approximate answer.

The analysis is performed using three AI applications, all given the same instruction, which was as

follows: Based on the position of the peaks in the following figure, identify the phases present in the copper radiation diffraction profile, using information from the web; the basis is oxygen, chromium, and iron. The figure shows the diffraction profile obtained from mechanical milling. The analysis was conducted in Spanish language.

1.1. GAMMA APP

The result provided by the Gamma application was as follows:
First, it identifies the observed peaks as shown in Figure 1. It then compares the results with the JCPDS data for the most common oxides in the Fe–Cr–O system, thereby determining the possible phases as shown in Fig. 2 (Remember that the use of applications were in Spanish).

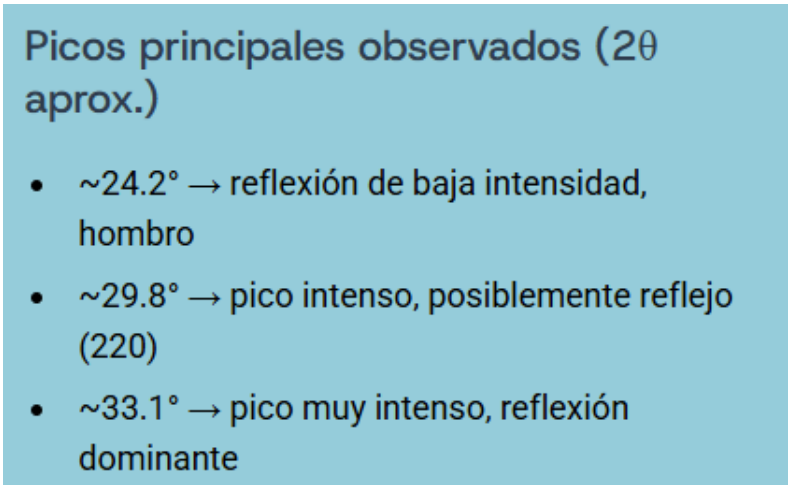


Fig 1.

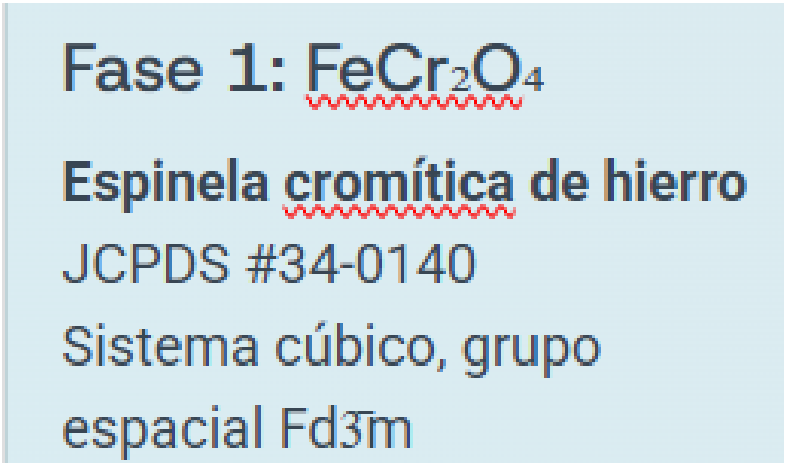


Fig 2

It then presents the relationship between the peaks, the angles, and their correspondence to phases in a table, ultimately proposing $FeCr_2O_4$, Fe2O3 and $(FeCr)_2O_3$ as possibilities, presenting an analysis

based on peak widths that indicate micro deformations and possible nanoscale sizes; this peak width analysis corresponds to the analysis performed on the particle size distribution and the mechanochemical process.

Table 1 presents a comparison of the identification of phases.

Tab. 1. Phases identified with Gamma

Gamma application	Analysis with EVA and database of ICDD
Fe_2O_3	Fe_2O_3
$FeCr_2O_4$	$(Fe_{0.6}Cr_{0.4})_2O_3$
$(FeCr)_2O_3$	

The results show a certain degree of approximation, though they lack precision, which suggests that a synthesis occurred that resulted in a phase of iron chromium oxide.

1.2. CLAUDE APP

The pattern analysis (Cu Kα radiation, λ ≈ 1.5406 Å) identifies main peaks around ~24–25°, 33–36°, 41°, 49–50°, 54–55°, 62–64°, and ~72–75° (2θ). These are identified as matching quite well with typical Fe and Cr oxides reported in databases such as ICDD/JCPDS, yielding the phases presented in Table 2.

Tab. 2. Phases identified with ChatGPT

ChatGPT application	Analysis with EVA and database of ICDD
Fe_2O_3	Fe_2O_3
Fe_3O_4	
$FeCr_2O_4$	$(Fe_{0.6}Cr_{0.4})_2O_3$
Cr_2O_3	

It can be seen that the analysis shows a greater deviation in the result, which leads to greater uncertainty and does not allow for a narrowing of the range of possible results.

1.3. CHATGPT APP

Using the collected data and a visual analysis of the diffractogram, the application was able to perform a detailed phase identification, presenting the results in an organized graph shown in Fig. 3, which lists the hkl planes. It also presented a table showing the relationship between angles, peaks, planes, and phases. Fig. 3 shows the table of phases identified using the Claude application.

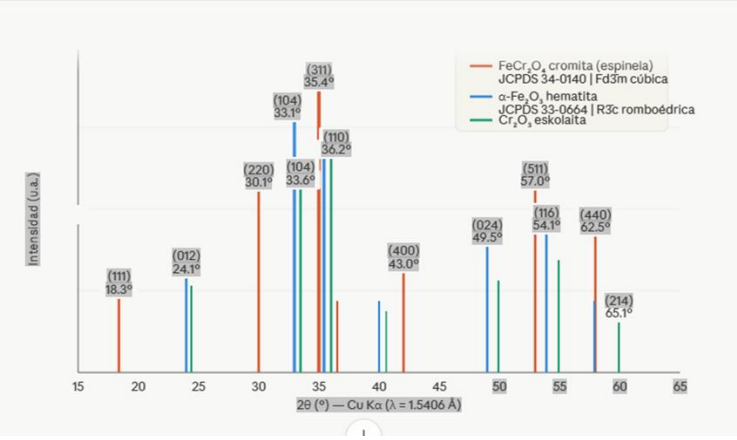


Fig. 3. XRD profile developed by Claude

Tab. 3. Phases identified with Claude

Claude Applicayion	Analysis with EVA and database of ICDD
Fe_2O_3	Fe_2O_3
Cr_2O_3	
$FeCr_2O_4$	$(Fe_{0.6}Cr_{0.4})_2O_3$

It can be observed that the analysis presents data indicating the presence of an iron chromium oxide phase, which is correct; however, there are deviations in the phase characteristics. There is a difference in the Cr_2O_3 phase compared to the gamma analysis, which does not show the $(FeCr)_2O_3$ phase, where the peaks identified as part of the Cr_2O_3 phase are present.

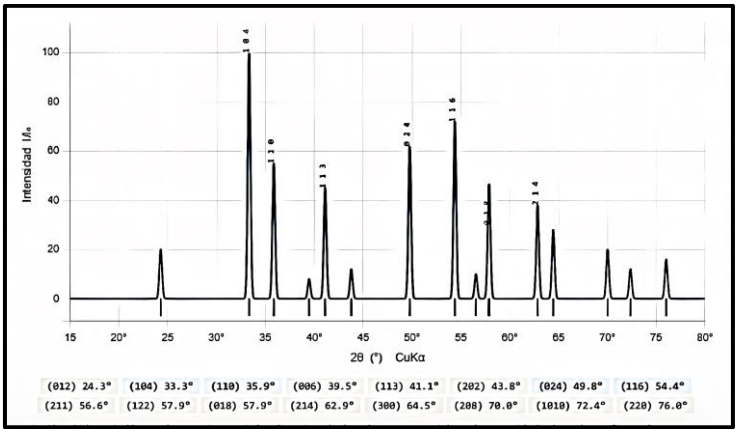


Fig. 4. XRD profile generated to $(Fe_{0.6}Cr_{0.4})_2O_3$ phase.

One of the reasons the applications do not detect the $(Fe_{0.6}Cr_{0.4})_2O_3$ phase is because it is not available on the web; apparently, this is why they default to the $FeCr_2O_4$ and Cr_2O_3 phases, where they find a similarity in the peak locations. However, in this study, while exploring Claude's solutions, its responses included the ability to generate phase diffraction profiles; therefore, it was requested to generate a

graph based on PDF-00-034-0412 and the structural information for $(Fe_{0.6}Cr_{0.4})_2O_3$. It calculated and constructed the graph, which is presented in Fig. 4.

This tool helps visualize and verify peaks when they are not available online, which aids in the characterization of materials; in the future, the databases available online will likely be expanded.

1.4. CHARACTERIZATION WITH SOFTWARE AND DATABASE

The following presents the results obtained from phase identification using the ICDD database and comparison software EVA, with the phase $(Fe_{0.6}Cr_{0.4})_2O_3$ shown in red and the Fe_2O_3 phase in red color, as can be seen in Fig. 5. It was the result obtained with traditional methodand the peaks. The result generated by Artificial Intelligence showed the peaks in same angle as were idntified in Fig. 4.

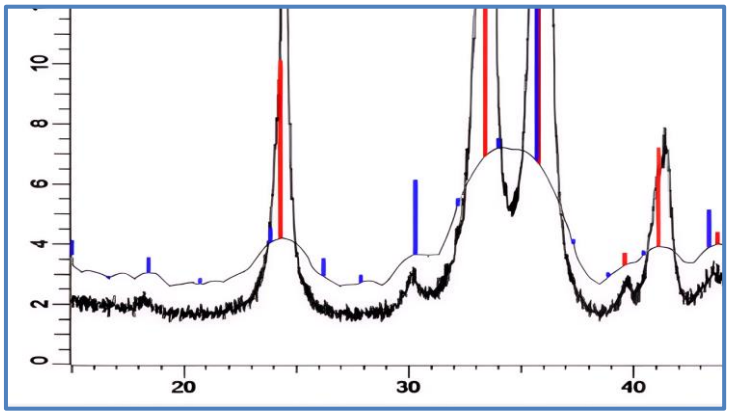


Fig. 5. XRD profile evaluated with basedata ICDD

CONCLUSION

The study showed that Gamma's artificial intelligence detected two phases that closely matched the obtained result, allowing the user to determine that the synthesis occurred in an iron-chromium-oxygen system $(FeCr)_2O_3$. With Claude, it was only possible to approximate the iron-chromium-oxygen phase; in both cases, Fe_2O_3 was identified, which is a step forward. One of Claude's advantages is the generation of diffraction profiles containing information on the phase's chemical composition, which represents a significant advancement in characterization by making that information available. Over time, analysis will likely become easier due to improved functions and greater availability of data. The information regarding nanoscale size and microdeformation provided by Gamma was accurate. Today there are variations due to limitations of profiles of XRD in the web, but the generation of profiles with Claud app is a good to explore with AI.

REFERENCES

[1] F. Rivero, (2023), "Estadísticas clave sobre inteligencia artificial que deberías conocer," DITRENDIA – Blog MKTefa, [en línea]. Disponible en: https://mktefa.ditrendia.es/blog/estadisticas-ia

[2] F. Rivero, (2023), Informe IA – Inteligencia Artificial en España y en el mundo 2023, DITRENDIA, Madrid, España. Disponible en: https://mktefa.ditrendia.es/blog/informe-ia-2023

[3] M. E. Cakar, (2025), "Artificial intelligence in materials characterization: methods, applications, and future perspectives," International Journal of Advances in Scientific Research and Engineering (IJASRE), vol. 11, no. 11, pp. 17–24. DOI: 10.31695/IJASRE.2025.11.3

[4] L, Chen, G. Zhu, Y. Chen, C. W. Lim, W. Chen, (2026) "AI-enabled high-throughput characterization of material properties based on indentation response", MechanoEngineering, vol. 1, no. 1, pp. 1, 01431-1 - 1, 01431-16, doi: 10.1063/5.0284635

[5] Madanchian, Mitra y Hamed Taherdoost. "El impacto de la inteligencia artificial en la eficiencia de la investigación". Results in Engineering, vol. 26, june 2025, 104743, https://doi.org/10.1016/j.rineng.2025.104743 .

[6] T. Cheng-Tek, (2020) "The impact of artificial intelligence on human society and bioethics", Tzu Chi Medical Journal, vol.32, no 4, p.p.:339–343. doi: 10.4103/tcmj.tcmj_71_20

[7] K. Choudhary, (2025), "DiffractGPT: Atomic structure determination from X-Ray Diffraction patterns using a generative pretrained transformer," The Journal of Physical Chemistry Letters, vol. 16, no. 8, pp. 2110–2119. DOI: 10.1021/acs.jpclett.4c03137 nih

[8] E. Lozić y B. Štular, (2023), "ChatGPT v Bard v Bing v Claude 2 v Aria v human-expert. How good are AI chatbots at scientific writing?" Future Internet, vol. 15, no. 10, art. 336. DOI: 10.3390/fi15100336 Occupational Safety and Health Administration

[9] D. Tshitoyan et al., (2025), "Exploring the expertise of large language models in materials science and metallurgical engineering," Digital Discovery (RSC Publishing), DOI: 10.1039/D4DD00319E nih

[10] F. J. Álvarez-Martínez et al., (2025), "There are significant differences among artificial intelligence large language models when answering scientific questions," Frontiers in Artificial Intelligence, vol. 8, art. 1664303. DOI: 10.3389/frai.2025.1664303 SERC

Pedro,Vera-Serna: https://orcid.org/0000-0001-7085-7374

SOFTWARE STANDARDS AND METRICS IN MODERN SYSTEM DEVELOPMENT: A QUANTITATIVE APPROACH TO QUALITY IMPROVEMENT

Graciela Margarita CETINA-QUIJANO, Pedro VERA-SERNA[1,2]

Universidad Politécnica de Tecámac, División de Ingenieros, Centro de Ingeniería Avanzada,[1,2]

Tecámac, Estado de México, México

univ.gcetina@gmail.com[1], *pedrovera.upt@gmail.com*[2]

ABSTRACT: Software quality is a critical factor in the development of modern systems, especially in areas such as mobile applications, web development, artificial intelligence, and the Internet of Things (IoT). The rapid growth of the technology industry has increased the need to ensure that software products are reliable, efficient, and maintainable from the earliest stages of their lifecycle. This study analyzes the implementation of international quality standards—specifically ISO/IEC 25010 and ISO/IEC/IEEE 15939—and software metrics as tools to improve product quality and the efficiency of the development process in academic projects. A quantitative model integrated into agile methodologies is proposed and evaluated using a correlational design. The results show a reduction in defects of up to 30% and an improvement in productivity of 15 to 20%, which supports the systematic adoption of these practices.

Key words: software metrics, software quality, ISO/IEC 25010, ISO/IEC/IEEE 15939, IEEE 1061, agile methodologies, DevOps, software engineering.

INTRODUCTION

Software development has evolved from a craft-based activity, typical of the 1950s, into a highly regulated and standardized engineering discipline [1][2].

The growing complexity of digital systems and recurring issues with quality, cost, and maintenance led to the so-called “software crisis,” a phenomenon that drove the adoption of systematic methodologies, quantitative metrics, and international standards [3][4][5].

Pressman points out that standards and metrics are the cornerstones of software quality assurance, as they make it possible to objectively measure attributes that would otherwise be subjective [5]. Similarly, Fenton and Bieman state that no process improvement is possible without prior measurement, since without quantitative data there is no basis for making informed decisions [6]. Despite this academic consensus, many organizations—particularly small and medium-sized enterprises—still lack structured models that integrate metrics and standards into their development cycle [7], [8].

In the educational setting, the situation is no different: students often do not perceive the practical relevance of quality and continuous improvement until they enter professional practice. This gap

between theory and application constitutes one of the main motivations for this study, which seeks to demonstrate—through case studies—how the systematic application of standardized metrics optimizes outcomes in both academic projects and professional development.

The primary objective of this study is to evaluate the impact of the joint implementation of process and product metrics, aligned with the ISO/IEC 25010 [9] and ISO/IEC/IEEE 15939 [12] standards, within agile methodologies (Scrum and Kanban), using a correlational and quantitative research design.

1. THEORETICAL FRAMEWORK

1.1 International standards for software quality and processes

ISO/IEC 25010 – Software Quality Model (SQuaRE)

The ISO/IEC 25010 standard is part of the SQuaRE (Software Quality Requirements and Evaluation) family and defines the software product quality model currently in use internationally [9].

It establishes eight quality characteristics: functional adequacy, reliability, performance efficiency, usability, security, compatibility, maintainability, and portability. Each characteristic is broken down into sub-characteristics that can be measured using quantifiable indicators, making it the most widely used framework for defining, evaluating, and certifying software quality [10], [11].

ISO/IEC/IEEE 15939 – Measurement Process

The ISO/IEC/IEEE 15939 standard specifies a structured process for defining, collecting, and analyzing software metrics [12]. It outlines the activities and tasks required to implement the measurement process in a systematic and repeatable manner, ensuring that the data obtained is valid, consistent, and comparable across projects. Its integration with ISO/IEC 25010 allows the quality attributes of the model to be directly linked to operationalizable metrics [6].

ISO/IEC 12207 – Software Life Cycle Processes

The ISO/IEC 12207 standard defines the overall process framework for the software life cycle, covering acquisition, development, maintenance, operation, and quality assurance [13]. It serves as the foundation upon which other quality standards and metrics are built, by providing the organizational and technical context in which these practices must be carried out.

ISO/IEC 14764 – Software maintenance

Complementing ISO/IEC 12207, the ISO/IEC 14764 standard regulates the software maintenance process in its four forms: corrective, adaptive, perfective, and preventive [14]. Measuring metrics such as the average time to fix errors and the density of residual defects is particularly relevant in this phase of the life cycle, where intervention costs tend to be higher if the software lacks adequate maintainability.

CMMI – Capability Maturity Model Integration

The CMMI model, developed by the Software Engineering Institute (SEI) at Carnegie Mellon University, assesses the maturity of an organization's development processes across five levels: Initial, Repeatable, Defined, Quantitatively Managed, and Optimized [16]. Level 4 (Quantitatively Managed) is particularly relevant to this study, as it establishes that statistical control of processes—through metrics such as defect density and test coverage—is a necessary condition for managing projects with predictable and consistent quality [16].

IEEE Standards

The IEEE publishes a set of complementary standards for software engineering. The IEEE 730 standard governs software quality assurance processes [17]; IEEE 829 establishes requirements for test documentation; and IEEE 1061 provides a structured methodology for evaluating software quality metrics, defining how to select, implement, analyze, and interpret these metrics within the context of a project [15].

1.2 National Standards: Mexico and Latin America

In Mexico, the MoProSoft model—formalized as the NMX-I-059-NYCE-2005 standard—was designed to improve the competitiveness of software development companies, with an emphasis on small and medium-sized organizations [7]. The model organizes processes into three levels (Senior Management, Management, and Operations) and establishes quality criteria adapted to the national context. Galvis-Lista and Sánchez-Torres document that models such as MoProSoft and its Latin American derivatives—including Competisoft (for Latin American SMEs) and MPS.BR (Brazil)—are regional variants based on international standards such as CMMI and ISO, but with a simplified structure that facilitates their adoption in organizations with limited resources [8].

1.3 Software Metrics: Classification and Key Indicators

Fenton and Bieman classify software metrics into three main categories based on the object of measurement [6]: product metrics, process metrics, and project metrics. This taxonomy, presented in

Table 1, is the most widely accepted in the software engineering literature and aligns with the measurement process defined in ISO/IEC/IEEE 15939 [12].

Table 1. Classification of software metrics according to Fenton and Bieman

Metric type	Description	Representative example
Product	Evaluates attributes of the finished software (functionality, reliability, usability)	Defect density per KLOC
Process	Measures the efficiency and performance of the development process	Delivery time, defect rate per sprint
Project	Manages compliance with scope, time, and cost	Effort in person-hours, cost per module

Within each category, there are specific indicators whose selection depends on the project's objectives. For this study, the metrics described in Table 2 were adopted, chosen for their proven validity in the literature [6], [5] and their alignment with the quality attributes of ISO/IEC 25010 [9]:

Table 2. Key metrics used in the model

Metric	Definition	Formula	Purpose
Defect density	Number of errors relative to the size of the software [6]	Defectos / LOC	Assess product quality
Test coverage	Percentage of code evaluated through automated tests [12]	Casos exec. / Total	Improve reliability
Productivity	Team efficiency in terms of code produced per unit of effort [6]	LOC / Horas	Measure efficiency
Cyclomatic complexity	Number of linearly independent paths through the source code [5]	V(G) = E − N + 2P	Maintainability
Team velocity	Story points completed per sprint in agile methodologies [18]	PH / Sprint	Agile productivity

Cyclomatic complexity, originally proposed by McCabe and formalized in the IEEE 1061 standard [15], measures the number of independent logical paths in a software module. Values between 1 and 10 are considered to indicate low complexity and high maintainability; values above 20 indicate code that is difficult to test and maintain [5]. Test coverage, on the other hand, is directly related to system reliability:

studies have documented that coverage levels above 70% are associated with significant reductions in defect density in production [6].

2. PROBLEM STATEMENT

Despite the availability of well-established international standards such as ISO/IEC 25010 [9] and ISO/IEC/IEEE 15939 [12], the systematic adoption of quality metrics in development teams remains limited. Pressman identifies that teams that do not implement formal measurement tend to repeat the same mistakes in each project, increasing correction costs as the software life cycle progresses [5]. Sommerville complements this perspective by noting that the absence of metrics generates projects with high variability in their outcomes, making both planning and quality assurance more difficult [18].

In the academic context, this problem is even more pronounced: software engineering students frequently develop projects without applying formal metrics or referencing international standards, which limits their preparation for professional practice. Cortés Rojas documents that even in educational technology prototyping projects, the absence of a quality framework structured under ISO/IEC 25010 leads to inconsistent and difficult-to-evaluate results [10].

Identified consequences:

- Increase in software defect density, with correction costs that can multiply tenfold between the development phase and the production phase [5].
- Recurring delays in project delivery due to the absence of process tracking metrics [6].
- Low system maintainability, resulting from a lack of control over cyclomatic complexity and other code structure indicators [15].
- Low end-user satisfaction with products that do not meet the usability and reliability attributes defined in ISO/IEC 25010 [9], [11].

Research question: What is the impact of the systematic implementation of software standards and metrics on product quality and the efficiency of the development process in modern systems projects?

3. HYPOTHESIS

General hypothesis: The systematic implementation of standards (ISO/IEC 25010 [9], ISO/IEC/IEEE 15939 [12]) and software metrics within agile methodologies significantly improves product quality and the efficiency of the development process.

H1: The use of process metrics reduces software defect density [6].

H2: The application of quality standards increases system maintainability and reduces the average error correction time [9], [15].

H3: The integration of metrics into agile cycles improves team velocity and end-user satisfaction [18].

4. METHODOLOGY

4.1 Research approach and design

The study adopts a quantitative approach, which allows the variables of interest to be objectively measured using numerical indicators. This approach is consistent with the recommendations of Fenton and Bieman, who establish that metrics must be evaluated through quantifiable methods to determine their real impact on software quality [6]. The design is descriptive-correlational: descriptive because it characterizes the behavior of metrics before and after their implementation; correlational because it analyzes the relationship between the application of metrics (independent variable) and software quality (dependent variable), using the Pearson correlation coefficient.

4.2 Variables and operationalization

Variables are defined based on the measurement framework established by ISO/IEC/IEEE 15939 [12] and the quality attributes of ISO/IEC 25010 [9]. Table 3 presents their operationalization.

Table 3. Operationalization of variables

Variable	Dimension	Indicator	Unit of Measurement
Software metrics (IV)	Process	Test Coverage	%
	Process	Team Productivity	LOC / hour
	Process	Deployment Frequency	Deployments / week
Software quality (DV)	Product	Defect Density	Defects / LOC
	Product	Error Correction Time	Hours
	Product	User Satisfaction	Scale 1–5

4.3 Proposed methodological model

The model is structured in five sequential phases, designed to be replicable in any type of software project, academic or professional:

- Phase 1 – Quality definition: the quality attributes relevant to the project are established, using the ISO/IEC 25010 model [9] as reference (functionality, reliability, usability, maintainability, performance efficiency).
- Phase 2 – Metrics selection: the most relevant process and product metrics for the project's objectives are identified, following the IEEE 1061 methodology [15] and the ISO/IEC/IEEE 15939 measurement process [12].
- Phase 3 – Data collection: data is obtained automatically through static analysis tools (SonarQube), version control systems (GitLab CI/CD), task managers (Jira Software), and unit test reports (JUnit, Selenium).
- Phase 4 – Data analysis: descriptive statistical techniques and Pearson correlation are applied to quantify the relationship between process metrics (coverage, velocity) and product quality indicators (defects, maintainability) [6].
- Phase 5 – Continuous improvement: based on the results, the process is fed back into each sprint, generating an iterative cycle of learning and adjustment aligned with agile practices [18].

4.4 Case studies and sample

The model was applied to six university projects distributed across four technology areas: web development, artificial intelligence (AI), Internet of Things (IoT), and mobile applications. Each team worked under Scrum or Kanban methodology for three two-week sprints, incorporating metrics into tracking boards and retrospective sessions.

This selection makes it possible to contrast the behavior of metrics across heterogeneous technology domains, highlighting both common patterns and the particularities of each area, as has been documented in previous studies on the application of ISO/IEC 25010 in educational technology projects [10], [11].

5. RESULTS

The results obtained across the six evaluated projects confirm the three hypothesis oriented forward. The systematic integration of metrics —aligned with ISO/IEC 25010 [9] and ISO/IEC/IEEE 15939 [12]— generated measurable and consistent improvements across all monitored quality and process indicators. Table 4 summarizes the average values obtained at the close of the third sprint.

Table 4. Improvement indicators obtained after implementing the proposed model

Evaluated indicator	Average value	Observed impact
Test coverage	78 %	Earlier error detection in each sprint
Defect reduction per KLOC	30 %	Lower failure rate in production after three iterations
Team velocity increase	+18 %	Increase in stories completed per sprint
Error correction time	2.5 h	Improvement in system maintainability
User satisfaction	4.6 / 5.0	High perception of quality and usability
Process–Quality correlation (r)	r = 0.82	Strong positive relationship: higher metrics integration, lower defect density

The Pearson correlation (r = 0.82) between defect reduction and productivity increase indicates a strong positive relationship: teams that adopted metrics more rigorously simultaneously showed lower defect density and higher development velocity. This result is consistent with the theoretical predictions of Fenton and Bieman [6], who argue that continuous measurement acts as a feedback mechanism that accelerates team learning.

Table 5 breaks down results by project and technology area, showing the reduction in defects before and after implementing the model in each case study:

Table 5. Defect reduction by project and technology area

Project	Technology Area	Defects Before	Defects After	Improvement (%)
Web Educational Platform	Web Development	45	20	55 %
Image classifier	Artificial Intelligence	60	30	50 %
Monitoring system	Internet of Things	50	25	50 %
School management App	Mobile Applications	38	17	55 %

Figures 1 and 2 visually illustrate the distribution of process and product metrics among the academic units studied, as well as a comparison of key indicators between the web development and artificial intelligence projects, which represented the extremes of the range of results obtained.

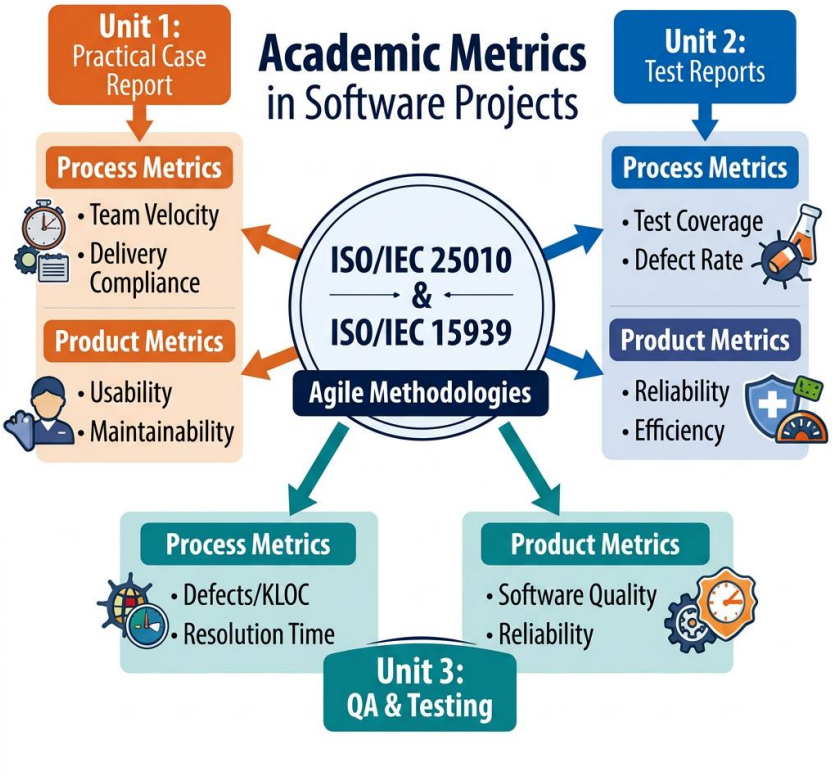


Fig. 1. Map of the relationship between process metrics, product metrics, and the academic units evaluated

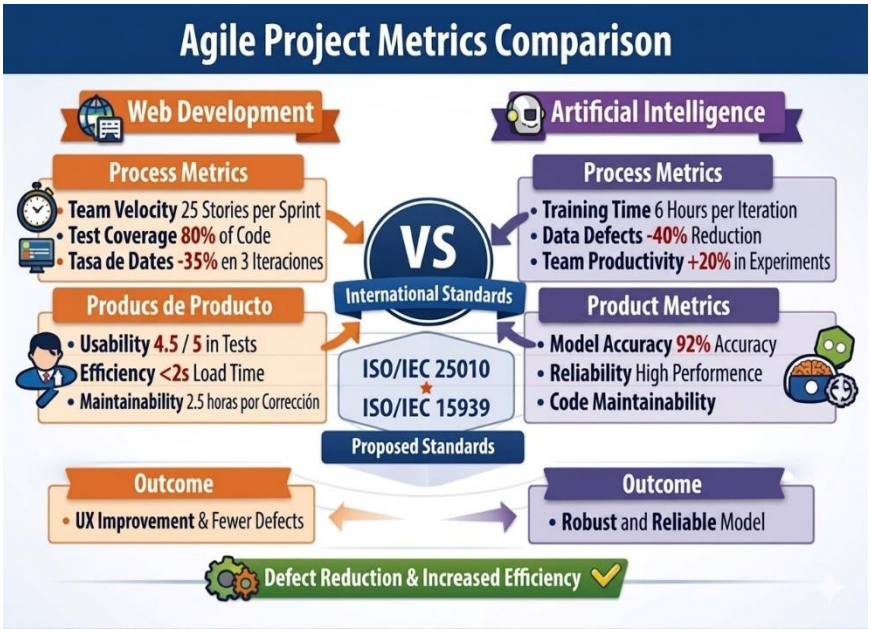


Fig. 2. Comparison of Key Metrics Between Web Development and Artificial Intelligence Projects

In web development projects, performance efficiency and usability metrics —sub-characteristics of ISO/IEC 25010 [9]— proved to be the most decisive indicators for user satisfaction (4.6/5.0). In AI projects, model reliability and code maintainability carried greater weight, which aligns with the findings of Fernández Jarrín on the application of ISO/IEC 25010 in complex computational systems [11].

6. DISCUSSION

The results presented are consistent with the body of evidence accumulated in the software engineering literature. The 30% reduction in defects per KLOC obtained in this study is comparable to the ranges reported by Fenton and Bieman in organizations that implement test coverage metrics above 70% [6]. Similarly, the 18% increase in team velocity aligns with the productivity improvements documented by Sommerville in agile environments that incorporate sprint tracking metrics [18].

CMMI Level 4 —Quantitatively Managed— establishes that statistical process control is a necessary condition for achieving predictable quality [16]. The results of this study suggest that even small teams, without formal certification, can achieve behaviors characteristic of that maturity level by adopting a bounded set of well-selected metrics integrated into their agile cycles.

Cortés Rojas documents, in the Mexican educational context, that structuring projects under ISO/IEC 25010 significantly increases the coherence between quality requirements and delivered artifacts [10]. The results of the present study reinforce that conclusion and add the quantitative dimension: not only does qualitative coherence improve, but defects are reduced and correction time is optimized.

It is important to note, however, that the implementation of metrics is not without challenges. Resistance to change, the lack of a measurement culture, and the difficulty of selecting the minimum set of relevant metrics are recurring obstacles identified in teams in training. As Fenton and Bieman warn, measurement overload —that is, the collection of data that is not used to make decisions— can erode productivity and generate skepticism toward quality practices [6]. For this reason, the proposed model prioritizes a reduced set of high-impact metrics, selected based on the specific objectives of each project.

From the perspective of technology trends, the integration of CI/CD (continuous integration and continuous deployment) tools with static analysis platforms —such as SonarQube— represents the most efficient path to automating metrics collection and making it transparent for the team. This trend is aligned with modern DevOps practices and the continuous improvement approach promoted by both ISO/IEC 12207 [13] and the CMMI model [16].

CONCLUSION

This study confirms that the systematic implementation of process and product metrics, aligned with ISO/IEC 25010 standards and ISO/IEC/IEEE 15939 , significantly improves software quality and the efficiency of the development team. The three hypotheses put forward were validated: a 30% reduction in defect density was obtained, a decrease in error correction time to an average of 2.5 hours as an indicator of improved maintainability, and an 18% increase in team velocity with a user satisfaction score of 4.6/5.0.

International standards —in particular ISO/IEC 25010 , ISO/IEC/IEEE 15939 and IEEE 1061— provide the structural framework needed to apply metrics in a viable, repeatable, and comparable manner across projects. Agile methodologies, in turn, offer the flexible environment that allows those metrics to be integrated into the development cycle without disrupting the team's workflow.

For the Mexican educational context, models such as MoProSoft represent an accessible entry point into a quality culture, complementary to international standards. The combination of both perspectives —local and international— offers students and teams in training a solid, gradual framework that can be applied from the very first academic projects.

As future work, it is proposed to extend the model to artificial intelligence and IoT environments with domain-specific metrics (model accuracy, latency, energy consumption), as well as to validate the model in organizations with different CMMI maturity levels with the aim of establishing graduated adoption guidelines.

REFERENCES

[1] Muhammad H. A., et al. ‘‘A COMPREHENSIVE REVIEW OF SOFTWARE DEVELOPMENT METHODOLOGIES: MODELS, MINDSET, AND MISUNDERSTANDINGS,” In Spectrum of Engineering Sciences, Zenodo. https://doi.org/10.5281/zenodo.16414601

[2] Rashid A. B., Md Ashfakul K. K.. (2024) “AI Revolutionizing Industries Worldwide: A Comprehensive Overview of Its Diverse Applications.” Hybrid Advances, vol 7, https://doi.org/10.1016/j.hybadv.2024.100277.

[3] Goel A., Masurkar S., Pathade G. R,, (2024), “An Overview of Digital Transformation and Environmental Sustainability: Threats, Opportunities, and Solutions.” Sustainability, vol 16,. https://doi.org/10.3390/su162411079.

[4] Dmitry P., Franke H., Netland T.H., (2022) “Digital Transformation: A Review and Research Agenda.” European Management Journal, vol. 41, https://doi.org/10.1016/j.emj.2022.09.007.

[5] R. S. Pressman, Software Engineering: A Practitioner's Approach, 7th ed. New York: McGraw-Hill, 2010.

[6] N. Fenton and J. Bieman, Software Metrics: A Rigorous and Practical Approach, 3rd ed. Boca Raton, FL: CRC Press, 2014.

[7] Secretaría de Economía, "NMX-I-059-NYCE-2005: Information Technology – Software Development Processes – Process Characteristics (MoProSoft)," Dirección General de Normas, Mexico, 2005.

[8] E. Galvis-Lista and J. Sánchez-Torres, "Review of software process improvement models applied in Latin American small and medium-sized enterprises," Revista Espacios, vol. 40, no. 28, p. 9, 2019.

[9] ISO/IEC 25010:2011, "Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models," International Organization for Standardization, Geneva, Switzerland, 2011.

[10] H. F. Cortés Rojas, "Use of the ISO 25010 standard to establish quality requirements in the design of an educational technology prototype based on augmented reality," Dialnet, 2024. Available at: https://dialnet.unirioja.es/descarga/articulo/10343572.pdf

[11] E. J. Fernández Jarrín, "Software quality evaluation metrics based on the ISO/IEC 25010 standard in the admissions system. UPEC case," repositorio.upec.edu.ec, Universidad Politécnica Estatal del Carchi, 2024.

[12] ISO/IEC/IEEE 15939:2017, "Systems and software engineering — Measurement process," International Organization for Standardization / Institute of Electrical and Electronics Engineers, 2017.

[13] ISO/IEC 12207:2008, "Software and systems engineering — Software life cycle processes," International Organization for Standardization, Geneva, Switzerland, 2008

[14] ISO/IEC 14764:2006, "Software engineering — Software life cycle processes — Maintenance," International Organization for Standardization, Geneva, Switzerland, 2006.

[15] IEEE Std 1061-1998, "IEEE Standard for a Software Quality Metrics Methodology," Institute of Electrical and Electronics Engineers, New York, USA, 1998.

[16] Software Engineering Institute, "CMMI for Development, Version 2.0," Carnegie Mellon University, Pittsburgh, PA, USA, 2018. Available at: https://cmmiinstitute.com

[17] IEEE Std 730-2014, "IEEE Standard for Software Quality Assurance Processes," Institute of Electrical and Electronics Engineers, New York, USA, 2014

[18] I. Sommerville, Software Engineering, 10th ed. Londres: Pearson, 2016.

Graciela Margarita Cetina Quijano https://orcid.org/0009-0007-0527-2215

Pedro,Vera-Serna: https://orcid.org/0000-0001-7085-7374

RECONSTRUCTION OF ENGINE BLOCKS THROUGH ON-SITE GRINDING OF HIGH-DEMAND ENGINE MONOBLOCKS IN MEXICO: 1.6, 2.5, AND 1.8.

ALCÁNTARA-SANDOVAL David[1], DÍAZ-MUÑOZ Daniel[1*] y VERA-SERNA Pedro[2]

1 Refacciones Servicio y Mantenimiento Automotriz STAR S. de R.L. de C.V., Tecámac, 55740, Estado de México, México.

2 Universidad Politécnica de Tecámac, Tecámac, 55740, Estado de México, México

servicioautomotrizstar@yahoo.com.mx , diazmunozdaniel18@gmail.com , pedrovera.upt@gmail.com

ABSTRACT: This paper presents the recovery of high-demand engines from vehicles in Mexico, showing the step-by-step process and equipment required to achieve high-performance functionality of the engines. The aim is to share the favorable results obtained, indicating the sequence of activities, and for this work to be of great help as a reference to contribute to reducing environmental impact, applying trends in giving new uses to products, contributing to the recycling of parts using low-cost equipment, and seeking to be accessible to SMEs. This project was developed in a practical manner by professional staff from the company Refacciones Servicio y Mantenimiento Automotriz STAR S de RL de CV. It is from this point of view that the authors, together with the organization, decided to develop this project for machine tool operators, which describes in detail, step by step, the process of one of the most common engine grinding jobs they perform, namely bed grinding (in-line cutting), implementing technical and empirical data as well as graphic examples of engines that have required this service. The purpose of this work was to provide a support guide for all those who are starting out in this area of the automotive sector. It is also hoped that it will serve as feedback for those who already have experience in this work and serve as a reference for them.

Key words: monoblock, grinding, rectificado, crankshaft, bancada, reconstruction

INTRODUCTION

La reconstrucción o rectificado de motores siempre ha sido un área clave en el sector automotriz, dado al servicio de mantenimiento correctivo que se brinda a los vehículos, pero desafortunadamente este tipo de trabajos no tienden a ser tan conocidos como lo es un servicio de mecánica general [1].

Personas que ya cuentan con cierto grado de conocimientos en la materia tiene destreza o bien noción de este trabajo y la función que realiza, es por eso que toda la información correspondiente a lo que conlleva los trabajos de rectificado no son tan común encontrarlos, si bien es cierto que las nuevas tecnologías han brindado más información [2], mejores equipos y datos más precisos correspondientes a esta área están disponibles, pero es una realidad que en muchos talleres rectificadores cuentan todavía con equipos, herramientas y formas de trabajo menos eficientes, transmitiendo los conocimientos de su trabajo de forma empírica a las nuevas generaciones [3].

El monoblock es una pieza de hierro o aluminio, el cual es fundamental en un motor de combustión interna, ya que este es el cuerpo que sostiene todo el tren alternativo, cuya función es transformar el movimiento lineal de un pistón en un movimiento giratorio del cigüeñal por medio de una reacción exotérmica, detonada de manera secuencial dentro de los cilindros en las cámaras de combustión [4].

Dado al tipo de partes que se aloja en el monoblock y la función que esta tiene, es normal que después de un determinado tiempo de uso algunas de las piezas sufran algún tipo de desgaste o deformación. Un mantenimiento inadecuado o incorrecto uso del vehículo también pueden ser causantes de un fallo en el motor, el cual conllevaría a un mantenimiento correctivo siendo más específico “corte en línea” y que se abordara de más adelante en este proyecto [5].

En un motor la bancada es la superficie circular donde se alojan los cojinetes y va montado el cigüeñal, este se ubica en la parte inferior del monoblock y está constituido por dos partes. Dividida por la mitad, la primera parte está unida al monoblock donde se montarán la mitad de cojinetes y el cigüeñal, posteriormente la segunda parte son tapas desmontables ya sea individuales o unidas esto dependiendo del tipo y marca del vehículo, aquí van a ir el resto de cojinetes y sostendrán el cigüeñal mediante unos tornillos que llevaran un determinado par de apriete (torque) dependiendo de la marca y modelo del vehículo [6].

El rectificado de la bancada del monoblock es un procedimiento de mecanizado que se realiza para restaurar la circunferencia y alineación del alojamiento de los cojinetes del cigüeñal; mediante la implementación de una mandrinadora lineal, con la cual se busca que las medidas del diámetro de

alojamiento de la bancada coincidan con las que estableció el fabricante, ya que en la bancada se forma una película de aceite la cual se encargara de reducir el desgate y la fricción que hay entre los cojinetes y cigüeñal, es por eso que las medidas de la bancada requieren ser muy precisas, ya que de no ser así pueden causar problemas graves en el funcionamiento del motor, presentar fugas de acetite, generar desgaste prematuro de piezas, observarse perdida de presión de aceite, con lo anterior es posible alcanzar una pérdida de eficiencia y hasta un posible daño del motor [7].

2. METODOLOGÍA, HERRAMIENTAS Y MAQUINARIA

A continuación, se presenta los pasos a seguir en la metodología implementada con el fin de dar claridad al proceso desarrollado, posterior a ello se presentan los elementos y equipos utilizados.

2.1 METODOLOGÍA

El trabajo fue desarrollado con el método científico, primero se observó, se identificó, se estableció el proceso a seguir, se ejecutó, se validó que fuera funcional, se repitió en varias ocasiones, permitiéndolo establecer como viable.

Este trabajo corresponde a la investigación experimental, la cual en ocasiones no es ampliamente difundida, se enfocan más a la investigación tradicional. Sin embargo, en este trabajo se inició con información de referencia por otras personas, pero se fueron encontrando hallazgos que permitieron establecer esta ruta de trabajo para culminar en una reparación exitosa en la Unidad Económica en repetidas ocasiones.

Para lo anterior se partió de la necesidad de documentar el proceso que lleva el rectificado de bancada, abarcando también métodos para el reconocimiento y valoración de monoblocks, partiendo de la recepción del monoblock y finalizando en su entrega [8].

A continuación, se presenta el diagrama de la metodología ocupada.

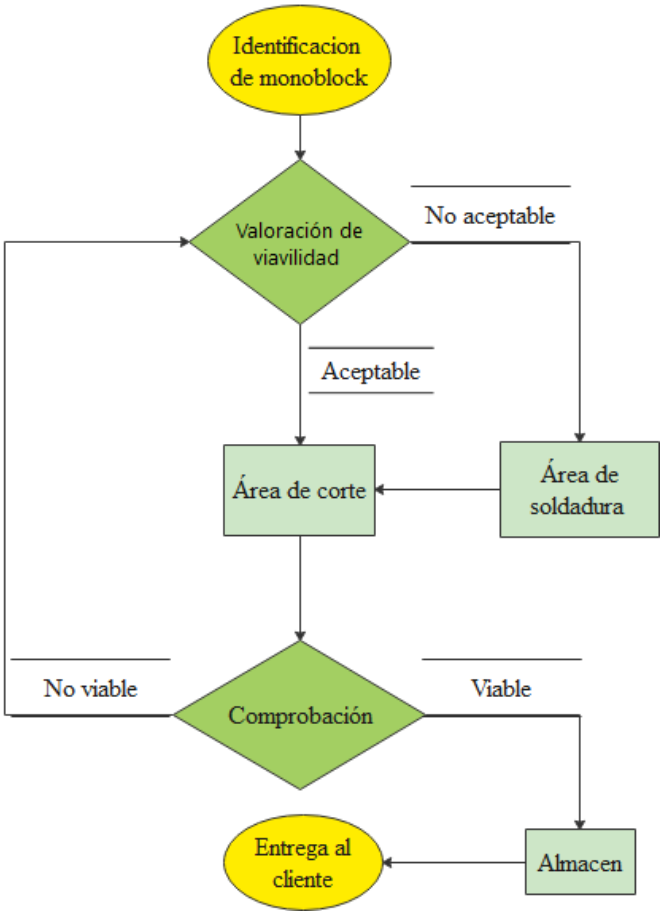


Figura 1. *Diagrama de la metodología empleada*.

Identificación del monoblock

Al inicio del proceso se recibe el monoblock, se realiza la identificación y la caracterización del mismo, apoyándose en la documentación disponible, obteniendo la información y datos requeridos para su maquinado de acuerdo a los datos del fabricante [8].

Valoración de viabilidad

Se determina con base a una inspección visual el estado en el que se encuentra el monoblock y si es requiere se traslada al área de soldadura para resanar partes perdidas de este, de lo contrario si estuviese en buenas condiciones pasa al área de corte en línea.

Área de corte

En este proceso se realizó preparación y montaje de la máquina de acuerdo con las especificaciones del monoblock a trabajar, para con ello alcanzar las dimensiones y cortes requeridos.

Área de soldadura

En caso de presentar desbastes excesivos y es viable su reparación mediante soldadura, se procede a integrar material mediante este proceso con estándares de alta resistencia y calidad.

Comprobación

En este paso se realiza una medición y comprobación de dimensiones y tolerancias, corroborando que estas estén bien de acuerdo a lo que marca el fabricante.

Almacén y entrega al cliente

En el almacén se reciben los productos validados y aprobados en espera de que llegue el cliente a recogerlo.

2.2. Instrumental

El saber reconocer bien tanto la máquina como sus herramientas es fundamental para el operador de máquina herramienta, ya que de eso depende el buen trabajo que se realiza, reduciendo tiempos y costos en el proceso de mecanizado a continuación se refiere el instrumental utilizado.

2.2.1. Barra

Como su nombre lo indica es una barra barrenada de acero in oxidable, cuya función es la de transportar nuestra pieza de corte y servir como una flecha/eje que trazara la dirección y sentido del mecanizado, contando con dos tipos de barra en el área de rectificado.

Tabla 1. Datos de barra pequeña.

Barra pequeña	
Distancia	160 cm
Diámetro	1.181”
No. Barrenos	10
Perno prisionero	Allen 5/32 (4 mm)

Figura 2. Barra de corte pequeña

Tabla 2. Datos de barra grande

Barra grande	
Distancia	186 cm
Diámetro	1.986”
No. Barrenos	13
Perno prisionero	Allen 5/32 (4 mm)

Figura 3. Barra de corte grande

2.2.2. CONOS

Son un tipo de centradores de hierro cuya función en el proceso de rectificado es al de sujetar y aproximar el centrado de la bancada con referencia a la barra.

El área de rectificado cuenta con 4 tipos de conos, los cuales serán implementados tomando en cuenta diámetro de la bancada a restaurar y el tipo de barra a utilizar.

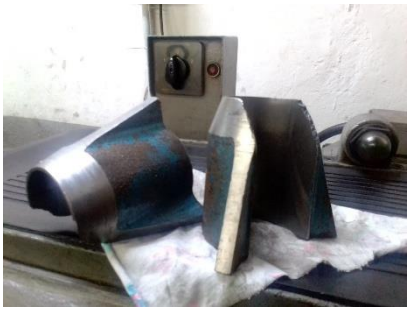


Figura 2. Conos tipo 2

Figura 1. Conos tipo 1

Figura 5. Conos tipo 3

Figura 6. Conos tipo 4

Tabla 3. Datos de los conos

CONOS CENTRADORES			
Tipo 1	Tipo 2	Tipo 3	Tipo 4
Barra pequeña Ø de bancada 1.800”-2.400”	Barra grande Ø de bancada 2.300”-2.900”	Barra grande Ø de bancada 2.800”-3.500”	Barra grande Ø de bancada 3.000”-5.200”

2.2.3. PALPADOR

Es un indicador de basa magnética con el cual se va a calibrar la altura del buril para realizar el debido corte.

Antes de usar el palpador se requiere ajustarlo al tipo de barra con la que se va a trabajar, ya que estas al tener un diámetro diferente se modifica la medida a la que se calibran los buriles.

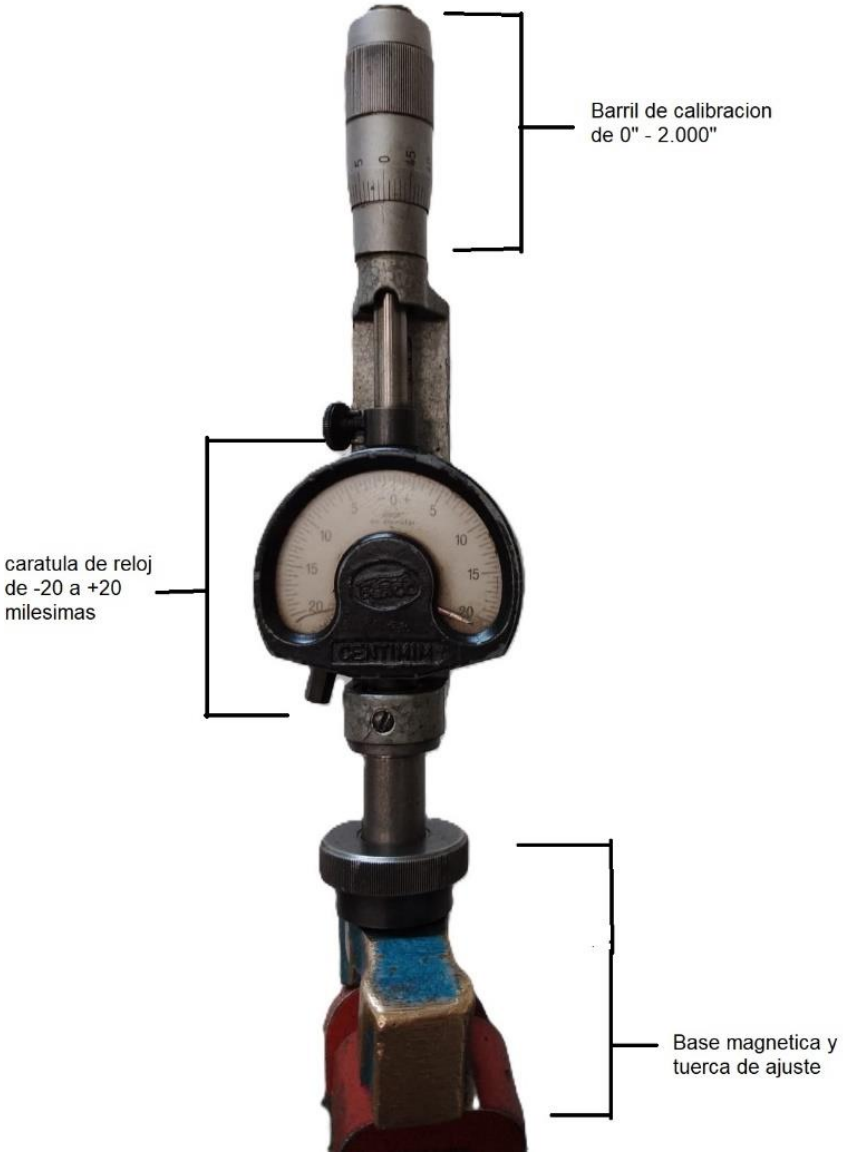


Figura 3. Partes del palpador

Ø de bancada - Ø de barra= Altura de buril
Ejemplo. Urvan 2.5 (*2.321" – 1.986"= 0.335"*)

Tabla 4. Fórmula para la altura de buril

2.2.4. BURILES

Nuestra principal herramienta de mecanizado. Son unas cuchillas mayor mente utilizadas en tornos, para realizar varios tipos de cortes y desbastes a piezas de metal, aunque en el área de rectificado se cuenta con varios tipos de buriles su uso en el rectificado de la bancada se implementara para dos cosas, el corte circular de la bancada y el careado o maquinado de los cajones axiales donde van montados los cojinetes axiales (*medialunas*).

Figura 4. Tipo de buriles

No.	buriles	
2	Buril con punta a 80°	Corte circular de la bancada
3	Buril con pastilla cuadrada	Careado del cajón del axial
2	Buril con corte izquierdo	Careado del cajón del axial

Tabla 4. Tipo de buriles

Figura 5. Buriles del taller

2.2.5 COJINETES PARA BARRA

Son unos cilindros huecos, que van montados en las columnas de soporte de la barra, los cuales se encargan de sostener la barra mientras esta se encuentra trabajando.

Se cuenta con dos pares de cojinetes para barra, esto debido a las dimensiones con las que cuenta cada barra.

Figura 7. Cojinetes para barra grande

Figura 6. Cojinetes para barra pequeña

Cojinetes para barra	Ø exterior	Ø interior
Barra grande	3.343”	1.971”
Barra pequeña	3.343”	1.185”

Tabla 5. Datos de cojinetes de barra

2.2.6. CENTRADORES DE RELOJ

De igual forma que el “Palpador”, son un tipo de indicadores de reloj con los cuales se va a centrar a precisión la bancada del motor, estos irán montados en la barra y por medio de una aguja se va a recorrer toda la circunferencia de las bancadas laterales, para que de esta manera el reloj indique que tan descentrado se encuentra.

Una vez que se termina de centrar el motor se procede a comprobar el centrado del buril; dando la medida completa y checamos que este toque en toda la bancada del motor, tanto las que están unidas al monoblock como la de las tapas (posteriormente rebajadas).

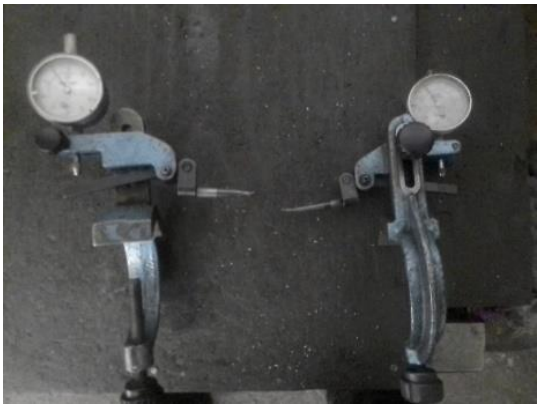


Figura 8. Centradores de reloj

2.2.7. MICRÓMETROS

Son instrumentos de medición con los cuales se pueden determinar el tamaño de un objeto con mayor precisión. Estos suelen venir en varias medidas y se dividen en dos tipos micrómetros, exteriores e interiores.

En el área de rectificado se implementa ambos tipos de micrómetros, ya que resulta indispensable su uso para la obtención de una correcta medición.

Micrómetro de exteriores (arco)

Esta herramienta ésta diseñada de una forma para obtener las dimensiones exteriores de una pieza, su lectura se da en milésimas de pulgada y se cuenta con varios micrómetros de diferentes medidas (0-1”, 1”-2”, 2”-3”, 3”-4” etc…)

Figura 10. Micrómetros exteriores del taller

Figura 9. Partes del micrómetro de exterior

Micrómetro de interiores

La función de este instrumento de medición es la obtención del diámetro de los orificios de las piezas mediante el alargamiento de uno o ambos extremos de la herramienta. En el área de rectificado se cuenta con tres tipos de micrómetros interiores.

Micrómetro de resorte

Este micrómetro es el implementado en el rectificado de bancada, ya que por su tamaño y modo de emplear resulta altamente práctico para la obtención de la medida de los cortes sin que estorbe la barra, está constituido por varias partes intercambiables y cuneta con un mecanismo de resorte que empuja un extremo de la herramienta hacia el otro punto de la bancada, obteniendo así su diámetro.

Micrómetro de trompo

Este micrómetro es implantado en el rectificado de cilindros, dado a su precisión y practicidad al momento de realizar múltiples mediciones. Consta de dos piezas verticales unidas por una rosca con una punta ajustable, obteniendo el diámetro mediante el desenroscado de las piezas hasta que ambos extremos terminen tocando la superficie del orificio.

Micrómetro T

Este micrómetro es el menos utilizado, ya que resulta más difícil obtener con él una medida exacta, pero su uso se emplea en orificios sumamente pequeños o de difícil acceso para los otros micrómetros, está constituido por tres extremos de manera perpendicular formando una T, esta al aflojarse del extremo más largo acciona un mecanismo de resorte expandiendo los otros dos extremos restantes, tocando así la superficie del orificio.

Figura 13. Micrómetro de interiores con resorte

Figura 12. Micrómetro de interiores en T

Figura 11. Micrómetro de interiores de trompo

2.2.8. ABRAZADERAS

Son cuatro accesorios complementarios de la máquina en forma de dona, cuya función principal es la de brindarle un tipo de aumento a nuestro buril, para poder realizar cortes a motores que cuenten con una bancada muy grande.

A la abrazadera se le coloca el buril y se calibra la medida que debe llevar correspondiente al diámetro de la abrazadera, posteriormente la abrazadera se monta encima de la barra abrazándola y sujetándose mediante un tornillo Allen.

Figura 14. Abrazaderas

abrazadera	1	2	3	4
Tornillo Allen	5mm	6mm	6mm	8mm
Diámetro	2.347”	2.756”	3.545”	4.532”

Tabla 6. Datos de las abrazaderas

2.2.9. EXTENSIÓN

Es una flecha de acero con conexiones en ambos extremos capaz de unirse por un lado con la barra y en el otro con la máquina, un accesorio complementario de la máquina, cuya función es la de ampliar el rango de distancia recorrido por el buril, cuando el brazo de la máquina ya no puede desplazar más lejos la barra, en el área de rectificado se cuenta con dos extensiones de diferente medida, la extensión más larga permite recorrer el buril a las bancadas 1 y 2 que son las que normalmente no se llegaría con el brazo de la máquina, la extensión más corta solo se emplea en el corte de la bancada 3 y en algunos casos pudiendo llegar a cortar a la bancada 2.

Nota: ambas extensiones son compatibles con la barra.

Figura 15. Extensiones

2.2.10. CARPETA DE DATOS

Es una carpeta con la recopilación de las medidas de la bancada de todos los motores que se han rectificado hasta ahora en el taller, para poder agilizar el proceso de rectificado al momento de buscar datos e información del motor que se ha identificado para trabajar, en caso de no tener registrado el motor en la carpeta, se puede buscar su información en alguna de las tablas técnicas disponibles.

Figura 16. Registro de motores del taller

2.3 MAQUINARIA

A continuación, se presentan las máquinas utilizadas en el proceso establecido en la metodología.

2.3.1. MANDRINADORA LINEAL

Una mandrinadora lineal es una máquina-herramienta, la cual se emplea para trabajar con piezas de gran envergadura y exactitud, cuya principal función es la realización de cortes en forma circular mediante un desplazamiento lineal, este tipo de máquinas normalmente son empleadas en diversos sectores industriales debido a su gran eficiencia, versatilidad y alta precisión.

En el sector automotriz son comúnmente ocupadas en el proceso de fabricación y reparación de piezas, mayormente enfocado en el monoblock.

Figura 17. Mandrinadora lineal

Para poder comprender de manera más amplia la mandrinadora, a continuación, se describen las partes más relevantes.

a) Base cremallera

Es el soporte fijo de todo el equipo de la mandrinadora lineal, su principal función es sostener por los extremos cada parte de la máquina, brindando la misma altura de partida para las piezas y una base plana para un buen desplazamiento mediante una cremallera situada a un costado de uno de los rieles de desplazamiento.

Figura 18. Base cremallera

b) Vigas

Son las estructuras que se encargan de soportar la carga y distribuir el peso del monoblock, dependiendo del tamaño de este se quitan o se dejan unas vigas de aumento con las que vienen estas estructuras. Pero en general su función es la de brindar un soporte en el cual se pueda cargar y fijar el monoblock, para evitar un posible movimiento inesperado.

Figura 19. Vigas

c) Puntillas

Estas son las encargadas de sujetar al monoblock e impedir su movimiento al momento de realizar el maquinado, de igual forma estas sirven para el centrado de la bancada del motor, a la vez que se van apretando para ir sujetando el monoblock.

Cuentan con una tuerca de apriete de 15/16 y un tornillo de centrado de 3/4 por cada puntilla.

Figura 21. Puntilla vista lateral

Figura 20. Puntilla

d) Columnas

Las columnas de soporte son las encargadas de cargar y mantener a una cierta altura de la barra, se encuentran situadas a los lados de las vigas sobre la base cremallera, estas también le brindan una lubricación a la barra al momento que esta se encuentra girando mediante unos dispensadores de aceite.

Las columnas cuentan con varias palancas que se encargan de apretar y mover las partes de la misma.

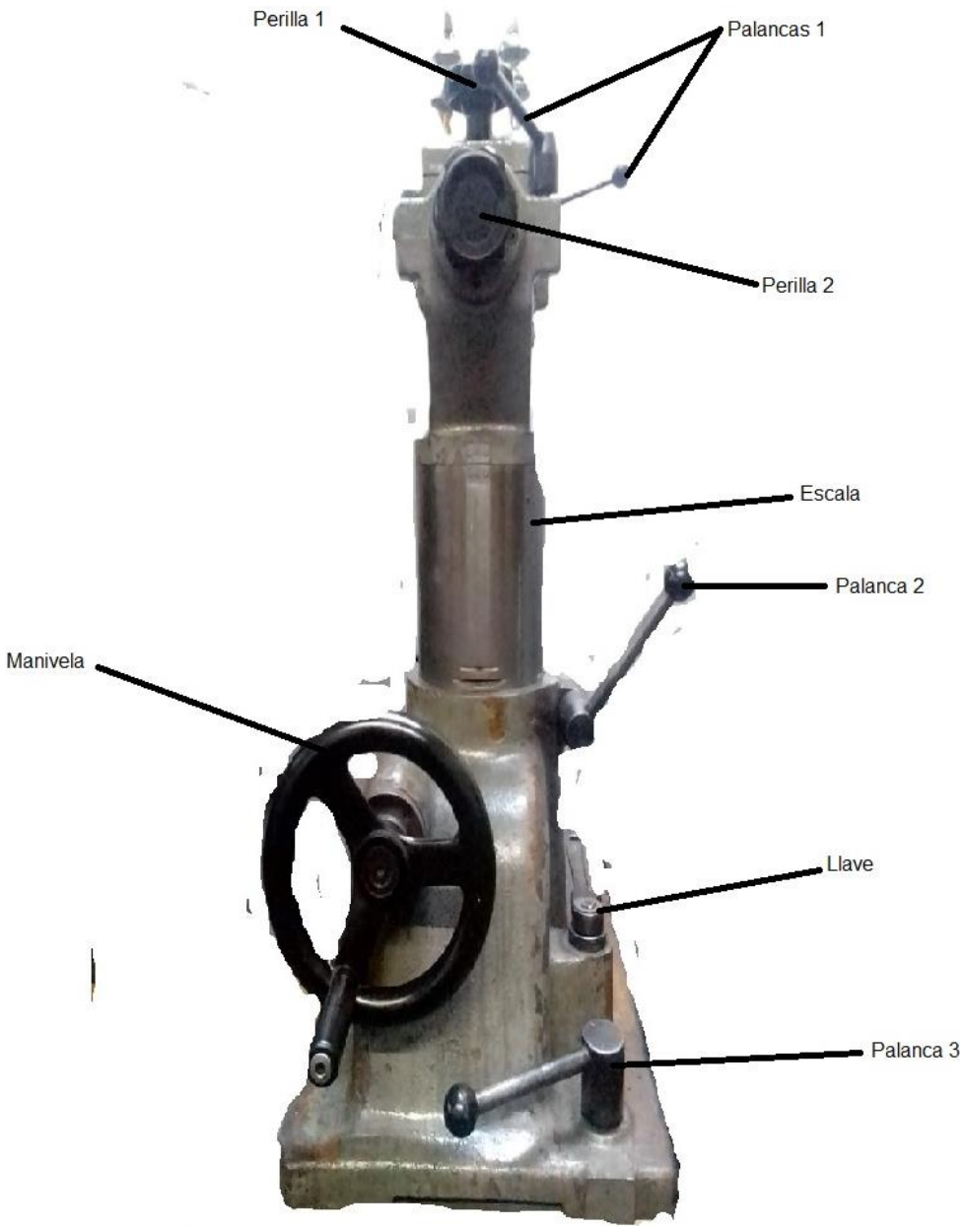


Figura 22. Partes de la columna

Parte	Función
Palancas 1	Aprietan la barra para no generar juego axial en el corte
Palanca 2	Aprieta la columna para que no varié la altura de la barra en el corte
Palanca 3	Fija la columna a la base y evita su desplazamiento
Perilla 1	Aprieta y sujeta los cojinetes de la barra
Perilla 2	Se encarga de desplazar de manera axial la barra para su centrado
Manivela	Es la encargada de ajustar la altura de las columnas
Llave	Se encarga de desplazar por toda la base a la columna
Escala	Sirve para medir la altura de la columna

Tabla 7. Partes de la columna

e) Motor eléctrico

La mandrinadora cuanta con un motor eléctrico de 220 volts, el cual es el encargado de transmitir la potencia de giro al sistema de embrague de la barra y mediante unas poleas de doble ranura se puede cambiar el número revoluciones al que va la barra.

Figura 24. Poleas del motor eléctrico

Figura 23. Motor eléctrico

Motor eléctrico	
Marca	FIMET
Modelo	M80A4
No.	1026132B
Tipo	Trifásico
Voltaje	220
RPM	1680

f) Sistema de embrague

Este es un mecanismo que permite la transmisión e interrupción de energía mecánica de un motor hacia un eje, controlando la marcha y la velocidad de giro.

En primera instancia el motor eléctrico transmite la fuerza de giro mediante una banda y dos poleas de doble ranura, con las cuales se selecciona el juego de revoluciones en el que se desea trabajar.

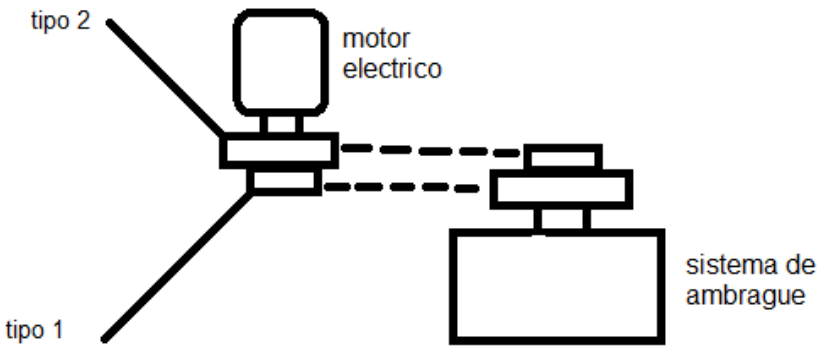


Figura 25. Diagrama del motor eléctrico

Posteriormente mediante una serie de palancas y manivelas se configura la altura, el tipo de avance, las revoluciones, el giro, la dirección y el acoplamiento de la fuerza del motor mediante un eje cardan hacia la barra, de esta manera dando inicio al procedimiento de corte en línea.

1. Altura

La altura se ajusta mediante una manivela que se encuentra debajo del sistema de embrague, unida a su misma columna de soporte y de igual forma que las otras columnas esta cuanta con su escala para medir la altura a la que se va a cortar la bancada.

2. Avance

La máquina cuenta con una manivela y una palanca con las cuales pueden controlar el avance de la barra dependiendo que opción escojamos. La manivela permitirá realizar un avance de forma manual, así que la barra se desplazará dependiendo los giros que le demos a la manivela y ya que está tiene un indicador de avance el cual está divido en 18 partes, cada una equivalente a 0.0040" en desplazamiento, determinamos que en una vuelta completa de la manivela la barra avanza 0.0720"

La palanca tiene la función de iniciar el avance automático, ya sea un avance lento o avance rápido dependiendo como lo pida el corte, esta se encuentra dividida en tres pasos (C, D, E).

C. Colocando la palanca a la izquierda la barra consigue un avance de 0.0030" por cada giro de la barra, obteniendo así un avance rápido.

D. Colocando la palanca en medio el avance de la barra pasa a ser manual, ocupando la manivela para realizar el desplazamiento.

E. Colocando la palanca a la derecha la barra consigue un avance de 0.0015" por cada giro de la barra, cortando por la mitad el avance rápido y obtenido así un avance lento.

	Avance por revolución
C	0.0030"
D	Neutral
E	0.0015"

Tabla 8. Avance de la mandrinadora

3. Revoluciones

Las revoluciones de la máquina se ajustan por medio de dos palancas, las cuales se dividen en seis pasos de forma gradual brindando un mejor control de las mismas y la primera palanca cuenta con dos extremos (A y B).

Al seleccionar el lado A comenzamos con las primeras tres revoluciones a las que podemos acceder, posteriormente cuando se cambia al lado B se puede trabajar con las revoluciones más altas.

La segunda palanca está dividida en tres pasos (I, II, III), de esta forma es posible terminar de seleccionar la revolución con la que vamos a trabajar.

Nota: es importante determinar antes en qué tipo de polea está trabajando el motor eléctrico, ya que de esta manera es posible conocer que juegos de revoluciones tendrá el sistema de embrague.

Palancas	Polea 1		Polea 2	
	A	B	A	B
I	90	340	100	390
II	140	515	155	600
III	216	816	252	960

Tabla 9. Velocidad de giro de la mandrinadora

4. El giro

Este es controlado por un interruptor eléctrico de tres pasos, el cual es el encargado de energizar el motor eléctrico e invertir el sentido del giro del mismo.

Si se coloca el interruptor en el número 1, el sentido del giro de la barra será hacia la derecha y si se coloca en el numero 2 el sentido del giro será hacia la izquierda, pero dejando el interruptor en el número 0 se des energiza la máquina y la barra no gira hacia ningún lado.

Figura 26. Interruptor de encendido

5. Dirección

Esta se ajusta mediante la implementación de una palanca de tres pasos y de manera muy sencilla seleccionamos la dirección del avance que tomara la barra, solamente moviéndola hacia la izquierda o a la derecha, colocando la palanca en medio la barra no determina dirección alguna evitando su avance.

Finalmente habiendo ajustado todas las partes de la mandrinadora, centrado el motor y configurado el sistema de embrague, procedemos a realizar el trabajo de rectificado de la bancada del motor.

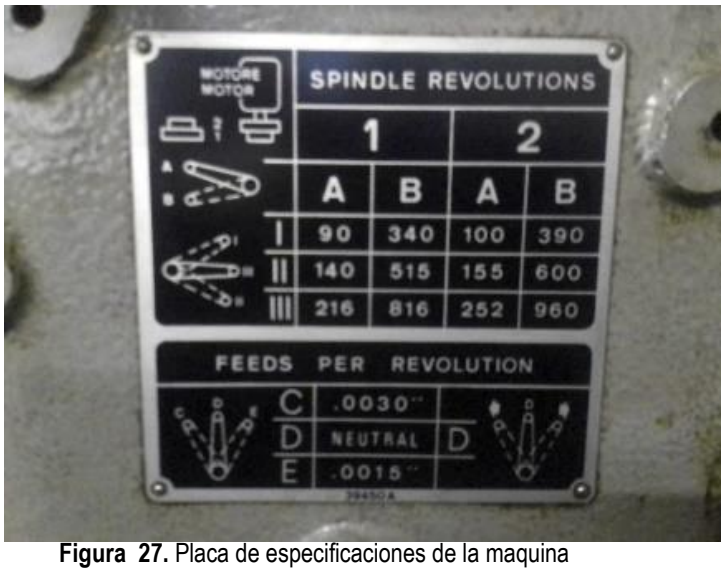


Figura 27. Placa de especificaciones de la maquina

3. DESARROLLO DE ACTIVIDADES

El presente trabajo presenta una guía que permite transmitir el conocimiento alcanzado después de haber comprobado que se alcanza un resultado favorable y funcional en el mantenimiento correctivo automotriz, validando el método científico, ya que repitiendo una y otra vez el resultado es funcional y se espera esté disponible para profesionales y personas que van iniciando en este sector. A continuación, se detalla el paso a paso que ha resultado adecuado para los motores analizados.

VALORACION Y RECONOCIMIENTO DEL MOTOR

Para un maestro mecánico o rectificador (operador de máquina-herramienta) es importante el poder determinar el estado en el que se encuentra el motor y las causas que originaron su fallo, para que de esta manera no se realizarle algún tipo de servicio que probablemente no requiere el motor, evitando gastos innecesarios y reduciendo tiempos de entrega, ya que al realizar un mantenimiento correctivo se tiene que tomar en cuenta ciertos puntos.

Generalmente cuando un motor se desvíela, se amarra o sufre algún tipo de daño interno, suele afectar a un conjunto de partes del tren alternativo, por lo que es común que la bancada del motor y el cigüeñal entre otros, sean partes que suelen dañarse de manera simultánea, generando el mantenimiento correctivo de ambas partes.

Valoración

En el momento que llega un monoblock el operador tiene que realizar una breve inspección de manera visual a este, en busca de algún tipo de desperfecto que no afecte al proceso de rectificado, ya que de ser necesario el trabajo podría aumentar en costo y tiempo para el mecánico.

Lo primero que se revisara va a ser el estado en el que se encuentra la bancada, ya que este es nuestro principal punto de interés siendo el área a rectificar, checamos que se vean bien físicamente el monoblock y que estos no cuenten con fisuras, abrasiones, deformaciones exageradas o ranuras por las cuales se tenga que mandar a soldar y/o rellenar el motor.

Con respecto a las tapas si presentan un problema similar, normalmente se cambian por unas de otro motor y en algunos casos también se mandan a soldar o rellenar.

Estos problemas normalmente suelen ser coaccionados por falta de una buena lubricación, mal mantenimiento y sometimiento de altas revoluciones al motor constantemente, esto repercutiendo directamente en los cojinetes, piezas móviles del monoblock y cabeza de cilindros.

Después el operador debe revisar el estado el que se encuentran los cajones de los axiales, ya que es muy común que estos sufran una abrasión con los mismos cojinetes axiales y tengan que ser rellenados para posteriormente ser maquinados en la mandrinadora.

Los cojinetes axiales tienen como función limitar el juego axial que se genera por la fuerza ejercida del embrague hacia el motor, al momento de desacoplar el disco de embrague y el volante de inercia.

Las principales causas que pueden ocasionar este problema, es un embrague en mal estado o mal armado, una mala lubricación y un montado incorrecto de los cojinetes axiales.

Figura 30. Cajón axial en buen estado

Figura 29. Cajón axial en mal estado

Figura 28. Comparación de tapas

Por último, el operador realizara una inspección a las tapas de bancada, checando que todas cuenten con una buena presión, ya que estas tapas requieren estar bien sujetas para evitar algún tipo de juego que se pueda generar por el movimiento del cigüeñal; de lo contrario se procede a soldar una orilla de la tapa para posteriormente quitar el excedente de material soldado e ir ajustando la presión de la tapa con respecto al monoblock.

Reconocimiento del motor

El distinguir el tipo de vehículo al cual le pertenece el motor que se va a trabajar es muy importante ya que cada marca y motor cuentan con sus propias medidas, a veces suele ser algo complicado debido a que generalmente solo contamos con el monoblock y el cigüeñal teniendo pocas referencias para su búsqueda, sin embargo podemos obtener los datos del motor con el diámetro de los muñones de centros y bielas del cigüeñal, los cuales por si solos suelen ser de mucha ayuda al momento de buscar el motor en las tablas de referencia que marca el fabricante, de igual forma con el diámetro de alojamiento es posible obtener los datos del motor con referencia a las tablas pero al ser un solo dato es un poco más complicado.

Algunos motores suelen traer el logo de la marca y el cilindraje grabados en el monoblock y cigüeñal, con estos datos reducimos el rengo de búsqueda para poder encontrar el motor más rápido, como lo son el caso de las marcas Ford, Volkswagen y Nissan en algunos de sus motores.

La manera más eficiente de encontrar un motor siempre será con el número de parte que viene en la caja de los cojinetes de bancada, se localiza de manera más rápida y precisa el motor, el único inconveniente son las marcas de los cojinetes ya que cada marca manejan sus propios códigos, catálogos y tablas, pero esto ya depende del mecánico y la marca que más le guste trabajar, siendo las más comerciales Clevite, Sealed Power, Glyco, Mopar y Moresa siendo esta última una marca que no solo trae su número de parte, sino también el nombre de los vehículos y sus marcas que las usan.

Figura 31. Tabla de especificaciones con sus respectivos cojinetes **"MORESA"**

Figura 32. Tabla de especificaciones con sus respectivos cojinetes **"CLEVITE"**

CENTRADO Y MAQUINADO

En el procesos de rectificado de bancada del motor es muy importante que los cortes de cada bancada estén lo más alineado posible entre sí y que la máquina esté preparada para el tipo de motor que va a cortar, de igual forma hay que revisar que tanto la máquina como todos sus accesorios se encuentren en buen estado, evitando así el riesgo de realizar un mal corte que posiblemente bloquee el cigüeñal al momento del montado y armado del motor.

Para esta parte del manual se tomará como ejemplo y evidencia tres tipos de motores, siendo los más comunes que llegan al taller para el rectificado de bancada.

- Nissan Tsuru 1.6 4 cil.

- Nissan Urvan 2.5 4 cil.

- Volkswagen Jetta 1.8 4 cil.

Llegando el motor el rectificador empezara por la valoración general, de la cual se determinara el estado en el que se encuentra el motor y así poder descartar problemas que conlleven a un trabajo externo al corte en línea.

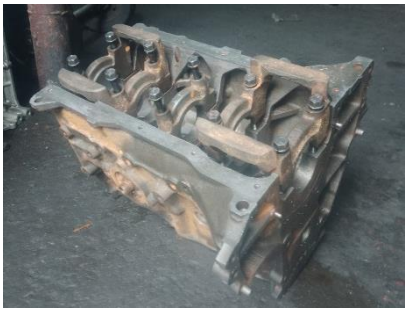


Figura 34. Monoblock de Tsuru 1.6

Figura 35. Monoblock de Jetta 1.8

Figura 33. Monoblock de Urvan 2.5

Un buen método para determinar si la bancada requiere el servicio correctivo es mediante una escala, colocándola de manera horizontal en la bancada del monoblock, posteriormente con el calibrador de lainas se procede a insertar las mencionadas de 0.0003" a 0.0005" entre la escala y cada una de las bancadas; en teoría no debe existir algún tipo de tolerancia pero si llegase a pasar libremente cualquiera de ellas, tal situación indicaría que el motor es considerado apto para un rectificado de bancada y en el caso de que una laina que pasara fuera de un valor mayor a 0.0010" la bancada requerirá de ser rellenada mediante soldadura, de lo contrario el cigüeñal podría tener vibraciones y perdida de presión de aceite, que a su vez generaría desgaste prematuro en los cojinetes de centros del cigüeñal.

La bancada debe traer un acabado liso sin ningún tipo de deformación, abrasión o ranurado tanto el monoblock como las tapas, formando una circunferencia casi perfecta.

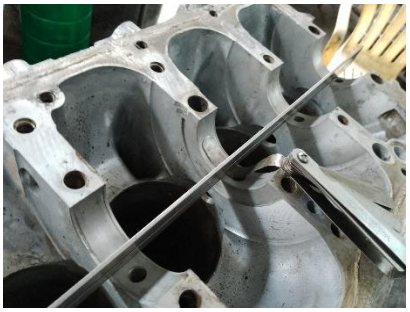


Figura 38. Revisión de Urvan 2.5

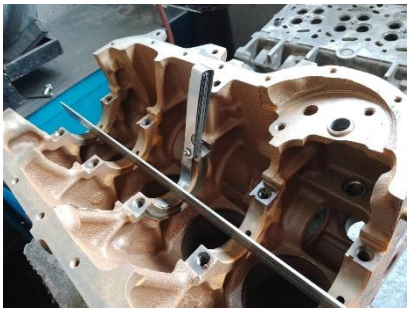


Figura 37. Revisión de Jetta 1.8

Figura 36. Revisión de Tsuru 1.6

Los cajones de los cojinetes axiales deben de estar en buen estado, ya que el fallo más común de los axiales es cuando sufren una abrasión con la bancada y las caras laterales del cigüeñal provocando la pérdida del cajón y un daño al cigüeñal, teniendo que rellenar mediante soldadura toda el área del cajón que se perdió, posteriormente será maquinado para la recuperación del cajón del cojinete axial y posible rellenado o refrentado del cigüeñal [9].

Figura 39. Cajón de Urvan 2.5 **antes/después**

Un método para determinar una buena holgura axial de la bancada es mediante la comprobación de luz (claro de lubricación) con lainas, se tomara como referencia las de dimensiones de 0.0002” y 0.0007”, insertando primero la de valor más pequeño justo entre la cara lateral del cigüeñal y el cojinete axial, esta tiene que entrar de manera libre de lo contrario si se atora o ni si quiera entrara significa que el juego axial es muy poco y se tendría que refrentar las caras del cigüeñal o el cajón, para que tenga la luz adecuada o en su defecto cambiar los cojinetes axiales, posteriormente se procederá a insertar la laina más grande y esta no tendrá que pasar entre el cojinete axial y la cara lateral del cigüeñal, ya que de pasar libremente esta nos indicaría que el juego axial es excesivo ocasionando desgaste prematuro en los cojinetes axiales, una mala lubricación de estos y un mayor riesgo de abrasión de los cojinetes axiales, teniendo que rellenar el cajón para darle mayor altura o en su defecto cambiar los cojinetes axiales.

Figura 40. Revisión de la holgura axial

En el proceso de rellenado de un monoblock de hierro se va a utilizar la soldadura por arco eléctrico SMAW (Shielded Metal Arc Welding) *"soldadura convencional"* con un electrodo (utp8 44222) de 1 a 4 electrodos dependiendo el área a rellenar, utilizando la mayor cantidad para una bancada, realizando varios cordones partiendo del centro de esta hacia los extremos para ir dando aumento de material, en el caso del cajón solo se realizara un cordón lo suficientemente grueso cubriendo toda el área dañada por el cojinete axial y se le colocaran dos puntos en la cara de la tapa dañada para evitar que los axiales se vayan a salir o girar [10].

Para un monoblock de aluminio se va a ocupar el proceso de soldadura TIG (Tungsten Inter Gas) utilizando de 1 a 2 varillas de aporte para soldadura (4043 1/8) ocupando la mayor cantidad para la bancada y el mínimo para rellenar el cajón del axial, de igual forma procedemos a colocar dos puntos en la tapa del medio para evitar que se giren los axiales utilizando un electrodo (utp8 44222) [10].

Figura 42. Cajón de Tsuru 1.6 rellenado

Figura 41. Cajón de Urvan 2.5 rellenado

Figura 43. Cajón de Jetta 1.8 rellenado

Una vez terminada la valoración y el rellenado de bancada del monoblock, procedemos a montarlo en las vigas y ajustamos la máquina a las dimensiones que tiene el monoblock.

El monoblock se colocará de manera invertida, dejando la cara de los cilindros a bajo, montada sobre las vigas y la bancada en la parte de arriba, colocando el frente del monoblock a lado de la primera columna.

En algunos casos el monoblock trae puestas las guías de la cabeza, esto en los Volkswagen y Nissan, dificultando su centrado y teniendo que retirarlas en ocasiones para tener un buen desplazamiento del monoblock sobre las vigas [11].

Figura 45. Guías de monoblock de Tsuru 1.6

Figura 44. Guías de monobloc de Jetta 1.8

Se ajusta la máquina de acuerdo al tipo de motor que se va a trabajar. Se tomará como dato principal el número de parte que traen grabados los cojinetes, de esa forma se encuentran más fácil los datos del motor en las tablas de especificación del fabricante, nuestro registro de motores ya trabajados o en los catálogos de las marcas [12], este dato nos brinda todo lo que necesitamos saber para poder cortar el motor.

Figura 47. Cojinetes de Urvan 2.5

Figura 46. Cojinetes de Jetta 1.8

Figura 48. Cojinetes de Tsuru 1.6

Al no contar con el número de parte se realiza la búsqueda en las tablas y catálogos en base a los datos que si podemos obtener del motor, como lo pueden ser el diámetro de alojamiento de la bancada, diámetros de los muñones de centro y bielas del cigüeñal, número/alineación de cilindros y en algunos motores se pueden traer grabado la marca y cilindrada del vehículo, posteriormente se procede a

realizar la comparación de datos obtenidos con los datos que nos arrojan las tablas y catálogos de cada uno de los motores, de esta manera el motor que más coincida con nuestros datos será el que se tomará como referencia para realizar el corte.

Figura 50. Tsuru 1.6 montado en vigas

Figura 51. Urvan 2.5 montado en vigas

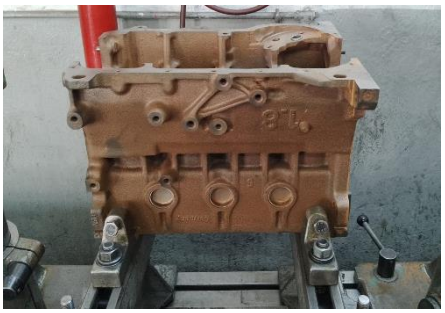


Figura 49. Jetta 1.8 montado en vigas

Una vez arriba el motor ajustamos la altura de las columnas, alineándolas con la bancada del motor, luego procederemos a insertar el extremo de la barra con la conexión al cardan por el cojinete de la primera columna, pasando por el medio de toda la bancada del motor hasta salir por el otro cojinete de la otra columna, posteriormente se procederá a dar el par de apriete (torque) a la primera y última tapa del motor esto en el caso del Nissan Tsuru 1.6 y Volkswagen Jetta 1.8, para el Nissan Urvan 2.5 serán todas las bancadas ya que las tapas están unidas en una sola pieza, de esta manera centraremos a los extremos del motor para dar la alineación que debe llevar el corte [12].

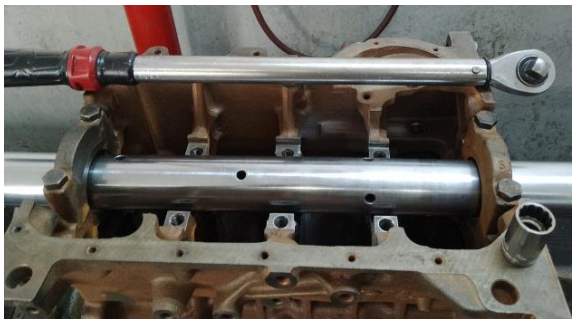


Figura 53. Montado de tapas de Jetta 1.8

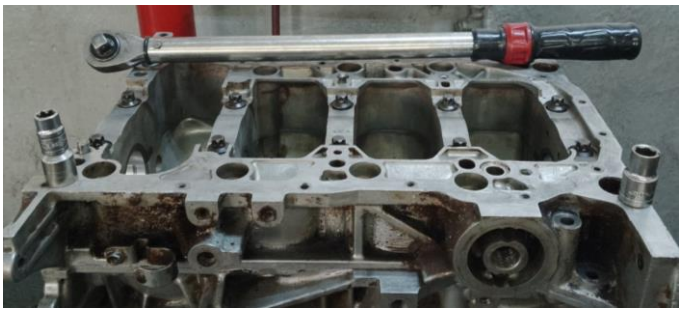


Figura 52. Montado de tapas de Urvan 2.5

Figura 54. Montado de tapas de Tsuru 1.6

	Dado para tapa de bancada	Par de apriete (Torque)
Nissan Tsuru 1.6	½ hexagonal	$34\ a\ 38^{lb/ft}$
Nissan Urvan 2.5	Int. E-14 (e-torx)	$29^{lb/ft}$+60°
	Ext. 12mm hexagonal	$20^{lb/ft}$
Volkswagen Jetta 1.8	14mm hexagonal	$47\ a\ 50^{lb/ft}$

Tabla 12. Datos técnicos de la bancada de motores

Los cojinetes de la barra se cambiarán dependiendo el motor, por ejemplo, el Volkswagen Jetta 1.8 y el Nissan Urvan 2.5 utilizan la barra grande para el corte, pero el Nissan Tsuru 1.6 tiene que ocupar la barra pequeña, de esta manera se evitan inconvenientes al cortar y medir por falta de espacio entre la barra y la bancada.

El motor que tenga el diámetro de su bancada menor a 2.300" requerirá el uso de la barra pequeña, pero de ser mayor el diámetro a este mismo valor se podrá implementar el uso de la barra grande para el rectificado.

	Urvan 2.5 2.321"	Tsuru 1.6 2.113"	Jetta 1.8 2.323"
Barra grande 1.968"	X		X
Barra chica 1.181"		X	

Tabla 10. Selección del tipo de barra

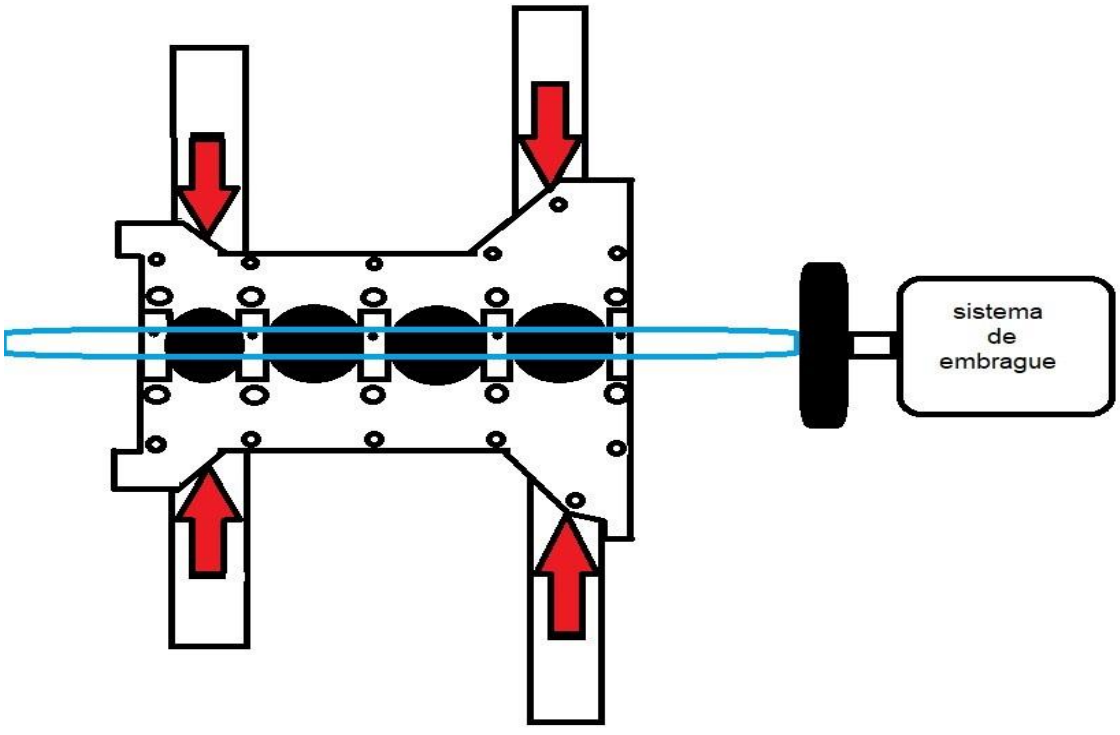
El cambio del cojinete para la barra es muy sencillo, aflojando la perilla 1 se suelta la mitad de la base que lo sujeta, posteriormente levantamos esa mitad dejándola recargada en el otro lado de la columna para así tener acceso a los cojinetes y poder realizar el cambio, estos cojinetes cuentan con una posición de uso ya que de no colocarse correctamente, se corre el riesgo de no tener una buena

lubricación en la barra al momento de estar trabajado, ya una vez colocados los cojinetes adecuados procederemos a seguir con el centrado.

Figura 55. Pasos para cambiar cojinete de barra

Una vez insertada la barra y con el torque adecuado las tapas se aflojan, las palancas 1 y 2, para liberar el juego de las columnas y poderlo centrar evitando que la misma barra se llegue a inmovilizar por esta acción.

Se debe cuidar en la barra que la distancia de los barrenos donde se monta el buril y el micrómetro sea la necesaria para pasar por cada bancada del motor, de lo contrario se tendría que desplazar tanto las columnas como las vigas hacia un lado, brindando la distancia que necesita para realizar el corte y medición de la bancada.



Posteriormente se colocan las cuatro puntillas en las vigas cuidando que estas estén apuntando una parte sólida y estable del monoblock, de ser necesario se le puede colocar calzas o placas en ciertas posiciones para que brinde estabilidad al momento de apretarlas y se tenga un mejor control del centrado, una vez identificado los puntos de apriete del monoblock se acercaran las puntillas y posteriormente después de realizar el primer centrado procederemos a apretarlos.

Dado que el material que manejan los motores de aluminio es muy frágil, se necesita colocar a los tornillos de centrado unos protectores para evitar dañar el monoblock al momento de centrarlo. En motores de fierro no son tan necesarios, ya que estos motores tienden a ser más resistentes a la fuerza ejercida por las puntillas.

Figura 57. Protectores de puntillas

Es necesario colocar en ambos extremos del motor montados sobre la barra los conos, dando de esta forma el primer centrado con el cual se buscara que la desalineación del motor con respecto a la barra se mínima y al momento de utilizar las puntillas demos el último centrado.

Figura 59. Conos montados en Urvan 2.5

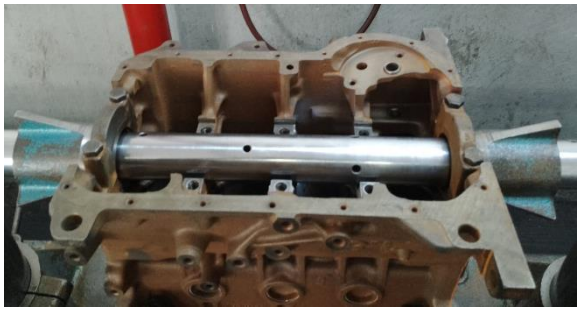


Figura 58. Conos montados en Jetta 1.8

Una vez que ya se colocaron los conos y se redujo la desalineación que hay entre bancada y barra procederemos a colocar las puntillas, cuidando que cada una de estas apoye en un lugar sólido y estable, de lo contrario será más difícil el centrado del monoblock y puede que este se mueva al momento de ser trabajado. Para los puntos donde no se cuenta con una buena superficie para recargar las puntillas, ahora hay que colocar unas pequeñas placas de acero para brindar una superficie sólida en la cual apoyar las puntillas.

Figura 61. Placas para brindar estabilidad a las puntillas

Cuando las puntillas ya están colocadas y los tornillos de centrado estas retrocedidos completamente se le darán 3 vueltas de apriete a estos, luego se apretarán la tuerca de sujeción de las puntillas con la llave 15/16 y después con la llave 3/4 se procede a apretar los tornillos de centrado de manera paralela e ir sujetando el monoblock.

Figura 62. Pasos para ajustar puntillas

Se procede a montar los centradores de reloj sujetándolos de la barra y colocándolos en la primera y última bancada del monoblock, posicionando la aguja en la orilla de la bancada para que de esta forma chequemos que tan desalineada esta la bancada con respecto a la barra.

Figura 63. Centradores de reloj

Mediante estos centradores buscaremos llegar a una alineación más precisa en los 4 puntos de la bancada (arriba, abajo y a los lados). Conforme apretemos una punta de centrado será necesario que se afloje la punta del lado opuesto para que de esa manera podamos desplazar el monoblock hacia el lado más descompensado de la bancada. Una vez teniendo los cuatro puntos de la bancada a una misma distancia con respecto a la barra, procederemos a bajar las columnas de la barra de 5 a 10 milésimas, con el objetivo de poder limpiar toda el área de la bancada que por deformación, desbaste o abrasión no alcanzaría a salir con la altura inicial de las columnas.

Posteriormente se procede a apretar todo el juego de palancas para inmovilizar el movimiento de las columnas, siempre checando que no se pierda el centrado a los lados de la bancada y el desplazamiento hacia abajo de 5 a 10 milésimas de las columnas, luego se corrobora que el centrado este bien mediante el uso del buril, se colocará en la barra con la altura correcta para cortar el diámetro de la bancada y se procederá a revisar que nuestro buril pase rallando a los lados de cada una de las bancadas del monoblock y en la parte de abajo se incruste la punta del buril, esto por el desplazamiento hacia bajo que tienen las columnas, checando que de esta manera el corte limpie toda área de la bancada unida al monoblock.

Figura 65. Pasos para inmovilizar columna

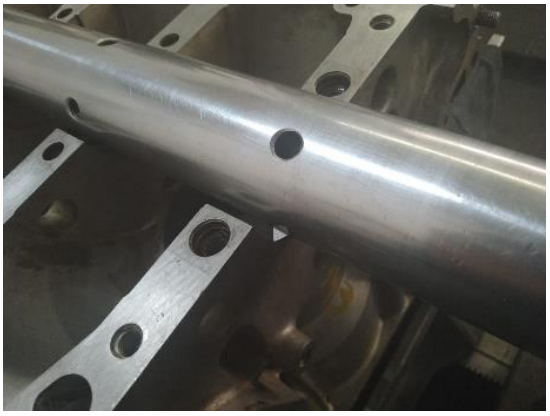


Figura 64. Comprobación del centrado

En esta etapa hay que rebajar todas las tapas de bancada en la máquina para desbastar tapas; en los motores Tsuru 1.6 y Jetta 1.8 es necesario rebajar mínimo 15 milésimas a cada una de las tapas, ya que de lo contrario se puede correr el riesgo de que el buril no corte bien toda el área de estas, dejando tapas a la mitad de corte o sin cortar.

Figura 67. Tapas de Tsuru 1.6

Figura 66. Tapas de Jetta 1.8

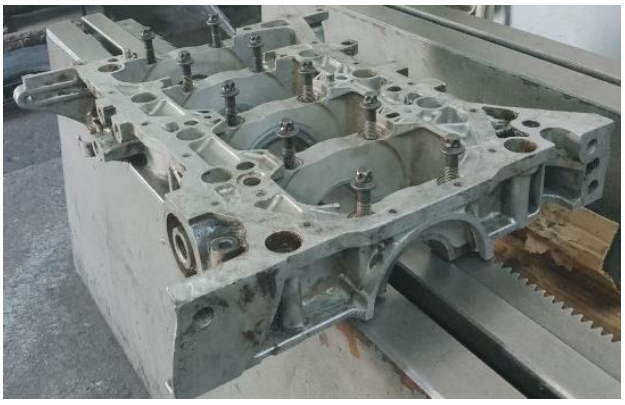


Figura 68. Tapa de Urvan 2.5

Para las tapas del Tsuru 1.6 y Jetta 1.8, se requiere del uso de una máquina para desbastar tapas de bancada, de esta forma logramos que nuestras tapas sean acortadas de manera uniforme, evitando un mal corte. Lo primero que vamos a revisar es la altura de todas las tapas mediante el uso del vernier, después se le quitaran 15 milésimas a cada tapa, mínimo para que el buril pueda limpiar toda el área de la tapa en el último corte.

Figura 69. Placa para tapa doble de Tsuru 1.6

Figura 70. Altura de una tapa de bancada

Se sujeta la tapa en una pequeña prensa la cual está enfrente de una piedra de desbaste, por medio de un indicador de caratula se alinean los extremos de la tapa para que su desbaste no termine descompensado y una vez ya preparada la tapa encendemos la máquina y le iremos metiendo dos milésimas por cada pasada de la piedra hasta conseguir las 15 milésimas.

Figura 72. Proceso de desbaste para tapas de Tsuru 1.6 y Jetta 1.8

Figura 71. Comparación de tapas rebajadas

Para la Urvan 2.5 es necesario realizar el proceso de careado o refrentado, ya que esta al ser una sola pieza requiere ser rebajada de manera uniforme, por lo cual se implementará el uso del torno para su maquinado. La tapa cuenta con unas guías las cuales cumplen la función de mantener la alineación de la bancada al momento de montarla en el monoblock, por lo que es necesario retirarlas antes de maquinar la pieza, para eso necesitaremos un machuelo de 1/2 el cual usaremos para hacerle una cuerda interna a las guías de la tapa y cuando parte del machuelo este adentro de la guía, con un tornillo de la bancada y un martillo de goma lo empujaremos por el lado contrario de la tapa hasta sacarlo y con cuidado desenroscaremos la guía del machuelo, realizando lo mismo a cada guía.

Figura 73. Tornillería y herramienta para la bancada de Urvan 2.5

Después se monta la tapa en el chuck del torno y con ayuda del indicador de caratula vamos a centrar la tapa, usaremos un martillo de goma para ajustar las esquinas de la tapa dejándolas a la misma distancia, posteriormente ponemos a girar la pieza y revisamos que esta no pierda el centrado, para después rebajar las 15 milésimas necesarias a la tapa de bancada.

Una vez ya rebajadas las tapas se procede a montarlas en el block para corroborar que nuestro buril si podrá cortarlas de manera completa.

Figura 74. Proceso para desbastar Tapa de Urvan 2.5

Figura 75. Tapa de Urvan 2.5 rebajada

Ya que todo éste preparado es necesario maquinar primero el cajón del cojinete axial, esto en caso de haber sido rellenado con soldadura.

Para este procedimiento se ocupan dos tipos de buriles, uno de pastilla cuadrada y otro de corte derecho, para la elaboración de la parte del cajón donde descansara el cojinete de juego axial, para ello se utiliza el buril de pastilla cuadrada, con este buril se retira todo el exceso de soldadura hasta llegar a la distancia que necesita el cigüeñal para tener una buena holgura axial. La medida ideal tiene que ser de 3 a 4 milésimas menor que la distancia que hay entre las caras laterales del muñón de bancada, esto tomando en cuenta el grosor de ambos cojinetes axiales, ya que estos serán los que estará directamente en contacto con las caras laterales del muñón y por lo tanto es importante que cuenten con una luz mayor a 2 y menor a 7 milésimas. Posteriormente se requiere montar el buril de pastilla cuadrada y mediante la ecuación se determina la altura que debe de llevar este para poder cortar el cajón de acuerdo a las medidas del cojinete axial.

Figura 78. Buril de pastilla cuadrada

Figura 77. Buril de corte derecho

Figura 76. Buril de punta a 80°

FORMULA	$\frac{\text{Ø Cojinete Axial} - \text{Ø Barra}}{2}$
EJEMPLO (Tsuru 1.6)	$\frac{2.790" - 1.181"}{2}$ ***Altura del buril= 0.804"***

Tabla 11. Fórmula para maquinar el cajón axial

En el sistema de embrague se usará la polea 1 seleccionando la primera palanca en el lado A y la segunda palanca en el primer paso, posteriormente se embraga la barra al eje cardan y seleccionamos un avance D (neutral), después escogeremos el giro de la barra de acuerdo al sentido de corte del buril, para finalmente ir dando el avance de forma manual y de manera lenta con cuidado vamos a ir metiendo cortes de 0.004".

Figura 79. Maquinado del cajón axial

Una vez ya maquinado el cajón se rebaja la parte exterior del mismo, se monta en la barra el buril con pastilla de corte derecho a la misma distancia que el resto del cordón de soldadura y de la misma forma

se van a ir metiendo cortes de 0.004” hasta que la mitad del cojinete axila pueda sobresalir del cajón, permitiendo un buen paso de lubricación en los canales de los cojinetes axiales y evitando que parte del cordón de soldadura tenga algún contacto con el cigüeñal provocando que este no cuente con giro libre.

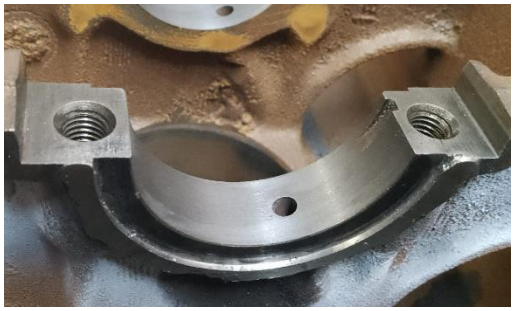


Figura 81. Cajón axial en monoblock de fierro

Figura 80. Cajón axial en monoblock de aluminio

Después de maquinar el cajón, se realiza un corte en línea, para esto es necesario cambiar los buriles de pastilla cuadrada y de corte derecho por un buril de punta de 80°, este se colocara en un barreno de la barra y con el calibrador se le dará la altura al buril, la cual tiene que ser 15 milésimas menor a la altura final para no empezar con un corte fuerte y posteriormente se subirá la altura 5 milésimas después de cada corte a las bancadas, hasta llegar a la medida completa y de esta forma meter el corte final.

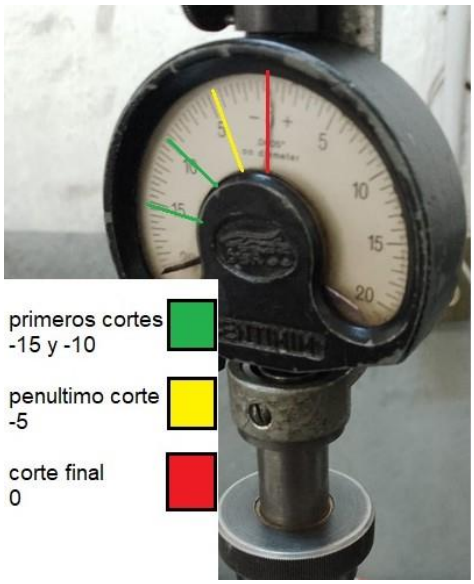


Figura 82. Escala de cortes en el palpador

La configuración del sistema de embrague será con la polea 1, seleccionando la primer palanca en la sección A y colocando la segunda en la velocidad 3, después se conecta la barra al eje cardan y embragamos el sistema, luego encendemos la máquina seleccionado el giro según el sentido de corte del buril, utilizando un avance rápido la tercera palanca se colocara en la sección C.

Al momento de estar cortando se ocuparán las extensiones para darle un mayor alcance a todas las bancadas del monoblock, la extensión grande, se ocupará cuando el corte se realice en las bancadas 1 y 2 ya que son los puntos más alejados del sistema de embrague; para la bancada 3 a veces se ocupará la extensión chica, ya que en ocasiones la distancia que se extiende la flecha es suficiente para cortar tanto las bancadas 5 y 4 como la 3.

En el primer corte el buril únicamente cortará la parte de en medio de las tapas y de esta manera conforme aumente 5 milésimas de altura a cada corte del buril, este se irá ampliando el área rectificada en las tapas. Una vez que se llegue a 5 milésimas de la medida final, el buril ya no solo cortara las tapas, si no que ya empezara a tocar en medio de la otra mitad de la bancada.

Figura 83. Escala de cortes en la bancada

En el momento que el buril ya empieza a cortar ambas partes de la bancada, se procede a medir con el micrómetro, determinado de manera exacta cuantas milésimas le faltan al buril para cortar la medida completa del diámetro de alojamiento de la bancada.

Ya para el último corte se procede a subirle las milésimas que le hagan falta a nuestro buril y cambiaremos de posición la palanca 2 a la velocidad 1, con la cual se va a dejar cortar un tramo de “0.160 a la bancada, mismo tramo donde se va a realizar la medición con nuestros 2 micrómetros, checando que el diámetro de la bancada sea el correcto y posteriormente se vuelve a dejar ir ese mismo corte en lo que resta de la bancada, en caso de que el corte no hay abierto completo el diámetro de la bancada se vuelve a pasar ese mismo corte en velocidad 3, de esta manera se logra abrir de 2 a 3 milésimas más al diámetro, de igual forma al mismo buril se le puede subir las milésimas que

necesite para terminar el corte. Se realiza de la misma manera este procedimiento a las bancadas restantes, terminado así el proceso de maquinado.

CALIBRACION Y MEDICION

En los buriles es posible utilizar dos instrumentos de medición para darle la altura requerida, esto dependiendo del tipo de maquinado que se va a realizar ya que cambia el método de calibración según el buril que se ocupe.

Buril de desbaste cuadrado o corte derecho

En primera instancia para el maquinado del cajón se requiere obtener la altura del buril, para lo cual primero tenemos que determinar cuál es el diámetro total del cajón o cojinete axial, estos datos se pueden encontrar en base a las tablas de especificaciones o catálogos de las marcas, de igual forma es posible medir el cajón o el cojinete mediante el uso de un instrumento de medición para conseguir el mismo dato, posteriormente con la implementación de la ecuación h= $(\frac{ØCA - ØB}{2})$ se obtiene la altura del buril.

Una vez que determinando la altura del buril, se debe utilizar el vernier situando al nonio a la medida que tenemos como altura, colocamos el seguro para evitar que se mueva la escala y luego ocupando el vástago de profundidad se va comparando con el buril mientras este se va sacando sacando del barreno de la barra hasta que iguale su distancia, posteriormente con mucho cuidado vamos a apretar el perno prisionero del barreno con la llave Allen 5/32 para fijarlo y ya meter corte.

Para el resto del cajón basta con incrementarle a nuestro buril las milésimas que sean necesarias para poder cubrir todo el cordón de soldadura sobrante y de esta forma poder cortar la parte exterior del cajón del cojinete axial.

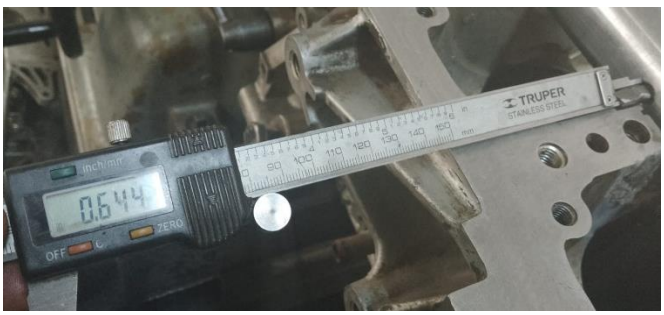


Figura 84. Calibración de buril para maquinado de cajón axial

Buril de corte punta 80°

De igual forma que en el punto anterior lo primero que se necesita es obtener la altura de nuestro buril, para esto es necesario saber cuál es el diámetro de alojamiento de nuestra bancada, el cual lo podemos conseguir mediante el uso de las tablas o catálogos y posteriormente utilizaremos la ecuación h2=$(\emptyset Ban - \emptyset Bar)$ para determinar la altura.

Antes de calibrar el buril hay que ajustar el palpador de acuerdo a la barra que se planea utilizar, posicionándolo el barril de calibración en 0 y aflojando la tuerca de ajuste, una vez que la tuerca esta floja y el palpador está libre vamos a bajar su punta recargándola en la barra hasta que la aguja del indicador de caratula se eleve a 0, de esta manera no se tendrán problemas al momento de usarlo, ya que el tener un palpador mal ajustado afecta a la calibración del buril, posteriormente se posicionara el barril de calibración con la altura que obtuvimos mediante la ecuación.

Insertaremos nuestro buril en un barreno de la barra y encima de este se colocará el palpador ya ajustado con la medida, con cuidado vamos a sacar el buril y mediante la caratula del palpador se deja a unas 15 milésimas debajo de la medida completa y se aprieta su peno para iniciar el corte, posteriormente se le van a subir de 5 milésimas hasta llegar al corte final.

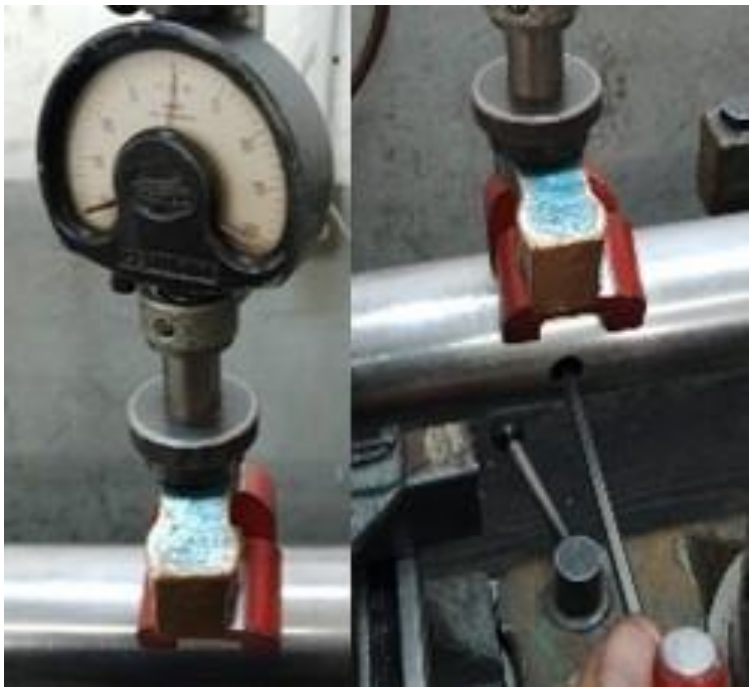


Figura 85. Calibración de buril para rectificado de bancada

Medición

Para realizar la comprobación del corte de la bancada se ocupa dos procesos de medición, con el primero se determina el diámetro total de la bancada una vez terminado el corte, mediante el uso de instrumentos de medición y el segundo se utilizará como comprobación de la medida, ya con el cigüeñal y metales montados obtendremos la luz que hay entre el muñón del cigüeñal y el metal de la bancada.

Micrómetros

En este proceso de medición es necesario apoyarse de dos micrómetros los cuales son de arco y resorte respectivamente, el micrómetro de arco debe contar con un rango de medición en el que pueda abarcar la medida de la bancada y para el micrómetro de resorte se necesita ajustarlo para poder realizar una buena medición.

Figura 86. Micrómetros utilizados en el rectificado de bancada

Una vez que nuestro buril ya pasó por una parte de la bancada con la medida final procedemos a medir, se introducirá el micrómetro de resorte en un barreno de la barra y se insertara en el área ya cortada de la bancada, hay que revisar que este toque la parte de abajo de la bancada con su extremo fijo, ya que de lo contrario se podría calzar con la misma barra y la medida sería errónea, luego es necesario aflojar el prisionero que mantiene sujeta la puntilla para que accione el resorte elevando esta misma a la parte superior de la bancada, después se vuelve a apretar el prisionero para fijarla y retirar el

micrómetro de la barra, se mide con el micrómetro de arco consiguiendo así la medida de la bancada recién cortada.

Figura 87. Proceso de medición de bancada

Plastigage

Para el uso de la herramienta plastigage es necesario que se monten los cojinetes de la bancada que se van a utilizar, ya que para el uso de esta herramienta se requiere del armado del cigüeñal.

En primera instancia se realizará la limpieza de todas las bancadas y sus tapas, estas no deberán de tener ninguna rebaba ya que esta puede causar problemas al momento de armar la bancada del motor, ocasionando que se dañen los metales o limitando la movilidad del cigüeñal al no contar con un giro libre.

Posteriormente se colocan los metales que van en la parte de la bancada del monoblock, estos se identifican por contar con un ranurado y un orificio por el cual la vena de lubricación va a suministrar el aceite formando una película que recubra todo el muñón del cigüeñal. En algunos casos para las tapas de bancada los metales no cambian, pero generalmente son lisos y no cuentan con algún tipo de canal, limitando su función a la de una pista donde circulara el aceite.

Figura 88. Proceso de armado de bancada

La mayoría de cojinetes también cuentan con una muesca de ubicación, la cual es importante para determinar la posición que llevan y evitan un desplazamiento axila o rotativo cuando el motor está en funcionamiento.

Una vez ya colocados los cojinetes que van en el monoblock, se procede a su lubricación con aceite para después montar el cigüeñal evitando así que los muñones tengan un contacto en seco con los cojinetes, ya que esto los puede dañar.

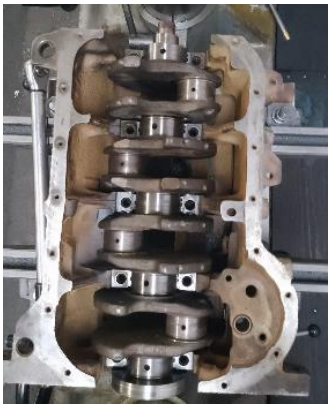


Figura 91. Armado de Jetta 1.8

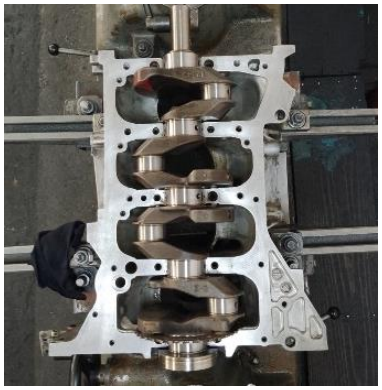


Figura 90. Armado de Urvan 2.5

Figura 89. Armado de Tsuru 1.6

Posteriormente ya montado el cigüeñal colocaremos un pedazo pequeño de la tira del platigage en cada muñón de centro, siempre tomando como guía la dirección del mismo cigüeñal para contar con una buena lectura, luego se colocaran los cojinetes restantes en las tapas y se procederá a ensamblar sin aceite en el conjunto de la bancada, siempre checando el orden y posición de las tapas, para esto se va a dar el torque requerido indicado en la ficha técnica, esta actividad se realiza con mucho cuidado ya que esta herramienta necesita que las piezas móviles mantengan una sola posición para obtener una buena medida del claro de lubricación.

Figura 92. Proceso de implementación del plastigage

En el punto siguiente se retiran las tapas y el plastigage se tendrá que haber deformado, en este punto es necesario comparar la escala que viene impresa en la envoltura del platigage con el ancho de la zona aplastada, obteniendo así la luz que tiene el motor.

La holgura aceptable que debe tener un motor es de 1 milésima por cada pulgada de diámetro del muñón, siendo generalmente de 0.0015” a 0.002” el claro de lubricación que llevan los vehículos a gasolina y uno que otro diésel.

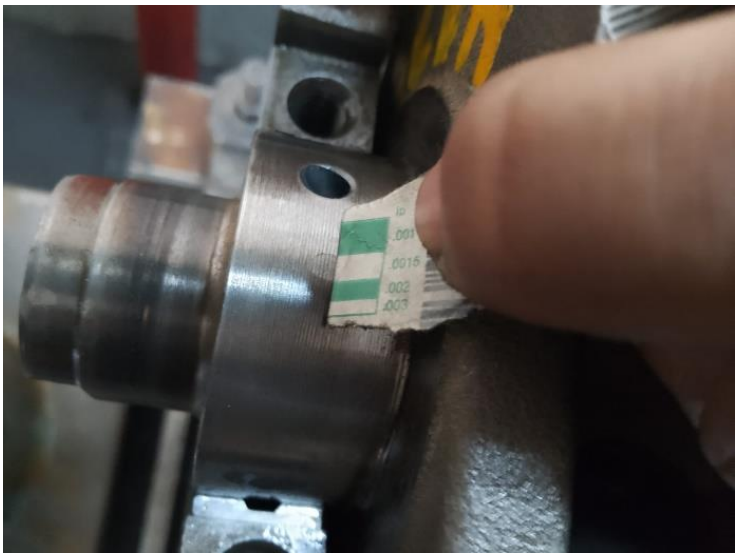


Figura 93. Proceso de medición del plastigage

TERMINADO

Una vez que ya se rectificó la medida y luz de nuestra bancada, se procede a realizar el armado de la bancada, para verificar que el trabajo sea adecuado y no muestre algún inconveniente.

Primero se retira el platigage marcado en los cojinetes y muñones del cigüeñal esto de una manera cuidadosa evitando dañar alguna de las partes, mediante unos trapos y solventes se van retirando, cuidando de no dejar algún residuo o rebaba.

Después se lubrican bien los lados expuestos de los muñones de centro del cigüeñal, para posteriormente montar las tapas de bancada en orden que lleva y con su respectivo torque.

Finalmente, la bancada debe contar con un giro libre del cigüeñal al momento de aplicarle una fuerza de torsión con nuestras manos y no debe tener un juego axial excesivo.

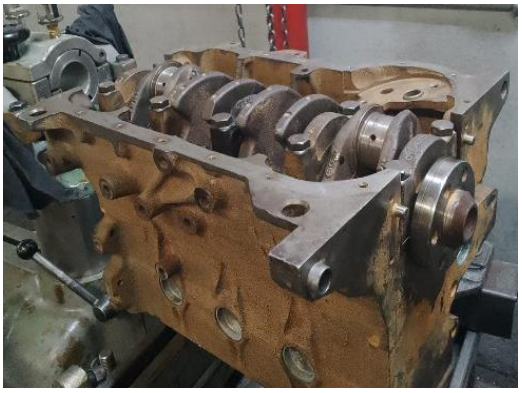


Figura 96. Monoblock de Jetta 1.8 terminado

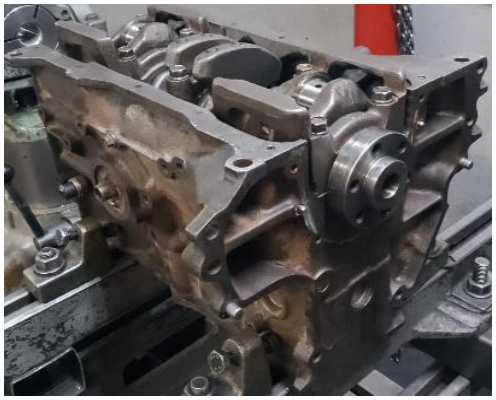


Figura 95. Monoblock de Tsuru 1.6 terminado

Figura 94. Monoblock de Urvan 2.5 terminado

De esta manera se concluye con el rectificado de una bancada (corte en línea) habiendo abarcado punto por punto todos los pasos que se deben de seguir al momento de realizar este tipo de servicios correctivos a un motor de combustión interna (Urvan 2.5, Tsuru 1.6 y Jetta 1.8) [13, 14].

Mediante los datos obtenidos en la investigación se comprobó y determino la viabilidad del proceso de rectificado, logrando recuperar las tolerancias originales de la bancada en repetidas ocasiones. Presentando los resultados del proceso en las tres bancadas en las tablas 14, 15 y 16.

Nissan Tsuru 1.6 (4 cil.)				
El vehículo presento problemas en el sistema de lubricación, siendo esta la principal razón que ocasiono el daño en la bancada				
	Ø Bancada	Ø Muñón de cigüeñal	Claro de lubricación (Luz)	Holgura axial
Tolerancias del fabricante	2.112”	1.967”	0.001”-0.003”	0.002”-0.004”
Desgates del Monoblock	2.114”	1.965”	0.005”	se perdió el cajón axial
Monoblock rectificado	2.112”	1.957” se rectificó en 0.010	0.002”	0.003”

Tabla 14. Resultados del monoblock Nissan **Tsuru 1.6**

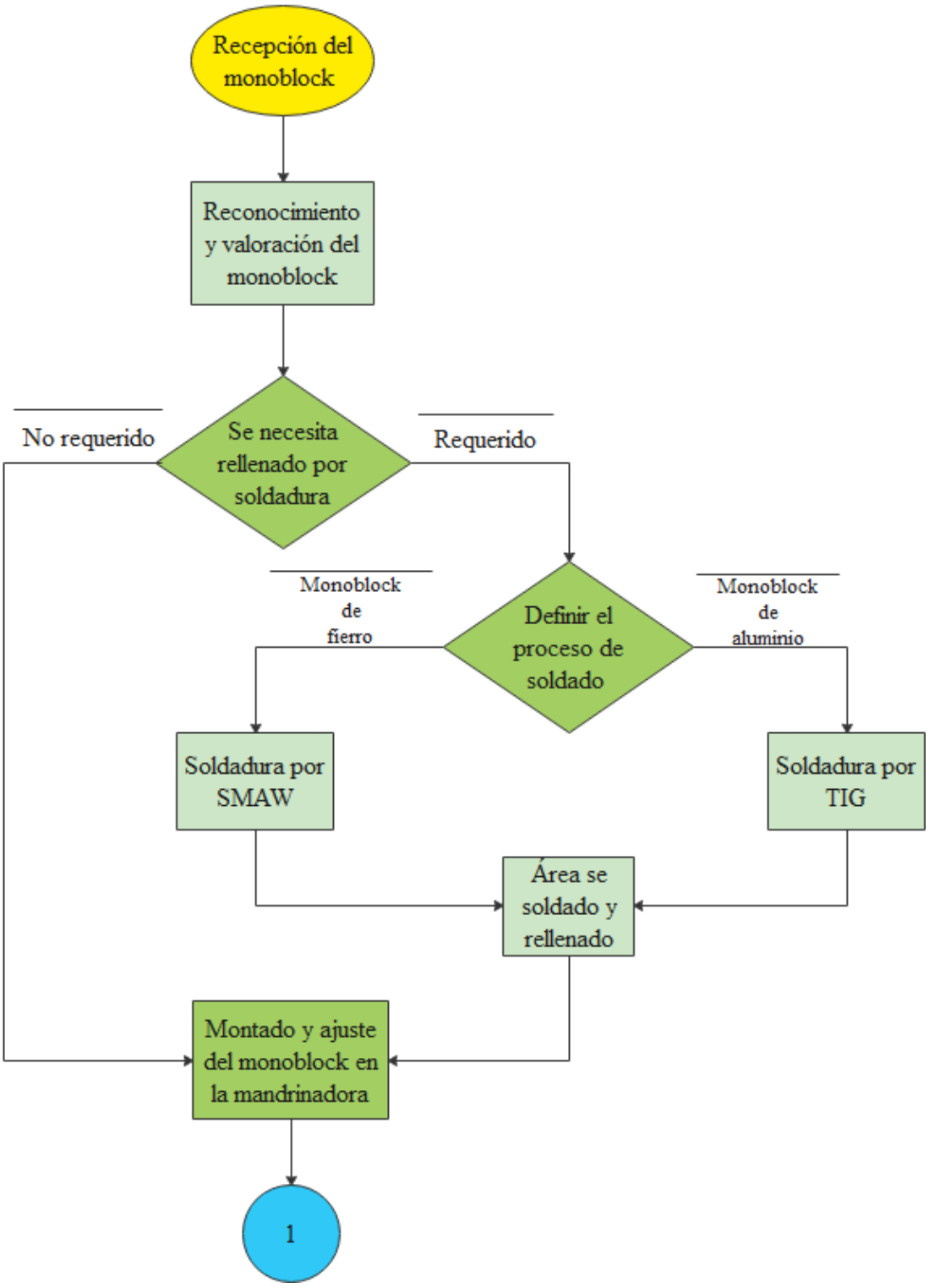
Nissan Urvan 2.5 (4 cil.)				
El vehículo presento un pronto desgaste en los cojinetes por el tipo de trabajo que realiza la unidad y un mal mantenimiento preventivo.				
	Ø Bancada	Ø Muñón de cigüeñal	Claro de lubricación (Luz)	Holgura axial
Tolerancias del fabricante	2.321”	2.164”	0.001”-0.002”	0.002”-0.005”
Desgates del Monoblock	2.322”	2.163”	0.004”	se perdió el cajón axial
Monoblock rectificado	2.321”	2.163”	0.002”	0.004”

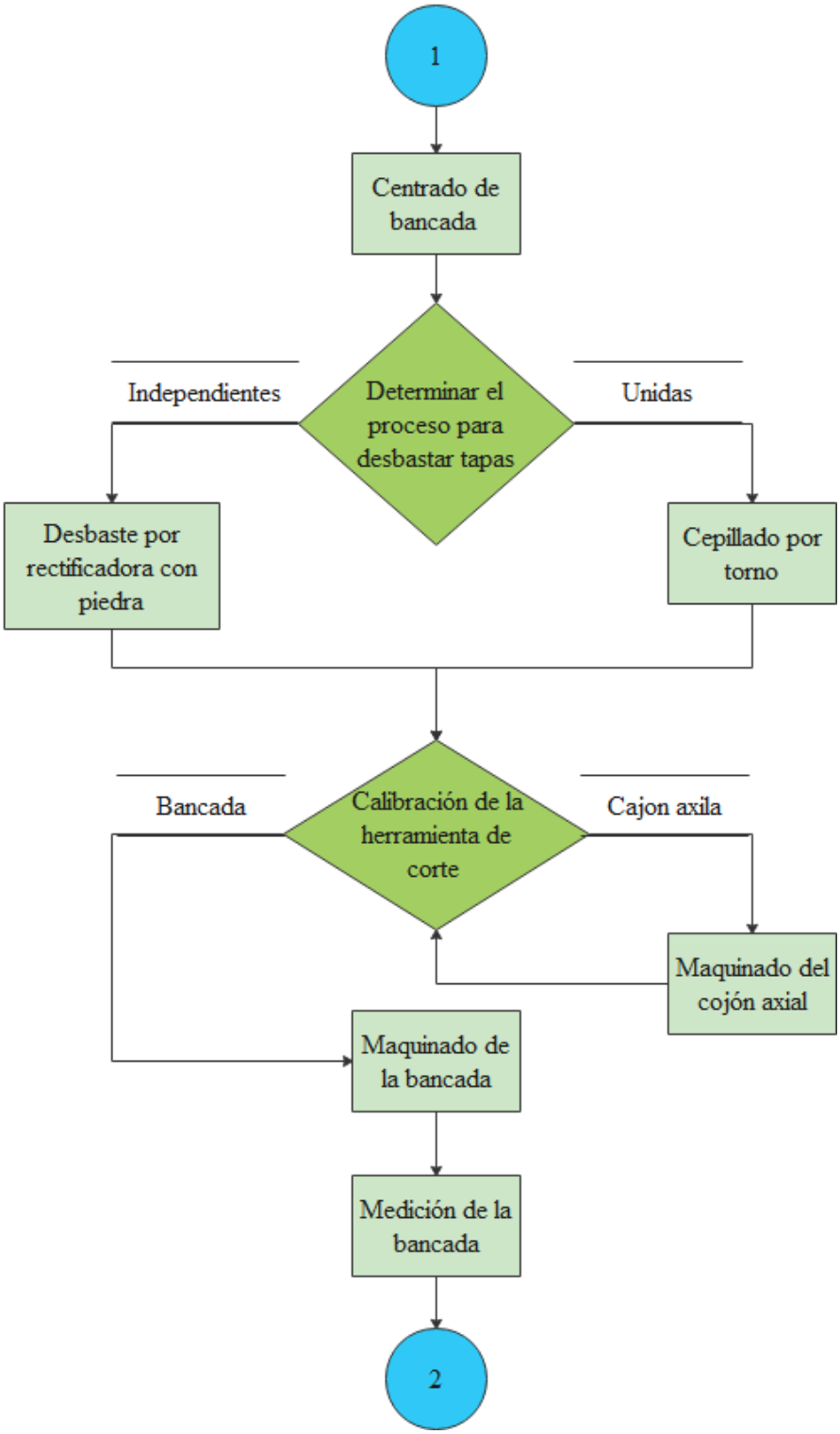
Tabla 15. Resultados del monoblock Nissan **Urvan 2.5**

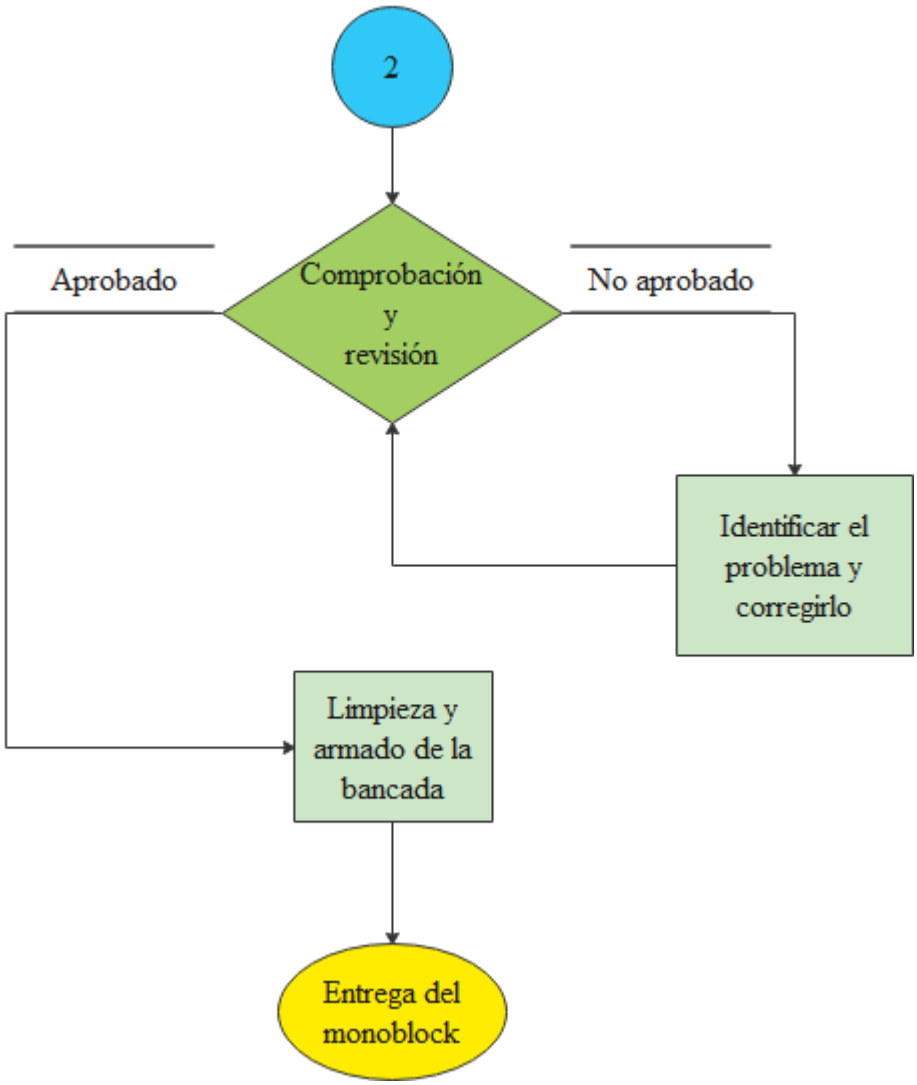
Volkswagen Jetta 1.8 (4 cil.)				
El vehículo presento un daño en el cojinete axial trasero, esto por un embrague en mal estado				
	Ø Bancada	Ø Muñón de cigüeñal	Claro de lubricación (Luz)	Holgura axial
Tolerancias del fabricante	2.323”	2.124”	0.001”-0.004”	0.003”-0.005”
Desgates del Monoblock	2.324”	2.112”	0.005”	se perdió el cajón axial
Monoblock rectificado	2.323”	2.104” se rectificó en 0.020	0.002”	0.003”

Tabla 15. Resultados del monoblock Volkswagen **Jetta 1.8**

DIAGRAMA DE FLUJO




1
Centrado de bancada
Determinar el proceso para desbastar tapas
Independientes
Unidas
Desbaste por rectificadora con piedra
Cepillado por torno
Calibración de la herramienta de corte
Bancada
Cajon axila
Maquinado del cojón axial
Maquinado de la bancada
Medición de la bancada
2


2
Comprobación y revisión
Aprobado
No aprobado
Identificar el problema y corregirlo
Limpieza y armado de la bancada
Entrega del monoblock

HOJA DE PROCESO			
1°	Recepción del motor	Se hace el registro del motor y se le realiza una orden de trabajo al cliente, posteriormente se traslada el motor a el área de maquinado correspondiente	
2°	Reconocimiento Y valoración del monoblock	Se determina el tipo de vehículo al que pertenece y se rectifica de acuerdo a las especificaciones del fabricante, posteriormente se revisa el estado físico de la bancada, cigüeñal y sus medidas.	
3°	Se necesita rellenado por soldadura	En base a la valoración física de la bancada determinamos si está o el cajón axial requiere del servicio de rellenado por soldadura para ser maquinado	El monoblock requiere el servicio: seguir paso 4°
			El monoblock no requiere el servicio: seguir paso 6°
4°	Definir el proceso de soldadura	Dependiendo del material en el que está fundido el monoblock se determinara el proceso de soldadura con el cual se va a rellenar	Fierro colado se usa el proceso de soldadura SMAW: seguir paso 5°
			Aluminio se usa el proceso de soldadura TIG: seguir paso 5°
5°	Área de soldado y rellenado	Ya identificados los puntos a soldar y el tipo de soldadura se procederá a rellenarlos mediante la aplicación de cordones, hasta conseguir la cantidad suficiente de material para ser maquinado	
6°	Montaje y ajuste de monoblock en la mandrinadora	Se sube el monoblock a la máquina y esta se ajusta de acuerdo a las especificaciones y dimensiones que tiene el monoblock	
7°	Centrado de bancada	Mediante una serie de pasos se procede a realizar el centrado de la máquina con respecto a la barra de corte e ir inmovilizando el monoblock	
8°	Determinar el proceso para desbastar tapas	La bancada de un monoblock puede contar con dos tipos de tapas, para estos casos el proceso de desbaste es diferente, dado por las dimensiones que manejan ambas se implementan dos máquinas diferentes	La bancada cuenta con tapas unidas: el desbaste se realizara en el torno: seguir paso 9°
			La bancada cuenta con tapas independientes: el desbaste se realizara en la rectificadora de tapas: seguir paso 9°
9°	Calibración de la herramienta de corte	Se preparara todo el material que se ocupa en el proceso de maquinado y también se calibra nuestro instrumental de corte dependiendo del área a maquinar	Se maquinara cajón axial: seguir paso 10°
			Se maquinara la bancada: seguir paso 11°
10°	Maquinado del cajón axial	Se realiza el maquinado del cajón del cojinete axial en el lado rellenado por soldadura mediante el uso de dos buriles de desbaste	Se terminó de maquinar el cajón axial: regresar paso 9°
11°	Maquinado de la bancada	Se da inicio al “corte en línea” mediante el uso de un buril de corte y posteriormente se le va aumentando la medida del buril hasta llegar al diámetro correspondiente de la bancada	

12°	Medición de la bancada	Previo y posterior a la realización del ultimo corte se medirá la bancada con los dos tipos de micrómetros para determinar si esta cuenta con el diámetro establecido por el fabricante	
13°	Comprobación y revisión	Se utilizara la herramienta plastigage para corroborar que la bancada cuente con la holgura requerida y se revisa que el cigüeñal cuente con un giro libre ya armado en el monoblock	Se presentó algún problema en la holgura y/o movilidad del cigüeñal: seguir paso 14°
			No se presentó algún problema en la holgura y/o movilidad del cigüeñal: seguir paso 15°
14°	Identificar el problema y corregirlo	Se realizara una inspección minuciosa de todas las partes que conforman la bancada para encontrar el problema que le está afectando y determinar su posible solución	Se realizó la respectiva corrección del problema: regresar paso 13°
15°	Limpieza y armado de la bancada	Se realizará una limpieza de toda la bancada y las partes que la conforman para finalmente armarla y mandar el monoblock a la sección de trabaos terminados	
16°	Entrega del monoblock	Se le notificara al cliente que el servicio correctivo de su motor a concluido, para que pueda recogerlo	

ANEXO

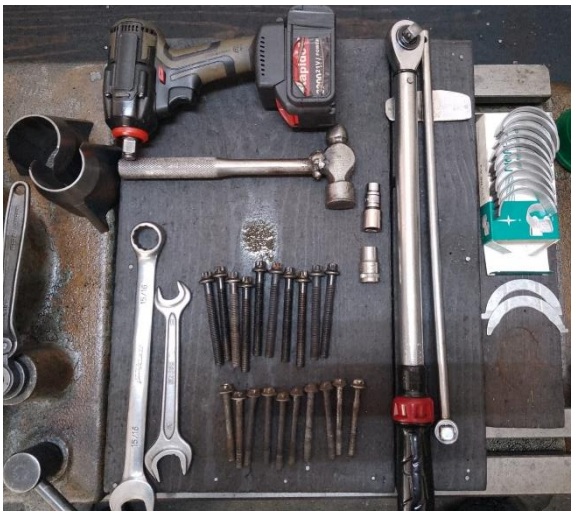


Figura 98. Herramienta para Urvan 2.5

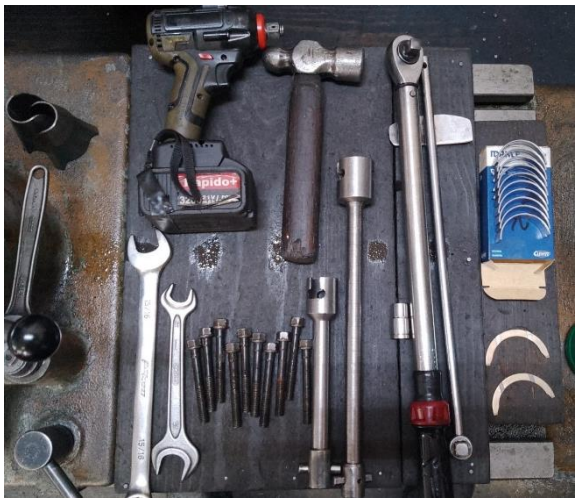


Figura 97. Herramienta para Tsuru 1.6

Figura 106. Desgaste del cajón axial de Urvan 2.5.

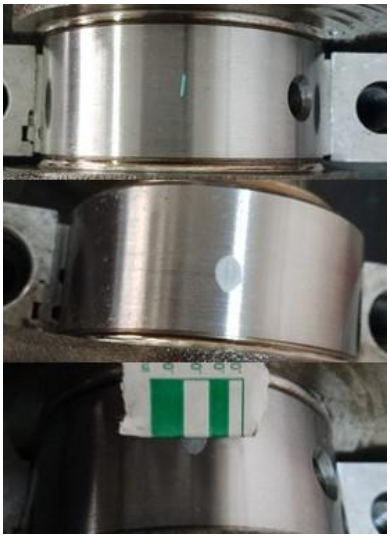


Figura 105. Funcionamiento de plastigage

Figura 104. Mandrinadora

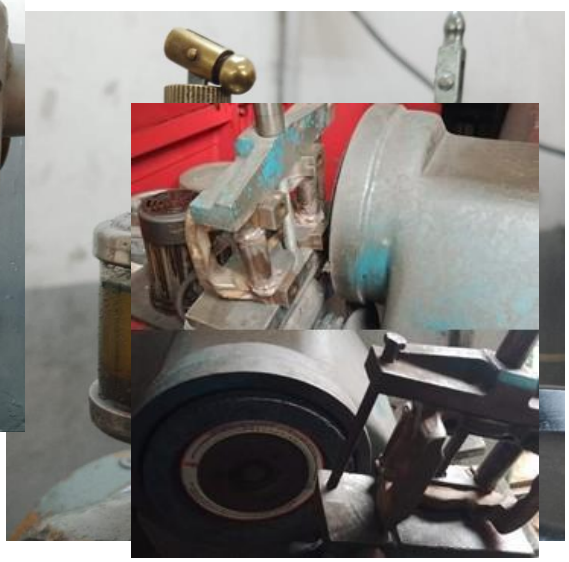


Figura 102. Herramienta para Jetta 1.8

Figura 10Figura 103. Tapa doble del Tsuru 1.6

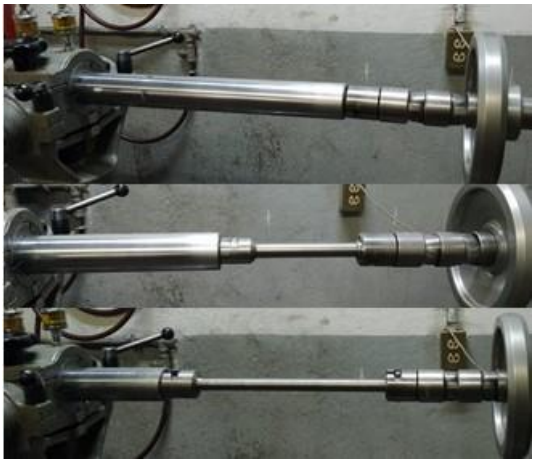


Figura 100. Extensiones

Figura 99. Motor V8 montado para corte en línea

CONCLUSIONES

El rectificado de motores sigue siendo muy poco conocido a comparación de otras áreas en el sector de la mecánica automotriz, se alcanzó a desarrollar una solución con imágenes de gran calidad permite lograr un mayor alcance y con ello se logra comprender de mejor manera el proceso del rectificado de bancada de un monoblock, para las personas que empiezan en esta área, de igual forma se generó un impacto positivo a lconocer la importancia de este tipo de trabajos y soluciones bien definidad y comprobadas para México y Latinoamérica principalmente.

El trabajo permitió identificar el proceso desde la recepción del monoblock, la valoración de las condiciones de inicio, logrando llevar al interesado a las siguientes etapas del proceso (soldadura y maquinado) describiendo el método correcto de la realización de estas actividades para el lector y finalmente se realizó la recopilación de resultados del proceso antes mencionado determinado su viabilidad como un tipo de mantenimiento correctivo para los vehículos que sufren un daño severo en la bancada del motor.

Es por eso que la realización y divulgación de este tipo de trabajos ayuda a dar una mayor exposición del rectificado automotriz, permitiendo formar una comunidad mejor preparada y con acceso más información. Se espera que este trabajo sirva de apoyo en la capacitación de nuevo personal para las empresas y talleres, ofreciendo un servicio con toma de descisiones informadas y capacitación confiable.

REFERENCES

[1] Nagy, J., & Lakatos, I. (2024). Predictive Maintenance and Predictive Repair of Road Vehicles—Opportunities, Limitations and Practical Applications. Engineering Proceedings, 27. https://doi.org/10.3390/engproc2024079027 Rededuca. (s. f.). *¿Qué son los conocimientos previos? I Contexto Educativo*. https://www.rededuca.net/contexto-educativo/c/conocimientos-previos

[2] Sebastião, L., Tirapicos, F., Payan-Carreira, R., & Rebelo, H. (2023). Skill Profiles for Employability: (Mis)Understandings between Higher Education Institutions and Employers. Education Sciences, 13(9), 905. https://doi.org/10.3390/educsci13090905

[3] Schalock, R. L. (2015). Las mejores prácticas de las organizaciones y los profesionales. Siglo Cero, 46(1), 7. https://doi.org/10.14201/scero2015461723

[4] Wong, V. W., & Tung, S. C. (2016). Overview of automotive engine friction and reduction trends–Effects of surface, material, and lubricant-additive technologies. Friction, 4(1), 1-28. https://doi.org/10.1007/s40544-016-0107-9

[5] Reitz, R. D., Ogawa, H., Payri, R., Fansler, T., Kokjohn, S., Moriyoshi, Y., Agarwal, A., Arcoumanis, D., Assanis, D., Bae, C., Boulouchos, K., Canakci, M., Curran, S., Denbratt, I., Gavaises, M., Guenthner, M., Hasse, C., Huang, Z., Ishiyama, T., . . . Zhao, H. (2019). IJER editorial: The future of

the internal combustion engine. International Journal Of Engine Research, 21(1), 3-10. https://doi.org/10.1177/1468087419877990

[6] Woodyard, D. (2009). Other Medium-Speed engines. En Elsevier eBooks (pp. 715-758). https://doi.org/10.1016/b978-0-7506-8984-7.00028-x

[7] Rectificado de bancada de block. (s. f.). Motor Mart. https://www.motormart.com.mx/pages/rectificado-de-bancada-de-block?srsltid=AfmBOopEiG5sZEgtZG_BqXg0O6RMd6aTz_FBGY09SI11oOv-Qi_q-K-j

[8] Billiet, W. (1979). Entretenimiento y reparacion de motores de automovil . España: Reverté S.A.

[9] Grupo CEAC. (2007). Manual CEAC del automóvil. España: ediciones CEAC S.A.

[10] Khoshnaw, F., Krivtsun, I., & Korzhyk, V. (2023). Arc welding methods. En Elsevier eBooks (pp. 37-71). https://doi.org/10.1016/b978-0-323-90552-7.00004-3

[11] Mike Caruso, S. F. (2009). Cylinder Head and Block Manual . U.S.A: AERA .

[12] Nissan Mexico. (1996). (NISSAN) Manual de Servicio Nissan Serie B12 . México: Nissan Mexicana S.A. de C.V.

[13] Santillán, D. P. (2011). Elaboración de un manual de procesos y procedimientos bajo estándares de calidad para rectificación de motores de vehículos livianos. Quito, Ecuador : Universidad Internacional de Ecuador .

[14] Torre, A. G. (2003). Ejecución de procesos de mecanizado, conformado y montaje 2° edición. Madrid, España : Ediciones Parafino S.A.

Daniel Díaz Muñoz: https://orcid.org/0009-0003-9516-5528

Pedro Vera Serna: https://orcid.org/0000-0001-7085-7374

David Alcántara Sandoval: https://orcid.org/0009-0003-3204-1712

FIRST PASSAGES: PARENTS, CHILDREN, AND THE TRANSITION INTO EARLY CHILDHOOD EDUCATION IN MALTA

Simon FARRUGIA
Malta Leadership Institute, School for Educational Studies, Malta
farrugia.sim@gmail.com

ABSTRACT: The transition into early childhood education is a significant emotional, social, and developmental process for both children and their families. This mixed-methods study, conducted in Malta, explored parental experiences of children's transition into early childhood education through the lenses of Van Gennep's rites of passage, Bronfenbrenner's ecological systems theory, and Vygotsky's sociocultural theory. Data were collected through semi-structured interviews with 15 parents and an online questionnaire completed by 51 parents. The findings show that emotional preparation, predictable routines, scaffolded support, and responsive communication with educators played an important role in helping children adjust across the phases of separation, liminality, and incorporation. This study also found that parents experienced their own parallel transition journeys, highlighting the shared and relational nature of early educational transitions. These findings offer practical insights for early years educators, policymakers, and family support professionals, particularly in culturally diverse and small-scale educational contexts.

Key words: early childhood education, educational transitions, parental experiences, rites of passage (Arnold Van Gennep), emotional adjustment, home-school collaboration

INTRODUCTION

The transition into early childhood education represents one of the first significant changes in a child's life. It requires children to adjust to new environments, routines, and relationships while simultaneously challenging parents to manage their own emotional responses and practical responsibilities (Peters & Roberts, 2015; Fabian & Dunlop, 2007). This process has increasingly become a central focus of research, policy, and practice in early childhood education as transitions are now recognised as not only developmental milestones for children but complex socioemotional processes involving multiple stakeholders (Dockett & Perry, 2007; Margetts, 2002).

Although many children demonstrate remarkable resilience, the transition to formal education can be accompanied by emotional distress, separation anxiety, behavioural regression, and varying degrees of adjustment difficulties (Rimm-Kaufman & Pianta, 2000; Theodotou, 2019). For others, transitions may become a gateway to positive socialisation, confidence building, cognitive stimulation, and long-term academic success when well supported (Fabian, 2012; Harju et al., 2023). What is increasingly

acknowledged is that transitions are not linear nor universally experienced. Rather, they are deeply influenced by cultural, contextual, and relational factors that shape children's and families' experiences (Dervin, 2020; Sollars, 2020).

Despite a growing body of literature, a notable gap remains: many studies focus heavily on child outcomes, institutional strategies, or educator perspectives, often overlooking the lived experiences of parents, who play a critical role both before and during transitions (Nicholson, 2018; O'Connor, 2013). This study directly addresses that gap, examining how parents perceive, interpret, and manage their children's transitions into early childhood education, guided by a robust theoretical framework.

1. THE CONTEXT OF EARLY CHILDHOOD EDUCATION IN MALTA

This study was conducted in Malta; a small island nation located in the Mediterranean Sea, characterised by its unique cultural, historical, and educational landscape. Early childhood education has gained significant importance in Maltese national policy over recent decades, reflecting both international trends and local priorities to ensure accessible, high-quality provision from an early age (Sollars, 2020). In Malta, formal education often begins as early as 3 years of age, with children entering childcare centres, kindergartens, or preschools before compulsory schooling.

Malta's early years sector is diverse, encompassing state, church, and private providers, each offering varied pedagogical approaches ranging from more structured academic curricula to play-based methodologies (Theodotou, 2020). Increasing cultural diversity, shifting family structures, and evolving societal expectations present both opportunities and challenges for educators and parents. While the country has invested in professional development for early years practitioners and curricular reforms aimed at promoting holistic development (Sollars, 2020), parents still navigate varying institutional practices, communication styles, and transition policies across settings. This makes Malta an instructive context for studying parental experiences of transition within a culturally complex yet relatively small national system.

2. DEFINING TRANSITIONS IN EARLY CHILDHOOD EDUCATION

Transitions are broadly defined as periods of change in which children move from one context or stage of life to another, requiring adjustment and adaptation to new environments, relationships, and routines (Dockett & Perry, 2007). As Lehman et al. (2001) argue, "transition must be conceptualized as a process that occurs over time, not as a short-term move from one environment to the next" (p. 6). In early childhood education, transitions may include moving from home to childcare, from preschool to

primary school, or even between classes within the same institution. These transitions are often accompanied by both developmental and emotional demands that require effective support mechanisms to foster positive adaptation (Fabian & Dunlop, 2007).

Children undergoing transitions must manage separation from familiar caregivers, adapt to different adult expectations, form new peer relationships, and learn unfamiliar rules and routines (Rimm-Kaufman & Pianta, 2000; Margetts, 2002). The process can therefore evoke emotional reactions such as anxiety, fear, or excitement while also triggering behavioural responses such as withdrawal, clinginess, or behavioural regression (Fabian, 2012; Pianta & Cox, 1999). As Perry et al. (2014) argue, transitions are not isolated events but processes that unfold over time and vary in duration and complexity for each child.

3. PARENTAL EXPERIENCES AND PERSPECTIVES ON TRANSITIONS

While much research focuses on children's emotional and behavioural responses during transitions, parents' experiences are equally important. As Lehman et al. (2001) explain, "the transition into public school kindergarten marks an important rite of passage for children and their parents and plays a critical role in later school success. Some of the positive consequences of successful adjustment include development of positive peer relationships, cooperative relationships with teachers, and long-term social competence and academic achievement" (p. 5). Parents often experience their own anxieties as they hand over caregiving responsibilities to teachers, question their child's readiness, and negotiate institutional expectations (Sollars, 2020; Peters, 2010). These emotional and cognitive processes influence how parents support their child's transition and shape the relational climate between home and school (O'Connor, 2013).

Several studies have found that parents who feel informed and supported by educators report greater confidence in their child's adjustment, while those who feel excluded or insufficiently informed may experience heightened anxiety and dissatisfaction with the transition process (Dockett & Perry, 2007; Price & Steed, 2016). Effective parental involvement prior to and during transitions has been associated with better child outcomes, suggesting that fostering strong home-school partnerships is a key factor in facilitating positive transitions (Fabian & Dunlop, 2007; Sylva et al., 2004).

Parental concerns often focus on children's emotional wellbeing, social integration, and developmental readiness for formal learning (Harju et al., 2023; Perry et al., 2014). In addition, cultural expectations and family circumstances, such as work commitments, family structure, or previous educational

experiences, can shape parental perspectives and capacities to support transitions (Dervin, 2020; Sollars, 2020).

4. THEORETICAL FRAMEWORKS GUIDING THE STUDY

4.1. ARNOLD VAN GENNEP'S RITES OF PASSAGE

> Ce ne sont pas les rites dans leur détail qui nous ont intéressé, mais bien leur signification essentielle et leurs situations relatives dans des ensembles cérémoniels, leur séquence […] afin de montrer comment les rites de séparation, de marge et d'agrégation, tant préliminaires que définitifs, se situent les uns par rapport aux autres en vue d'un but déterminé. […] Leur disposition tendancielle est partout la même, et sous la multiplicité des formes se retrouve toujours, soit exprimée consciemment, soit en puissance, une séquence type : le schéma des rites de passage.
> (Van Gennep, 1909, p. 275)
> It is not the rites in their details that have interested us, but rather their essential significance and their relative positions within ceremonial ensembles, their sequence […] in order to show how the rites of separation, of margin, and of aggregation, both preliminary and definitive are situated in relation to one another toward a specific goal. […] Their typical arrangement is everywhere the same, and beneath the multiplicity of forms, one always finds, whether consciously expressed or latent, a typical sequence: the schema of the rites of passage.
> (English translation by the author)

The principal theoretical framework guiding this study is Dutch–German–French ethnographer and folklorist Arnold Van Gennep's (1909) rites of passage, originally formulated to describe cultural rituals marking transitions between life stages. Van Gennep conceptualised transitions as a tripartite process consisting of separation, liminality, and incorporation: separation involves detachment from a familiar setting or role, liminality represents an in-between period of uncertainty and adjustment, and incorporation signifies the eventual integration into the new role or environment. Van Gennep (1960) emphasised, "interest lies not in the particular rites but in their essential significance and their relative positions within ceremonial wholes that is, their order," noting that "a typical pattern always recurs: the pattern of the rites of passage" (p. 191).

Several scholars have adapted this model to educational transitions, finding its structure useful for analysing the emotional and social challenges children and parents face during periods of change (Abeliovich, 2018; Fabian, 2012; Harju et al., 2023). This framework allows for a dynamic understanding of transitions as evolving rather than instantaneous events.

In this study, Van Gennep's model is applied to examine how parents perceive their own and their children's experiences as they navigate the transition into formal education. Each phase provides an analytical lens to explore preparation, adjustment, and integration processes.

4.2. BRONFENBRENNER'S ECOLOGICAL SYSTEMS THEORY

Bronfenbrenner's (1979) ecological systems theory further informs this research by situating transitions within interconnected environmental layers that influence child development: the microsystem which encompasses direct relationships between children, parents, and educators; the mesosystem which reflects interactions between these microsystems as, for example, parent-teacher communication; the exosystem which includes external factors such as institutional policies or workplace demands; and the macrosystem which encompasses cultural norms, values, and societal expectations.

Transitions are thus not isolated within the child's experience but shaped by a complex web of relational and institutional influences. Parental experiences are situated within both microsystemic interactions and broader macrosystemic expectations, including national education policies, cultural beliefs about readiness, and institutional transition practices (Nicholson, 2018; Sollars, 2020).

4.3. VYGOTSKY'S SOCIOCULTURAL THEORY

Vygotsky's (1978) sociocultural theory adds a developmental dimension to understanding transitions by highlighting the role of social interaction and scaffolding in learning. Children do not navigate transitions independently; rather, they rely on guidance from more knowledgeable others, such as parents, educators, or peers, who support their emotional regulation, language acquisition, and social integration (Dockett & Perry, 2007; Price & Steed, 2016).

The concept of scaffolding is particularly relevant to transition processes as parents and teachers co-construct supportive strategies that help children interpret and adapt to unfamiliar routines and expectations (Theodotou, 2020). These interactions facilitate meaning-making and serve to reduce uncertainty during the liminal phase, strengthening children's capacity for successful incorporation into new educational settings.

5. CURRENT GAPS IN RESEARCH

Despite extensive scholarship on transitions, much of the literature remains child-centred or institution-centred, often neglecting parental voices as key agents in the transition process (Nicholson, 2018; O'Connor, 2013). While policies frequently stress parental involvement, the emotional experiences, coping strategies, and perceptions of parents remain under-explored, particularly in specific cultural contexts such as Malta.

Moreover, although theoretical frameworks such as Van Gennep's, Bronfenbrenner's, and Vygotsky's are increasingly referenced, few studies offer integrated analyses that simultaneously apply these lenses to both parental and child experiences. This study contributes to filling that gap by offering an empirically grounded, theoretically rich examination of parental perspectives on early childhood transitions.

6. AIM AND OBJECTIVES OF THE STUDY

The aim of this study is to explore how parents experience and interpret their children's transitions into early childhood education using Arnold Van Gennep's rites of passage framework as the primary analytical lens.

The study addresses the following objectives:

- To examine how parents prepare their children for the transition into early educational settings.
- To explore parents' perceptions of children's emotional and behavioural responses during the adjustment phase.
- To identify when and how parents perceive their children have fully adapted.
- To investigate the support strategies employed by parents and the role of educators during transitions.

7. METHODS

7.1. RESEARCH DESIGN

This study employed a convergent mixed-methods design (Creswell & Plano Clark, 2017), integrating both qualitative and quantitative data to provide a comprehensive exploration of parental experiences during children's transitions into early childhood education. The use of mixed methods allowed for in-depth exploration of parents' personal narratives while also identifying broader patterns and trends across a larger sample.

A mixed-methods approach was particularly suited to this study's aim of capturing both the subjective emotional experiences of parents and the more generalisable aspects of preparation, adjustment, and support during transitions (Bryman, 2012). The qualitative strand allowed for rich, narrative accounts of parents' experiences, while the quantitative strand complemented these insights by offering descriptive data regarding common experiences, emotional responses, and support strategies.

7.2. PARTICIPANTS

The study involved two distinct but related samples. The qualitative sample consisted of 15 parents of children aged between 3 and 7 years who participated in semi-structured interviews. The sample included both mothers and fathers, representing diverse family structures, employment statuses, and educational backgrounds. All participants were parents of children who had recently experienced a transition into early childhood education settings, such as childcare centres, preschools, or kindergartens. The children attended various types of institutions, including state, church, and independent schools in Malta.

The quantitative sample comprised 51 parents who completed an online questionnaire distributed via JotForm. Participants were recruited through convenience sampling, primarily via parental networks and social media platforms related to early years education. The questionnaire collected demographic information, including parental age, gender, level of education, child's age at transition, and the type of early years institution attended.

While the sample was not intended to be fully representative of the Maltese population, the combination of diverse backgrounds allowed for the identification of common themes and variations across family contexts.

8. DATA COLLECTION PROCEDURES

8.1. QUALITATIVE DATA COLLECTION

Semi-structured interviews were conducted using a flexible interview guide specifically designed to reflect the theoretical focus on transition stages derived from Van Gennep's (1909) rites of passage. The interview guide explored parental experiences across four thematic areas:

1. Separation (Preparation and First Days): How parents prepared their children and themselves for the transition.
2. Liminality (Adjustment Period): Emotional and behavioural responses observed in children and experienced by parents.
3. Incorporation (Settling in and Belonging): Indicators of adaptation and integration into the educational setting.
4. Support Systems: Perceptions of the role of parents and educators in facilitating successful transitions.

Interviews were conducted either face-to-face or via online video conferencing platforms, depending on participant preference and COVID-19 restrictions at the time. Each interview lasted approximately 30 to 45 minutes. All interviews were audio-recorded with participants' consent and subsequently transcribed verbatim for analysis.

8.2. QUANTITATIVE DATA COLLECTION

The quantitative component of the study involved a structured online questionnaire developed using JotForm. The instrument included both closed and open-ended questions and was designed to gather data across several key areas: demographic information, including the child's age, parental background, and type of institution attended; parental emotions prior to the transition; children's emotional and behavioural responses during the transition period; the duration of the adjustment phase; strategies employed by parents to support their children; and parents' perceptions of support provided by the school or educators.

The questionnaire was pilot tested with a small sample of parents (n=5) prior to full distribution to ensure clarity and relevance. The final instrument included Likert-scale items, multiple-choice questions, and open comment boxes, enabling both quantitative analysis and limited qualitative elaboration.

9. DATA ANALYSIS

9.1. QUALITATIVE ANALYSIS

Qualitative data were analysed using thematic analysis, following Braun and Clarke's (2017) six-phase approach. The analysis began with familiarisation with the data, followed by the generation of initial codes. Subsequently, themes were identified through the process of searching for patterns across the coded data. These themes were then reviewed and refined to ensure coherence and relevance. After this stage, themes were defined and named to accurately represent the data's core meanings. Finally, a complete report was produced, integrating the identified themes into the overall analysis.

Both inductive and deductive coding strategies were employed. Inductive coding allowed themes to emerge naturally from participants' narratives, while deductive coding applied Van Gennep's (1960) framework to organise and interpret parental experiences systematically across the three phases of transition. Coding was completed manually to ensure close engagement with the data.

9.2. QUANTITATIVE ANALYSIS

Quantitative data were analysed descriptively using frequency distributions and percentages to identify common patterns across the sample. Given the exploratory nature of the study and the sample size, inferential statistics were not conducted. Instead, descriptive statistics served to triangulate findings from the qualitative interviews and offer an overview of parental experiences across the larger sample. All quantitative analysis was performed manually; no statistical software (e.g., SPSS, R, Excel) was used.

10. RESULTS

The findings are presented following Van Gennep's (1909) three-phase model of transition, separation, liminality, and incorporation, alongside an emergent theme of support systems, which arose prominently in both qualitative and quantitative data.

10.1. SEPARATION: PREPARATION AND INITIAL EMOTIONAL RESPONSES

Parents described multiple preparation strategies before their children's entry into early childhood education settings. These ranged from structured activities such as school visits and information sessions to more informal methods such as reading books, role-play, and creating predictable routines at home (Participants 1, 3, 6, 7, 9, 10, 15). Several parents emphasised their awareness of the importance of preparing their children emotionally and practically:

> We visited the school a few times and introduced her to her teacher beforehand. We also started talking about school in a very positive way to make it seem exciting rather than scary. (Participant 7)

Despite these efforts, many parents reported experiencing their own anxieties about the transition. These emotional responses ranged from concern over whether their children would adjust socially to worries about being physically separated for the first time (Participants 1, 3, 5, 6, 10). As one mother shared:

> Although I kept telling my daughter how wonderful school would be, deep inside I felt quite anxious about how she would manage without me. (Participant 1)

The quantitative data aligned with these narratives. According to questionnaire responses, 47% (n=24) of parents reported feeling anxious prior to the transition, while 35% experienced mixed emotions of excitement and concern.

From the children's perspective, emotional reactions during the first days also varied. While some children demonstrated confidence and enthusiasm, others exhibited distress, particularly at morning drop-off. Parents reported frequent instances of crying, clinginess, and refusal to separate from caregivers (Participants 3, 7, 9, 10, 13). Survey results showed that 58% of parents stated their child was "a little nervous but adjusted quickly," whereas 29% reported their child was "upset or anxious for a while." A small minority indicated ongoing distress.

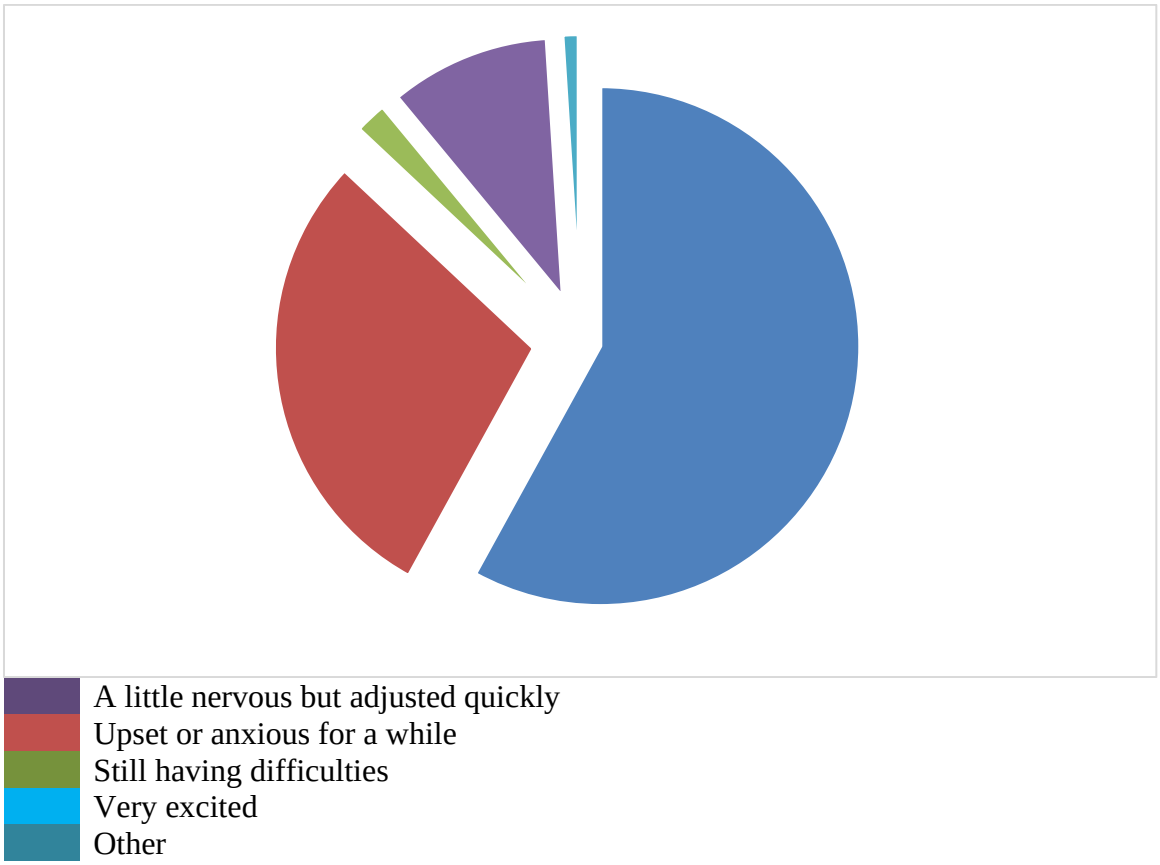


Fig. 1. Children's first-day emotional reactions.

The initial separation phase was described by many parents as emotionally intense but relatively brief, especially when preparation had been thorough and when children were gradually introduced to their new setting.

10.2. LIMINALITY: THE ADJUSTMENT PERIOD

The liminal phase, the ambiguous period between initial entry and full adaptation, revealed wide variation in children's emotional, behavioural, and physiological responses. Parents reported both observable behaviours and emotional shifts during this phase.

Many parents described inconsistent behaviours at home, such as sudden outbursts, mood swings, regressions in previously mastered behaviours, for example sleep disruptions and bedwetting, and increased dependence on caregivers (Participants 1, 4, 6, 7, 11, 13).

> In the mornings, he would be excited to see his friends, but by the evening, he became very sensitive, needed constant cuddles, and sometimes would cry for no apparent reason. (Participant 4)

Other parents mentioned somatic symptoms that appeared to coincide with the transition period, including stomach aches, headaches, and frequent complaints of being "sick" before school (Participants 3, 6, 9, 12).

> There were mornings when she complained of stomach pain; at first, we thought it was medical, but the doctor reassured us it was linked to stress. (Participant 6)

Quantitative data supported these qualitative findings. Survey responses indicated that:

- 74% of parents observed mood changes.
- 56% reported disruptions in sleep patterns.
- 48% noted heightened clinginess and emotional dependence.
- 39% observed somatic complaints such as stomach aches or headaches.

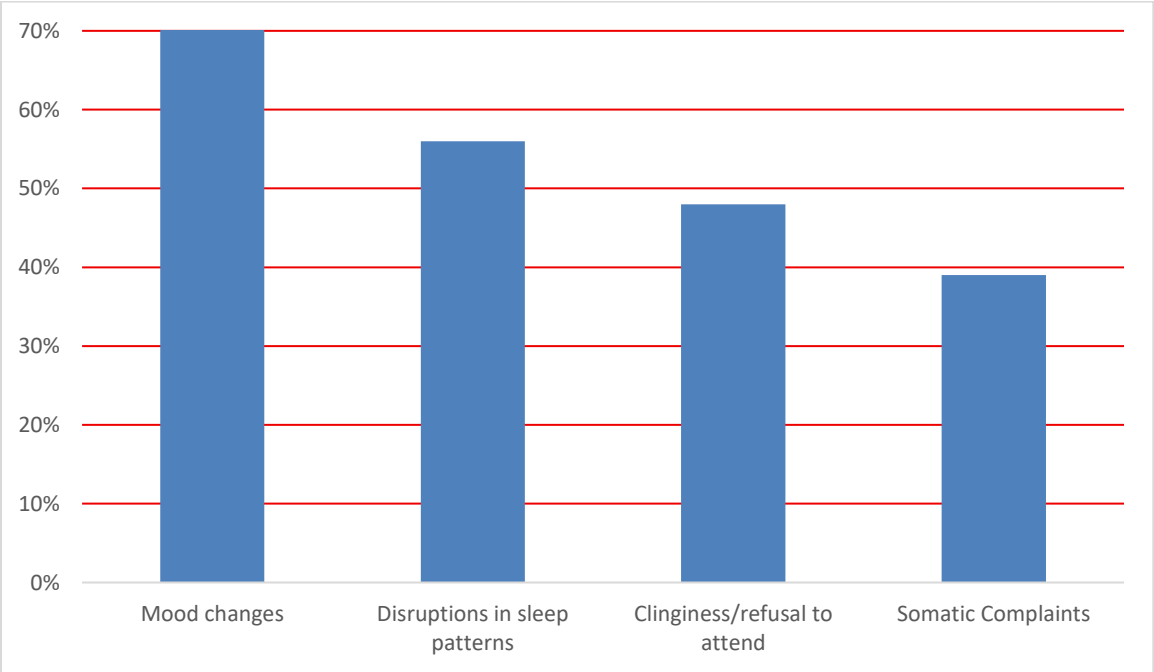


Fig. 2. Adjustment difficulties during the liminal phase.

The duration of this liminal adjustment phase varied significantly. For some families, adaptation occurred relatively quickly, within a few weeks, while others required several months to reach a sense of stability. According to the questionnaire results, 62% of parents reported that their children settled within 1–2 months, while approximately 28% indicated that their child needed 3 or more months to fully adjust.

10.3. INCORPORATION: SETTLING IN AND ADAPTATION

In the incorporation phase, most parents identified clear indicators that their children had achieved emotional stability, social integration, and comfort in their educational environment. Key signs included enthusiastic discussions about school activities, growing independence during morning drop-offs, and the development of peer relationships.

> By the end of the first term, he had completely settled. He would wake up excited to go to school and was always eager to share what he had learned that day. (Participant 2)

For many parents, the formation of friendships was particularly reassuring and strongly associated with their child's sense of belonging (Participants 3, 5, 8, 11, 13).

> It was when she started talking about her friends and inviting them for playdates that I realised she felt fully comfortable. (Participant 5)

Quantitatively, 82% of parents reported observable improvements in their child's mood and confidence after the transition period, suggesting a high degree of eventual successful incorporation for most children in the sample.

10.4. SUPPORT SYSTEMS: PARENTAL AND EDUCATOR STRATEGIES

Across all transition stages, the presence of effective support systems emerged as a crucial factor in shaping both parental and child experiences. Parents highlighted the importance of establishing predictable routines at home, offering daily conversations about school, and employing creative tools such as visual schedules, countdown calendars, or storybooks about school life.

> We used a calendar where she would cross out the days until school started. That seemed to give her some sense of control. (Participant 9)

Survey responses revealed that:

- 79% of parents used daily conversations to prepare or reassure their child.
- 51% created predictable routines to ease the transition.

- 42% utilised visual supports such as calendars, storybooks, or school visits.

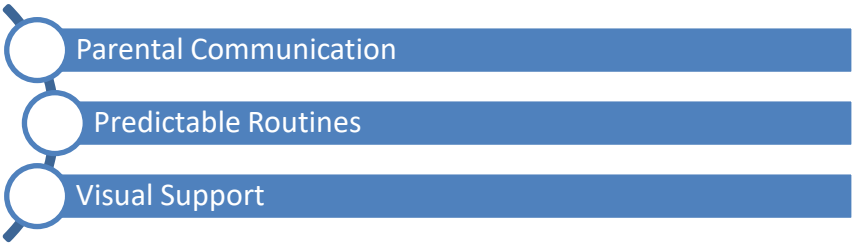


Fig. 3. Parental support strategies used during early childhood transition.

The role of educators was consistently emphasised as central to a smooth transition. Parents valued teachers who demonstrated sensitivity to children's emotions, maintained flexible drop-off routines, and communicated regularly with families.

> The daily updates from her teacher helped tremendously — even a simple photo or short message gave me peace of mind. (Participant 1)

However, several parents also expressed that communication varied considerably across institutions. Some felt that more structured induction programmes or clearer communication protocols would have further reduced parental anxiety (Participants 5, 7, 10).

> The school could have offered more scheduled orientation sessions for parents, not just for children. (Participant 5)

These comments suggest variability in how schools and educators involve parents during the transition period.

10.5. EMERGENT REFLECTIONS FROM PARENTS: TRANSITION AS A SHARED EXPERIENCE

A notable finding across many interviews was that parents often described the transition not only as their child's experience but as a shared family transition. Parents highlighted their own simultaneous emotional adaptation:

> I had to let go in stages, just like she had to. We both needed time to adjust. (Participant 11)

This mutual adaptation echoes emerging literature that recognises transitions as dynamic processes affecting all family members (Fabian & Dunlop, 2007; Harju et al., 2023). Several parents described a

parallel process of developing trust in the educational institution, confidence in their child's capabilities, and acceptance of new routines as part of their own transition journey.

11. DISCUSSION

The aim of this study was to explore how parents experience and interpret their children's transitions into early childhood education settings in Malta, guided by Van Gennep's rites of passage framework (1909) and further informed by Bronfenbrenner's ecological systems theory (1979) and Vygotsky's sociocultural theory (1978). This section interprets the findings through these theoretical lenses while situating the results within existing literature.

11.1. UNDERSTANDING TRANSITIONS AS MULTILAYERED EMOTIONAL PROCESSES

The findings of this study underscore that transitions are not singular, one-time events but ongoing emotional, behavioural, and cognitive processes affecting both children and parents. The separation phase revealed parental preparation strategies designed to ease children's entry into new educational settings, yet parents themselves often described significant internal anxieties. This echoes previous research noting that transitions involve parallel emotional adjustments for parents who often struggle with the simultaneous demands of supporting their child while managing their own sense of loss, uncertainty, and relinquished control (Fabian & Dunlop, 2007; Nicholson, 2018; Peters & Roberts, 2015).

Parental concerns observed in this study, including fears about separation, social adjustment, and institutional practices, are widely documented in transition literature (Perry et al., 2014; O'Connor, 2013). As Rimm-Kaufman and Pianta (2000) suggest, parents' emotional states may directly influence how they support their children's adjustment, with highly anxious parents potentially transmitting unintentional emotional signals that can shape children's own responses.

11.2. VAN GENNEP'S RITES OF PASSAGE: PHASED BUT FLUID TRANSITIONS

Van Gennep's (1960) tripartite model provided a useful organising framework for analysing parental experiences across separation, liminality, and incorporation phases. However, the findings also reveal that transitions may not always follow a clear, linear sequence. This is consistent with Harju et al. (2023) who observed that "children's transitions in ECEC comprise various permutations of discontinuities and continuities" (p. 735). Some parents described cycles of adaptation and regression where children displayed progress followed by temporary setbacks, especially when external stressors, such as illness and changes in routine, disrupted emerging stability.

The liminal phase was particularly rich in emotional complexity. Parents observed fluctuating behaviours, sleep disturbances, mood swings, and somatic complaints that align with existing research highlighting this period of ambiguity as emotionally taxing for both children and caregivers (Fabian, 2012; Margetts, 2002; Perry et al., 2014). Parents' own coping strategies during this phase were critical in shaping children's capacity to tolerate uncertainty and gradually settle into new routines (Fabian & Dunlop, 2007).

Meanwhile, the incorporation phase was marked by growing social engagement, independence, and confidence, echoing prior research suggesting that positive transitions often coincide with the development of peer relationships and secure attachments to new educators (Harju et al., 2023; Theodotou, 2020). Importantly, parents often linked their own sense of security with observable signs of their child's successful incorporation.

11.3. BRONFENBRENNER'S ECOLOGICAL SYSTEMS: THE INFLUENCE OF CONTEXT

The findings also strongly reflect Bronfenbrenner's (1979) ecological systems perspective, demonstrating how children's transitions unfold within multilayered environmental systems. The microsystem, involving parents, children, and educators, played a pivotal role in shaping transitional experiences. Direct interactions between parents and educators were central to parents' perceptions of safety and competence in the school environment.

At the mesosystem level, the quality of communication between home and school emerged as a consistent theme. Parents who received regular feedback and emotional reassurance from teachers reported greater confidence and reduced anxiety; a finding consistent with Price and Steed (2016) and Sylva et al. (2004). Conversely, some parents who felt excluded or insufficiently informed expressed ongoing worries about their child's wellbeing.

The macrosystem, including national policies, cultural expectations, and societal norms, also subtly shaped parental expectations. In Malta, where formal schooling begins relatively early, parents expressed concerns about developmental readiness and varying pedagogical approaches across different providers (Sollars, 2020; Theodotou, 2019). Some parents noted that greater standardisation of transition protocols across settings could reduce inconsistencies in parental experiences.

11.4. VYGOTSKY'S SOCIOCULTURAL LENS: SCAFFOLDING THE TRANSITION

Vygotsky's (1978) emphasis on scaffolding was evident in both parental and educator roles throughout the transition process. Parents employed numerous scaffolding strategies, including structured routines,

daily conversations, and visual tools like calendars and storybooks, which provided children with both emotional and cognitive preparation. These findings align with Dockett and Perry's (2007) argument that parents act as 'bridging agents' who mediate children's adjustment by interpreting new contexts and buffering emotional uncertainty.

Teachers also acted as significant scaffolding figures. Parents frequently praised educators who demonstrated flexibility, emotional sensitivity, and proactive communication. The value of personalised support in reducing transition anxiety supports Fabian and Dunlop's (2007) assertion that educator responsiveness is central to successful adaptation. Importantly, scaffolding was most effective when parents and educators worked collaboratively, reinforcing one another's support strategies; a key principle in Vygotsky's model of socially mediated learning (Sylva et al., 2004; Theodotou, 2020).

11.5. PARENTS AS CO-NAVIGATORS IN TRANSITION JOURNEYS

One of the emergent contributions of this study is its clear positioning of parents as active co-navigators, rather than passive observers, of their child's transition journey. Several parents described their own parallel transitions, negotiating not only their child's wellbeing but their own adaptation to new institutional structures, roles, and relationships with educators. This is echoed by Peltoperä et al. (2023) who found that "the frames used by the parents to discuss the children's transitions were stabilising the children's lives, balancing between staying at home and attending ECEC and adjusting to norms and rules" (p. 306).

This dual transition experience reflects newer transition research that challenges the view of transitions as solely child-focused (Fabian & Dunlop, 2007; Harju et al., 2023). Instead, transitions involve entire family systems that collectively negotiate emotional, social, and practical changes.

> We both transitioned together. As she got used to school, I got used to not being with her. It wasn't easy for either of us at first, but we adapted side by side. (Participant 11)

Such narratives suggest that future transition practices should explicitly acknowledge the family unit as an interconnected system requiring mutual support during transitional periods.

12. PRACTICAL IMPLICATIONS FOR EARLY CHILDHOOD EDUCATION

This study offers several practical recommendations that may enhance transition practices for both families and institutions:

- Structured orientation programmes: Gradual induction days for both children and parents may reduce anxieties prior to formal entry.
- Home-school communication protocols: Consistent, scheduled updates from educators help reassure parents and foster trust during the early weeks.
- Parental engagement in preparation: Encouraging parents to use routines, visual supports, and child-led discussions can provide children with a greater sense of predictability and agency.
- Professional development for educators: Training teachers in emotionally sensitive transition practices may improve both child and parent experiences.
- Culturally responsive approaches: Institutions should remain flexible to account for cultural, linguistic, and family-structure diversity (Dervin, 2020; Sollars, 2020).

These recommendations mirror best-practice models advocated in prior literature (Price & Steed, 2016; Sylva et al., 2004) while adding new emphasis on the shared, emotionally complex nature of parental participation.

13. LIMITATIONS OF THE STUDY

Although the mixed-methods design allowed for both depth and breadth, certain limitations should be acknowledged. First, the sample was geographically specific to Malta, which may limit the generalisability of the findings to wider international contexts. Second, the reliance on self-reported data introduces potential subjectivity and bias in both the interview narratives and questionnaire responses. Third, while the sample size was sufficient for the exploratory nature of this study, it remains relatively small for making broad national claims. Finally, educator perspectives were not included, which may have offered further insight into institutional transition practices. Nonetheless, these limitations also present opportunities for future research to complement and extend the present findings.

14. DIRECTIONS FOR FUTURE RESEARCH

Several avenues for further investigation emerge from this study. Future research could include the perspectives of educators, exploring how teachers experience and manage transitions, including their views on parent-school partnerships. Longitudinal studies may also prove valuable, offering insights into both short-term adaptation and longer-term educational outcomes as children progress through their educational trajectories. Further research could focus on vulnerable groups, including children with special educational needs, multilingual backgrounds, or those from migrant and minority families, whose transition experiences may differ significantly. Consistent with Bakopoulou (2022) notes, “the combination of existing variability in settings’ transition practices and the impact of the Covid-19

pandemic may have a disproportionately negative impact on children with identified SEND [Special Educational Needs and Disabilities] and would therefore intensify existing educational inequalities and endanger children's school readiness even further" (p. 649). Furthermore, cross-national comparisons could explore how different national policy frameworks shape transition practices in varying cultural contexts. Expanding research in these directions would strengthen the knowledge base required to support families and institutions in designing effective, inclusive, and contextually responsive transition practices.

15. CONCLUSION

This study has offered a detailed exploration of how parents experience and interpret their children's transitions into early childhood education, applying Van Gennep's rites of passage framework alongside Bronfenbrenner's ecological systems theory and Vygotsky's sociocultural theory. By adopting a convergent mixed-methods approach, combining both in-depth interviews and questionnaire data, the research has produced a nuanced account of how transitions unfold as emotional, relational, and developmental processes.

The findings illustrate that transitions are not isolated or one-dimensional events, but rather complex, ongoing experiences shaped by both children's individual developmental needs and the emotional states of their parents. Parents actively prepare their children through structured routines, emotional reassurance, and practical preparations, yet they simultaneously experience their own anxieties and adaptations as part of this shared journey.

Van Gennep's (1909) framework effectively captured the phased nature of these experiences, while also revealing that transitions may not always follow a strict linear sequence. Instead, as seen in the liminal phase, children and parents often experience periods of emotional fluctuation before achieving full incorporation into new educational settings.

Bronfenbrenner's (1979) ecological model further emphasised that children's transitions are embedded within interconnected systems. The critical role of home-school communication emerged consistently across participants, with parental confidence closely tied to the level of educator sensitivity, responsiveness, and transparency. These mesosystemic relationships were often decisive in either easing or complicating transition experiences.

Vygotsky's (1978) sociocultural theory illuminated how both parents and educators act as scaffolding agents, guiding children's emotional regulation and adjustment through intentional support strategies.

Parents' use of visual aids, storytelling, and consistent routines reflects this scaffolding process, as does the emotional attunement displayed by sensitive educators.

Importantly, the study highlights that transitions affect not only children but entire families. Parents described their own parallel transition journeys, gradually developing confidence in the new educational system as their children adapted. Recognising transitions as shared family experiences rather than solely child-centred events offers an important shift in how practitioners, policymakers, and researchers might approach transition planning.

Several practical implications emerge from these findings: the value of gradual induction programmes, structured communication protocols, culturally sensitive support strategies, and continued professional development for educators. These recommendations offer actionable insights that may enhance transition experiences across diverse educational contexts.

While the study's scope was geographically specific to Malta, the emerging themes are highly relevant for broader international discussions on transition practices. By centring parental voices, often underrepresented in transition research, this study adds depth to the growing recognition that successful transitions depend on collaborative, sustained, and emotionally attuned partnerships between families and educational institutions.

Ultimately, supporting transitions in early childhood requires attention not only to children's developmental readiness but also to the emotional journeys of parents who stand beside them. When both children and parents are supported as active participants in these transitions, the foundation for long-term educational engagement, wellbeing, and resilience is considerably strengthened.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all the participants who contributed to this study. I am especially thankful to the parents who participated in the interviews, both face-to-face and via telephone, for sharing their personal experiences and insights. I also wish to thank the parents who generously spared their time to complete the online questionnaire through JotForm. Their valuable contributions have made this research possible.

REFERENCES

[1] R. Abeliovich, (2018), "Reconsidering Arnold Van Gennep's Les Rites de Passage from the perspective of performance studies," Journal of Classical Sociology, vol. 18, no. 4, pp. 283–298.

[2] I. Bakopoulou, (2022), "The impact of the COVID-19 pandemic on early years transition to school in the UK context," European Early Childhood Education Research Journal, vol. 32, https://doi.org/10.1080/1350293X.2023.2265087, no. 3, pp. 638–654.

[3] U. Bronfenbrenner, (1979), The ecology of human development: Experiments by nature and design. Harvard University Press.

[4] Clarke, V., & Braun, V. (2017). Thematic analysis. The journal of positive psychology, 12(3), 297-298.

[5] A. Bryman, (2012), Social research methods, 4th ed. Oxford University Press.

[6] A. Butler and M. Ostrosky, (2018, September). Reducing challenging behaviours during transitions: Strategies for early childhood educators to share with parents. Naeyc.org. Retrieved from https://www.naeyc.org/resources/pubs/yc/sep2018/reducing-challenging-behaviors-during-transitions

[7] P. Cantor and D. Osher, (2021), The science of learning and development. Routledge.

[8] J. W. Creswell and V. L. Plano Clark, (2017), Designing and conducting mixed methods research, 3rd ed. SAGE Publications.

[9] F. Dervin, (2020), Interculturologies: Moving forward with interculturality in research and education. Springer Nature.

[10] S. Dockett and B. Perry, (2007), Transitions to school: Perceptions, expectations, experiences. UNSW Press.

[11] H. Fabian, (2012), "Children's transitions in early childhood education and care: Various combinations of dis-/continuities," Early Years, vol 44, no. 3–4, pp. 735–750.

[12] H. Fabian and A. W. Dunlop, (2007), Transitions in the early years: Debating continuity and progression for young children in early education. Routledge.

[13] K. Harju, M. Vuorisalo, M. Paananen and N. Rutanen, (2023), "Children's transitions in early childhood education and care: Various combinations of dis-/continuities," Early Years, vol. 44, https://doi.org/10.1080/09575146.2023.2232951, no. 3–4, pp. 1–16.

[14] G. Karrby, (2020), "Cultural perspectives on transitions in early childhood education," International Journal of Early Years Education, vol. 28, no. 3, pp. 312–329.

[15] J. Leffler, (2017), Early childhood education: Stakeholders' perspectives about kindergarten readiness in Mississippi. Scholars Junction. https://scholarsjunction.msstate.edu/cgi/viewcontent.cgi?article=2597&context=td

[16] C. Lehman, E. Brennan and B. Friesen, (2001), “Early childhood transitions,” Focal Point: Early Childhood Research & Practice, vol. 15, https://pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?article=1021&context=socwork_fac, no. 1, pp. 5–7.

[17] K. Margetts, (2002), “Transition to school—Complexity and diversity,” European Early Childhood Education Research Journal, vol. 10, https://doi.org/10.1080/13502930285208981, no. 2, pp. 103–114.

[18] C. McLachlan, M. Fleer and S. Edwards, (2013), Early childhood curriculum. Cambridge University Press.

[19] P. Nicholson, (2018), “Play-based pedagogy under threat? A small-scale study of teachers’ and pupils’ perceptions of pedagogical discontinuity in the transition to primary school,” Education 3–13, vol. 47, https://doi.org/10.1080/03004279.2018.1496124, no. 4, pp. 450–461.

[20] T. G. O’Connor, (2013), “Early experiences and psychological development: Conceptual questions, empirical illustrations, and public policy implications,” Journal of Child Psychology and Psychiatry, vol. 54, no. 4, pp. 395–408.

[21] M. Peltoperä, K. Harju, M., Vuorisalo and N. Rutanen, (2023), “Stabilise, balance and adjust: Framing the relational work of parents during children's transitions in early childhood education and care,” Early Years, vol. 44, https://doi.org/10.1080/09575146.2023.2232951, no. 3–4, pp. 301–318.

[22] B. Perry, S. Dockett and A. Petriwskyj, (2014), Transitions to school: Contemporary issues and new directions. Springer.

[23] S. Peters, (2010), Literature review: Transition from early childhood education to school. ResearchGate. https://www.researchgate.net/publication/45646711_Literature_review_Transition_from_early_childhood_education_to_school

[24] S. Peters, C. Hartley and P. Rogers, (2009), “Supporting the transition from early childhood education to school: Insights from one Centre of Innovation project,” Early Childhood Folio, vol. 13, https://doi.org/10.18296/ecf.0177

[25] S. Peters and J. Roberts, (2015), “Transitions from early childhood education to primary school: An interview with Sally Peters,” Set: Research Information for Teachers, vol. 2, https://doi.org/10.18296/set.0012, pp. 3–8.

[26] R. C. Pianta and M. J. Cox, (1999), The transition to kindergarten. Brookes Publishing.

[27] C. L. Price and E. Steed, (2016), “Culturally responsive strategies to support young children with challenging behaviour,” YC Young Children, vol. 71, https://www.researchgate.net/publication/313902813_Culturally_responsive_strategies_to_support_young_children_with_challenging_behavior, no. 5, pp. 36–43.

[28] D. A. Prykanowski, J. R. Martinez, B. Reichow, M. A. Conroy and K. Huang, (2018), "Measurement of young children's engagement and problem behaviour in early childhood settings," Behavioral Disorders, vol. 43, no. 3, pp. 519–529.

[29] S. E. Rimm-Kaufman and R. C. Pianta, (2000), "An ecological perspective on the transition to kindergarten: A theoretical framework to guide empirical research," Journal of Applied Developmental Psychology, vol. 21, no. 5, pp. 491–511.

[30] S. Scholarworks and T. Randolph, (n.d.), Early childhood stakeholders' perspectives on parent empowerment to successfully transition children to formal school. Walden University. https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=13799&context=dissertations

[31] V. Sollars, (2020), Reflecting on "quality" in early childhood education: Practitioners' perspectives and voices. Early Years. https://doi.org/10.1080/09575146.2020.1849034

[32] K. Sylva, E. Melhuish, P. Sammons, I. Siraj-Blatchford and B. Taggart, (2004), The Effective Provision of Pre-School Education (EPPE) Project: Final Report. Institute of Education, University of London.

[33] E. Theodotou, (2019), "An empirical study comparing different art forms to develop social and personal skills in early years education," Education 3–13, vol. 48, https://doi.org/10.1080/03004279.2019.1618890, no. 4, pp. 471–482.

[34] E. Theodotou, (2020), "An empirical study comparing different art forms to develop social and personal skills in early years education," Education 3–13, vol. 48, no. 4, pp. 471–482.

[35] A. Van Gennep, (1909), Les rites de passage. Paris: Émile Nourry.

[36] A. Van Gennep, (1960), The rites of passage (M. B. Vizedom & G. L. Caffee, Trans.). University of Chicago Press.

[37] L. S. Vygotsky, (1978), Mind in society: The development of higher psychological processes. Harvard University Press.

[38] C. Webster-Stratton, (1999), How to promote children's social and emotional competence. SAGE Publications.

Simon, Farrugia: https://orcid.org/0009-0005-2039-2827

MAKING MUSIC, MAKING MEANING: EDUCATOR INSIGHTS ON AUTISM AND MUSICAL ENGAGEMENT

Simon FARRUGIA[1], Kim CRAUS[2]
Malta Leadership Institute, School for Educational Studies, Malta
farrugia.sim@gmail.com kimcraus247@gmail.com

ABSTRACT: This qualitative study explores how primary music educators in Malta engage with the inclusion of students with autism through semi-structured interviews. Drawing on qualitative data from semi-structured interviews with fifteen teachers across state, church, and private schools, the research examines teacher preparedness, instructional strategies, observed student outcomes, collaborative practices, and perceived institutional challenges.

Findings reveal that while most educators support inclusive values, few have received specific training on teaching music to students with autism. Many rely on adaptive strategies developed through experience, including the use of structured routines, visual aids, movement-based activities, and flexible participation. Teachers reported a range of positive outcomes, particularly in emotional expression, focus, and social engagement. However, these outcomes were often achieved in spite of systemic limitations, such as insufficient planning time, lack of resources, and limited collaboration with support staff.

The study is informed by the neurodiversity paradigm, Universal Design for Learning, and sociocultural theory. It contributes to an under-researched area by highlighting the perspectives of music educators within a small-state, policy-driven inclusion system. The findings underscore the need for targeted training, institutional support, and a broader recognition of music's role within inclusive education.

[1] Simon Farrugia is an author and lecturer at the Malta Leadership Institute where he teaches courses related to the field of education. His research interests lie in ethnomusicology, particularly Maltese musical traditions, audiovisual research, and the semiotics of music, and in education, with an emphasis on inclusion, creativity, and the sociology of schooling. His academic work includes a television documentary series on world music and a co-authored book on Maltese historical anthropology, in addition to several publications in music education and ethnomusicology. His most recent publication is the monograph The Maltese Wind Band: A Musical Tradition and Its Practice Today (McFarland, 2023) as well as the ethnographic film Sounds of Weeping: Funeral Marches in Maltese Society and Culture which premiered at the 48th International Council for Traditions of Music and Dance world conference in January 2025 in Wellington, New Zealand. He also serves on the Malta National Committee of RILM (Répertoire International de Littérature Musicale).

[2] Kim Craus is a dedicated Learning Support Educator who has been working in the field since 2016. She holds a degree in Psychology and recently earned a Bachelor's degree in Education, specializing in facilitating and adapting educational programs for students with diverse learning needs. In addition to her professional interests in education and psychology, Kim is also passionate about music. She has been playing the piano since the age of eight and has completed her musical studies up to Grade 8. Kim also serves as the church organist at the St. Venera Parish.

While focused on the Maltese context, the study raises issues relevant to wider international discussions on the inclusion of neurodivergent learners in specialist subjects. Music can offer students with autism meaningful opportunities for engagement, but such outcomes require structured, sustained, and well-supported teaching practice.
Key words: Inclusive music education; autism spectrum disorder; Teacher perspectives; Neurodiversity in classrooms; Universal Design for Learning (UDL)

INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition characterised by differences in social interaction, communication, sensory processing, and repetitive behaviours. The prevalence of autism has risen globally, leading to growing emphasis on inclusion within mainstream education. In this context, music education offers potential for meaningful engagement, particularly for children on the autism spectrum, many of whom display strong auditory memory and sensitivity to sound patterns (Ockelford, 2013; Schlaug et al., 2005).

Malta's education system promotes inclusion through national policy frameworks that encourage the integration of learners with diverse needs into mainstream classrooms. Learning Support Educators (LSEs) are assigned to work with students who require individualised support. Music education is compulsory in the primary curriculum and is delivered by either generalist teachers or itinerant music specialists who may work across multiple schools. Although national policy encourages inclusive practice, there is limited research on how music educators in Malta adapt their methods to accommodate neurodivergent students, particularly those with autism (Zammit-Mangion, 2020).

Internationally, studies suggest that learners with autism often respond positively to music due to its predictability, repetition, and capacity for non-verbal expression (Geretsegger et al., 2014; Lim, 2010). In classroom settings, music can offer structured yet flexible opportunities for participation and interaction. However, research also highlights challenges in implementing inclusive practices in music education, such as a lack of training, limited resources, and insufficient collaboration between educators and support staff (Draper et al., 2024; Hourigan, 2009).

The majority of music education literature in autism contexts is clinical or therapeutic in nature, focusing on individual or small-group interventions delivered by trained music therapists. Less is known about how generalist or school-based music teachers engage students with autism within typical classroom conditions. This gap is particularly apparent in smaller jurisdictions like Malta, where structural and logistical constraints influence how inclusion is enacted on a day-to-day basis.

This study addresses that gap by exploring the experiences of 15 music educators teaching students with autism in Maltese primary schools. It examines their preparedness, teaching strategies, observations of student outcomes, and perceptions of institutional support. The study addressed the

following research questions: How do Maltese primary music educators perceive and enact inclusion for students with autism?; What instructional strategies and institutional supports do they report using?; and What challenges and outcomes do they observe in inclusive music teaching? The aim is to understand how inclusion is operationalised in music classrooms from the perspective of the teachers involved. Ten additional interviews were conducted for this paper, building on an earlier dissertation study, thereby strengthening the data pool and allowing for deeper thematic analysis.

By foregrounding teacher voice, this paper contributes to the broader discourse on inclusive education and highlights music's role as a potential medium for meaningful engagement. While the findings are context-specific, they raise issues of relevance to inclusive practice more broadly, including the need for clearer policy implementation, specialised training, and a more systematic approach to resource provision in music education. In this context, *making music* becomes inseparable from *making meaning*, as students and teachers engage in shared experiences that support connection, expression, and inclusive participation.

1. CONTEXT AND LITERATURE REVIEW

Malta's commitment to inclusive education is reflected in its national education policy, which encourages the full participation of learners with disabilities in mainstream classrooms (Ministry for Education and Employment, 2019). Students diagnosed with Autism Spectrum Disorder (ASD) are typically supported by Learning Support Educators (LSEs) and, where applicable, Inclusion Coordinators (INCOs), who assist teachers in developing and delivering differentiated instruction. Despite this policy framework, the implementation of inclusive practices remains uneven, particularly in specialist subjects such as music (Zammit-Mangion, 2020).

Music is a compulsory subject in the Maltese primary curriculum and is delivered either by generalist class teachers or itinerant music specialists. The nature of music teaching, frequently involving movement, improvisation, auditory stimuli, and group interaction, can offer both opportunities and challenges for children on the autism spectrum. While music often holds intrinsic appeal for learners with autism due to its structured and repetitive elements (Heaton, 2009; Ockelford, 2013), its delivery in group formats without sufficient adaptation may also lead to sensory overload or social withdrawal (Darrow, 2008; Lim, 2010).

1.2. MUSIC AND AUTISM: EVIDENCE FOR ENGAGEMENT

International literature provides substantial evidence that music can positively affect communication, behaviour, and emotional expression in autistic children. Studies such as Geretsegger et al. (2014) and

Boso et al. (2007) highlight the neurological responsiveness of children with autism to musical input. These responses may be attributed to the way music engages multiple brain regions simultaneously, promoting integration across auditory, motor, and affective domains (Koelsch, 2014; Schlaug et al., 2005).

Kern and Aldridge (2006) demonstrated that structured music activities improved engagement and reduced off-task behaviour in students with autism, while Brownell (2002) noted music's capacity to support alternative modes of communication, especially for children with limited verbal skills. These findings are echoed in more recent work by Ruiz et al. (2023), who explored music therapy's role in enhancing brain connectivity and social motivation among children on the spectrum.
Although many of these studies focus on music therapy or clinical interventions, there is growing recognition of music's role within educational contexts. Adamek and Darrow (2018) argue that inclusive music education distinct from therapy, has pedagogical value, especially when designed around student strengths. Similarly, Bunt and Stige (2014) emphasise the socio-musical dimensions of participation, noting that group music-making supports identity and connection in ways that extend beyond therapeutic outcomes.

1.3. TEACHER PERSPECTIVE AND CLASSROOM PRACTICE

Despite growing interest in inclusive music, research focused specifically on teacher perspectives remains limited. A notable exception is Draper et al. (2024), who conducted interviews with US-based elementary music teachers. Their findings reveal that most teachers supported inclusion but lacked autism-specific training and struggled with inconsistent access to classroom support.

These concerns are mirrored in Hourigan's (2009) study, which found that music educators often rely on informal strategies and trial-and-error when teaching students with autism. Similarly, the "Grooving in My Body" project (2023) showed how teachers intuitively modified activities, such as movement exercises and rhythmic games, to accommodate sensory and attentional needs. However, without consistent training or institutional backing, such adaptations remain fragmented and reliant on individual effort.
The issue of preparedness is central to effective inclusion. Teachers frequently report receiving little or no training on how to support students with autism in music classrooms (Draper et al., 2024; Lim, 2010). Many also highlight the lack of inclusive materials, such as visual aids, simplified scores, or sensory-friendly instruments (Drye, 2024). A small number of studies, such as Borg (2020), which focused on

music therapy in a Maltese hospital context, suggest that educators can benefit from collaborative models involving therapists or support staff.

1.4. PEDAGOGICAL FRAMEWORKS FOR INCLUDSION

To understand how music can support inclusion, it is helpful to examine broader educational frameworks. One such model is Universal Design for Learning (UDL), which encourages flexible approaches to teaching by offering multiple means of engagement, representation, and expression (CAST, 2018). Music, by its nature, aligns well with these principles: it can be auditory, visual, kinaesthetic, and emotional.

Furthermore, the neurodiversity paradigm challenges deficit-based views of autism and instead emphasises difference as a natural part of human variation (Singer, 1999; Armstrong, 2010). In music education, this approach encourages teachers to build on the unique cognitive and sensory profiles of their students, rather than viewing autism solely as a barrier to participation.
Incorporating these frameworks into teacher training may help bridge the gap between policy and practice. However, as Zammit-Mangion (2020) and others note, structural factors, including timetabling, staffing models, and resource availability, continue to affect the consistency of inclusive implementation in small-state settings such as Malta.

1.5. SUMMARY

The reviewed literature supports the notion that music holds unique potential for engaging students with autism, both therapeutically and educationally. However, realising this potential in mainstream classrooms requires more than enthusiasm; it depends on well-trained educators, accessible resources, and institutional systems that facilitate collaboration. While Malta's inclusive policy landscape is well-established, further empirical research is needed to understand how music educators interpret, adapt, and enact these policies in their daily teaching practice.

This study contributes to that effort by exploring the voices of music educators in Malta who have taught students with autism within inclusive settings. Their insights offer a grounded view of how inclusion operates on the classroom floor, illuminating both the opportunities music provides and the gaps that remain.

2. CONCEPTUAL FRAMEWORK

This study draws on three interrelated frameworks to guide its analysis of inclusive music teaching for students with autism: the neurodiversity paradigm, Universal Design for Learning (UDL), and sociocultural theory.

2.1. NEURODIVERSITY PARADIGM

The neurodiversity framework challenges deficit-based models of autism, instead recognising neurological differences as part of human variation (Singer, 1999; Armstrong, 2010). From this perspective, autistic traits, such as sensory sensitivity, focused interests, and pattern recognition, are viewed not solely as impairments but as cognitive differences that can be strengths in specific contexts. Music, with its structural regularity, repetition, and potential for non-verbal expression, may align particularly well with some of these strengths (Ockelford, 2013; Heaton, 2009).

This paradigm encourages educators to reframe their understanding of student behaviour and learning styles. Rather than aiming to 'normalise' students with autism, teachers are invited to create learning environments that respect diverse ways of thinking and processing information. In the context of music education, this may involve flexible participation structures, sensory adaptations, or valuing alternative forms of expression and communication.

2.2. UNIVERSAL DESIGN FOR LEARNING (UDL)

UDL is an educational framework developed to support all learners by offering multiple means of engagement, representation, and action (CAST, 2018). Its application in inclusive education is well established, yet its integration into music teaching, particularly for students with autism, remains underexplored.

In music classrooms, UDL may be realised through differentiated activities, varied instrument choices, visual supports, and opportunities for movement and improvisation. These adaptations allow learners to access the curriculum in ways that align with their individual profiles. Recent studies (e.g. Draper et al., 2024; Drye, 2024) suggest that many teachers already use such strategies informally, even if they are not explicitly trained in UDL.

By aligning pedagogical choices with UDL principles, music educators can reduce barriers to participation while maintaining high expectations for all students. This also supports learners with autism who may need alternative ways of demonstrating understanding or engaging in classroom routines.

2.3. SOCIOCULTURAL THEORY

Sociocultural theory, rooted in the work of Vygotsky (1978), emphasises the social nature of learning and the importance of cultural tools, such as language, symbols, and artefacts, in mediating development. Music, as both a cultural practice and a form of social interaction, can be considered a powerful mediational tool in inclusive education.

For students with autism, musical participation may serve as a bridge for social connection, joint attention, and shared meaning-making. Collaborative music-making activities, such as group rhythm exercises or call-and-response tasks, create structured social contexts where learners can engage at different levels of complexity (Geretsegger et al., 2014; Bunt & Stige, 2014). Within these settings, teachers play a critical role in scaffolding interaction and modelling inclusive participation.
Together, these three frameworks offer a foundation for understanding the pedagogical and relational dynamics at play in inclusive music classrooms. They support an approach that is not only adaptive but affirming, one that recognises and builds upon the capacities of all learners, including those on the autism spectrum. During the data analysis process, these frameworks informed the interpretation of themes, particularly in identifying teacher practices aligned with UDL, neurodiversity-informed approaches, and socially mediated learning.

3. METHODOLOGY

3.1. RESEARCH DESIGN

This study employed a qualitative research design to explore the experiences of music educators who have taught students with autism in Maltese primary schools. A phenomenological approach was adopted to capture how teachers perceive and make sense of their inclusive practice (Creswell and Poth, 2018). The aim was not to generalise findings, but to provide a detailed account of educators' interpretations, strategies, and concerns when teaching music to learners with additional needs in mainstream settings.

3.2. PARTICIPANTS

Fifteen music educators participated in the study. Five were drawn from a previous dissertation study, and ten were newly recruited. Participants were selected using purposive sampling to ensure that all had experience teaching at least one student with autism in a mainstream primary school. They represented a range of teaching backgrounds, including class teachers with musical responsibilities,

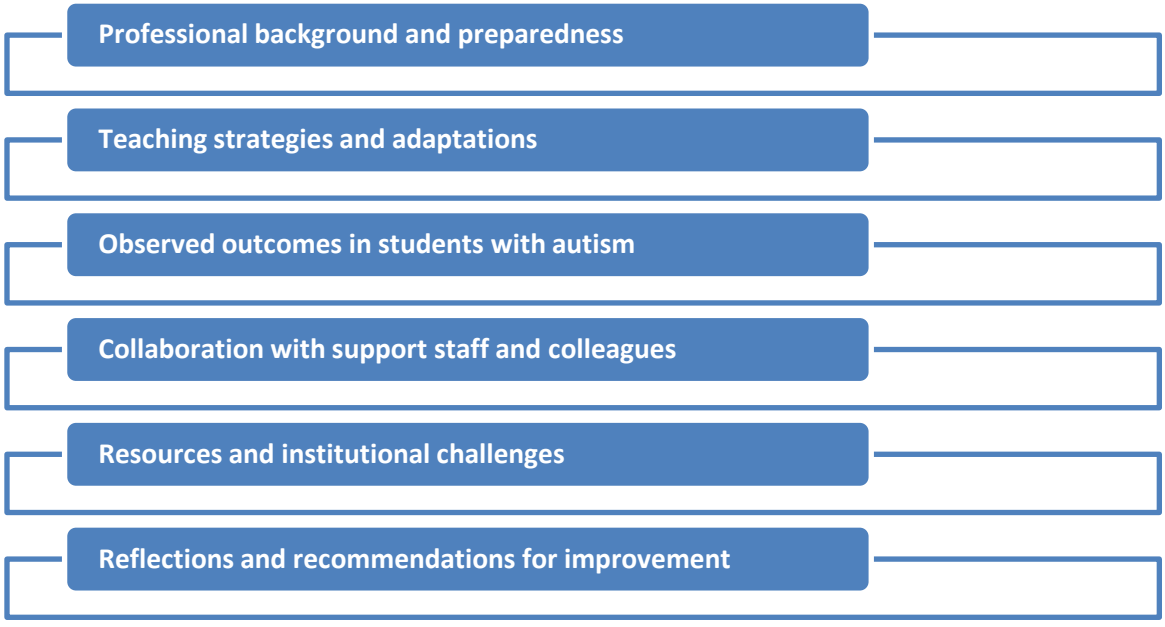
itinerant music specialists, and freelance music educators working within school programmes. Teaching experience ranged from three to over 20 years.

Participants were drawn from state, church, and private schools across Malta. Their teaching contexts varied in terms of institutional support, class sizes, and available resources. Most had some experience collaborating with Learning Support Educators (LSEs), but the frequency and depth of this collaboration differed across cases.

3.3. DATA COLLECTION

Semi-structured interviews were used as the primary method of data collection. Interviews lasted approximately 30 to 45 minutes and were conducted either in person or via online platforms, depending on the participant's preference. Each interview was audio-recorded with consent and later transcribed verbatim for analysis.

The interview guide included open-ended questions organised under six thematic domains:



These domains were informed by both the previous study and existing literature on inclusive music education (e.g. Draper et al., 2024; Hourigan, 2009).

3.4. ETHICAL CONSIDERATIONS

Ethical approval was obtained from the relevant ethics committee prior to data collection. All participants received an information sheet outlining the aims of the study and signed informed consent forms.

Participation was voluntary, and individuals could withdraw at any stage without consequence. Anonymity was ensured, and all identifying details were removed from transcripts and reporting.

3.5. DATA ANALYSIS

Data were analysed using thematic analysis, following the six-step model proposed by Braun and Clarke (2006). Transcripts were read multiple times for familiarity before being coded inductively. Codes were then grouped into potential themes, which were refined through iterative comparison and alignment with the research questions.

The final themes included:

Although the analysis was interpretative, care was taken to remain close to participants' language and meaning. Quotes were selected to illustrate key patterns while preserving individual voice and variation.

3.6. TRUSTWORTHINESS

Several strategies were employed to ensure trustworthiness, including data triangulation between the new and original samples, reflexive journaling by the primary researcher to address potential bias, and peer debriefing with the second author to test and refine emerging themes.

This approach ensured that the findings were both grounded in the data and interpreted through a transparent and reflective process.

4. FINDINGS AND DISCUSSION

This section presents the findings from interviews with 15 music educators, structured under six thematic categories. The themes emerged inductively and were consistent across both the original and new interview datasets. The discussion integrates relevant literature to contextualise teacher perspectives and illuminate implications for inclusive music practice.

4.1. TEACHER PREPAREDNESS AND EXPERIENCE

Most participants reported limited formal training related to autism or inclusive music pedagogy. While several had attended general inclusion workshops or brief continuing professional development sessions, few had received music-specific training tailored to working with students on the autism spectrum.

> "I've picked up things from experience and talking with LSEs… but no real training" (Teacher 1).

This mirrors the findings of Hourigan (2009) and Draper et al. (2024), who note that music teachers often rely on informal learning and on-the-job adaptation. Several teachers in the current study stated that their confidence increased through direct experience, yet they also acknowledged gaps in their initial preparedness.

Some participants expressed a desire for training that was practical, classroom-oriented, and sensitive to the realities of mixed-ability group teaching. Their reflections align with research by Adamek and Darrow (2018), who argue for targeted professional development that integrates pedagogical and neurodevelopmental insights.

4.2. TEACHING STRATEGIES AND ADAPTATIONS

Teachers described a wide range of strategies used to engage students with autism in music lessons. These included:

- Structured routines (e.g., predictable lesson formats)
- Visual supports (e.g., charts, cards, colour-coded instruments)
- Tactile and movement-based activities (e.g., drumming, dancing, scarf work)
- Simplification of content (e.g., adapting rhythms or using repetition)
- Choice-based participation (e.g., allowing students to select instruments)

> "Repetition and rhythm games are great. Some students love using boomwhackers or colour-coded notes" (Teacher 7).

These strategies are consistent with principles of Universal Design for Learning (CAST, 2018), which promote multiple means of engagement and representation. They also reflect literature on music's

accessibility for learners with autism (Geretsegger et al., 2014; Lim, 2010). For example, movement-based learning aligns with findings from the "Grooving in My Body" study, which emphasised the role of embodied engagement in sustaining attention and regulating behaviour (2023).

A key observation across interviews was the importance of flexibility. Teachers noted that rigid lesson plans often failed with neurodivergent learners. Instead, they advocated for adaptive teaching that balanced structure with responsiveness.

4.3. STUDENT ENGAGEMENT AND OBSERVED OUTCOMES

Educators reported that music had positive effects on many students with autism, particularly in terms of emotional expression, focus, and confidence. Several described breakthroughs where students who typically avoided verbal interaction engaged through music.

> "One student who barely spoke started humming along in class. It felt like a breakthrough" (Teacher 2).

Teachers also observed that music supported self-regulation and reduced anxiety for some learners, echoing the findings of Ruiz et al. (2023) and Boso et al. (2007), who link musical activity with neural integration and affective stability. Others described how individual students found 'their voice' in music, often through instruments or solo tasks.

"I had a student who hated group work but loved solo singing. That's where he excelled" (Teacher 6).

These accounts suggest that music provides a structured yet flexible environment in which students with autism may engage on their own terms. Such findings indicate that music can support inclusive goals when educators are empowered to adapt practice responsively. The findings support earlier research on music as an alternative communication channel (Brownell, 2002; Kern and Aldridge, 2006).

4.4. INSTITUTIONAL SUPPORT AND COLLABORATION

Teachers offered mixed accounts of the support they received. In some cases, collaboration with LSEs and Inclusion Coordinators was described as productive and regular. Elsewhere, participants reported feeling isolated, especially when teaching music in a 'specialist' capacity with limited contact with class teams.

> "Honestly, I feel a bit on my own. I try to ask for help, but support is limited" (Teacher 4).

Time constraints and scheduling conflicts were commonly cited as barriers to collaboration. Participants noted that while LSEs were essential partners, joint planning was rarely feasible due to systemic pressures. These findings echo those of Draper et al. (2024) and Drye (2024), who highlight structural rather than attitudinal barriers to inclusion.

Teachers also raised concerns about inconsistent policies between schools, with some institutions placing strong emphasis on inclusion and others appearing less committed. As in Zammit-Mangion's (2020) study, the variability in local implementation was seen as a source of tension.

4.5. RESOURCES AND CHALLENGES

Resource availability emerged as a critical factor influencing teachers' capacity to implement inclusive music lessons. Frequently mentioned resources included:

- Visual aids (e.g., charts, symbols)
- Sensory-sensitive instruments (e.g., soft drums, shakers)
- Digital tools (e.g., iPads, music apps)

However, many educators stated that they sourced these materials themselves, as schools did not consistently provide them.

> "I get creative … beanbags, drums. It's tricky when I don't know much about a student beforehand" (Teacher 5).

Participants also reported difficulties in managing group dynamics, especially when balancing the needs of neurodivergent learners with the rest of the class. Some expressed concern about fairness or burnout when support was limited.

The challenges identified are well documented in the literature, with previous studies pointing to inadequate training, lack of tailored materials, and insufficient planning time as persistent barriers (Lim, 2010; Hourigan, 2009; Darrow, 2008).

4.6. REFLECTIONS AND RECOMMENDATIONS

Despite the challenges, many participants expressed a strong belief in the potential of music as an inclusive practice. Their reflections suggested a shared understanding that music could offer a unique and valuable entry point for engagement.

"Music gives structure and freedom at the same time. That's why it works" (Teacher 6).

The reflections outlined above align with the recommendations presented in the following section. They highlight recurring priorities across interviews, including the need for targeted training, institutional planning time, and resourcing that reflects inclusive aims.

These suggestions align closely with proposals in the wider literature. For instance, Adamek and Darrow (2018) argue for the embedding of inclusion training in music education courses, while CAST (2018) and Armstrong (2010) advocate for systemic adoption of UDL principles.

4.7. SUMMARY

Across all six themes, a consistent narrative emerges: music has clear potential to engage students with autism, but effective inclusion depends on training, resources, and institutional support. Teachers show willingness and creativity, yet too often operate in systems that do not prioritise or adequately support inclusive practice. The findings echo international concerns while providing specific insight into the Maltese context.

5. IMPLICATIONS AND RECOMMENDATIONS

The findings of this study highlight both the promise and the limitations of inclusive music education for students with autism in Maltese primary schools. While teachers demonstrated commitment and adaptability, their efforts were often constrained by structural, institutional, and pedagogical factors. Drawing on the themes discussed, several implications and recommendations can be identified for policy, practice, and future research.

5.1. PROFESSIONAL TRAINING AND INITIAL TEACHER EDUCATION

A clear and recurring issue was the limited training received by music educators on how to support learners with autism and other additional needs. Most participants had not encountered inclusion-specific content during their initial teacher education, and few had accessed targeted in-service training. There is therefore a strong case for embedding autism-specific strategies within both pre-service and ongoing professional development. This should include:

- Practical, classroom-based approaches to adapting music lessons
- Collaboration techniques with Learning Support Educators and families
- Awareness of sensory regulation and communication needs in music contexts

These recommendations are supported by international studies (e.g. Draper et al., 2024; Adamek and Darrow, 2018) and align with calls for more consistent and specialised training within the field of inclusive arts education.

5.2. EMBEDDING UNIVERSAL DESIGN FOR LEARNING IN MUSIC

While many teachers intuitively adopted inclusive practices, their approaches were often piecemeal and unsupported by formal frameworks. Universal Design for Learning (CAST, 2018) offers a suitable model for building consistent inclusion into the structure of music education.

Training and curriculum development should encourage teachers to:

- Provide multiple entry points to music-making (e.g. movement, listening, visual symbols)

- Use flexible assessment practices
- Incorporate student voice and choice within lesson design

This approach would benefit not only students with autism but all learners, by reducing barriers to participation and broadening the definition of musical success.

5.3. INSTITUTIONAL COLLABORATION AND TIME ALLOCATION

A recurring theme in the data was the fragmented nature of collaboration between music educators, LSEs, and classroom teachers. When collaboration occurred, it was typically informal and reactive, rather than planned or embedded in school systems.

There is a need for:

- Scheduled planning time between music teachers and LSEs
- Clear communication channels across teaching teams
- School-level leadership to prioritise inclusion across the curriculum

As shown in studies such as Drye (2024) and Zammit-Mangion (2020), inclusion cannot rely solely on individual goodwill. Systemic structures must facilitate coordinated and sustained support.

5.4. RESOURCES AND ENVIRONMENT

Participants frequently described a lack of accessible resources, including visual aids, adapted instruments, and sensory-friendly materials. This often led to frustration and personal expense. Policy-makers and school leaders should prioritise the provision of:

- Visual and tactile learning tools
- Noise-sensitive instruments
- Digital applications for music and communication

Schools should also review the physical environment of music spaces, ensuring they are appropriate for learners with sensory sensitivities. These adjustments do not require major investment, but they do require planning and commitment.

5.5. VALUING MUSIC WITHING INCLUSIVE POLICY

Finally, there was a perceived disconnect between inclusive education policies and the status of music within schools. Several participants felt that music was undervalued or treated as an extra-curricular concern, limiting its potential to contribute meaningfully to inclusion.

Given the evidence that music can support emotional regulation, communication, and social participation (Geretsegger et al., 2014; Ruiz et al., 2023), school and national policies should:

- Recognise music education as a legitimate component of inclusion

- Include arts education in national guidelines for supporting neurodivergent learners
- Encourage research and development within inclusive music pedagogy

5.6. SUMMARY

Music educators are well placed to contribute to inclusive education, particularly for students with autism, yet they require clearer support, better training, and stronger systemic alignment. By addressing these areas, schools can enhance not only access but also the quality of engagement and learning outcomes for all students.

6. CONCLUSION

This study explored the experiences of fifteen music educators teaching students with autism in Maltese primary schools. Through semi-structured interviews, it examined how teachers navigate inclusion in their practice—considering their training, strategies, observations of student outcomes, institutional support, and challenges encountered.

The findings indicate that while teachers often demonstrate creativity, flexibility, and commitment, they operate in systems that do not consistently prioritise or resource inclusive music education. Most participants reported limited formal training related to autism, and while some adapted intuitively to learners' needs, their approaches were typically developed through trial and error rather than structured guidance. Collaboration with Learning Support Educators and classroom staff was described as valuable, yet inconsistent—often hindered by lack of time and systemic coordination.

Despite these limitations, teachers observed positive outcomes for many of their students with autism. These included improved engagement, emotional expression, and social interaction. Music was seen as a medium that offered both structure and openness—a space where students could participate in ways not always possible in other classroom settings.

The study's findings align with international literature that emphasises music's potential as a tool for inclusion, particularly for neurodivergent learners (Geretsegger et al., 2014; Ockelford, 2013). However, realising this potential requires more than isolated acts of goodwill. It calls for system-level changes: embedded training, consistent access to adapted resources, planned collaboration time, and a clear recognition of music's role within inclusion policy.

This research contributes to an under-researched area by examining the intersection of autism, music education, and teacher experience in the specific context of a small island education system. By foregrounding the voices of educators, it provides grounded insights into the practical realities of inclusive teaching. While the findings are based on the Maltese context, they hold relevance for broader discussions on inclusive practice in specialist subjects.

Future research may build on this study by examining student perspectives or exploring the impact of specific training interventions on teaching confidence and classroom outcomes. Comparative studies across small states could also offer valuable insight into how local systems mediate the implementation of inclusive education policies.

Music offers learners with autism a structured space for communication and participation. Realising this potential requires sustained systemic investment in educator training, support structures, and inclusive resources.

REFERENCES

[1] Adamek, M. S., & Darrow, A. A. (2018). *Music in special education* (3rd ed.). American Music Therapy Association.

[2] Armstrong, T. (2010). *The power of neurodiversity: Unleashing the advantages of your differently wired brain*. Da Capo Lifelong Books.

[3] Borg, A. (2020). The impact of music therapy on children and adolescents with cancer in a hospital setting. *[Unpublished Dissertation]*. University of Malta.

[4] Boso, M., Emanuele, E., Minazzi, V., Abbamonte, M., & Politi, P. (2007). Effect of long-term interactive music therapy on behaviour profile and musical skills in young adults with severe autism. *The Journal of Alternative and Complementary Medicine, 13*(7), 709–712.

[5] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

[6] Brownell, M. D. (2002). Musically adapted social stories to modify behaviours in students with autism: Four case studies. *Journal of Music Therapy, 39*(2), 117–144.

[7] Bunt, L., & Stige, B. (2014). *Music therapy: An art beyond words* (2nd ed.). Routledge.

[8] CAST. (2018). *Universal Design for Learning Guidelines version 2.2*. CAST, Inc. https://udlguidelines.cast.org

[9] Darrow, A. A. (2008). Music educators' perceptions regarding the inclusion of students with disabilities in the music classroom. *Update: Applications of Research in Music Education, 26*(2), 23–30.

[10] Draper, E. A., Runfola, M., & Barrett, J. R. (2024). Identifying elements of inclusion: Interviews with elementary music teachers about their students with disabilities. *Journal of Research in Music Education*, Advance online publication. https://doi.org/10.1177/00224294241238607

[11] Drye, J. B. (2024). Children's social preference for teachers versus peers in autism inclusion classrooms: An eye-tracking study. *Autism Research, 17*(3), 482–495. https://doi.org/10.1002/aur.3158

[12] Geretsegger, M., Elefant, C., Mössler, K. A., & Gold, C. (2014). Music therapy for people with autism spectrum disorder. *Cochrane Database of Systematic Reviews*, (6), CD004381. https://doi.org/10.1002/14651858.CD004381.pub3

[13] Heaton, P. (2009). Assessing musical skills in autistic children who are not savants. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1522), 1443–1447.

[14] Hourigan, R. M. (2009). Preservice music teachers' perceptions of fieldwork experiences in a special education classroom. *Journal of Research in Music Education, 57*(2), 152–168.

[15] Kern, P., & Aldridge, D. (2006). Using embedded music therapy interventions to support outdoor play of young children with autism in an inclusive community-based child care program. *Journal of Music Therapy, 43*(4), 270–294.

[16] Koelsch, S. (2014). Brain correlates of music-evoked emotions. *Nature Reviews Neuroscience, 15*(3), 170–180.

[17] Lim, H. A. (2010). Effect of "Developmental Speech and Language Training Through Music" on speech production in children with autism spectrum disorders. *Journal of Music Therapy, 47*(1), 2–26.

[18] Ministry for Education and Employment. (2019). *A national inclusive education policy*. Government of Malta.

[19] Ockelford, A. (2013). *Music, language and autism: Exceptional strategies for exceptional minds*. Jessica Kingsley Publishers.

[20] Ruiz, S. M., Honzel, N., & Dvorak, R. (2023). Music therapy and connectivity in autism: New evidence from EEG and parent report. *Frontiers in Psychology, 14*, Article 1256771. https://doi.org/10.3389/fpsyt.2023.1256771

[21] Schlaug, G., Norton, A., Overy, K., & Winner, E. (2005). Effects of music training on the child's brain and cognitive development. *Annals of the New York Academy of Sciences, 1060*(1), 219–230.

[22] Singer, J. (1999). *Why can't you be normal for once in your life? From a "problem with no name" to the emergence of a new category of difference*. In M. Corker & S. French (Eds.), *Disability discourse* (pp. 59–67). Open University Press.

[23] Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

[24] Zammit-Mangion, D. (2020). Inclusion in a small island state: Reflections from Maltese teachers. *International Journal of Inclusive Education, 24*(5), 486–501.

 Simon, Farrugia: https://orcid.org/0009-0005-2039-2827

PEDAGOGICAL BOUNDARIES AND COMPETENCE PRESERVATION IN AI-ASSISTED LEARNING ENVIRONMENTS

Elek TÓTH

Independent Researcher, Eötvös Loránd University community, Budapest, Hungary

jv5rrk@inf.elte.hu

ABSTRACT: The integration of artificial intelligence into education has become one of the most significant pedagogical transformations of recent years. AI-based systems can provide personalized support, rapid feedback, and reduce certain routine cognitive burdens. However, current discourse primarily emphasizes efficiency and accessibility benefits, while significantly less attention is devoted to the long-term effects of excessive cognitive delegation on competence development.

This paper proposes a pedagogical framework for designing competence-preserving AI-assisted learning environments. Drawing on the literature of cognitive offloading, productive struggle, desirable difficulties, and metacognition, the study examines which cognitive processes may involve pedagogical risks when excessively delegated. Particular attention is given to problem decomposition, abstraction, independent debugging, reflective verification, sustained attention, and productive cognitive effort. The paper argues that the central challenge of educational AI integration is not the rejection or restriction of technology, but the conscious definition of pedagogical delegation boundaries. In particular, the study distinguishes between AI functioning as scaffolding, which supports learner participation and cognitive development, and AI functioning as replacement, which may bypass cognitively formative learning processes. Consequently, educational systems should be designed not only for performance optimization, but also for the long-term sustainability of human competencies.

Key words: Artificial Intelligence in Education; Cognitive Offloading; Competence Preservation; Metacognition; Productive Struggle

INTRODUCTION

Artificial intelligence is rapidly reshaping not only educational technologies, but also the cognitive structure of learning itself. Large language model–based systems such as ChatGPT, Gemini, and Copilot no longer function merely as information retrieval tools, but increasingly operate as active cognitive partners capable of generating text, solving problems, constructing arguments, and simulating complex reasoning processes. Current educational discourse, however, primarily emphasizes the efficiency-related benefits of AI integration, including personalized support, rapid feedback, increased accessibility, and the automation of various learning processes [1].

At the same time, growing attention has been directed toward the long-term implications of delegating cognitive processes to AI systems. Generative artificial intelligence does not merely store or display information; it is increasingly capable of supporting or partially replacing higher-order cognitive operations such as analysis, abstraction, reasoning, and evaluation [2]. In this sense, the current transformation differs qualitatively from earlier forms of technological offloading such as note-taking, calculators, or search engines. While previous technologies primarily externalized memory or information access, generative AI systems increasingly externalize cognitive processing itself. Earlier debates surrounding digital technologies already raised concerns regarding the weakening of sustained attention, deep reading, and reflective thinking in online environments [3].

Knowing the correct answer does not necessarily mean that meaningful learning has taken place. Research in educational psychology and cognitive science has long emphasized that deep learning fundamentally depends on productive cognitive effort, error correction, reflective verification, and independent problem-solving. The theory of desirable difficulties argues that learning conditions which initially appear slower, more effortful, or less efficient often produce stronger long-term retention and transfer [4]. Consequently, cognitive effort should not necessarily be interpreted as an obstacle to learning, but rather as one of its essential developmental conditions. At the same time, cognitive load theory suggests that excessive and unstructured mental overload may hinder learning efficiency, particularly for novice learners [5]. Therefore, the pedagogical challenge is not the elimination of all cognitive support, but the preservation of cognitively formative effort while reducing unnecessary overload.

Similarly, the pedagogical concept of productive struggle highlights that meaningful mathematical and problem-solving activity requires a controlled level of cognitive challenge through which learners remain capable of progressing autonomously [6]. The purpose of productive struggle is not the maximization of frustration, but the creation of cognitively formative situations in which learners actively construct meaning, reflect on their reasoning processes, and gradually build durable competencies [7].
From this perspective, the pedagogical problem of AI integration extends beyond technological or ethical concerns and becomes deeply connected to the nature of cognitive development itself. If learners consistently delegate processes that previously served as primary spaces of competence formation, there is a risk that cognitively formative activity itself becomes diminished. Recent research increasingly suggests that excessive AI dependence may contribute to weaker critical thinking, reduced metacognitive control, and declining independent problem-solving abilities [1].

This issue becomes particularly important in relation to metacognition. Effective learners are capable of monitoring their own understanding, identifying gaps in knowledge, and regulating their learning strategies [8]. AI systems, however, often shorten or bypass precisely these reflective processes by providing immediate answers, pre-structured reasoning, and automated problem solutions. Consequently, AI may reduce not only cognitive workload, but also the learner's metacognitive engagement.

The aim of this paper is therefore to examine which forms of cognitive delegation within AI-assisted learning environments may involve pedagogical risks. The study argues that the central challenge of educational AI integration is not whether AI should be used, but rather how pedagogical boundaries of delegation should be defined. Drawing on the literature of cognitive offloading, productive struggle, desirable difficulties, and metacognition, the paper seeks to propose a competence-preserving pedagogical perspective that supports the long-term maintenance of autonomous cognitive activity.

I. COGNITIVE DELEGATION IN EDUCATIONAL CONTEXTS

The concept of cognitive offloading refers to processes in which individuals delegate certain cognitive operations to external tools, technologies, or environmental supports in order to reduce mental workload [9]. The phenomenon itself is not new. Human cultural development can partly be interpreted as the progressive externalization of cognitive burdens. Some philosophical approaches have even argued that external cognitive tools may become functional extensions of the human mind itself [10]. Writing systems, books, calculators, maps, and digital calendars all function as cognitive technologies that extend human capabilities [11].

Traditional forms of cognitive offloading, however, were predominantly passive in nature. They stored, organized, or displayed information, but did not directly replace reasoning processes themselves. Generative AI systems differ substantially in this regard, functioning increasingly as active cognitive mediators [12]. Rather than merely retrieving information, these systems generate arguments, synthesize sources, produce texts, decompose problems, and propose alternative solutions. In this sense, generative AI changes not only how people access information, but also how cognitive work itself is distributed.

The pedagogical significance of this shift becomes particularly apparent when AI systems begin to replace not only routine tasks, but also higher-order cognitive operations. Generative AI is already capable of writing essays, generating program code, solving mathematical tasks, and summarizing scientific literature [2]. Consequently, delegation increasingly affects processes such as abstraction,

analytical reasoning, problem decomposition, and evaluation rather than merely memory-based operations.

At the same time, the literature on cognitive offloading emphasizes that externalization itself should not automatically be considered harmful. Under certain conditions, offloading may be highly adaptive because it frees cognitive resources for more complex forms of thinking [12]. External memory aids and organizational supports can significantly improve performance, particularly under conditions of high cognitive load. Certain forms of AI assistance may become pedagogically counterproductive precisely because they remove cognitively formative stages from the learning process.

Problems emerge when offloading becomes excessive and persistently replaces cognitively formative processes that previously functioned as central mechanisms of competence development. Several studies suggest that learners tend to over-rely on external supports, particularly when they possess lower confidence in their own competencies [12]. Over time, such dependence may contribute to the weakening of internal cognitive strategies and autonomous reasoning abilities.

Metacognition plays a particularly important role in this process. Metacognitive monitoring enables learners to recognize knowledge gaps, regulate learning strategies, and determine when external support is genuinely necessary [13]. AI systems, however, may easily create an illusion of competence in which learners feel they “understand” a problem while merely accessing externally generated solutions. This aligns with recent studies discussing the phenomenon of the “illusion of learning” within AI-assisted environments [14].

The classical pedagogical concept of the Zone of Proximal Development (ZPD) defines learning as occurring most effectively within a space where learners can progress with temporary external support [15]. Recent scholarship suggests, however, that permanently available AI assistance may unintentionally undermine this developmental dynamic. dos Santos Jr. and Birdwell describe this condition as the “Zone of No Development” (ZND), a learning state in which constant assistance suppresses productive cognitive struggle and weakens genuine developmental growth [14]. In this sense, the ZND concept represents a critical inversion of the original ZPD framework. While scaffolding within the ZPD is intended to gradually transfer cognitive responsibility back to the learner, permanently available AI assistance may interrupt this transition. As a result, learners may remain in a state of externally supported performance without fully internalizing the underlying cognitive processes themselves.

One of the most important pedagogical challenges of contemporary AI integration is therefore determining which forms of delegation support competence development and which may weaken it. Educational systems must decide not only how to increase efficiency and performance, but also how to preserve autonomous reasoning as an active human practice within learning processes.

II. COMPETENCIES POTENTIALLY AT RISK

The pedagogical implications of cognitive delegation become particularly significant when AI systems begin to replace not only routine operations, but also cognitively formative learning activities. Importantly, the concern is not that human competencies suddenly disappear through AI usage, but rather that opportunities for their active development and reinforcement may gradually diminish. From a pedagogical perspective, competence development depends not only on successful task completion, but also on repeated engagement in cognitively demanding processes. When such processes are consistently externalized, the developmental function of learning activities may weaken over time.

A. Foundational Procedural Competencies

Certain foundational competencies are especially vulnerable to excessive cognitive delegation because they rely heavily on repeated practice and procedural reinforcement. These include mental calculation, symbolic manipulation, syntactic fluency, and the automation of basic disciplinary operations. While AI systems can significantly accelerate task completion in these areas, overreliance on automated assistance may reduce opportunities for procedural consolidation.

This concern is not entirely new within educational research. Similar debates emerged previously around calculators, spelling correction systems, and digital navigation technologies. However, generative AI extends delegation beyond isolated operations and increasingly performs entire procedural sequences autonomously. As a result, learners may complete tasks successfully without sufficiently engaging in the underlying cognitive operations themselves.

The theory of desirable difficulties suggests that procedural effort and repeated retrieval are not obstacles to learning, but essential mechanisms of durable memory consolidation and transfer. Learning conditions that initially appear slower or more effortful often produce stronger long-term retention precisely because they require sustained cognitive engagement [4]. Consequently, the complete automation of procedural activities may unintentionally reduce opportunities for strengthening foundational competencies.

B. Higher-Order Cognitive Competencies

Beyond procedural skills, AI-assisted environments may also affect higher-order cognitive competencies such as abstraction, analytical reasoning, problem decomposition, model construction, and independent debugging. These competencies are particularly important because they enable learners not merely to reproduce knowledge, but to structure, interpret, and transform information autonomously.

Generative AI systems increasingly perform many of these operations directly. They can propose argumentative structures, decompose complex tasks into smaller components, generate explanations, and even suggest reasoning pathways. While such support may improve short-term performance and accessibility, it also raises important pedagogical questions concerning the learner's active cognitive participation.

Research on productive struggle emphasizes that meaningful cognitive development often emerges through sustained engagement with uncertainty, partial failure, and iterative reasoning processes [6]. Higher-order competencies develop not simply through exposure to correct solutions, but through active participation in the reasoning process itself. If learners consistently bypass these cognitively formative stages through automated assistance, opportunities for developing autonomous reasoning strategies may decrease.

This issue becomes especially relevant in domains requiring complex problem-solving, including mathematics, programming, scientific reasoning, and academic writing. In such contexts, the developmental value of learning frequently lies not only in the final solution, but in the cognitive processes required to construct it.

C. Metacognitive Competencies

Metacognitive competencies may represent one of the most sensitive areas affected by excessive AI dependence. Metacognition involves the learner's ability to monitor understanding, evaluate reasoning quality, identify knowledge gaps, and regulate learning strategies [8]. These reflective processes are essential for autonomous learning and long-term intellectual independence.

AI systems, however, frequently provide highly fluent, coherent, and authoritative responses that may reduce the learner's need to engage in reflective verification. As a consequence, learners may increasingly rely on externally generated reasoning without sufficiently evaluating its validity, limitations, or underlying assumptions. A student using AI-generated mathematical solutions, for example, may follow

the procedural steps without recognizing conceptual misunderstandings. Similarly, AI-assisted academic writing may produce structurally coherent texts that learners are unable to independently defend, revise, or critically evaluate. In programming contexts, automatic debugging tools may correct errors successfully while bypassing the learner's own diagnostic reasoning process.

The problem is therefore not limited to factual errors or inaccurate outputs. Even accurate AI-generated outputs may weaken metacognitive engagement if learners gradually shift from active evaluation toward passive acceptance of externally generated reasoning. Over time, learners may become less likely to question whether they genuinely understand a concept, whether alternative interpretations exist, or whether a solution strategy is appropriate in a different context.

The literature on effort monitoring further suggests that learners often misinterpret cognitive ease as evidence of successful learning [16]. In AI-assisted environments, the immediate availability of polished explanations and solutions may strengthen this illusion of learning, particularly when learners no longer experience the productive cognitive effort typically associated with deeper understanding. The fluency and confidence of AI-generated responses may therefore create a misleading sense of mastery despite relatively shallow cognitive processing.

Recent discussions surrounding the "Zone of No Development" similarly warn that permanently available assistance may gradually reduce self-regulated reasoning and cognitive autonomy [14]. From a pedagogical perspective, this raises important questions regarding how AI systems should be designed to preserve the learner's reflective role within the learning process. AI-supported environments may therefore require mechanisms that actively encourage verification, justification, uncertainty evaluation, and reflective self-assessment rather than merely optimizing immediate task completion.

D. Cognitive Endurance and Sustained Attention

Another potentially affected domain involves cognitive endurance, including sustained attention, deep processing, persistence during uncertainty, and tolerance for cognitively demanding tasks. Many educationally valuable activities require learners to maintain attention over extended periods while working through incomplete understanding, temporary confusion, or iterative problem-solving processes.

Generative AI systems may unintentionally reduce exposure to these cognitively effortful situations by continuously shortening reasoning pathways and minimizing friction within the learning process.

Although this often improves efficiency and user experience, it may simultaneously reduce opportunities for developing persistence and cognitive resilience.

The desirable difficulties framework highlights that learning effectiveness is frequently associated with conditions that feel effortful and cognitively demanding [4]. Similarly, research on productive struggle argues that controlled cognitive difficulty is not merely tolerable, but developmentally valuable [4][6]. If educational systems increasingly optimize learning exclusively for speed, fluency, and immediate task success, there is a risk that cognitively formative forms of sustained mental effort become progressively marginalized.

From this perspective, the pedagogical challenge of AI integration extends beyond questions of technological capability or instructional efficiency. It increasingly concerns the preservation of those cognitive conditions through which durable competencies, reflective autonomy, and intellectual resilience are developed.

III. PEDAGOGICAL BOUNDARIES OF DELEGATION

The preceding discussion suggests that the pedagogical challenge of artificial intelligence in education is not simply a question of technological adoption, but rather one of instructional boundary-setting. From this perspective, the central issue is not whether cognitive delegation should occur at all, but which processes may be delegated without undermining competence development, and under what pedagogical conditions such delegation remains developmentally beneficial.

Delegation itself is not necessarily pedagogically harmful. Educational systems have always relied on forms of scaffolding, mediation, and external support. In many cases, AI systems may significantly improve accessibility, reduce unnecessary cognitive overload, and support learners during complex tasks. The pedagogical problem therefore does not emerge from assistance itself, but from the possibility that assistance becomes permanent, excessive, or developmentally indiscriminate.

One of the most important distinctions in this context is the difference between AI functioning as scaffolding and AI functioning as replacement. Scaffolding traditionally refers to temporary support structures that assist learners while gradually transferring cognitive responsibility back to them. Within Vygotskian pedagogy, scaffolding is considered effective precisely because it is intentionally faded as competence increases [15]. AI systems, however, often risk becoming persistent cognitive substitutes rather than temporary developmental supports.

This distinction fundamentally changes the pedagogical role of technology. When AI primarily scaffolds learning, it may support comprehension, feedback, exploration, and differentiated instruction while preserving the learner's active cognitive participation.

At the same time, AI-assisted learning environments may also support deeper conceptual engagement under appropriate pedagogical conditions. By reducing excessive procedural burden, AI systems may allow learners to allocate greater cognitive resources toward higher-order reasoning, conceptual exploration, and creative problem-solving. Particularly for advanced learners, certain forms of cognitive delegation may therefore enhance rather than diminish meaningful intellectual activity.

In contrast, when AI consistently replaces reasoning processes, learners may increasingly disengage from cognitively formative activities themselves. The developmental risk therefore lies not in technological assistance alone, but in the gradual displacement of productive cognitive effort.
From a pedagogical perspective, several dimensions become especially important when defining appropriate boundaries of delegation.

First, the timing of delegation matters significantly. Certain forms of AI assistance may be pedagogically beneficial only after foundational competencies have already been established. Early automation of procedural or reasoning processes may reduce opportunities for initial competence consolidation. In contrast, more advanced learners may benefit from AI support that allows them to focus on higher-order conceptual challenges.

Second, the developmental level of the learner must be considered. Younger learners and novices typically require stronger engagement in foundational cognitive activities in order to develop stable internal strategies. Excessive delegation during early developmental stages may therefore carry different risks than delegation within expert-level learning environments.

Third, the pedagogical purpose of AI usage must remain explicit. AI may support accessibility, feedback, exploration, creativity, or differentiated instruction, but these goals do not necessarily justify the full replacement of cognitively formative processes. Educational design should therefore distinguish between support that enhances learning and support that bypasses learning processes altogether.

Fourth, learners should remain in a position of evaluative and metacognitive control. Even when AI systems generate solutions, explanations, or suggestions, the learner should still actively engage in verification, interpretation, and reflective judgment. Without such involvement, the learning process risks shifting from cognitive participation toward passive consumption of externally generated reasoning.

These considerations suggest that educational AI integration requires a competence-sensitive approach rather than a purely efficiency-oriented one. Current educational discourse often evaluates AI systems primarily according to performance outcomes such as speed, accuracy, accessibility, or productivity. While these dimensions are undoubtedly important, they may not adequately capture the developmental quality of learning processes themselves.

From the perspective of productive struggle and desirable difficulties, certain forms of cognitive effort possess intrinsic pedagogical value precisely because they contribute to durable competence formation [3]. Consequently, not every cognitively demanding activity should necessarily be optimized away. Some forms of difficulty may remain developmentally necessary even when technological shortcuts are available.

This suggests that competence preservation should become an explicit criterion of educational AI design. Educational systems must increasingly determine not only what technology can do, but also what learners should continue doing themselves as part of meaningful intellectual development.

IV. TOWARDS COMPETENCE-PRESERVING AI INTEGRATION

In the context of this paper, competence preservation refers to maintaining cognitively formative learner participation within educational processes despite technological assistance. The concept does not imply resistance to automation itself, but rather the preservation of developmental cognitive engagement.

The pedagogical challenge of artificial intelligence is therefore not adequately addressed through simple opposition between “AI use” and “AI avoidance.” Instead, educational AI integration should increasingly be evaluated according to whether it preserves meaningful cognitive participation alongside technological support.

A competence-preserving approach does not reject cognitive delegation altogether. Rather, it seeks to ensure that AI-assisted learning environments continue to support autonomous reasoning, reflective judgment, and durable cognitive competencies. Importantly, AI systems should support cognitive activity without fully eliminating productive cognitive effort. Research on desirable difficulties and productive struggle consistently demonstrates that certain forms of difficulty remain pedagogically valuable because they require active mental engagement [4][6]. While excessive cognitive overload may hinder learning, the complete removal of challenge may also weaken developmental growth.

A second principle concerns the preservation of learner agency and evaluative control. Even when AI systems provide explanations or generate solutions, learners should remain actively involved in interpretation, verification, and critical assessment. Otherwise, opportunities for metacognitive regulation and reflective monitoring may gradually decline.

The distinction between AI as scaffolding and AI as replacement becomes particularly important in this context. Competence-preserving integration favors forms of AI assistance that guide and structure reasoning processes while still requiring meaningful learner participation. In contrast, systems that fully replace analytical or reflective activity may produce short-term efficiency gains at the cost of reduced long-term cognitive engagement. Some illustrative examples of competence-preserving and competence-replacing forms of AI support are summarized in Tab. 1.

Tab. 1. Examples of competence-preserving and competence-replacing AI support in educational contexts

Cognitive Process	**Potentially Competence-Preserving AI Support**	**Potentially Competence-Replacing AI Support**
Procedural practice	graduated hints requiring learner completion	full automation
Problem decomposition	guided scaffolding	complete solution generation
Metacognitive reflection	reflective prompts	passive answer consumption
Academic writing	structural suggestions	full text generation
Debugging	error localization	automatic correction without reasoning

Educational AI systems should therefore increasingly be evaluated according to questions such as:

- Does the system support or bypass productive cognitive effort?
- Does the learner remain in a reflective and evaluative role?
- Which competencies continue to require active human participation?
- Which processes are scaffolded, and which are fully automated?
- Does the system preserve opportunities for uncertainty, reasoning, and independent problem-solving?

These questions suggest the need for a more pedagogically differentiated understanding of AI integration. Current discourse often frames AI primarily in terms of capability and performance: how quickly systems generate answers, how accurately they solve problems, or how efficiently they personalize instruction. However, pedagogical quality cannot be reduced entirely to efficiency metrics. Educationally valuable learning processes may sometimes require slower reasoning, iterative struggle, reflective correction, and sustained cognitive effort.

What counts as pedagogically appropriate AI support may therefore differ considerably across disciplines, learner populations, and developmental stages. Forms of delegation that are pedagogically appropriate for advanced learners may be problematic for novices. Similarly, some domains may tolerate higher levels of automation without substantial developmental loss, while others rely heavily on active reasoning and procedural participation for competence formation.

Importantly, competence preservation should not be interpreted as resistance to innovation. Rather, it represents an attempt to align technological integration with broader educational aims. The purpose of education is not solely the production of immediate task success, but the development of learners capable of autonomous reasoning, critical reflection, and independent intellectual activity. AI systems may support these goals, but only if pedagogical design consciously preserves the developmental role of human cognitive engagement.

From this perspective, the future of educational AI may depend less on how powerfully systems can replace cognition, and more on how effectively they can support learning while maintaining the learner's active cognitive role within the educational process.

CONCLUSION

Drawing on the literature of cognitive offloading, productive struggle, desirable difficulties, and metacognition, the study proposed that the central pedagogical challenge of AI integration concerns the preservation of meaningful cognitive engagement. From this perspective, educationally valuable learning processes should not automatically be optimized away simply because technological shortcuts become available. Certain forms of cognitive effort remain developmentally significant precisely because they contribute to autonomous reasoning, reflective judgment, and long-term intellectual resilience [4][6].

The paper therefore introduced a competence-preserving perspective on educational AI integration. Rather than framing AI as either inherently beneficial or inherently harmful, this approach emphasizes the importance of defining pedagogical boundaries of delegation. In this context, the distinction between

AI as scaffolding and AI as replacement becomes particularly significant. While scaffolding-oriented AI systems may guide reasoning, support reflection, and preserve learner participation, replacement-oriented systems risk bypassing cognitively formative stages of learning altogether [14][15]. In the long term, the educational value of AI may depend less on immediate performance gains and more on whether learners continue to participate actively in the reasoning processes behind those gains.

Importantly, the purpose of competence-preserving AI integration is not to resist technological innovation, but to align technological capabilities with broader educational aims. Education ultimately concerns not only successful task completion, but the development of learners capable of independent thought, critical reflection, and self-regulated reasoning. From this perspective, the future of educational AI may depend less on replacing cognition and more on supporting human development without diminishing the formative role of cognitive engagement itself. Not every cognitively expensive process is pedagogically unnecessary.

Future research should further investigate how competence-preserving principles may be operationalized across different educational levels, disciplines, and instructional contexts. In particular, empirical studies are needed to examine how varying degrees of cognitive delegation influence metacognitive regulation, problem-solving autonomy, and the development of durable competencies over time [8][9]. The present study intentionally adopts a conceptual and theoretical perspective rather than an empirical one. Future work may therefore examine competence-preserving AI integration through classroom-based case studies, longitudinal observations, and discipline-specific instructional experiments.

REFERENCES

[1] M. Ramos-Benitez, S. Keller and Y. Huang, (2026), “A pedagogical framework for safeguarding cognitive skills in AI-assisted learning environments,” Education Sciences, vol. 16, no. 2, pp. 144–162.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal et al., (2020), “Language models are few-shot learners,” Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901.

[3] N. Carr, (2010), The Shallows: What the Internet Is Doing to Our Brains. New York: W. W. Norton & Company.

[4] R. A. Bjork and E. L. Bjork, (2020), “Desirable difficulties in theory and practice,” Journal of Applied Research in Memory and Cognition, vol. 9, no. 4, pp. 475–479.

[5] J. Sweller, (1988), “Cognitive load during problem solving: Effects on learning,” Cognitive Science, vol. 12, no. 2, pp. 257–285.

[6] J. Young, R. Smith and M. Taylor, (2024), “Productive struggle in action: Supporting meaningful mathematical thinking in contemporary classrooms,” Journal of Mathematics Education, vol. 17, no. 1, pp. 44–63.

[7] K. Baker, N. Ng-A-Fook and A. Hargreaves, (2020), “Productive struggle in mathematics education: A systematic review,” Mathematics Education Research Journal, vol. 32, no. 4, pp. 567–589.

[8] J. H. Flavell, (1979), “Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry,” American Psychologist, vol. 34, no. 10, pp. 906–911.

[9] E. F. Risko and S. J. Gilbert, (2016), “Cognitive offloading,” Trends in Cognitive Sciences, vol. 20, no. 9, pp. 676–688.

[10] A. Clark and D. Chalmers, (1998), “The extended mind,” Analysis, vol. 58, no. 1, pp. 7–19.

[11] I. E. Dror and S. Harnad, (2008), “Offloading cognition onto cognitive technology,” in Cognition Distributed: How Cognitive Technology Extends Our Minds, I. E. Dror and S. Harnad, Eds. Amsterdam: John Benjamins Publishing, pp. 1–23.

[12] K. Ngai and S. J. Gilbert, (2026), “Cognitive offloading and its implications for student learning,” Educational Psychology Review, vol. 38, no. 1, pp. 55–78.

[13] G. Desvaux, S. J. Gilbert and E. F. Risko, (2026), “Metacognitive training facilitates optimal cognitive offloading,” Memory & Cognition, vol. 54, no. 2, pp. 211–229.

[14] V. B. dos Santos Jr. and T. Birdwell, (2025), “The unspoken crisis of learning: The Zone of No Development in AI-assisted education,” arXiv preprint arXiv:2511.12822.

[15] L. S. Vygotsky, (1978), Mind in Society: The Development of Higher Psychological Processes. Cambridge, MA: Harvard University Press.

[16] A. B. H. de Bruin, J. Roelle, S. K. Carpenter and M. Baars, (2023), “Synthesizing cognitive load and self-regulation theory: Effort monitoring and regulation during learning,” Educational Psychology Review, vol. 35, no. 3, pp. 1–27.

ASSESSMENT OF PASSENGER RIDE COMFORT IN A DMU WITH A MODIFIED SUSPENSION SYSTEM

Ján DIŽO, Alyona LOVSKA, Miroslav BLATNICKÝ, Martin BUČKO

Department of Transport and Handling Machines, Faculty of Mechanical Engineering, University of Žilina, Univerzitná 8215/1, 010 26 Žilins, Slovak Republic

jan.dizo@fstroj.uniza.sk, alyona.lovska@fstroj.uniza.sk, miroslav.blatnicky@fstroj.uniza.sk, martin.bucko@fstroj.uniza.sk

ABSTRACT: A modifications of urban railway vehicles is one of possible manners to improve features of existing vehicles at the proper standard for travelling and with the acceptable spent costs. Passenger ride comfort in urban railway vehicles with an independent traction (DMU) is very important point of view to assess the quality of urban railway vehicles. Body oscillations, which arise during vehicle movements on a railway track, affect the level of passenger ride comfort. Accelerations are the main consequence of oscillations. The goal of the presented research is to asses passenger ride comfort in an urban railway vehicle. The point is, that an original suspension system equipped by steel coil spring is replaced by air spring suspesnion system. The research is performed by means of simulation computations. It is suggested to replace this suspension system in a DMU railway vehicle during its modernization process. The passenger ride comfort is assessed by means of the ride comfort index N_{MV}. Fifieen points in a vehicle floor was chosen and two running speeds of 60 km/h and 80 km/h were chosen. Simulation computations were performed on a railway track section. It was found out, that the chosen air suspesnion system contributes to higher level of passenger ride comfort.

Key words: railway vehicle, ride comfort, simulation computations, suspension system

INTRODUCTION

Passenger railway transport is one mode of transport, which has certain advantages in comparison with road transport. There are mainly lower air resistance, rolling resistance and higher axle load. Passenger railway transport is popular mainly due to its comfort, safety as well as relatively high speed of travelling [1,2]. One of the main objectives in the field of passenger railway transport is developing the railway vehocle, which are designed within the higher standards form the performance pont of view, passenger ride comfort point of view as well as protection of environment point of view. Hence, public interest in passenger railway transport also offers safety and comfort of urban railway vehicles [3,4]. The aim of this research is to assess the effects of replacement of original coil springs of the DMU railway vehicle by the

air suspension system. Passenger ride comfort is evaluated by means of the selected ride comfort index marked as N_{MV} ride comfort index on the vehicle body floor.

1. ASSESSMENT OF PASSENGER RIDE COMFORT

There are recognized two methods, which allow to assess passenger ride comfort from the vehicle mechanical vibration point of view. There are namely the indirect method and the direct method. The indirect methods allow to analyze passenger ride comfort, when accelerations are measured in the chosen points of the vehicle body. On the other hand, the direct method uses experimental tests and subjective feelings of passengers in the vehicle body. In principle, both methos need to know accelerations [5]. In case of the indirect method, accelerations in the selected points are measured by means of simulation computations. A great advantage of the indirect method is, that it is not necessary to perform expensive and time-consuming real tests. The European standard EN 12299:2009 [6] contains more ride comfort indices. They can be evaluated in a railway vehicle. The N_{MV} index of passenger ride comfort was chosen for this research. It is passenger ride comfort index calculated on the vehicle body floor. The chosen DMU urban railway vehicle is described in more detail in the next section. There were chosen fifteen points for identification of accelerations. These locations were chosen for the first vehicle of the DMU in the running direction. The N_{MV} passenger ride comfort index, which is mentioned above, is a number calculated based on the defined procedure. It expresses the level of passenger ride comfort. It is calculated based on the following formula [6]:

$$N_{MV} = 6 \cdot \sqrt{\left(a_{xp95}^{W_d}\right)^2 + \left(a_{yp95}^{W_d}\right)^2 + \left(a_{zp95}^{W_b}\right)^2} \tag{1}$$

where: $a_{xp95}^{W_d}$, $a_{yp95}^{W_d}$, and $a_{zp95}^{W_b}$ – the 95-percentage of accelerations measured in the x, y and z directions, respectively, W_d – the weighting function for the x and y directions, W_b – the weighting function for the z direction.

Tab. 1. A list of ranges of the calculated values of the passenger ride comfort index N_{MV} [6].

Range of ride comfort index	Comfort level
$N_{MV} < 1.5$	Very comfortable
$1.5 \leq N_{MV} < 2.5$	Comfortable
$2.5 \leq N_{MV} < 3.5$	Average comfortable
$3.5 \leq N_{MV} < 4.5$	Uncomfortable
$N_{MV} \geq 4.5$	Very uncomfortable

The acceleration signals are proecessed in the Simpack PostProcessor module. It is done by means of a special implemented funcionality, at which, the needed mathematical functions are automaticly loaded to

the selected acceleration signal. The resulting values of the N_{MV} index are compared with the values introduced in the EN standard [6] and these values are listed in Table 1.

2. A SIMULATION MODEL OF THE DMU RAILWAY VEHICLE

The main task of the research is assessment of passenger ride comfort in the DMU railway vehicle. The point of the research is to assess, how a modified suspension system including an air spring will contribute to passenger ride comfort. The DMU railway vehicle known as the 813 class was investigated. It is a railway vehicle, which is an object of interest to modernized it in order to improve its reliability, efficiency and passenger ride comfort. As it is equipped with an independent traction system, it operates on non-electrified railway lines. The original railway vehicle and the modernized railway vehicle are shown in Fig. 1, at which, the original vehicle is equipped with a steel coil spring. It is supposed, that the future modernization of the vehicle could include an air spring suspension.

a) b)

Fig. 1. The DMU railway vehicle: a) an original model [7], b) the latest modernization of the vehicle [8].

The computational multibody model included only rigid bodies, which were interconnected by massless visco-elastic elements. Moreover, some special modelling elements, such as wheel/rail contact, bushing and similar were defined in the model. Visco-elastic elements simulated suspension system. Fig. 2 shows an illustration of steel coil spring and an air spring and a scheme of a non-linear air spring model considered in the Simpack software [9].

A simplified mathematical description of the non-linear air spring modelled in the Simpack software is as follow is defined in the software documentation. The total volume consists of the airbag volume and the constant volume of the mounting [9]:

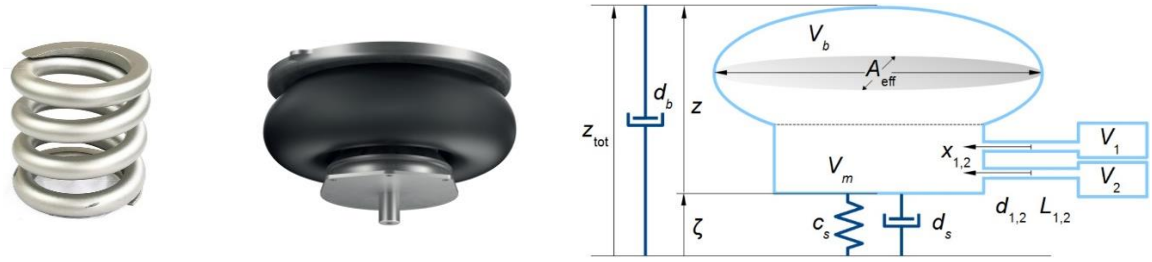
$$V_t = V_t(z) = V_b(z) + V_m \tag{2}$$

where: V_b – the airbag volume, V_m – the mounting volume.

The polytropic equation describes the relationship between a change in volume and a change in pressure:

$$\frac{p}{p_{0_t}} = \left(\frac{V_{t0}}{V_t}\right) \tag{3}$$

where: p – the current value of pressure in the airbag, p_0 – the initial pressure initially determined from the nominal force and effective area, V_{t0} – the initial volume of the airspring [9].



a) b) c)

Fig. 2. An illustration of springs: a) a steel coil spring, b) an air spirng, c) a scheme of a non-linear air spring in the Simpack software [9].

3. RESEARCH RESULTS AND DISCUSSION

The simulation computations were performed for defined running conditions of the analysed DMU railway vehicle. It was running on a railway track, which geometry corresponds to a real Slovak railway section. Running speed of the DMU railway vehicle was set to the value of 60 km/h and 80 km/h. The results are presented in a form of bar graphs. The DMU railway vehicle was simulated for coil spring and for air spring. It means, that four graphs of the results are presented. It is an average running speed, which is for the chosen railway track. These results are depicted in Fig. 3. As it cam be seen for the running conditions corresponding to the running speed of 60 km/h and coil springs the highest value of the NMV index is at the ends parts of the vehicle. The highest value is of 1.47 measured in the from left part of the vehicle (Fig. 3a). In contrast, the lowest values are in the middle part of the vehicle, and the value is of 1.16 in the right side (Fig. 3a). When the results for air springs are observed, it can be seen, that all values are lower in comparison with coil springs (Fig. 3b). It is also possible to observe, that the distribution of ride comfort indices tends to similar trends as for the coil springs. Further, the higher running speed led to higher values of the ride comfort index. This is possible to observe in Fig 3c and Fig. 3d. As it can be seen, the coil spring suspension system is not able to eliminate negative effects of vibrations as effectively as air springs. This is obvious by higher values of ride comfort indices for coil springs (Fig. 3c) in comparison with air springs (Fig. 3d).

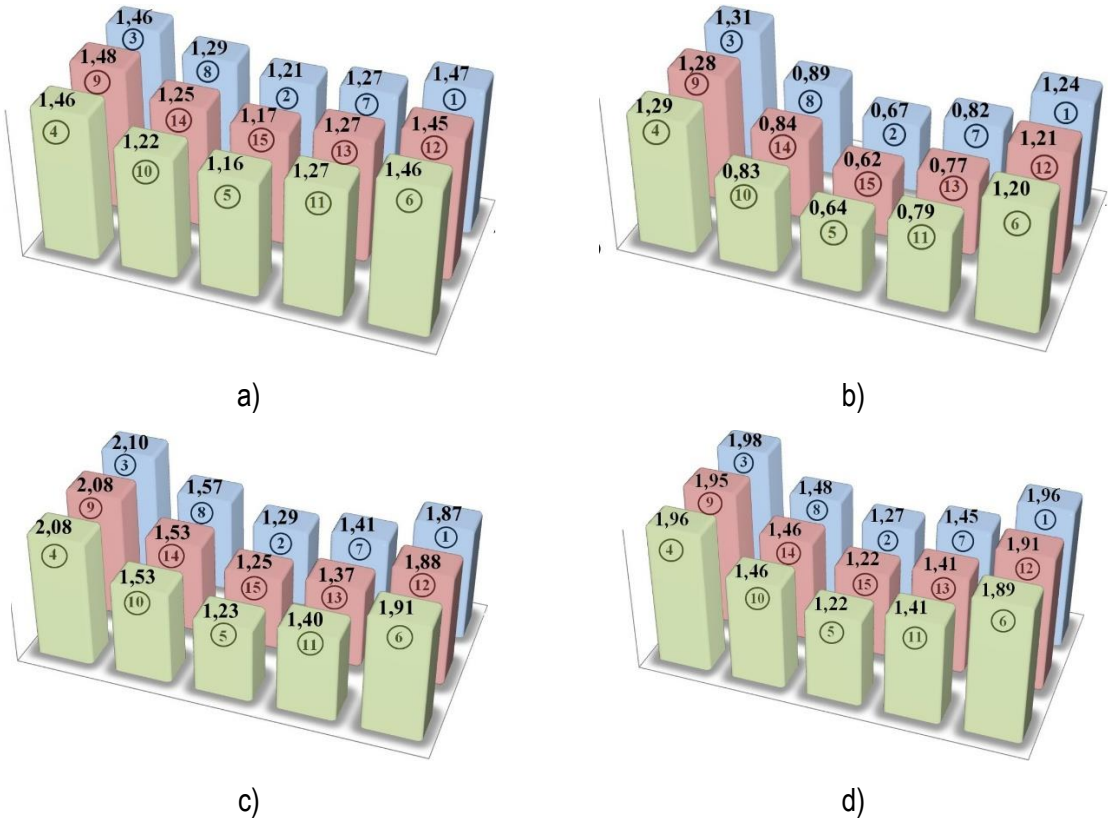


Fig. 3. The achieved results of simulation computations: a) coil springs, 60 km/h, b) air spring, 60 km/h, c) coil springs, 80 km/h, d) air spring, 80 km/h.

Simulation computations showed that replacements of original steel coil springs by air springs in the bogie of the DMU railway vehicle has a potential to improve running properties regarding to passenger ride comfort. The comfort index N_{MV} reached lower values for air springs for all cases in comparison with coil springs. The future research will be focused on other dynamical simulation of the vehicle, mainly form the safety point of view. It is also necessary to investigate more running conditions, i.e. for more running speeds, railway sections as well as for various mass and load of the vehicle. Commissioning of the proposed modifications will require to perform number of real testing of the modified vehicle.

CONCLUSION

Railway vehicles should be able to provide sufficient level of passenger ride comfort as well as sufficient running safety during many-years operation. This, the main task is to design suitable technical solutions of modern railway vehicles. The presented research also belonged to such task. Its main task was to assess passenger ride comfort in a DMU railway vehicle, in which, original coil springs were considered to be replaced by air springs. This suitability of this idea was assessed based on the level of passenger ride comfort. The examined DMU railway vehicle was considered in simulation computations. The results

of the solved study prove a possible positive and potential of improvement of running properties by replacement of original steel coil springs by air spring suspension system.

REFERENCES

[1] M. Vojtek, J. Matuska, J. Siroky, J. Kugler and M. Kendra, (2021), “Possibilities of railway safe-ty improvement on regional lines,” Transportation Research Procedia, vol. 53, pp. 8–15, DOI https://doi.org/10.1016/j.trpro.2021.02.001

[2] M. Kostrzewski, Y. Abdelatty, A. Eliwa and M. Nader, (2022), “Analysis of modern vs. conventional development technologies in transportation—the case study of a last-mile delivery process,” Sensors, vol. 22, 9858, DOI .https://doi.org/10.3390/s22249858

[3] R. Melnik and B. Sowinski, (2013), “Application of the rail Vehicle's monitoring system in the process of suspension condition assessment,” Communications Scientific Letters of the University of Zilina, vol. 15, pp. 3–8, DOI https://doi.org/10.26552/com.C.2013.4.3-8

[4] S. K. Szürke, G. Kovács, M. Sysyn, J. Liu and S. Fischer, (2023), “Numerical optimization of bat-tery heat management of electric vehicles,” Journal of Applied and Computational Mechanics, vol. 9, pp. 1076–1092, DOI https://doi.org/10.22055/jacm.2023.43703.4119

[5] Kardas-Cinal, E., (2020), “Statistical analysis of dynamical quantities related to running safety and ride comfort of a railway vehicle,” Scientific Journal of Silesian University of Technology Series Transport, vol. 106, pp. 63–72, DOI https://doi.org/10.20858/sjsutst.2020.106.5

[6] EN 12299:2009. Railway applications - Ride comfort for passengers - Measurement and evaluation.

[7] Diesel motor coach, class 810 (in Slovak).
Available on: https://sk.wikipedia.org/wiki/Motorov%C3%BD_voze%C5%88_810

[8] ZSSK modernizes motor units: "Mravce" will be added to "Bagety" (in Slovak). Available on: https://www.railpage.net/zssk-modernizuje-motorove-jednotky-k-bagetam-pribudnu-mravce/

[9] Simpack User Manual, 2024.

Ján Dižo: https://orcid.org/0000-0001-9433-392X

Alyona Lovska: https://orcid.org/0000-0002-8604-1764

Miroslav Blatnický: https://orcid.org/0000-0003-3936-7507

Martin Bučko: https://orcid.org/0009-0006-0566-4430

STRENGTH ANALYSIS OF AN OPEN WAGON BODY WITH STIFFENERS IN THE FRAME

Alyona LOVSKA, Juraj GERLICI, Ján DIŽO

Department of Transport and Handling Machines, Faculty of Mechanical Engineering, University of Žilina, Univerzitná 8215/1, 010 26 Žilina, Slovak Republic

alyona.lovska@fstroj.uniza.sk, juraj.gerlici@fstroj.uniza.sk, jan.dizo@fstroj.uniza.sk

ABSTRACT: The article highlights the features of determining the strength of an open wagon body with braces in the frame during shunting collision. The presence of braces ensures a reduction in the load on the backbone beam when absorbing longitudinal loads. They are located between the rear stops of the automatic couplers and the vertical sheet of the pivot beam. The results of the calculations have proven that such an improvement is advisable. At the same time, the stress in the backbone beam is 2.7% lower than in a typical design. The conducted research will contribute to the creation of measures aimed at improving the efficiency of the operation of railway cars.

Key words: open wagon, design improvement, structural strength, body stress state, railway transport

INTRODUCTION

The accelerated pace of Ukraine's integration into the system of international transport corridors necessitates the adaptation of railway veicles, as one of the leading transport industries, to work in market conditions. To increase the competitiveness of freight wagons in operation, it is necessary to select the optimal parameters of structural elements at the stage of their design and calculation, while observing the conditions of strength and reliability. The degree of replenishment of the wagons fleet in recent years is insignificant. This necessitates the introduction into operation of new technical solutions for improving wagon bodies, technologies for their maintenance and repair, etc., which will allow maintaining the technical condition of railway vehicles with the existing repair base.

The current state of the issue of improving the bodies of wagons was identified by analyzing publications in this area. Thus, in work [1], in order to improve the strength of the open wagon body, its manufacture from high-strength steels was proposed. The advantages of using such steel in wagon construction are presented. The authors presented the results of optimization calculations regarding the choice of the thickness of the wagon structural elements, provided that this grade of steel is used for their manufacture.

However, the work does not contain an analysis of the stress-strain state of the wagon body, which would allow to identify the loading of the body shell taking into account the solution proposed by the authors.

In the study highlighted in [2], the use of beams with corrugated walls as a profile for the backbone beam of a car is justified. The results of the calculation of the wagon body for strength, as well as its modal analysis, have proven that such an implementation is justified. However, this study was conducted on the example of a passenger car frame. That is, this work did not pay attention to the feasibility of implementing such beams in the frame of an open wagon. To improve the strength of the freight wagon frame, the work [3] proposed creating its structure in the form of “egg crates”.

The justification of the proposed solution and the prospects for its use in freight wagon structures are presented. However, the authors did not consider the possibility of its use in the structures of open wagon frames. In work [4], to improve the strength of freight wagons, it is proposed to introduce new materials with improved characteristics as the material of their structural components. The study was conducted on the example of magnesium alloys. It was proved that the use of such material contributes not only to reducing the weight of the wagon, but also to improving its strength characteristics.

However, the authors did not consider the feasibility of introducing such materials for the manufacture of the wagon frame. An analysis of the structural materials of new generation wagon bodies is given in [5]. The paper indicates the advantages of using new progressive materials for individual components of wagon structures. These works do not consider the issues of improving the load-bearing structures of wagon bodies to ensure their strength during shunting collisions. Measures to improve the load-bearing structure of the universal open wagon body are given in [6].

To reduce the maximum equivalent stresses in the node of interaction of the backbone beam with the pivot, it is proposed to install reinforcing linings on the lower shelves of the backbone beam. It is important to note that such a solution allows for stress reduction in the backbone beam of the car frame in its cantilever part, and not in the node of interaction with the pivot beam. The analysis of scientific publications proves that the issue of improving wagon bodies to improve their durability in operation requires further research.

The purpose of the research is to create structural solutions to improve the strength of the open wagon body in operation:

- to propose solutions to improve the open wagon body to reduce its load in operation;
- calculate the strength of the open wagon body.

1. MATERIALS AND METHODS OF THE STUDY

The main hypothesis of the study is that the use of braces in the frame will help reduce the load on the backbone beam of the open wagon frame (Fig. 1). To confirm this hypothesis, a calculation was made for the strength of the open wagon body with braces in the frame. The open wagon model 12-757 was chosen as the prototype. The braces were made of I-beams No. 14.

The strength calculation was performed using the finite element method in the SolidWorks Simulation software package. The graphical-analytical method was used to determine the number of mesh elements. Isoparametric tetrahedra were used as finite elements. Number of mesh elements was 493357 and number of nodes was 159910.

At the same time, the maximum element size was 80 mm, the minimum was 16 mm, the maximum aspect ratio was 680.33, the percentage of elements with an aspect ratio less than 3 – 25.7, more than 10 – 26.3. The minimum number of elements in a circle was 9, the ratio of increasing the size of elements of the finite element mech 1.8.

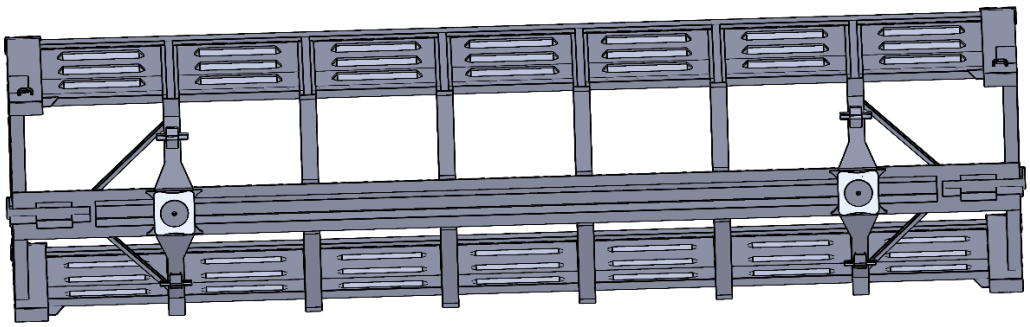


Fig. 1. A spatial model of the open wagon body.

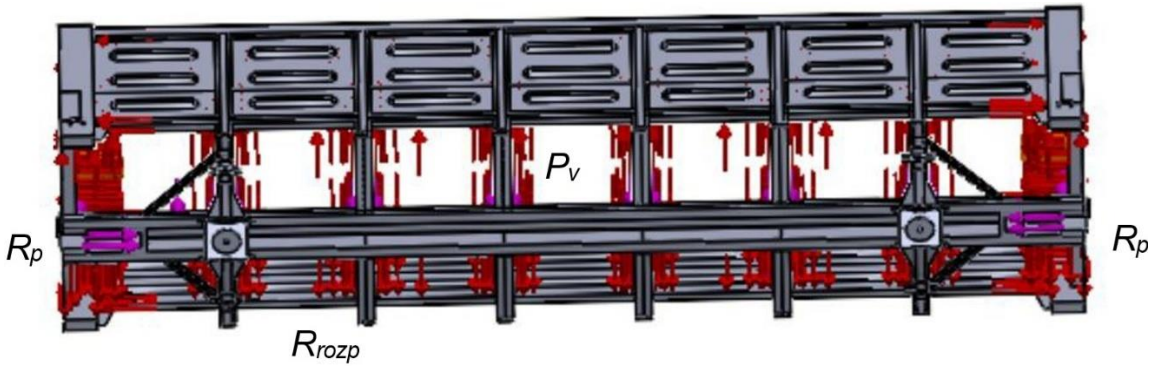


Fig. 2. A calculation diagram of the open wagon body.

When compiling the model, welds in the interaction zones of individual parts were not considered. body structure elements between each other. The design scheme of the open wagon body (Fig. 2) considers the following loads: vertical R_v, longitudinal R_p on the vertical wall of the rear stop of the automatic coupler, and the reaction R_p to the action of the load R_p, pressure from the strut bulk cargo R_{rozp} (coal).

The model was fixed in the areas where the body rests on the running gear. The material for manufacturing the body is steel grade 09G2S, which has the permissible stress at this design mode of 310.5 MPa [7].

2. RESEARCH RESULTS

The results of the calculation of the strength of the open wagon body are shown in Fig. 3. From the calculations performed, it can be concluded that the maximum The stresses in the body were 298.4 MPa. The dislocation of these stresses occurs in the backbone beam of the frame, namely between the rear stops of the automatic couplers and the pivot beams.

These stresses are 3.9% lower than the permissible ones, which are 310.5 MPa, and also 2.7% lower than those operating in a typical body design. The maximum displacements occur in the middle part of the body and are 2.57 mm (Fig. 4). Therefore, the proposed implementation is effective.

It is important to note that its implementation is possible not only at the stage design, and modernization of wagons.

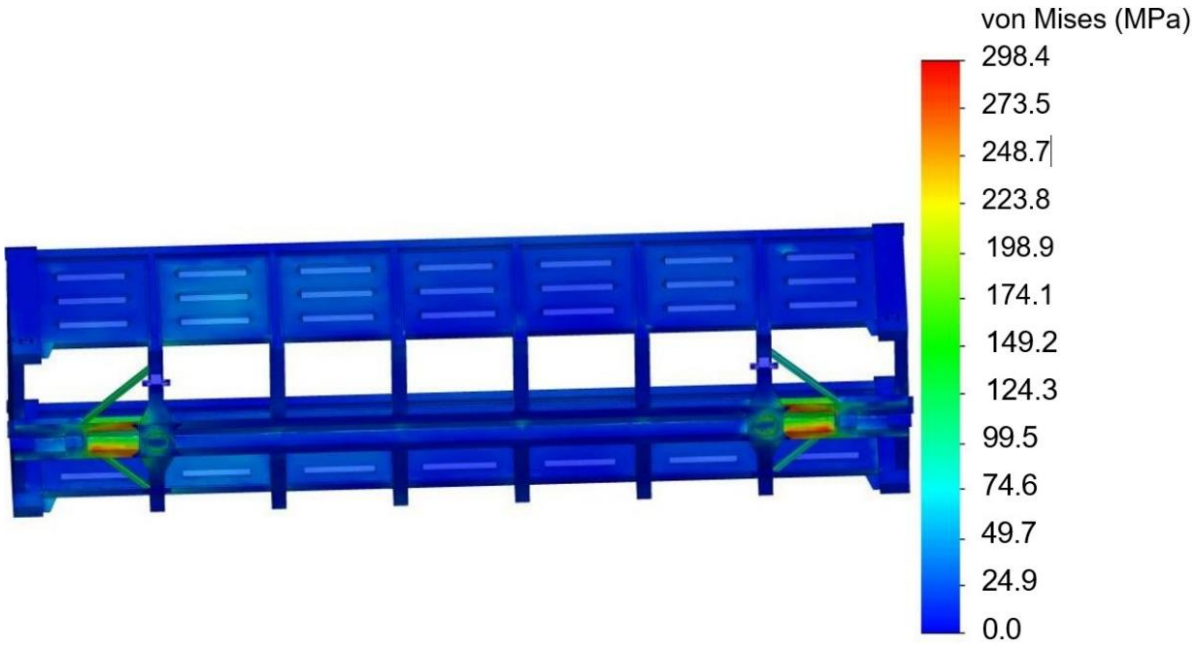


Fig. 3. The stressed state in the open wagon body.

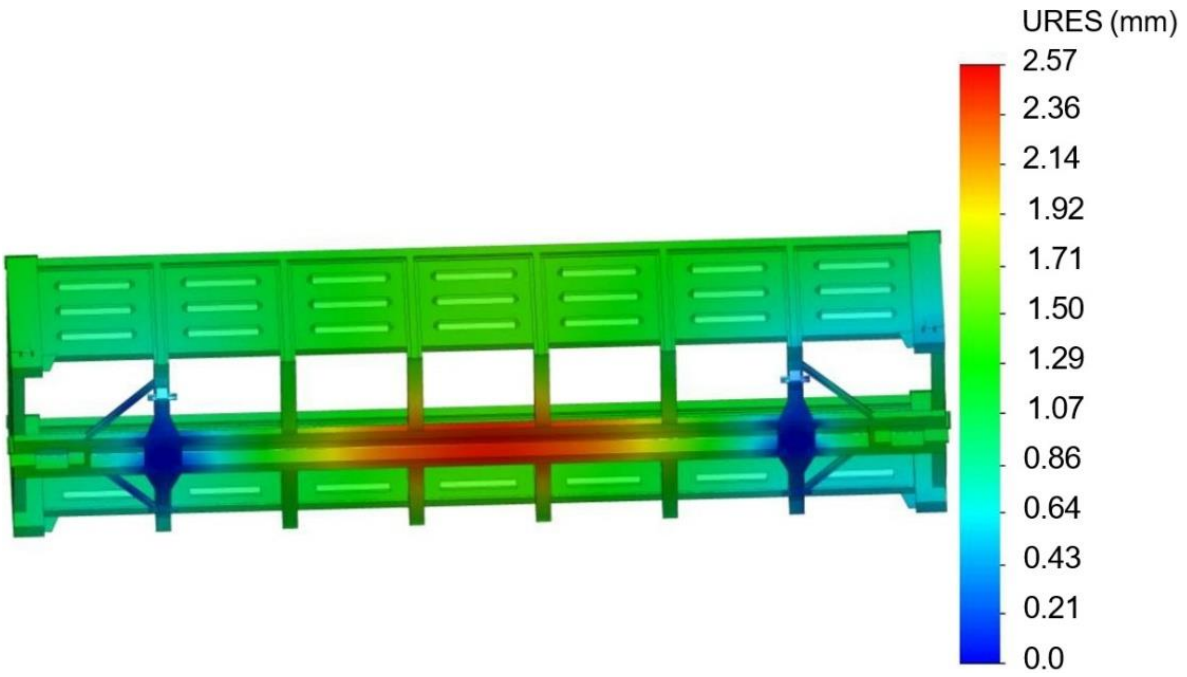


Fig. 4. Deflections of the open wagon structure.

CONCLUSION

1. A solution is proposed to improve the open wagon body to reduce its load in operation. It is proposed to equip its frame with braces that connect the rear stops of the autocouplers with the pivot beams. This will allow partially unloading the backbone beam under the most unfavorable operating conditions.

2. A calculation was made for the strength of the open wagon body with braces in the frame. The maximum stresses in the open wagon body were 298.4 MPa and concentrated in the backbone beam, namely in the areas between the rear stops and the pivot beam. The calculated stresses are 3.9% lower than the permissible ones and 2.7% lower than those acting in a typical body structure.

The maximum displacements occur in the middle part of the body and are 2.57 mm. The results of the calculations proved the effectiveness of the implementation proposed improvement.

REFERENCES

[1] F. Galimova, Y. Khurmatov, M. Abdulloev, B. Jumabekov, D. Sultonaliev and D. Ergeshova, (2021), “Modern Open wagon with Lightweight Body,” Lecture Notes in Networks and Systems, vol. 247, DOI https://doi.org/10.1007/978-3-030-80946-1_94, pp. 1043–1050.

A. Lovska, J. Gerlici and J. Dižo, (2025), “Research of the possibility of using beams with corrugated walls in a passenger rail car frame,” Scientific Reports, vol. 15, DOI https://doi.org/10.1038/s41598-025-12783-0, 26833.

[2] D. Jeong, D. Tyrell, M. Carolan and A. Benjamin Perlman, (2009), “Improved Tank Car Design Development: Ongoing Studies on Sandwich Structures,” Proceedings of 2009 ASME Joint Rail Conference, DOI https://doi.org/10.1115/JRC2009-63025

[3] W. G. Lee, J.-S. Kim, S.-J. Sun and J.-Y. Lim, (2018), The next generation material for lightweight railway car body structures: Magnesium alloys,” Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, vol. 232, DOI https://doi.org/10.1177/095440971664614, no. 1, pp. 25–42.

Freight cars major metals. Trains, 2015, Marts. 20 p.

[4] A. Lovska, Yu. Yu. Shafunov, (2015), “Improvement of the supporting structure of the open wagon body to ensure its strength during shunting collisions (in Ukrainian),” Collection of scientific works of UkrDUZT, no. 158, pp. 29–35.

[5] DSTU 7598:2014. Freight wagons. General requirements for calculations and design of new and modernized 1520 mm gauge wagons (non-self-propelled). 250 p.

Alyona Lovska: https://orcid.org/0000-0002-8604-1764

Juraj Gerlici: https://orcid.org/0000-0003-3928-0567

Ján Dižo: https://orcid.org/0000-0001-9433-392X

DEEP LEARNING WITH ATTENTION MECHANISMS FOR SMOG FORECASTING UNDER EXTREME CONDITIONS

Aneta WIKTORZAK[1], Ryszard SZCZEBIOT[1], Leszek GOŁDYN[1]

[1] Faculty of Computer Science and Technology, University of Lomza, ul. Akademicka 14, 18-400 Lomza, Poland

awiktorzak@al.edu.pl, rysbiot@al.edu.pl, lgoldyn@al.edu.pl

ABSTRACT

Accurate short-term forecasting of particulate matter concentrations remains a challenging task due to the nonlinear and dynamic nature of atmospheric processes. This study proposes a hybrid deep learning framework that integrates Long Short-Term Memory (LSTM) networks with an attention mechanism to improve the prediction accuracy of PM10 and PM2.5 under rapidly changing and extreme environmental conditions.

The model operates on multivariate time-series data that combine air quality measurements and meteorological variables, enabling the extraction of complex temporal dependencies. The attention mechanism dynamically assigns weights to input time steps, allowing the model to focus on the most informative temporal patterns and mitigate the limitations of standard recurrent architectures.

The proposed approach was evaluated against baseline models, including classical LSTM and NARX. Experimental results demonstrate that the LSTM-Attention model achieves superior predictive performance, reducing RMSE for PM10 from 17.1 (standard LSTM) to 13.5, and for PM2.5 from 20.3 to 16.2. Additionally, the model improves the coefficient of determination (R^2) from 0.81 to 0.88 for PM10 and from 0.78 to 0.85 for PM2.5. Under extreme pollution conditions, the maximum relative error is reduced from 21.3% to 5.6% for PM10 and from 25.1% to 8.4% for PM2.5, highlighting the robustness of the proposed method.

The results confirm that incorporating attention mechanisms significantly enhances both accuracy and stability of air quality forecasting models, while also improving interpretability through the analysis of attention weights. The proposed framework provides a scalable and effective solution for real-time environmental monitoring systems and supports the development of advanced decision-support tools in Smart City applications.

The novelty of this work lies in the integration of a temporally adaptive attention mechanism with stacked LSTM layers to enhance sensitivity to abrupt environmental changes, resulting in superior performance in extreme pollution scenarios.

Keywords: air quality forecasting, PM10, PM2.5, deep learning, Long Short-Term Memory (LSTM), attention mechanism, time-series prediction, environmental data, smog prediction, explainable artificial intelligence (XAI).

1. INTRODUCTION

Air pollution forecasting has become a critical component of modern environmental management and public health protection strategies. Elevated concentrations of particulate matter, particularly PM10 and PM2.5, are strongly associated with increased risks of respiratory and cardiovascular diseases, as well as premature mortality. According to the World Health Organization, long-term exposure to air pollution remains one of the leading environmental causes of death worldwide. Consequently, the development of reliable forecasting models is essential for early warning systems and informed decision-making at both governmental and municipal levels.

In recent years, the rapid advancement of artificial intelligence, especially within the field of deep learning, has opened new possibilities for modeling complex environmental processes. Unlike traditional statistical approaches, deep learning models are capable of capturing nonlinear relationships and temporal dependencies present in multivariate time-series data. This capability is particularly important in air quality prediction, where pollutant concentrations depend on a combination of emission sources and dynamic meteorological conditions [1].

Among deep learning techniques, recurrent neural networks (RNNs), and in particular Long Short-Term Memory (LSTM) networks, have proven highly effective in processing sequential data [2]. LSTM architectures are specifically designed to address the vanishing gradient problem, enabling the model to retain information over longer time horizons. This makes them well-suited for forecasting tasks in which historical environmental conditions influence future pollutant levels.

However, despite their advantages, conventional LSTM models may struggle to accurately capture sudden fluctuations in air quality caused by rapid meteorological changes, such as temperature inversions or abrupt shifts in wind patterns. These limitations have motivated the integration of attention mechanisms, originally introduced in the context of natural language processing, into time-series forecasting models

[3]. The attention mechanism allows the model to dynamically assign different levels of importance to various time steps in the input sequence, thereby improving both prediction accuracy and interpretability.

The aim of this study is to investigate the effectiveness of a hybrid LSTM model enhanced with an attention mechanism for short-term forecasting of PM10 and PM2.5 concentrations under dynamic and extreme atmospheric conditions. The proposed approach is evaluated against baseline models, including a standard LSTM model and the NARX model architectures, with a particular focus on performance during high-risk smog episodes.

Despite the growing interest in advanced deep learning approaches for air quality forecasting, the effectiveness of different modeling strategies varies significantly depending on data characteristics and environmental conditions. Therefore, a critical review of existing methods is necessary to identify their strengths and limitations, and to justify the need for further methodological improvements.

2. RELATED WORK

Air quality forecasting has been extensively studied using a wide range of statistical and machine learning techniques. Early approaches were primarily based on classical regression models and time-series analysis methods, such as ARIMA and autoregressive models. While these techniques provided baseline predictive capabilities, they were often limited in their ability to capture nonlinear dependencies and complex interactions between environmental variables.

With the development of machine learning, more advanced models such as Nonlinear Autoregressive Network with Exogenous Inputs (NARX) were introduced for air quality prediction. NARX models incorporate both past values of the target variable and external inputs, such as meteorological data, making them suitable for modeling dynamic environmental systems. However, their performance is often constrained by limited scalability and difficulties in capturing long-term temporal dependencies [4].

The emergence of deep learning significantly improved forecasting performance. In particular, Long Short-Term Memory (LSTM) networks have become a dominant approach for modeling time-series data due to their ability to retain long-term information and mitigate the vanishing gradient problem [2]. Numerous studies have demonstrated the effectiveness of LSTM models in predicting PM2.5 and PM10 concentrations, especially when trained on large-scale environmental datasets [5].

To further enhance feature extraction, hybrid architectures combining convolutional and recurrent layers have been proposed. For example, CNN-LSTM models leverage convolutional neural networks (CNNs)

to capture spatial or local temporal patterns, which are then processed by LSTM layers to model sequential dependencies. Such architectures have shown improved performance in scenarios where both spatial and temporal correlations are significant [6].

More recently, attention mechanisms have been incorporated into deep learning models to improve both predictive accuracy and interpretability. Originally developed for natural language processing tasks, attention allows models to focus selectively on the most relevant parts of an input sequence. In the context of air quality forecasting, attention-based LSTM models have demonstrated superior performance compared to standard recurrent architectures, particularly in capturing abrupt changes in pollutant concentrations [3], [7].

Furthermore, recent research trends indicate a shift toward more advanced architectures, such as Transformer-based models, which rely entirely on self-attention mechanisms and eliminate recurrence altogether [8]. These models offer improved scalability and parallelization capabilities, making them suitable for large-scale environmental data analysis.

Despite these advancements, challenges remain in accurately predicting air pollution under extreme and rapidly changing conditions. Many existing models still struggle with sudden spikes in pollutant levels and require further improvement in robustness and interpretability. This study addresses these limitations by proposing a hybrid LSTM model with an attention mechanism, specifically designed to enhance performance in high-risk smog scenarios.

The reviewed approaches highlight the evolution from classical statistical models to advanced deep learning architectures. However, they also reveal persistent challenges related to robustness, interpretability, and performance under extreme conditions. In response to these limitations, the following section presents the data sources and methodological framework adopted in this study.

3. DATA AND METHODOLOGY

3.1. DATA SOURCES AND DESCRIPTION

The proposed study is based on the integration of heterogeneous environmental and meteorological datasets, enabling a comprehensive representation of air quality dynamics. The primary data sources include measurements from distributed air quality monitoring stations and publicly available meteorological databases.

The environmental dataset consists of time-series measurements of key pollutants, including PM10 and PM2.5 concentrations, as well as auxiliary variables such as sulfur dioxide (SO_2) and nitrogen dioxide (NO_2). These measurements are typically collected at high temporal resolution, allowing the analysis of short-term fluctuations and rapid pollution events.

In parallel, meteorological data were incorporated to capture atmospheric conditions influencing pollutant dispersion. These variables include air temperature, relative humidity, atmospheric pressure, wind speed, and wind direction. The integration of meteorological and environmental variables is essential, as air pollution formation and accumulation are strongly dependent on weather dynamics [4].

While the integration of heterogeneous data sources provides a comprehensive representation of environmental dynamics, raw datasets are often inconsistent and require appropriate preprocessing before they can be effectively used in machine learning models.

3.2. DATA PREPROCESSING

To ensure consistency and model readiness, the collected datasets underwent a series of preprocessing steps. First, all data streams were temporally aligned to a uniform time grid. Missing values, which are common in sensor-based systems, were handled using interpolation techniques, including linear interpolation for short gaps.

Subsequently, all features were normalized to a common scale using min-max normalization, defined as:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

This transformation ensures numerical stability during model training and prevents dominance of features with larger value ranges.

A sliding window approach was applied to construct input sequences. Each input sample consists of a fixed number of consecutive time steps, representing both historical pollutant levels and meteorological conditions. The model output corresponds to the predicted pollutant concentrations at the next time step, enabling short-term forecasting.

Once the data are properly aligned, cleaned, and transformed, they become suitable for input into predictive models. The following section describes the architecture of the proposed deep learning model designed to exploit these processed time-series data.

3.3. MODEL ARCHITECTURE

The predictive model is based on a hybrid deep learning architecture that combines recurrent neural networks with an attention mechanism. The core component is the Long Short-Term Memory (LSTM) network, which is well-suited for modeling sequential dependencies in time-series data.

The architecture consists of the following components:

- **Input layer**: accepts a three-dimensional tensor representing batch size, time steps, and feature dimensions.
- **Stacked LSTM layers**: two consecutive LSTM layers are used to capture both short-term and long-term temporal dependencies. Each layer processes the sequential input and propagates hidden states through time.
- **Attention layer**: implemented following the concept introduced by Attention Mechanism, this component assigns adaptive weights to different time steps, allowing the model to focus on the most informative parts of the input sequence [3].
- **Pooling layer**: aggregates the attention-weighted outputs into a compact representation.
- **Dense output layer**: generates final predictions for PM10 and PM2.5 concentrations.

Architecture of the proposed hybrid LSTM-Attention model (Fig. 1). The input time-series data are processed through stacked LSTM layers to capture temporal dependencies. The attention mechanism assigns weights to relevant time steps, and the aggregated representation is passed to a dense layer for predicting PM10 and PM2.5 concentrations.

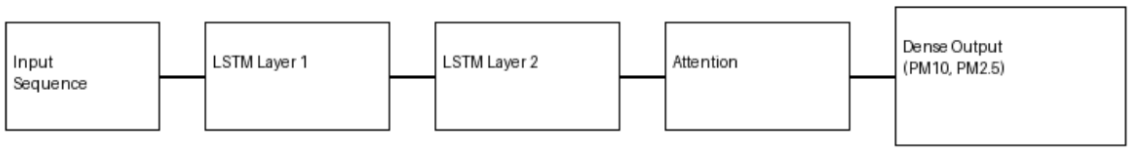


Figure 1. Architecture of the proposed hybrid LSTM-Attention model.

The attention mechanism can be conceptually expressed as:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}$$

where α_t represents the attention weight for time step t, and e_t is a learned relevance score. This formulation enables the model to dynamically emphasize critical temporal features.

LSTM Layer

The LSTM network processes sequential data using gated mechanisms. At each time step t, the following operations are performed:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \cdot \tanh(C_t)$$

where: f_t – forget gate, i_t – input gate, o_t – output gate, C_t – cell state, h_t – hidden state.

ATTENTION MECHANISM

To enhance temporal feature selection, an attention layer is applied over hidden states:

$$e_t = v^T \tanh(W_h h_t + b_h)$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}$$

$$c = \sum_{t=1}^{T} \alpha_t h_t$$

where: α_t – attention weights, c – context vector.

The final prediction is computed as:

$$\hat{y} = W_c \cdot c + b_c$$

The above equation represents the final output layer of the proposed model, where the predicted pollutant concentrations are computed based on the context vector obtained from the attention mechanism.

- $\hat{y} \in \mathbb{R}^2$ - denotes the predicted output vector, typically representing the forecasted concentrations of PM10 and PM2.5.

- $c \in \mathbb{R}^H$ - is the **context vector**, which is a weighted sum of hidden states produced by the LSTM layers. It aggregates the most relevant temporal information selected by the attention mechanism.
- $W_c \in \mathbb{R}^{2 \times H}$ - is the **weight matrix** of the fully connected (dense) layer, mapping the context vector to the output space.
- $b_c \in \mathbb{R}^2$ - is the **bias vector**, allowing the model to adjust predictions independently of the input features.

Having defined the structure of the proposed model, the next step involves specifying the training strategy, including the optimization process and learning configuration used to estimate model parameters.

3.4. TRAINING PROCEDURE

The model was trained using supervised learning with historical data. The objective function is the mean squared error (MSE), defined as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Optimization was performed using the Adam optimizer, which combines adaptive learning rates with momentum to accelerate convergence [1].

To ensure robustness and generalization, the dataset was divided into training, validation, and test subsets. Additionally, multiple training runs were conducted to reduce the impact of random initialization and improve reproducibility.

To objectively assess the effectiveness of the trained model, appropriate evaluation metrics must be defined. These metrics allow for quantitative comparison between different approaches and provide insight into prediction accuracy.

3.5. EVALUATION METRICS

Model performance was evaluated using standard regression metrics, including:

- Root Mean Squared Error (RMSE), which measures prediction accuracy.
- Coefficient of Determination (R^2), indicating the proportion of variance explained by the model.

These metrics provide complementary insights into both absolute error magnitude and overall model fit. With the methodology and evaluation criteria established, the next section presents the experimental results obtained using the proposed model and compares them with baseline approaches.

4. EXPERIMENTAL RESULTS

4.1. EVALUATION SETUP

To assess the effectiveness of the proposed hybrid model, a comprehensive experimental framework was established. The dataset was divided into three subsets: training (70%), validation (15%), and testing (15%). This split ensures both proper model learning and unbiased evaluation.

All models were trained multiple times to reduce the impact of random initialization and improve reproducibility. The experiments were conducted using a GPU-accelerated environment, ensuring efficient training of deep neural networks.

The following models were compared:

- the NARX model,
- a standard LSTM model,
- the proposed LSTM with an attention mechanism.

Based on this experimental setup, the following metrics were used to quantitatively evaluate model performance.

4.2. EVALUATION METRICS

The predictive performance of the models was evaluated using standard regression metrics:

- **Root Mean Squared Error (RMSE):**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **Coefficient of Determination (R^2):**

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

These metrics provide insight into both prediction accuracy and the model's ability to explain variance in the data.

4.3. QUANTITATIVE RESULTS

Table 1 summarizes the performance of all evaluated models.

Model	RMSE PM10	RMSE PM2.5	R^2 PM10	R^2 PM2.5
NARX	21.6	25.2	0.72	0.69
LSTM	17.1	20.3	0.81	0.78
LSTM + Attention	**13.5**	**16.2**	**0.88**	**0.85**

The results clearly indicate that the proposed hybrid model outperforms baseline approaches across all metrics. The reduction in RMSE demonstrates improved prediction accuracy, while higher R^2 values confirm better model fit.

While aggregated metrics provide an overall assessment of model performance, a more detailed analysis of prediction dynamics is necessary to understand how the models behave over time.

4.4. TIME-SERIES PREDICTION ANALYSIS

To further evaluate model performance, predicted pollutant concentrations were compared with actual measurements over selected time intervals.

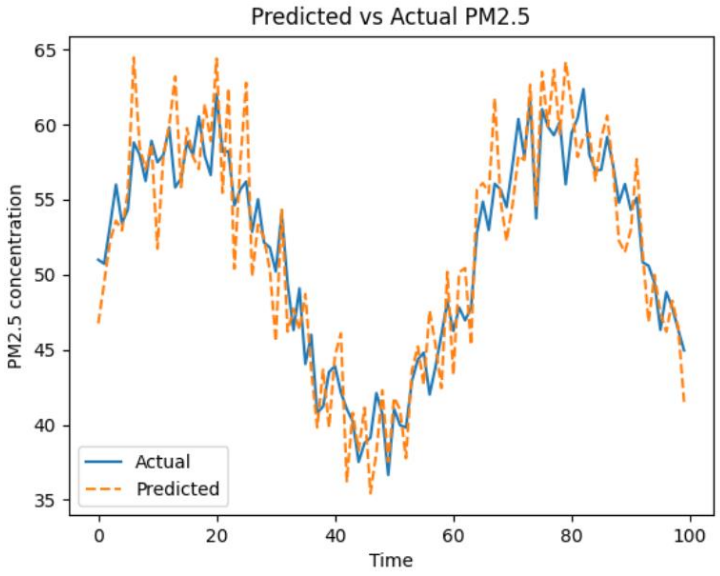


Figure 2. Presents a representative prediction sequence for PM2.5 over a 24-hour period.

The hybrid model closely follows real-world fluctuations, effectively capturing both gradual trends and rapid spikes in pollution levels. In contrast, the standard LSTM model exhibits noticeable lag during sudden changes.

In addition to visual comparison of predicted and actual values, it is important to examine the statistical distribution of prediction errors to evaluate model reliability.

4.5. ERROR DISTRIBUTION ANALYSIS

To assess model reliability, prediction errors were analyzed using histogram-based visualization.

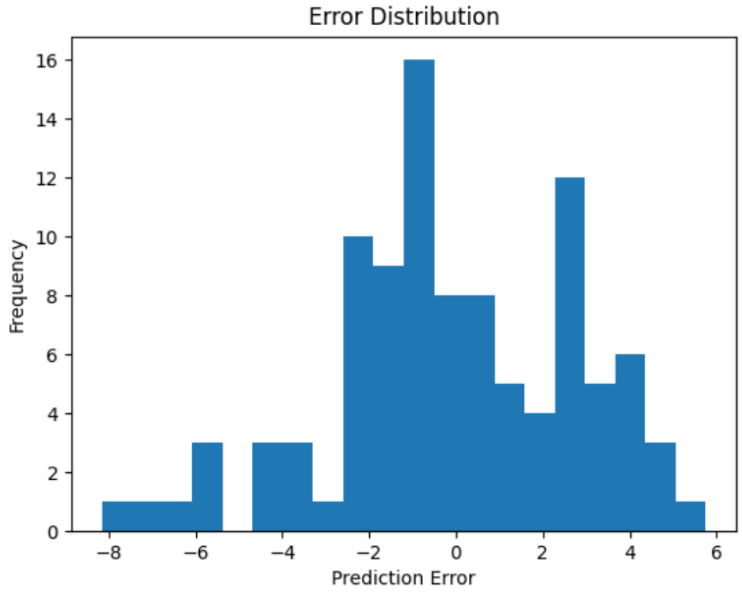


Figure 3 presents the distribution of prediction errors for PM2.5.

The distribution is approximately symmetric and centered around zero, indicating that the model does not systematically overestimate or underestimate pollutant concentrations. This suggests good generalization performance.

4.6. PERFORMANCE UNDER EXTREME CONDITIONS

A key objective of this study was to evaluate model performance during high-risk smog episodes. Selected time periods with elevated pollution levels (PM10 > 100 µg/m³) were analyzed.

The results demonstrate a substantial improvement achieved by the attention-based model. The reduction in maximum error highlights its ability to respond effectively to sudden environmental changes.

Table 2 presents the maximum relative error for extreme conditions.

Model	Max Error PM10	Max Error PM2.5
LSTM	21.3%	25.1%
LSTM + Attention	**5.6%**	**8.4%**

Beyond accuracy under extreme conditions, understanding how the model arrives at its predictions is equally important, particularly in safety-critical applications.

4.7. INTERPRETATION OF ATTENTION MECHANISM

An additional advantage of the proposed model is its interpretability. By analyzing attention weights, it is possible to identify which time steps contribute most to the prediction.

The analysis revealed that higher weights are typically assigned to nighttime and early morning observations, which are known to play a critical role in smog formation due to atmospheric stability.

4.8. DISCUSSION OF RESULTS

The experimental results confirm that incorporating the Attention Mechanism significantly enhances the performance of LSTM models.

Key findings include:

- improved prediction accuracy (lower RMSE),
- better model fit (higher R^2),
- superior robustness under extreme conditions,
- increased interpretability of predictions.

These advantages make the proposed model particularly suitable for real-world deployment in air quality monitoring and early warning systems.

The results presented above provide a comprehensive evaluation of the proposed model. The following section discusses their implications in a broader research and application context.

5. DISCUSSION

The experimental results presented in the previous section clearly demonstrate that the integration of the Attention Mechanism into the LSTM architecture leads to a substantial improvement in predictive

performance for air quality forecasting tasks. This enhancement is particularly evident in scenarios characterized by rapid fluctuations in pollutant concentrations, where traditional models often fail to respond adequately.

One of the key advantages of the proposed hybrid model is its ability to dynamically assign importance to different time steps within the input sequence. Unlike standard LSTM networks, which treat all temporal inputs with equal significance, the attention-based approach selectively emphasizes the most relevant observations. This capability is crucial in environmental systems, where certain periods—such as nighttime or early morning hours—play a disproportionately important role in smog formation due to atmospheric stability and reduced air circulation.

The quantitative improvements observed in RMSE and R^2 metrics confirm that the attention mechanism enhances both accuracy and model generalization. Moreover, the significant reduction in prediction error under extreme conditions highlights the robustness of the proposed approach. In practical terms, this means that the model is more reliable during critical pollution events, which is essential for real-time air quality monitoring and early warning systems.

Another important aspect is the improved interpretability of the model. The analysis of attention weights provides insight into the internal decision-making process, aligning with the broader research trend toward explainable artificial intelligence (XAI). In applications where model predictions influence public health decisions, transparency is of paramount importance. The ability to identify which time steps contributed most to a given prediction allows domain experts to validate and trust the model's outputs.

From a computational perspective, the hybrid LSTM-Attention model maintains a reasonable level of complexity. Although the addition of the attention layer increases the number of parameters, the overall training time remains acceptable when using modern hardware accelerators such as GPUs. This makes the approach suitable for practical deployment in real-world systems, including urban monitoring platforms and Smart City infrastructures.

Despite these advantages, several limitations should be acknowledged. First, the model relies heavily on the quality and availability of input data. Sensor errors, missing values, or inconsistencies in measurement frequency can negatively impact prediction accuracy. Although preprocessing techniques mitigate these issues, they do not fully eliminate the risk of data-related bias.

Second, the current model focuses on short-term forecasting. While this is sufficient for immediate decision-making, extending the prediction horizon remains a challenge. Longer forecasting intervals typically require more complex architectures or additional contextual data.

Finally, although the attention mechanism improves interpretability, it does not fully explain causal relationships between variables. The model identifies correlations rather than direct cause-and-effect relationships, which should be considered when interpreting results.

Future research directions may include the exploration of Transformer-based architectures, which rely entirely on attention mechanisms and offer improved scalability [8]. Additionally, incorporating multimodal data—such as satellite imagery, traffic data, or user-generated sensor data—could further enhance predictive performance. Another promising direction is the development of hybrid predictive-reactive systems that not only forecast pollution levels but also trigger automated responses, such as adaptive traffic control or ventilation management.

In summary, the findings confirm that the proposed hybrid model represents a significant step forward in air quality forecasting. By combining high predictive accuracy, robustness, and interpretability, it provides a solid foundation for the development of next-generation environmental monitoring systems.

Based on the insights gained from both experimental evaluation and theoretical analysis, the final section summarizes the key contributions of this study and outlines directions for future research.

6. CONCLUSION AND FUTURE WORK

This study presented a hybrid deep learning approach for short-term air quality forecasting, combining Long Short-Term Memory networks with an Attention Mechanism. The proposed model was designed to address the limitations of traditional time-series forecasting methods, particularly in scenarios characterized by rapid environmental changes and extreme pollution events.

The experimental results demonstrated that the integration of the attention mechanism significantly improves prediction accuracy, as reflected by lower RMSE values and higher R^2 coefficients. Moreover, the model exhibited superior robustness under extreme conditions, where accurate forecasting is most critical for public health protection. These findings confirm that attention-based architectures are well-suited for modeling complex and dynamic environmental systems.

An additional contribution of this work is the improved interpretability of the model. By analyzing attention weights, it is possible to identify the most influential time steps in the prediction process. This feature aligns with the growing demand for transparency in artificial intelligence systems, particularly in domains where model outputs support decision-making processes affecting human health and safety.

From an application perspective, the proposed model can serve as a core component of intelligent air quality monitoring systems. Its ability to provide accurate short-term forecasts makes it suitable for integration into early warning systems, Smart City platforms, and environmental decision-support tools. Such systems could enable proactive responses, including public alerts, traffic regulation, and adaptive control of ventilation systems in public buildings.

Despite its advantages, the study also highlights several areas for further improvement. First, the current model focuses on short-term forecasting horizons. Extending the prediction range to medium- and long-term intervals remains an open research challenge and may require the incorporation of more advanced architectures or additional contextual data.

Second, future work should explore the integration of multimodal data sources, such as satellite imagery, traffic flow information, and data from mobile sensors. The inclusion of such heterogeneous data could provide a more comprehensive representation of environmental conditions and further enhance prediction accuracy.

Another promising direction involves the adoption of Transformer-based models, which rely entirely on self-attention mechanisms and offer improved scalability and parallelization capabilities [8]. These architectures may outperform recurrent models, especially when dealing with large-scale datasets.

Finally, the development of predictive-reactive systems represents an important step toward practical deployment. Such systems would not only forecast air pollution levels but also automatically initiate mitigation strategies based on predicted conditions. This could significantly reduce population exposure to harmful pollutants and improve overall urban air quality management.

In conclusion, the proposed hybrid LSTM-Attention model provides an effective and scalable solution for air quality forecasting. Its combination of accuracy, robustness, and interpretability makes it a strong candidate for real-world applications and a valuable contribution to the field of environmental data science.

BIBLIOGRAPHY

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.

[2] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, 1997.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," 2014.

[4] X. Li et al., "Deep learning architecture for air quality predictions," *Environmental Science and Pollution Research*, 2020.

[5] A. Heydari et al., "Air pollution forecasting using deep learning models," *Clean Technologies and Environmental Policy*, 2022.

[6] D. Qin et al., "CNN-LSTM-Attention model for PM2.5 forecasting," *IEEE Access*, 2019.

[7] D. R. Liu et al., "Attention-based LSTM for air pollution prediction," *Expert Systems with Applications*, 2021.

[8] A. Vaswani et al., "Attention Is All You Need," *NeurIPS*, 2017.

Aneta Wiktorzak: https://orcid.org/0000-0002-3885-1340

Ryszard Szczebiot: https://orcid.org/0000-0002-9084-915X

Leszek Gołdyn: https://orcid.org/0000-0002-0689-8590

COMPARATIVE ANALYSIS OF THE PERFORMANCE OF CLASSICAL PID CONTROLLERS WITH VARIABLE PI_D STRUCTURE CONTROLLERS

Leszek GOŁDYN[1], Ryszard SZCZEBIOT[1], Aneta WIKTORZAK[1]

[1] Faculty of Computer Science and Technology, University of Lomza, ul. Akademicka 14, 18-400 Lomza, Poland
rysbiot@al.edu.pl, lgoldyn@al.edu.pl, awiktorzak@al.edu.pl

ABSTRACT: This paper compares the performance of PID controllers with variable structure in terms of their performance in controlling a second-order static object with a transport delay, compared to the performance of classical PI control.

Based on simulation models of control systems with classical PID controllers and simulations of a system with a PI_D controller with variable structure, it can be observed that using the modified structure controllers yields waveforms with shorter control times. It can also be concluded that the system with the PI_D controller has lower overshoot and a much shorter settling time than the system with the PI controller. The system with the PI controller exhibits large decaying oscillations, while the waveform with the variable structure controller exhibits minimal and rapidly decaying oscillations.

After conducting simulation models of automatic control systems with PID controllers with variable structure (PI_D), it can be concluded that using these controllers significantly improves the quality of the control system in terms of both settling time and overshoot, compared to classical PI control. Even several times smaller values of the regulation time were observed here for both PI_D.

Key words: automatic control systems, PID controllers, PI_D controllers, controllers with a changed structure, programmable logic controllers

INTRODUCTION

Controllers are a crucial and essential device in industrial control systems. Processes often require maintaining pre-defined, specific operating parameter values with the required accuracy and time. Skillful engineering use of controllers in production processes allows for increased efficiency, reduced energy consumption through optimal utilization, and increased process automation. In continuous signal control systems, not only in industry, the most commonly used type of controller is the PID controller. It could be said that controllers based on the PID algorithm are eternally young. The first automatic stabilizing device was used by James Watt in 1788. Since then, the control algorithm, which has been modified many times, essentially operates on the same principle. A PID controller consists of three functional parts. They can be described in a simplified manner. The first part is the proportional term (P), followed by the integral

term (I), and the derivative term (D). The proportional term influences the signal rise time, the integral term reduces the steady-state error to zero, and the derivative term ensures the system's ability to respond to rapid signal changes. During the operation of a facility and the operation of the control system, certain elements related to the facility may change, affecting the initial values of the facility parameters. In such cases, controllers may not optimally affect the process or even cause system instability. Therefore, many modified PID-based algorithms have been developed, creating variable-structure controllers with unconventional algorithms. These are not widely used, but they can be just as effective. In certain situations, PI_D controllers are significantly superior to the commonly used classic PID control law [1].

Alphonsus [2, 3] and Namekar [4] presents a review on the applications of programmable logic PLC controllers.

The aim of this work and analysis is to examine an automatic control system with a variable-structure PI_D controller in terms of its performance under changes in the object's parameters, represented by changes in its transfer function parameters, and under disturbances. Variable-structure PI_D controllers. Variable-structure PI_D controllers are PID controllers whose structure differs somewhat from that of a classic PID controller, causing them to operate differently. Such controllers are used in specific cases where conventional PID controllers are unable to provide adequate control. This article focuses on one variable-structure PID controller: the PI_D controller.

All simulations of the control systems studied in this article were conducted in the Matlab/Symulik environment [5]. The PI_D controller exhibits PI properties when the setpoint changes and PID properties when the controlled variable changes [1]. Similar to a PID controller, the PI_D controller consists of three parts. The proportional and integral parts are connected as in a PI/PID controller. The derivative part, however, is connected directly to the feedback signal, not to the error, as in a standard PID controller. Furthermore, the derivative part is subtracted, not added, as in a standard PID controller [1].

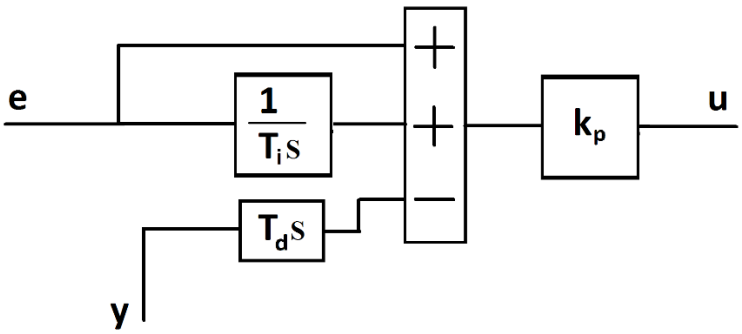


Fig. 1. PI_D controller structure.

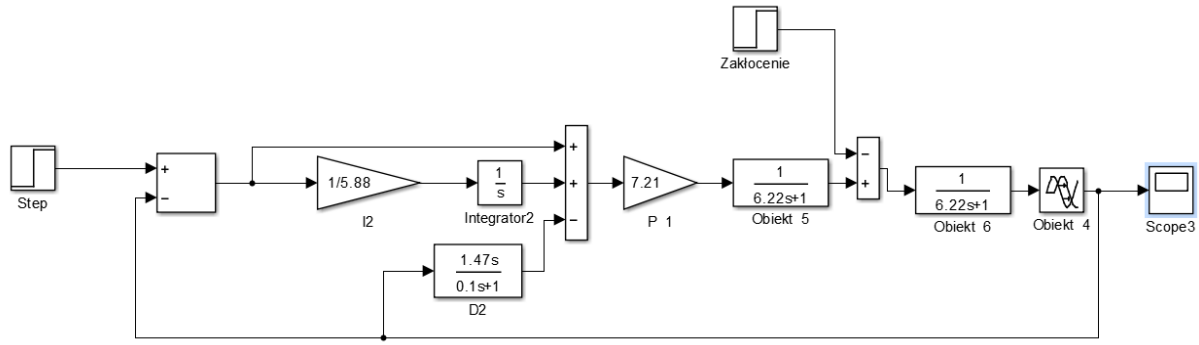


Fig. 2. Automatic control system with a PI_D controller.

The structure of the PI_D controller is shown in Fig. 1, and Fig. 2 shows the tested (simulated) automatic control system for the object described by equation (1) using the PI_D controller in the Matlab/Simulink environment.

1. CONTROL OBJECT AND ITS IDENTIFICATION

To investigate models of automatic control systems using the variable structure controllers and unconventional control algorithms proposed in this article, a two-chamber closed pneumatic cascade was selected as the control object. This object belongs to the group of static automation objects and is of second order. Parametric identification was performed using the Strejc method [6, 7].

$$G(s) = \frac{e^{-1.09s}}{(6.22s+1)^2} \tag{1}$$

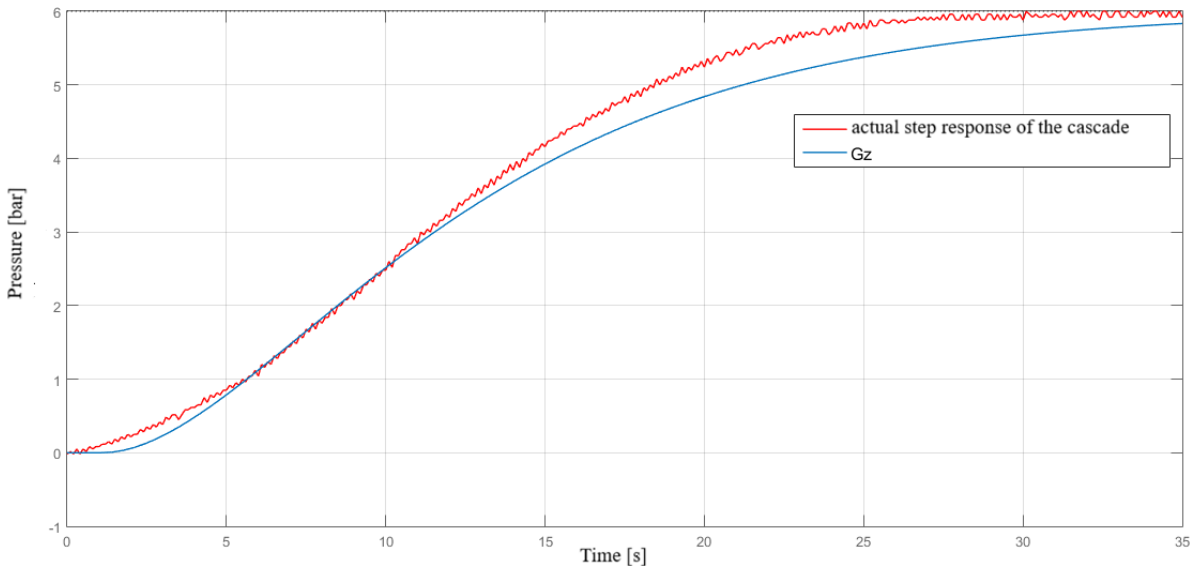


Fig. 3. Comparison of the actual step response of the cascade with the step response determined based on the identified equivalent transfer function of the object.

Fig. 3 compares the dynamic response of the identified inertial object, plotted from the model using equation (1), with the experimental response obtained from tests of a second-order pneumatic cascade.

2. TESTING THE AUTOMATIC CONTROL SYSTEM WITH A PI_D CONTROLLER AND CONTROLLER SETTINGS

The PI and PI_D controller settings will be determined from the Ziegler-Nichols second law (stability limit) [6]. Fig. 4 shows the Bode characteristic with the determined magnitude and phase margins.

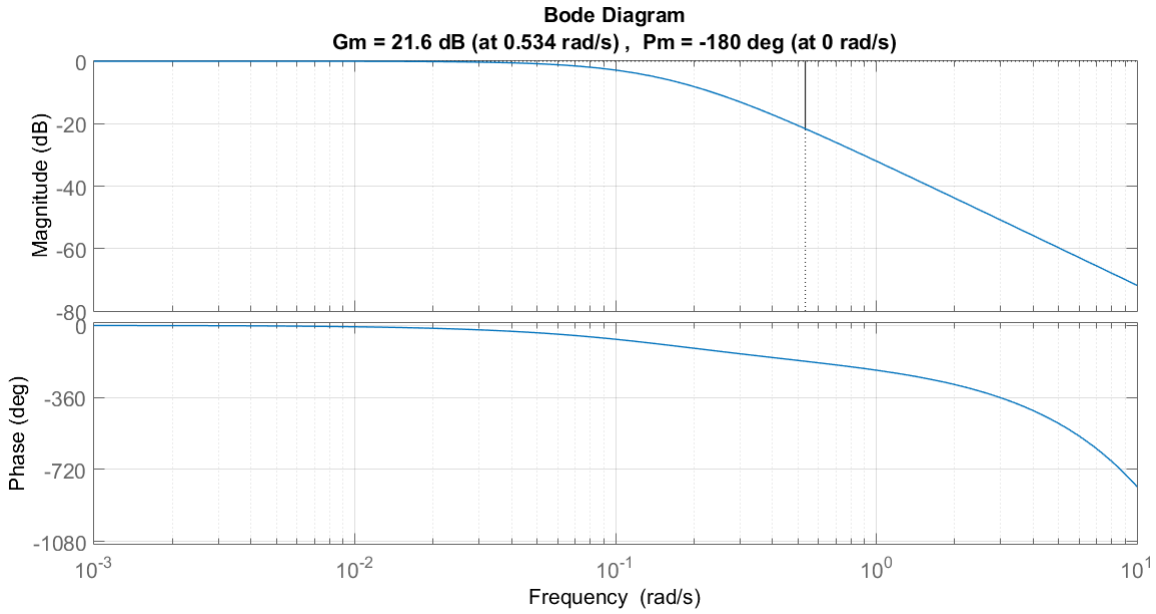


Fig. 4. Bode characteristic for the tested object.

Using the determined modulus margin (Fig. 4), K_{KR} was calculated:

$$20\log K_{KR}=21.6[\text{dB}],$$

$$K_{KR}=10^{\frac{21.6}{20}}=10^{1,08}=12.023.$$

Knowing the critical gain K_{KR}, we determine the oscillation time T_{osc} from the controlled variable $y(t)$ of the automatic control system with the P controller (Fig. 5) with gain $k_p=K_{KR}$. Based on the obtained oscillations (non-quenching), we read the oscillation period.

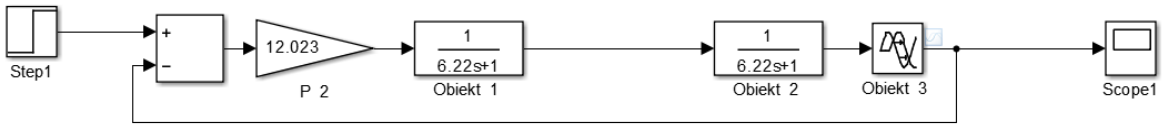


Fig. 5. Diagram of an automatic control system with a P controller.

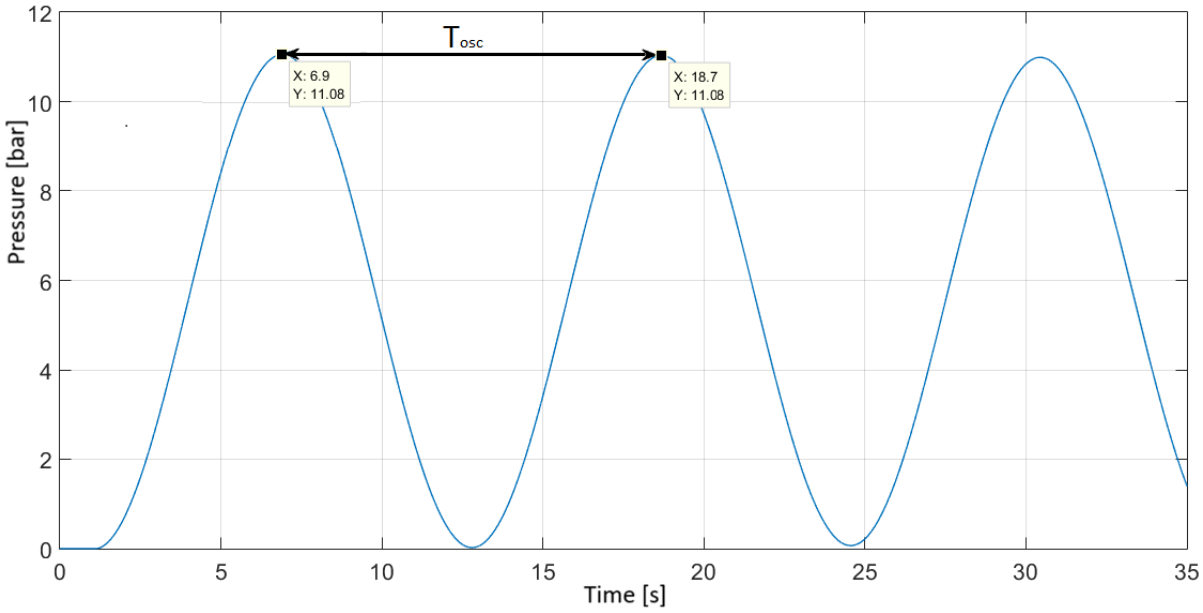


Fig. 6. Unquenched oscillations for a system with a P controller.

The oscillation time T_{osc} determined from Fig. 6 is 11.77 s. Then, knowing K_{KR} and T_{osc} and using the second Ziegler-Nichols method [7], the controller settings were determined:

P: $k_P = 0.5 \cdot K_{KR} = 6.01$;

PI: $k_p = 0.45 \cdot K_{KR} = 5.41$, $T_i = T_{osc} \cdot 0.85 = 9.8$s;

PID: $k_P = 0.6 \cdot K_{KR} = 7.21$, $T_i = 0.5 \cdot T_{osc} = 5.88$s, $T_d = 0.125 \cdot T_{osc} = 1.47$s.

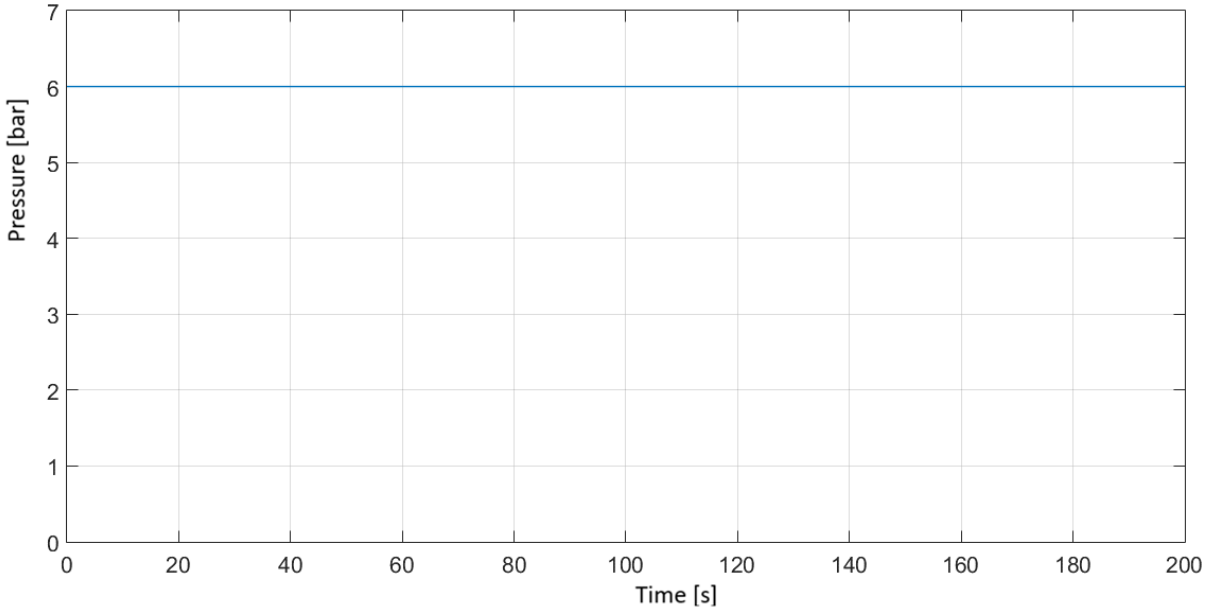


Fig. 7. Continuous input signal.

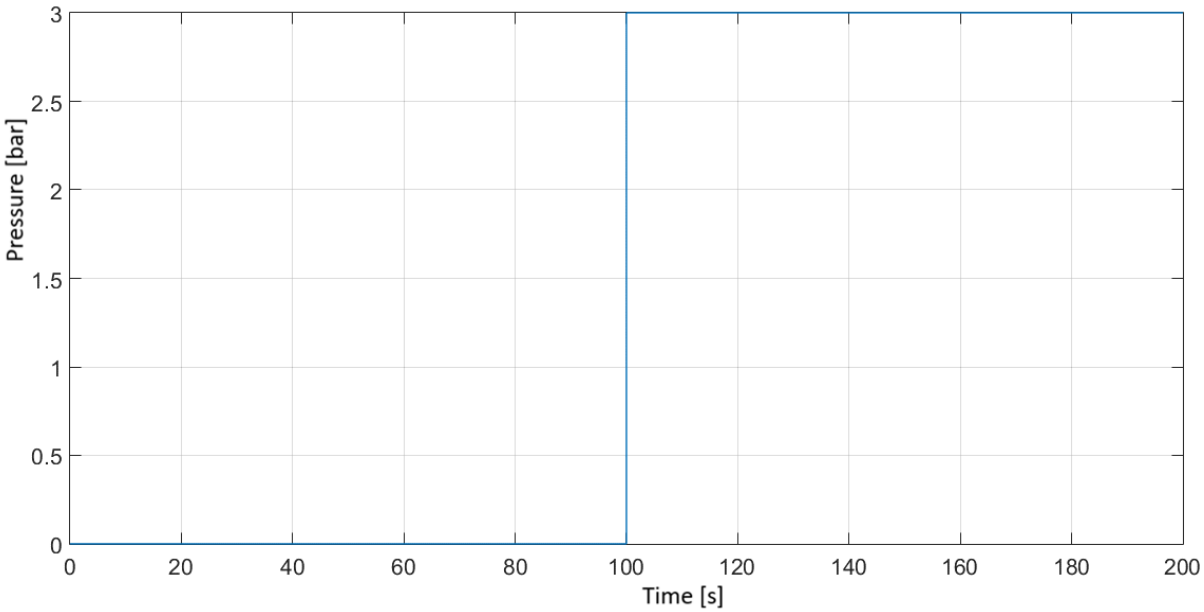


Fig. 8. Step-type disturbance signal.

Figures 7 and 8 show two classic input signal waveforms used in testing the control system, as described in Figures 5, 9, and 10. Figure 7 shows the forcing signal at the input to the system. Figure 8 shows the disturbance signal acting directly on the system.

As shown in Figures 9 and 10, the control systems with both the PI_D and PI controllers were subject to a step change in the setpoint and a step change in the disturbance. The setpoint step amplitude was 6 bar (Figure 7), and the disturbance amplitude was 3 bar (-3 bar) (Figure 8). The step change in the setpoint value occurred at t=0 s (at the start of the simulation), and the disturbance signal (disturbance) occurred at t=100 s. The final simulation time for the systems was t_{stop}=200 s.

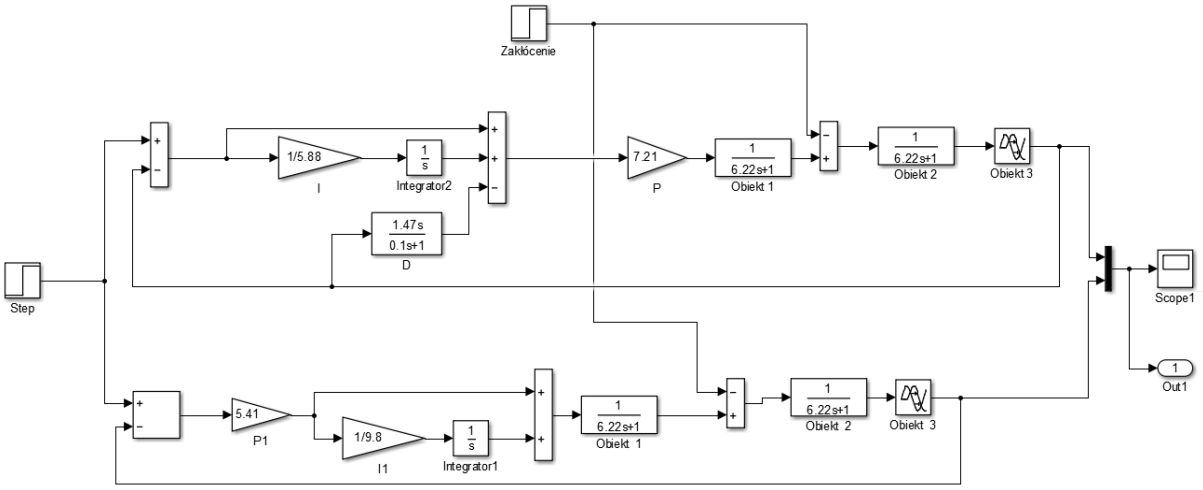


Fig. 9. Diagram of an automatic control system with a PI_D controller (upper part of the diagram) and a PI controller (lower part of the diagram).

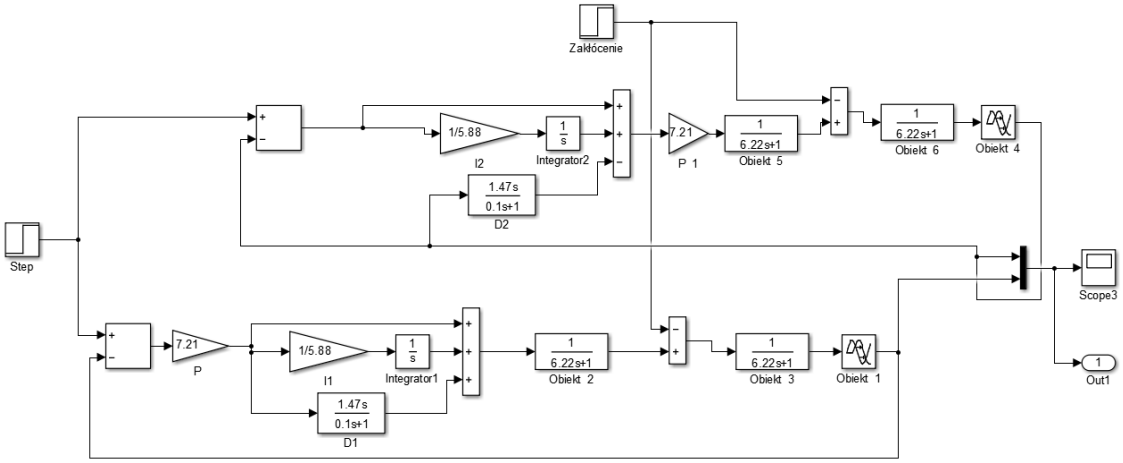


Fig.10. Diagram of an Automatic Control System with a PI_D Controller (upper part of the diagram) and a PID Controller (lower part of the diagram).

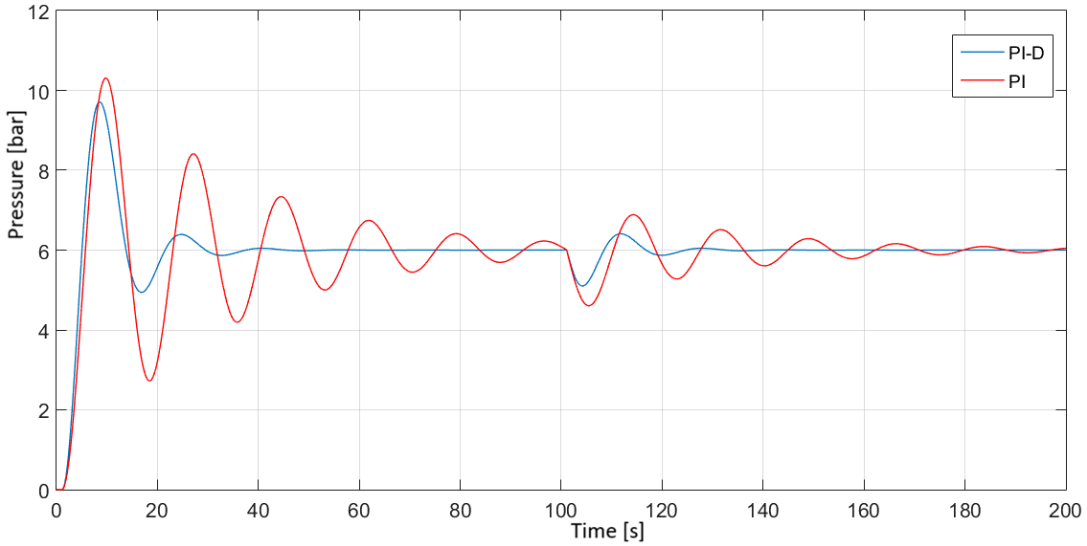


Fig. 11. Comparison of Automatic Control Systems with PI_D and PI Controllers.

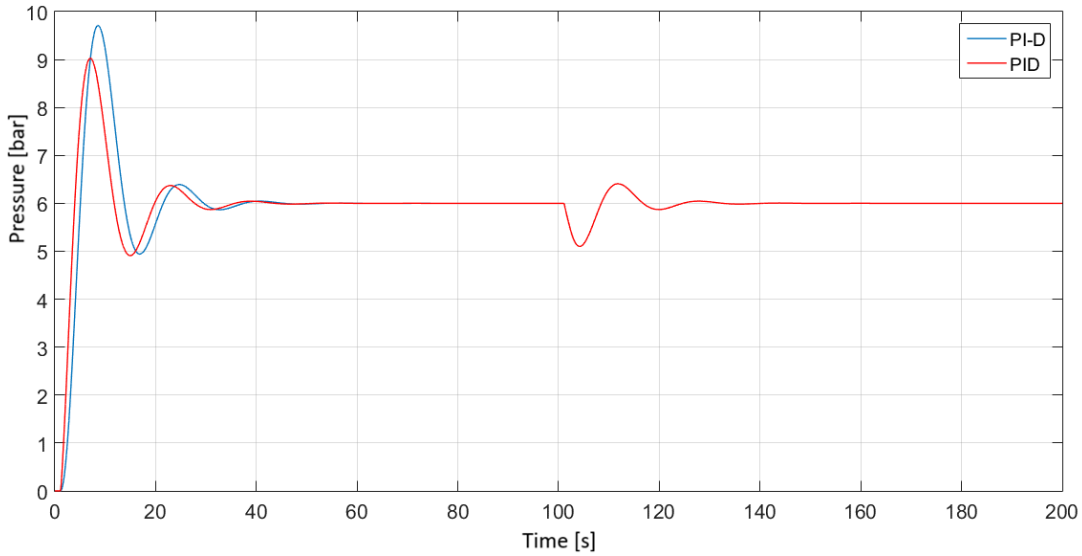


Fig. 12. Comparison of Automatic Control Systems with PI_D and PID Controllers.

Fig. 11 compares the control results using PI_D controllers and the classic PI controller as shown in Fig. 9.

Figure 12 compares control systems with a PI_D and PID controller. The PI_D controller has a larger overshoot than the PID controller, while the settling time of both systems is very similar. After reaching the setpoint, both controllers operate identically when a disturbance occurs.

CONCLUSION

This paper compares the performance of PID controllers with variable structure in terms of their performance when controlling a second-order static object with a transport delay, compared to the performance of classical PI control. The comprehensive analysis does not include the effect of changes in the parameters of the controlled object. The effect of changing the parameters of the controlled object on the dynamic behavior of the regulated variable is of interest, as these values are used in controllers with unconventional algorithms.

Based on simulation models of control systems with a classical PID controller and simulations of the system with a PI_D variable structure controller, it can be observed that using controllers with modified structures resulted in shorter settling times. Similar conclusions can be drawn when assessing the magnitude of the overshoot. Based on Fig. 11, among other things, it can be concluded that the system with a PI_D controller has lower overshoot and a much shorter settling time than the system with a PI controller. The system with a PI controller exhibits large decaying oscillations, while the waveform with a variable structure controller exhibits minimal and rapidly decaying oscillations.

After simulating automatic control systems with PID controllers with a variable PI_D structure, it can be concluded that using this type of controller significantly improves the quality of the control system's performance in terms of both settling (regulation) time and overshoot, compared to conventional PI control (Fig. 11). A direct conclusion from this article suggests another direction for further analysis and simulation. These should involve determining the impact of the plant parameters on the characteristics of systems using unconventional controllers.

Based on Fig. 11, it can be concluded that the system with a PI_D controller has lower overshoot and a much shorter settling time than the system with a PI controller. The system with a conventional PI controller exhibits large decaying oscillations.

After conducting simulation models of automatic control systems with PID controllers with a variable PI_D structure, it was found that using these controllers significantly improved the quality of the control system's operation, both in terms of settling time (regulation) and overshoot, compared to classic PI control. The regulation time for the PI_D controller was even several times shorter.

REFERENCES

[1] Brzózka J.: Regulatory i układy automatyki, MIKOM, Warszawa, 2004.

[2] ALPHONSUS, Ephrem Ryan; ABDULLAH, Mohammad Omar. A review on the applications of programmable logic controllers (PLCs). *Renewable and Sustainable Energy Reviews*, 2016, 60: 1185-1205.

[3] ALPHONSUS, Ephrem Ryan; ABDULLAH, Mohammad Omar. A review on the applications of programmable logic controllers (PLCs). *Renewable and Sustainable Energy Reviews*, 2016, 60: 1185-1205.

[4] NAMEKAR, Swapnil Arun; YADAV, Rishabh. Programmable Logic Controller (PLC) and its applications. *International Journal of Innovative Research in TechnologY (IJIRT)*, 2020, 6.11: 372-376.

[5] Łysakowska B., Mzyk G.: Komputerowa symulacja układów regulacji w środowisku Matlab/Simulink, Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław, 2005.

[6] Siemieniako F., Gosiewski Z.: Automatyka, Tom I, Modelowanie i analiza układów, Wydawnictwo Politechniki Białostockiej, Białystok, 2006.

[7] Siemieniako F.: Podstawy automatyki z zadaniami, Wydawnictwo Politechniki Białostockiej, Białystok, 1996.

Leszek Gołdyn: https://orcid.org/0000-0002-0689-8590
Ryszard Szczebiot: https://orcid.org/0000-0002-9084-915X
Aneta Wiktorzak: https://orcid.org/0000-0002-3885-1340

APPLICATION OF GROVER'S ALGORITHM TO THE 3SAT PROBLEM

TRONCZYK Piotr[1]

Akademia Łomżyńska, Wydział Nauk Informatyczno-Technologicznych

ptronczyk@al.edu.pl[1]

ABSTRACT: The 3-SAT problem is a core challenge in computational complexity. With the rise of quantum computing, new methods are being developed to apply quantum algorithms to NP-hard problems. This paper explores an approach using Grover's algorithm to address Boolean satisfiability. We focus on the design of a quantum oracle capable of recognizing satisfying assignments. The research also examines the limitations imposed by quantum gate implementation and the impact of quantum speedup on system scalability.

Key words: 3SAT, Grover[1], Quantum computing, speedup

INTRODUCTION:

Developed in 1996 by Lov Grover at Bell Labs, Grover's algorithm is a cornerstone of quantum computing, providing a quadratic speedup for searching unstructured databases. While classical search algorithms require $\mathrm{O}(N)$ time, averaging $N/2$ operations. Grover's algorithm achieves a complexity of $\mathrm{O}\left(\sqrt{\mathrm{N}}\right)$. This efficiency is fundamentally driven by the mechanism of quantum amplitude amplification, which enhances the probability of measuring the target state.

This approach can be adapted to the SAT problem by mapping the Boolean formula to an oracle function, where the search space is treated as an unsorted database of potential assignments.

1. QUANTUM ORACLE

The algorithm employs an oracle that identifies the target state by performing a phase flip. It should be noted that while the algorithm assumes the existence of such an oracle, it does not specify its construction; the oracle's implementation depends entirely on the specific problem instance. The process begins with the preparation of a uniform superposition of all possible qubit configurations. This initialization method is characteristic of various quantum algorithms, including the Deutsch[2] and Shor[3] algorithms.".

2. CLASSICAL OPERATIONS ON QUANTUM COMPUTER

Quantum computers are universal in the sense of the Turing machine, meaning they are capable of computing any function that is classically computable. Consequently, they are subject to the same fundamental theoretical constraints, such as the halting problem. While classical programming relies on high-level abstractions, quantum programming currently remains largely at the gate level; because high-level quantum programming languages have not yet reached the maturity of their classical counterparts, quantum computers programming is primarily conducted through the direct assembly of quantum gate sequences.

The simplest example is the natural numbers $\mathrm{n} \in \mathbb{N}$. Each such number can be represented in binary as the sum of powers of 2. For example:

$$12 = 8 + 4 = 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 0 \times 2^0.$$

The binary representation of the number 12 is a sequence of bits that represent the coefficients of each power of two:

$$12_{10} = 1100_2.$$

In this case, four bits are needed to represent the number 12. In general, using a register of length n bits, we can represent 2^{n} different values ranging from 0 to $2^{\mathrm{n}} - 1$.

An analogous scheme can be used to represent numbers in a quantum register. We adopt the convention that the least significant bit (LSB) is on the right. In the graphical representation of a quantum circuit, the top line corresponds to the bit with index 0. Thus, the number 12 (in binary notation 1100_2) will be encoded in four qubits and written in keta notation as $|1100\rangle$

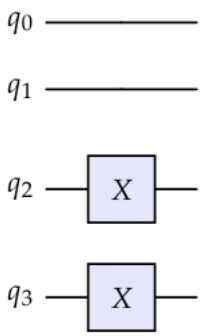


Figure 1 Quantum circuit encoding 12 (in binary notation 1100_2)

There are at least two fundamental differences between quantum and classical (Boolean) logic. First, quantum gates must be reversible. This means that there must be a unique correspondence between

the input and output, which consequently requires that the number of input qubits equal the number of output qubits.

This contrasts with classical logic, where many operations are irreversible – for example, most logic gates take two arguments and return only one, making it impossible to reconstruct the input state from the output.

Another significant difference is the possibility of changing the basis. Quantum gates allow for transformations of the state basis; for example, a computational basis $|0\rangle$, $|1\rangle$ can be transformed into a basis $|+\rangle$, $|-\rangle$. This concept does not exist in classical logic, because logical values are rigid and limited to the set $\{0,1\}$, meaning that a similar change of basis is impossible.

When constructing quantum logic circuits, we can use a similar scheme, starting with the simplest single-qubit gate and building quantum logic based on it.

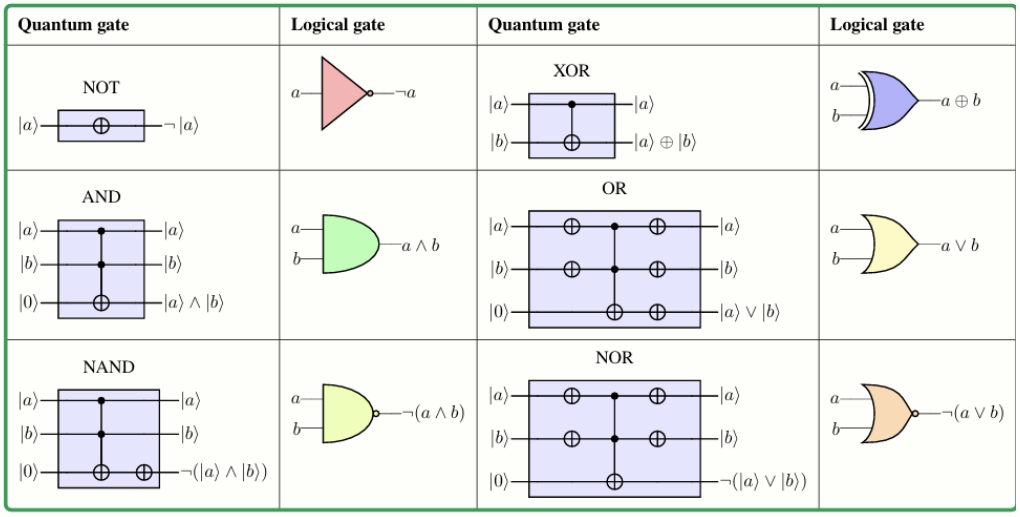


Figure 2 Quantum gates

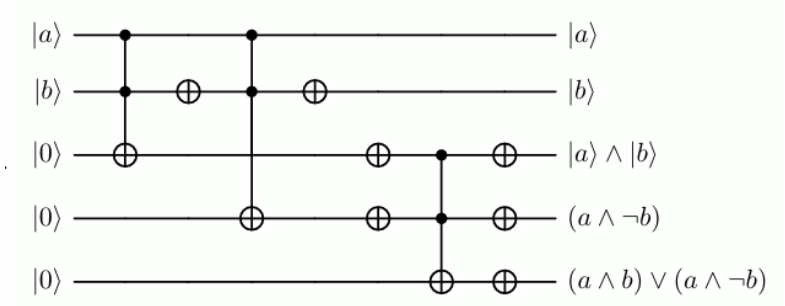


Figure 3 Quantum circuit for logical expression

Consider a simple problem: finding the satisfying assignments for an OR function. Out of the four possible input combinations $\{00,01,10,11\}$,the solutions are $\{01, 10, 11\}$. We shall construct a quantum circuit consisting of an oracle and a diffusion operator.

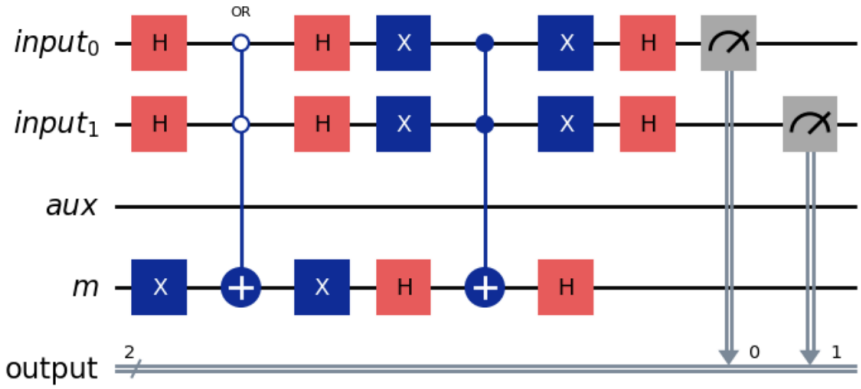


Figure 4 Grover circuit for OR problem

Executing the circuit 1000 times we obtain following results:

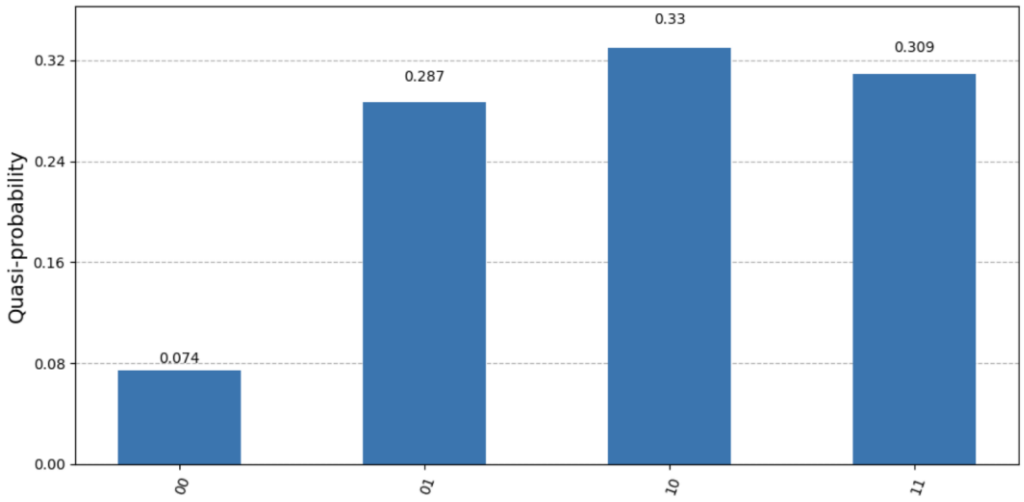


Figure 5 Probability distribution

Thus, the probability of obtaining the correct result in a single trial is 91%. In classical brute force method we have 75% probability of success.

3. HOW TO CALCULATE THE OPTIMAL NUMBER OF ITERATIONS IN GROVER'S ALGORITHM

Grover's algorithm works by rotating a state vector within a two-dimensional subspace. This subspace is spanned by two vectors:

1. $|\mathrm{bad}\rangle$: The superposition of all non-target states.
2. $|\mathrm{good}\rangle$: The superposition of all target (correct) states.

Let N be the total number of states and M be the number of target states. The initial state of the system $|s\rangle$ (the uniform superposition), can be expressed in terms of the angle θ it makes with the "bad" states: $\sin(\theta) = \frac{M}{N}$. Where θ represents the amplitude of the target states in the initial superposition.

Each Grover iteration (consisting of the Oracle and the Diffusion Operator) performs a rotation of the state vector by an angle of 2θ in this 2D plane.

After the preparation of the state, the vector is at an angle θrelative to the "bad" axis. After k iterations, the angle of the state vector becomes: $\theta_{final} = (2k + 1)$.

The goal is to rotate the state vector as close as possible to the target axis ($|\mathrm{good}\rangle$), which corresponds to an angle of $\pi/2$ (90 degrees). To find the optimal number of iterations k, we set the final angle to $\pi/2$:

$$(2k + 1) \approx \frac{\pi}{2}$$

Now, we solve for k:

$$2\mathrm{k} + 1 \approx \frac{\pi}{2\theta}$$
$$2\mathrm{k} \approx \frac{\pi}{2\theta} - 1$$
$$\mathrm{k} \approx \frac{\pi}{4\theta} - \frac{1}{2}$$

In most practical applications of Grover's algorithm, the number of target states M is much smaller than the total number of states N ($\mathrm{M} \ll \mathrm{N}$). This means the angle θ is very small. For small angles, we can use the approximation $\sin(\theta) \approx \theta$.

Since $\sin(\theta) = \sqrt{\mathrm{M/N}}$, we can substitute $\theta \approx \sqrt{\mathrm{M/N}}$ into our equation:

$$k \approx \frac{\pi}{4\sqrt{M/N}} - \frac{1}{2}$$

By simplifying, we arrive at the standard formula for the optimal number of iterations:

$$k \approx \frac{\pi}{4}\sqrt{\frac{N}{M}}$$

4. SAT PROBLEM

The Boolean Satisfiability Problem (**SAT**) is the problem of determining if there exists an assignment of truth values (True or False) to a set of variables that makes an entire logical formula evaluate to True.

The "3" in 3-SAT refers to the specific structure of the constraints. In 3-SAT, the formula must be in Conjunctive Normal Form (CNF), consisting of:

1. Variables: $\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \ldots, \mathrm{x}_\mathrm{n}$ (each can be True or False).
2. Literals: A variable (x_1) or its negation ($\neg \mathrm{x}_1$).
3. Clauses: A group of literals joined by OR (∨). In 3-SAT, each clause must contain exactly three literals.
4. The Formula: All clauses are joined by AND (∧).

A 3-SAT formula looks like this: $(x_1 \vee x_2 \vee \neg x_3) \wedge (\neg x_1 \vee \neg x_2 \vee x_3)$ Let $\mathrm{x}_1 = \mathrm{True}, \mathrm{x}_2 = \mathrm{False}, \mathrm{x}_3 = \mathrm{True}$.

x_1	x_2	x_3	$x_1 \vee x_2 \vee \neg x_3$	$\neg x_1 \vee \neg x_2 \vee x_3$	Formula
True	False	True	$T \vee F \vee F = T$	$\mathrm{F} \vee \mathrm{T} \vee \mathrm{T} = T$	$T \wedge T = T$

3-SAT is a cornerstone of computer science because it is NP-Complete. This has two massive implications:

- **Hardness**: There is no known "fast" (polynomial-time) algorithm to solve 3-SAT. As the number of variables n grows, the time required to solve it grows exponentially (in the worst case) using current methods.
- **Universality**: If anyone ever finds a fast way to solve 3-SAT, they will have simultaneously found a fast way to solve every problem in the NP class.

5. SOLUTION WITH GROVER ALGORITHM

Consider the following problem $(x_1 \vee x_2 \vee \neg x_3) \wedge (\neg x_1 \vee \neg x_2 \vee x_3) \wedge (x_1 \vee \neg x_2 \vee x_3)$

x_1	x_2	x_3	$x_1 \vee x_2 \vee \neg x_3$	$\neg x_1 \vee \neg x_2 \vee x_3$	$x_1 \vee \neg x_2 \vee x_3$	Res
True	True	True	$T \vee \mathrm{T} \vee F = T$	$\mathrm{F} \vee \mathrm{F} \vee \mathrm{T} = T$	$\mathrm{T} \vee \mathrm{F} \vee \mathrm{T} = T$	T
True	True	False	$T \vee \mathrm{T} \vee \mathrm{T} = T$	$\mathrm{F} \vee \mathrm{F} \vee \mathrm{F} = F$	$\mathrm{T} \vee \mathrm{F} \vee \mathrm{T} = T$	F
True	False	True	$T \vee \mathrm{F} \vee \mathrm{F} = T$	$F \vee \mathrm{T} \vee \mathrm{T} = T$	$T \vee \mathrm{T} \vee \mathrm{T} = T$	T
True	False	False	$T \vee \mathrm{F} \vee \mathrm{T} = T$	$F \vee \mathrm{T} \vee \mathrm{F} = T$	$T \vee \mathrm{T} \vee \mathrm{F} = T$	T
False	True	True	$F \vee \mathrm{T} \vee \mathrm{F} = T$	$T \vee \mathrm{T} \vee \mathrm{T} = T$	$F \vee \mathrm{F} \vee \mathrm{T} = T$	T
False	True	False	$F \vee \mathrm{T} \vee \mathrm{T} = T$	$T \vee \mathrm{F} \vee \mathrm{F} = T$	$F \vee \mathrm{F} \vee \mathrm{F} = F$	F
False	False	True	$F \vee \mathrm{F} \vee \mathrm{F} = F$	$T \vee \mathrm{T} \vee \mathrm{T} = T$	$F \vee \mathrm{T} \vee \mathrm{T} = T$	F
False	False	False	$F \vee \mathrm{F} \vee \mathrm{T} = T$	$T \vee \mathrm{T} \vee \mathrm{F} = T$	$F \vee \mathrm{T} \vee \mathrm{T} = T$	T

The probability of guessing the result is 5/8, which is 63%. With the quantum approach, for one step of Grover's algorithm, this probability increases to 74%.

For two iterations of Grover's algorithm we can increase the probability of guessing the correct result to 86%.

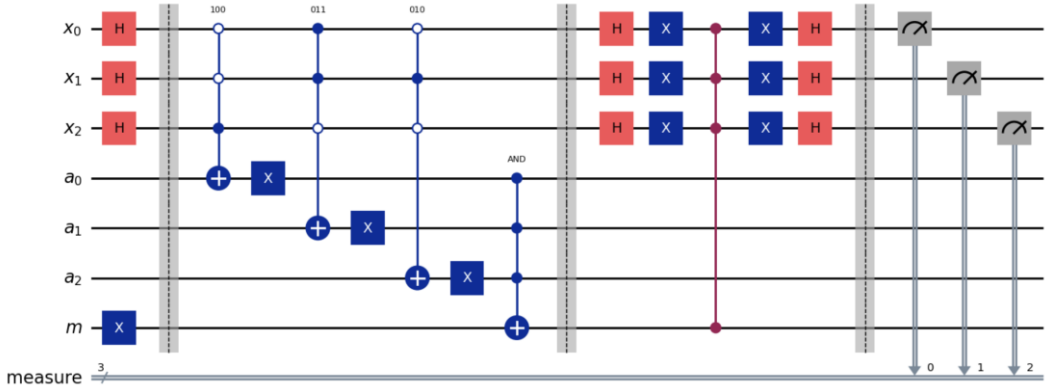


Figure 6 Quantum Circuit with one iteration

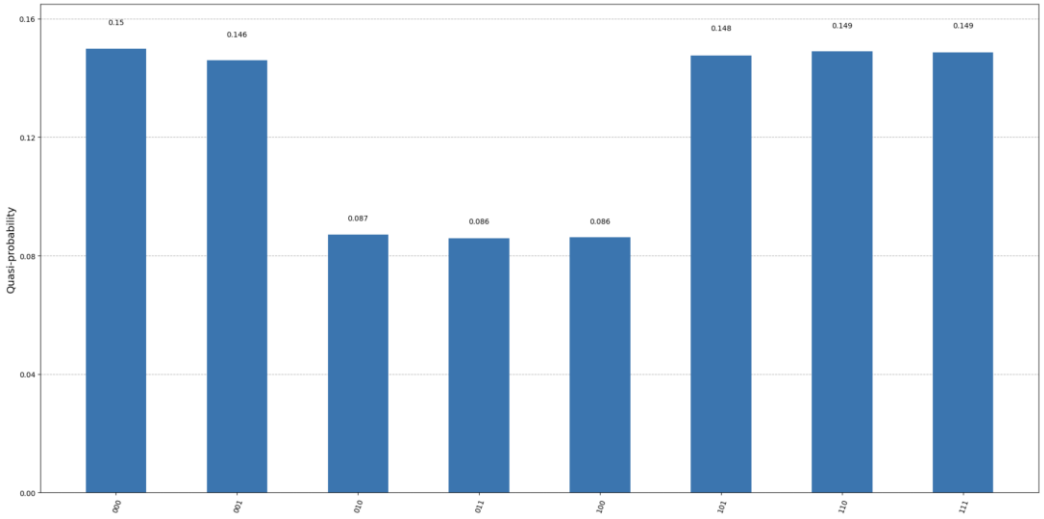


Figure 7 Probability distribution after 1000 shots

CONCLUSIONS

The primary conclusion is that Grover's algorithm provides a quadratic speedup rather than an exponential one. While a classical brute-force search for a 3-capture assignment takes $O(2^n)$ time, Grover's algorithm reduces this to $O(2^{n/2})$. This means that while the quantum approach is significantly faster, it does not "break" the complexity of 3-SAT in the same way that Shor's algorithm breaks RSA (which provides an exponential speedup). The problem remains fundamentally difficult even for quantum computers.

A critical bottleneck is the cost of the Oracle. To use Grover's algorithm, one must implement the 3-SAT constraints as a quantum circuit (the Oracle).

- As the number of variables (n) and clauses (m) increases, the complexity of the quantum circuit required to evaluate the 3-SAT formula grows significantly.

- If the overhead of constructing and executing this complex circuit exceeds the time saved by the quadratic speedup, the quantum advantage may be negated in practical applications.

This study explores the implications of deviating from the theoretically optimal number of Grover iterations in the presence of environmental decoherence. We investigate whether a "sub-optimal" execution characterized by fewer iterations and a reduced success probability yields a lower total computational overhead when accounting for the cumulative error rates inherent in NISQ (Noisy Intermediate-Scale Quantum) devices.

In an ideal, noise-free environment, the complexity of Grover's algorithm is governed by the requirement to maximize the success probability P, where $\mathrm{P} \rightarrow 1$ as the number of iterations k approaches $\mathrm{k}_{\mathrm{opt}} \approx \frac{\pi}{4}\sqrt{\mathrm{N}}$. The total computational cost C can be modeled as:

$$C = k\frac{1}{P(k)}$$

Where k is the number of iterations and $1/\mathrm{P}(\mathrm{k})$ represents the expected number of repetitions required to achieve a single successful measurement. Mathematically, in a unitary (noise-free) system, C is minimized when k is chosen to maximize $\mathrm{P}(\mathrm{k})$, leading to the well-known $\mathrm{O}(\sqrt{\mathrm{N}})$ complexity.

REFERENCES

[1] Lov K. Grover, (1986), "A fast quantum mechanical algorithm for database search" in Title of the arXiv. https://arxiv.org/abs/quant-ph/9605043

[2] Deutsch D., (1985), "Quantum theory, the Church–Turing principle and the universal quantum computer," Proceedings of the Royal Society A. pp 97–117.

[3] Peter W. Shor,., (1997), "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer," SIAM Journal on Computing, Society for Industrial & Applied Mathematics, DOI 10.1137/s009753979529317

2D–3D FUSION IN MEDICAL IMAGING: METHODS, ALGORITHMS AND CHALLENGES

Izabela LESZCZYŃSKA

Polish-Japanese Academy of Information Technology,
Department of Mechanics, Computer Science and Robotics, Warsaw, Poland
ilesz@pjwstk.edu.pl

ABSTRACT: This article reviews image processing methods for 2D–3D medical image fusion in surgical and interventional guidance. The main objective of the fusion is to relate current two dimensional procedural images, such as ultrasound, fluoroscopy, endoscopy, radiography, or angiography, to three dimensional data obtained from CT, MRI, 3D ultrasound, segmented structures, surface models, or anatomical reconstructions. The review focuses on the computational aspects of fusion.

The work first summarises the characteristics of the most relevant 2D and 3D medical imaging modalities. It then presents a methodological taxonomy of 2D–3D fusion approaches, including manual correspondence selection, semi-automatic methods based on landmarks and features, sensor and probe tracking, markerless surface registration, intensity based registration, projection based registration, model based fusion, deformable fusion, and machine learning methods. Particular attention is given to the strengths and limitations of each group in relation to surgical use.
The review shows that no single method is suitable for all image pairs and clinical scenarios. Projection based methods are strong in X-ray-to-CT registration, feature based and landmark based methods remain important in multimodal fusion, deformable methods are required for soft tissue and histology related applications, and machine learning methods can improve speed and representation learning. However, all approaches remain affected by practical limitations such as initialisation, modality differences, tissue deformation, restricted field of view, computational cost, and validation difficulties.
Key words: 2D–3D fusion, medical image registration, surgical navigation, image guided surgery, multimodal imaging

1. INTRODUCTION

Modern surgery increasingly relies on the integration of complementary imaging data acquired before and during an intervention. Preoperative three-dimensional imaging modalities, such as computed tomography (CT) and magnetic resonance imaging (MRI), provide detailed anatomical information,

including the spatial relationships between target structures and adjacent vessels, nerves, organs, and other critical tissues [1-3]. In contrast, intraoperative two-dimensional imaging modalities, such as ultrasound, fluoroscopy, endoscopy, or radiography, provide real-time or near-real-time information about the current surgical field [1, 4, 5]. The fusion of 2D and 3D medical images aims to combine these complementary sources of information into a coherent spatial representation that can support spatial orientation during the surgery and hence surgical decision making.

The motivation for 2D–3D medical image fusion is directly related to the limitations of intraoperative perception – visual and ultrasound. During surgical procedures, parts of the anatomy may be obscured by overlying tissue, limited field of view, bleeding, organ deformation, or the restricted perspective imposed by minimally invasive access [1, 4, 5]. By aligning intraoperative 2D images with preoperative 3D data, fusion systems can provide the surgeon with additional information about hidden or partially visible anatomical structures. This may help reduce unnecessary dissection, avoid inadvertent vessel injury, and support more selective tissue manipulation. As a result, accurate 2D–3D fusion may improve surgical precision, reduce collateral tissue damage, and contribute to safer interventions.
In this review, 2D–3D fusion is understood as the computational process of spatially integrating two-dimensional medical images with three-dimensional anatomical data. In most cases, this process involves 2D–3D registration, that is, the estimation of the spatial transformation that aligns a 2D image with a 3D volume, model, or surface. Depending on the imaging modalities and clinical context, the 2D data may represent an ultrasound frame, fluoroscopic image, endoscopic image, histological section, angiographic image, or radiographic projection [1,4,5,6,7]. The 3D data may consist of CT, MRI, 3D ultrasound, anatomical models, segmented structures, or reconstructed surfaces [1–4,8]. Thus, 2D–3D fusion is not a single algorithmic task, but rather a broad class of image-processing problems involving projection geometry, feature correspondence, similarity estimation, spatial transformation, and uncertainty management.
This review focuses on the image processing methods that enable 2D–3D fusion. It covers approaches based on manual and semi-automatic correspondence selection, classical image processing, and methods based on machine learning. Fully automatic 2D–3D fusion remains challenging because of modality differences, tissue deformation, limited field of view, and the ambiguity caused by projecting 3D anatomy into 2D images. For this reason, manual or semi-automatic selection of corresponding anatomical points remains an important strategy in clinical practice. Although manual correspondence picking requires user interaction, it can be clinically feasible. In many workflows, expert-selected landmarks or clearly defined anatomical correspondences provide a robust initialisation for subsequent point tracking and reduce the risk of convergence to an incorrect solution.

The article is organised as follows. Section 2 introduces the main types of 2D and 3D medical imaging data used in fusion pipelines and discusses their image processing characteristics. Section 3 presents the mathematical foundations of 2D–3D registration. Section 4 introduces a methodological taxonomy of 2D–3D fusion approaches, beginning with manual correspondence selection, then discussing semi-automatic methods based on landmarks, sensor-assisted and markerless registration, classical methods based on intensity and finally frameworks based on machine learning. Section 5 addresses technical challenges, and future directions. The article concludes by discussing the balance between algorithmic automation, clinical reliability and deployability in surgical image guidance systems.

2. IMAGING MODALITIES AND DATA CHARACTERISTICS

2D–3D fusion methods operate on imaging data that differ not only in dimensionality, but also in resolution, contrast mechanism and clinical availability. These differences strongly influence the design of registration algorithms, the choice of image features, the type of transformation model, and the expected robustness of the fusion. A method that performs well for X-ray-to-CT registration, for example, cannot usually be transferred directly to ultrasound-to-MRI or endoscopy-to-CT fusion, because each modality encodes different anatomical information.

This section summarises the main characteristics of 2D and 3D medical imaging data relevant to 2D–3D fusion. Emphasis is placed on image processing properties rather than clinical applications.

2.1. 2D MEDICAL IMAGING DATA

Two-dimensional medical images are commonly acquired intraoperatively or during diagnostic procedures because they can often be obtained quickly, repeatedly, and using relatively simple equipment. In the context of 2D–3D fusion, they usually provide the current procedural view that must be aligned with a 3D representation.

2.1.1. X-RAY RADIOGRAPHY AND FLUOROSCOPY

X-ray radiography and fluoroscopy are imaging modalities based on projection. They represent the accumulated attenuation of X-rays along projection rays passing through the patient's body. This makes these techniques highly suitable for 2D–3D registration with computer tomography (CT) based on projection. The latter is a 3D reconstruction also based on X-ray attenuation [1,2].

Fluoroscopy is an X-ray imaging technique that produces real-time 2D images of internal anatomy. In medical procedures, it is a real-time view used to continuously visualise structures such as bones,

vessels, catheters, needles, and other surgical instruments. It is practical for intraoperative guidance, but exposes the patient to ionising radiation.

X-ray and fluoroscopic images are characterised by [1,2]:

- overlapping anatomical structures;
- high contrast for bone and metallic instruments;
- lower contrast for soft tissues;
- possible noise from low dose X-ray acquisition;
- geometric distortion depending on the imaging system;
- availability during intervention.

In fusion pipelines, X-ray images are often compared with digitally reconstructed radiographs generated from CT volumes. This makes X-ray-to-CT registration one of the most technically mature forms of 2D–3D fusion. However, the projection nature of the data creates inherent ambiguity, especially in single view registration, where different 3D poses may produce similar 2D appearances [1].

2.1.2. ULTRASOUND IMAGES

Two-dimensional ultrasound (US) is widely used because it is portable, real-time, safe, and relatively inexpensive [4,5]. It provides dynamic information about soft tissues and can be used during surgical or diagnostic procedures. However, it is also one of the most challenging modalities for 2D–3D fusion [1,4].

Ultrasound images are formed from reflected acoustic waves, and their appearance depends on acoustic impedance differences, probe orientation, contact pressure, angle, and tissue properties [4,5]. Unlike CT or MRI, ultrasound image intensities do not have a simple and stable anatomical meaning across acquisitions [4,5].

Typical characteristics include [4,5]:

- speckle noise;
- shadowing and acoustic enhancement;
- limited field of view;
- anisotropic resolution;
- variable image quality;
- deformation caused by pressure;
- weak or missing boundaries for some structures;
- real-time acquisition.

In 2D–3D fusion, ultrasound may be registered with CT, MRI, 3D ultrasound, or anatomical models. The task of registration is difficult, because the 2D ultrasound plane captures only a thin section of the anatomy of the patient's body. Ultrasound fusion often requires landmarks, segmentation, learned descriptors, and tracking [1,4].

2.1.3. ENDOSCOPIC AND OPTICAL IMAGES

An endoscopic image is a 2D optical image acquired using an endoscope – an instrument ending with a camera, inserted into the body through a natural opening or a small surgical incision. It shows the visible surface of internal anatomy, such as mucosa, organs, vessels, or surgical instruments.

Endoscopic and optical images provide direct visual information about exposed or visible anatomical surfaces. They are especially relevant in minimally invasive surgery, robotic surgery, and image-guided navigation based on camera views.

Unlike X-ray or ultrasound, optical images primarily encode surface appearance rather than internal anatomy. Their intensities depend on illumination, reflectance, camera properties, tissue texture, blood, smoke, specular reflections, and instrument occlusion.
Important characteristics include [1]:

- surface only visibility;
- strong dependence on lighting conditions;
- specular highlights;
- occlusion by tools or tissue;
- deformation of visible soft tissue;
- limited field of view;
- availability of video sequences;
- possible use for structure-from-motion or photogrammetry [8].

Endoscopic images can also be used for local 3D reconstruction when multiple views, stereo information, or shading cues are available. Cao et al. described stereo vision and shape from shading methods based on endoscope imaging, showing how optical image sequences may provide surface geometry useful for navigation and registration [9].

For 2D–3D fusion, optical images may be aligned with CT- or MRI-derived anatomical models, preoperative segmentation, or reconstructed 3D surfaces. Since internal structures are not directly

visible, optical-to-3D fusion often depends on surface geometry, anatomical landmarks, or indirect registration through reconstructed 3D models from camera data [1,8].

2.1.4. OTHER 2D IMAGING METHODS

Other methods of two-dimensional image acquisition that are performed in real time during the surgical or diagnostic procedure are mentioned in this article. These include more specialised techniques with a narrower range of use cases.

Digital subtraction angiography (DSA) is a 2D X-ray-based vascular imaging method. It visualises blood vessels by subtracting imaging without contrast from imaging with contrast. It suppresses background anatomy and emphasises vascular structures. 2D-3D registration often relies on vascular centrelines, bifurcations, and projected vessel geometry rather than raw intensity matching [6].

Echocardiography is an imaging method based on ultrasound used to visualise cardiac structures [4,10]. Histological imaging produces high resolution images of thin tissue sections after tissue preparation, staining, and scanning [7].

2.2. 3D IMAGING DATA

Three-dimensional medical imaging is **commonly** acquired preoperatively or intraoperatively to provide a spatial representation of anatomical structures. In the context of 2D–3D fusion, they usually serve as the reference data containing volumetric information about organs, vessels, bones, lesions, or other clinically relevant structures. Compared with 2D images, 3D data provide more complete anatomical context and allow the spatial localisation of structures that may be hidden, partially visible, or outside the field of view during the procedure. However, they are often not available in real time, may require longer acquisition or reconstruction, and may not fully reflect intraoperative anatomical changes caused by patient repositioning, tissue deformation, organ motion, or surgical manipulation.

2.2.1. COMPUTED TOMOGRAPHY

Computed tomography (CT) provides **volumetric** images based on X-ray attenuation. CT is widely used in 2D–3D fusion because it offers high spatial resolution, strong contrast for bone and calcified structures, and the possibility to generate synthetic X-ray projections [1,2].

CT volumes are characterised by [1,2]:

- isotropic spatial resolution in modern scanners;
- high contrast for bone, air, and metallic objects;

- moderate to limited contrast for soft tissues;
- suitability for segmentation and surface extraction.

CT is particularly useful for generating digitally reconstructed radiographs for X-ray-to-CT registration. It is also frequently used as a preoperative anatomical reference during surgery. However, CT acquired before surgery may not accurately represent soft tissue deformation during surgery.

2.2.2. MAGNETIC RESONANCE IMAGING

Magnetic Resonance Imaging (MRI) provides volumetric data with excellent soft tissue contrast. It is commonly used when soft tissue anatomy are the subject of the imaging, e.g. lesions, vessels, or nerves. In prostate interventions, for example, multiparametric MRI provides information used for lesion detection, localisation, and risk assessment, and can support targeted biopsy workflows [11]. Unlike CT, MRI intensities do not have a single standardised physical scale across scanners and sequences, which complicates direct intensity-based registration [1,3].

Relevant image-processing characteristics include [3]:

- high contrast for soft tissue;
- anisotropic resolution in some acquisitions;
- lower visibility of bone compared with CT [2,3];
- rich anatomical and functional information.

MRI is frequently used as the 3D reference modality for ultrasound or intraoperative image fusion. However, the difference between MRI contrast and intraoperative modalities such as ultrasound or endoscopy makes multimodal registration challenging. Methods often rely on structural descriptors, segmentation, landmarks, or learned representations rather than raw intensities alone [1,4].

2.2.3. OTHER 3D IMAGING METHODS

Other methods of three-dimensional image acquisition or reconstruction are also mentioned in this article. These include more specialised techniques that are used in narrower clinical, experimental, or computational contexts. In 2D–3D fusion, such methods usually provide volumetric, surface-based, functional, or model-based representations that complement standard CT, MRI, and 3D ultrasound data. Examples include cone-beam CT, micro-CT, PET, SPECT, 3D digitisation, photogrammetric reconstruction, and dynamic anatomical models [1,7,8,12,13].

3D ultrasound (3D US) is an ultrasound modality that provides volumetric acoustic data instead of a single 2D imaging plane. It may be acquired using a dedicated 3D probe, a mechanically swept probe, or reconstruction from tracked 2D ultrasound frames. 3D US preserves more spatial information than 2D US and can support volume-to-volume or slice-to-volume registration. However, it is still affected by ultrasound-specific limitations such as speckle, shadowing, limited field of view, and operator dependence [4,5].

Cone-beam computed tomography is a volumetric X-ray imaging technique that uses a cone-shaped X-ray beam and a flat panel detector to reconstruct 3D anatomy. Cone-beam CT is often used intraoperatively or in interventional suites [12]. Compared with diagnostic CT, it may have lower soft tissue contrast and more artifacts, but it provides useful 3D anatomical information during procedures. In 2D-3D fusion, cone-beam CT can serve as an intraoperative 3D reference or as an intermediate modality between preoperative CT and 2D fluoroscopy [1,12].

Micro-computed tomography (μCT) is a high resolution form of computed tomography used mainly for small specimens, ex vivo tissue samples, and preclinical imaging. It provides a much higher resolution than clinical computed tomography. In fusion with histology, μCT can be registered with 2D histological sections. The main challenges include soft tissue contrast limitations and deformation of the histology slide relative to the original specimen [7].

Three-dimensional echocardiography is an cardiac imaging technique based on ultrasound that provides volumetric images of the heart. 3D transesophageal echocardiography, or 3D TEE, is acquired using a probe placed in the esophagus and can provide detailed volumetric views of cardiac structures such as valves and the aortic root. In the literature, 3D TEE is discussed in relation to comparison with CT and cardiovascular multimodality registration [10, 14].

Positron emission tomography (PET) is a 3D functional imaging modality that visualises metabolic or functional activity using radioactive tracers. PET has lower spatial resolution than CT and MRI but provides functional information that anatomical imaging cannot directly show. In multimodal imaging, PET is often combined with CT or MRI, producing hybrid PET/CT or PET/MRI data [13].

Single-photon emission computed tomography (SPECT) is a 3D nuclear medicine imaging modality that visualises the distribution of gamma-emitting radiotracers. Like PET, SPECT provides functional rather than primarily anatomical information. It is commonly combined with CT in hybrid SPECT/CT imaging to relate tracer uptake patterns to anatomical structures [13].

3D vascular models are geometric representations of blood vessels reconstructed from volumetric imaging or angiographic **data**. They may consist of vessel surfaces, centerlines, bifurcations, or graph-like vascular trees. In 2D–3D fusion, 3D vascular models can be registered with 2D DSA images by matching projected 3D vessel structures to observed 2D vascular patterns [15].

3D surface scanning or 3D digitisation creates a three-dimensional representation of visible external or exposed anatomy. This can be obtained using photogrammetry, structured light, depth cameras, laser scanning, or other surface acquisition methods. In 2D–3D fusion, the resulting surface can be registered to CT/MRI-derived surfaces or used for patient-to-image registration [8].

Photogrammetry is a technique for reconstructing 3D geometry from multiple 2D camera images. Structure from motion is a related computational method that estimates both camera motion and 3D structure from image sequences. In medical imaging fusion, photogrammetry may provide a 3D surface representation that supports registration with preoperative CT or MRI [8].

3D anatomical models are computational representations of anatomical structures, often derived from CT, MRI, ultrasound, segmentation, or statistical modelling. They may represent structures such as organs, bones, vessels or tumours. In 2D–3D fusion, they can serve as simplified geometric references instead of full image volumes.

2.3. SUMMARY OF IMAGING MODALITIES

The diversity of imaging modalities creates both the value and the difficulty of 2D–3D fusion. Two-dimensional images can be acquired in real time but are limited by restricted field of view, lower accuracy and artifacts specific to their modality [1,4,5]. Three-dimensional imaging provides more comprehensive information but cannot be obtained in real time. Consequently, successful 2D–3D fusion requires careful consideration of data characteristics before algorithm design.

3. METHODOLOGICAL TAXONOMY OF 2D-3D FUSION APPROACHES

In 2D–3D medical image fusion, the central computational problem is to estimate the spatial relationship between a two-dimensional procedural image and a three-dimensional representation of the same or related anatomy. The 2D data may come from ultrasound, fluoroscopy, radiography, endoscopy, or angiography, while the 3D data may be derived from CT, MRI, 3D ultrasound, segmented structures, surface models, or other anatomical reconstructions. Registration provides the transformation or geometric correspondence needed to relate structures visible in the 2D image to structures represented

in the 3D data, allowing information from the 3D dataset to support interpretation of the current procedural view.

Several methodological strategies have been proposed to solve this problem. Conventional fusion systems frequently rely on external tracking of the imaging device, especially in fusion based on ultrasound, where the position and orientation of the probe are tracked relative to the patient and to the preoperative CT or MRI volume [16,17]. However, such systems may require optical or electromagnetic tracking, fiducial markers, calibration procedures, and additional hardware, which can increase workflow complexity and limit usability in clinical practice [16,17].

For this reason, markerless methods based on features and similarity have been investigated. These approaches aim to reduce dependence on external markers or tracking systems by using landmarks, local structural descriptors, self-similarity features, or volumetric image patches [18,19]. Nevertheless, many of these methods remain limited by the need for expert initialisation, or carefully validated multimodal datasets [16,18–21]. The following taxonomy groups the main 2D–3D fusion approaches according to the source of spatial information used for registration: manual correspondences, semi-automatic features, external tracking, markerless surface registration, similarity based on intensity, projection geometry, deformable modelling, and learning-based estimation.

3.1. FUSION BASED ON MANUAL CORRESPONDENCE

Fusion based on manual correspondence relies on the operator to select anatomically corresponding points or structures in the 2D and 3D data. These correspondences may include vessel bifurcations, bone landmarks, organ boundaries, vertebral centroids, skin points, lesion margins, or other clearly identifiable anatomical features.

From a practical perspective, this remains one of the most feasible solutions in clinical practice. Although it requires user interaction, it can be performed quickly when the landmarks are well defined. It also benefits from the surgeon’s or radiologist’s anatomical expertise, which leads to more reliable correspondence selection than fully automatic feature detection.

An example of this approach is spine registration based on features that relies on manually identified vertebral centroids or vertebral body corners in preoperative MR/CT and intraoperative X-ray images [22]. Similarly, CT–ultrasound registration approaches based on landmarks can use anatomical points defined by the user as registration constraints or initialisation [23].

The main advantages of manual fusion based on correspondence are real-time or near-real-time applicability, interpretability of selected landmarks, robustness when anatomical landmarks are clear, suitability for initialisation of automatic registration. The main limitations are operator dependence, limited reproducibility, and reduced autonomy. Therefore, manual correspondence selection is often best understood not as a competing alternative to automatic methods, but as a clinically robust initialisation or correction mechanism.

3.2. SEMI-AUTOMATIC REGISTRATION BASED ON LANDMARKS AND FEATURES

Semi-automatic methods reduce manual workload by combining features selected by the user and detected through image processing. The user may define a small number of initial points, contours, or regions, after which the algorithm estimates the final transformation.

Typical detected features include:

- anatomical landmarks;
- contours and boundaries;
- vessel centrelines;
- bifurcation points;
- bone corners;
- organ surfaces;
- segmented structures.

This category is especially useful when direct intensity comparison is unreliable because of multimodal image contrast, as in ultrasound–MRI, ultrasound–CT, or endoscopy–CT fusion. Feature-based registration can be more possible to interpret than dense intensity based methods because the transformation is estimated from explicit anatomical structures.

Several studies follow this logic. Yang et al. proposed a local structure orientation descriptor for multimodal liver ultrasound–MRI registration, using structural information rather than direct intensity correspondence [19]. Wang et al. used a weighted self-similarity structure vector for ultrasound–MRI registration, also addressing the difficulty of multimodal appearance differences [18]. In vascular imaging, dual view DSA-to-3D vascular model registration uses projected vascular structures and point correspondences as the basis for alignment [24].

However, the performance of methods based on features depends strongly on the quality and repeatability of feature extraction. If the selected or detected structures are ambiguous, partially visible,

deformed, or inconsistently represented across modalities, registration may be unstable. In addition, many semi-automatic methods still require expert initialisation, which limits full automation [19,22,23].

3.3. FUSION ASSISTED BY SENSOR AND PROBE TRACKING

Sensor-assisted fusion uses external tracking hardware to estimate the position and orientation of the 2D imaging device relative to the patient or reference coordinate system. This is common in ultrasound–CT/MRI fusion and image-guided interventions.

Common tracking technologies include:

- electromagnetic tracking;
- optical tracking;
- tracked ultrasound probes;
- tracked surgical instruments;
- fiducial markers – artificial reference points placed on or inside the patient, instrument, or imaging setup so that the same points can be identified in different images or coordinate measurements;
- calibration phantoms – physical reference models with precisely known geometry and material properties, designed to be imaged by a medical imaging system in order to measure, correct, or validate its spatial and imaging characteristics, used to determine parameters such as image scale, geometric distortion, resolution, probe position, or the transformation between image coordinates and external tracking coordinates.

The main advantage of fusion assisted by sensors is that it provides explicit information about the imaging device, thereby reducing the registration complexity. In some workflows, once the probe or instrument is calibrated and tracked, the 2D image plane can be placed directly in the 3D coordinate system [4].

This approach is widely represented in procedures guided by multimodal image fusion, where tracking systems and fiducial markers are often used to connect intraoperative images with preoperative CT or MRI data [4]. Paccini et al. also describe the practical limitations of US–MR/CT fusion systems based on tracking, including line-of-sight constraints for optical tracking and metal sensitivity for electromagnetic tracking [16].

MRI/US fusion biopsy platforms in prostate cancer provide another clinical example of this workflow, where preoperative MRI information is combined with ultrasound guidance during the procedure [25].

The limitations are mainly practical. Optical tracking requires line of sight. Electromagnetic tracking is sensitive to metallic objects and field distortions. Both approaches add hardware, calibration steps, cables, and workflow complexity. For this reason, methods assisted by sensors may be accurate, but they are not always ideal for clinical deployment [16,17].

3.4. MARKERLESS REGISTRATION BASED ON SURFACE

Markerless registration based on surface aims to avoid fiducial markers and external tracking by using a 3D scan of the surface of the skin or internal organs. In such methods, the patient's skin surface or visible anatomical surface is extracted and aligned with a surface derived from CT, MRI, or reconstruction done using a camera [16,26].

This approach is attractive because it reduces the need for artificial markers. It may also simplify workflow when a surface scan can be acquired quickly using optical cameras, depth sensors, or photogrammetry.

Paccini et al. proposed a proof-of-concept approach for US–MRI/CT fusion based on markerless skin registration, directly addressing the problem of simplifying fusion without external markers [16]. Photogrammetry and 3D digitisation can also support this class of methods by reconstructing external surface geometry that may be registered with volumetric medical data [26].

Its main limitation is that external surface alignment does not always guarantee accurate alignment of internal anatomy. Soft tissue deformation, breathing, patient repositioning, and differences between external and internal structures can reduce accuracy. Therefore, registration based on surface is often useful for initialisation, but may require additional internal landmarks [16,26].

3.5. REGISTRATION BASED ON INTENSITY

Registration based on intensity estimates alignment by directly comparing image intensities between the 2D image and a slice or projection generated from the 3D data. Instead of relying on explicit landmarks, the algorithm searches for the transformation that maximises an image similarity measure.

This approach is especially effective when the compared images have similar contrast mechanisms, as in X-ray-to-CT registration with digitally reconstructed radiographs. In cardiac imaging, 2D-echocardiography-to-CT registration has been performed iterative optimisation [27]. In histology-to-CT or histology-to-μCT fusion, intensity based methods may also be used as baselines, although deformation and modality differences often limit their accuracy [28].

Registration based on intensity becomes difficult in strongly multimodal cases, such as ultrasound–MRI or endoscopy–CT fusion. In these cases, corresponding anatomical structures may have very different intensity patterns. The method is also often sensitive to initialisation, artifacts, occlusion, and local optima. As a result, registration based on intensity is frequently combined with optimisation in different scales, manual initialisation, segmentation, or learnt descriptors [19,27,28].

3.6. REGISTRATION BASED ON PROJECTION

Methods based on projection are used when the 2D image represents a projection of the 3D anatomy, as in X-ray, fluoroscopy, or digital subtraction angiography. The main idea is to generate synthetic 2D projections from the 3D data and compare them with the acquired image, iteratively searching for the projection with minimal error.

Registration based on projection is geometrically well defined and widely used in X-ray-to-CT fusion. Gopalakrishnan et al. proposed intraoperative 2D/3D registration using differentiable X-ray rendering, which makes the rendering operation compatible with gradient-based optimisation [29]. Chen et al. proposed a fully differentiable correlation-driven 2D/3D registration approach for X-ray-to-CT fusion, combining differentiable registration with learned image representations [30]. Jaganathan et al. used synthetic X-ray projections generated from CT volumes in a self-supervised X-ray-to-CT registration framework [31].

The main limitation of this approach is its computational cost, especially when many projections must be generated during iterative optimisation. Single view registration based on generating projections is also affected by depth ambiguity, because different 3D poses may produce similar 2D projections. Multi view methods, such as dual view DSA-to-3D vascular model registration, can reduce this ambiguity, but they increase computational complexity [24].

3.7. FUSION BASED ON MODEL

2D–3D fusion based on model uses an explicit anatomical, geometric, or physical model to constrain the registration process. Instead of aligning only raw image intensities or independently detected landmarks, these methods incorporate prior knowledge about the expected shape, motion, or deformation of the target structure. For example, Luo et al. proposed registration of intraoperative 2D ultrasound with a dynamic 3D aortic model for transcatheter aortic valve implantation, where the model represented the expected geometry and motion of the aortic anatomy during the procedure [32]. In such approaches, the

model acts as an intermediate representation between sparse or noisy 2D observations and the more complete 3D anatomical model.

The main strength of fusion based on model is that it can improve robustness when the 2D image contains limited, noisy, or partially visible information. This is particularly relevant for ultrasound, endoscopy, and fluoroscopy, where the field of view is restricted and only selected anatomical structures may be visible. It can also support tracking over time, because temporal changes can be interpreted relative to a deformable or dynamic anatomical representation rather than as independent frame-by-frame registrations [32].

The main weakness is that the result depends strongly on the validity of the model. If the model does not accurately represent patient specific anatomy, intraoperative deformation, respiratory or cardiac motion, or pathological variation, it may bias the registration toward an incorrect solution. Model construction may also require segmentation, parameter tuning, biomechanical assumptions, or imaging data, which increases workflow complexity. Therefore, fusion based on model is most useful when the anatomy to be observed is sufficiently well defined and when the model assumptions are compatible with the expected intraoperative changes.

3.8. DEFORMABLE FUSION

Deformable fusion is required when the anatomy changes shape between 2D and 3D acquisitions. This is common in soft tissue surgery, abdominal imaging, cardiac procedures, interventions guided by ultrasound, and histology-to-volume registration.

In histology-to-volume fusion, Chen et al. proposed a deformable 2D–3D registration method that estimates the corresponding plane in the 3D volume and then refines deformation of out the plane [28].

The main advantage of deformable fusion is improved anatomical realism. The main limitation is increased complexity. Deformable models introduce many degrees of freedom, require stronger regularisation, and are more difficult to validate than rigid transformations. Recent deformable registration based on deep learning methods attempt to address this complexity, but they introduce additional requirements related to training data and generalisation [33,34]. Hering et al. further showed that label driven deep deformable registration can be improved by local distance metrics in cardiac motion tracking, illustrating the role of task specific constraints in deformable registration [35]. Chen et al. proposed a deep discontinuity preserving registration network, addressing the problem that deformation fields should not always be spatially smooth across anatomical boundaries [36].

3.9. FUSION BASED ON MACHINE LEARNING

Methods based on machine learning, especially deep neural networks, estimate image features, correspondences, similarity measures, transformation parameters, or deformation fields. Their main motivation is to reduce dependence on handcrafted similarity metrics and iterative search while improving runtime during deployment. In 2D–3D fusion, they are particularly attractive for multimodal registration, where direct intensity comparison is often unreliable [30,37,38].

The methods can be divided into the following groups:

- deep feature extraction followed by geometric estimation;
- end-to-end transformation regression methods;
- differentiable registration based on features;
- dual-view and multi-view learning-based registration;
- self-supervised and weakly supervised learning methods;
- machine-learning trends and systematic perspectives.

Methods based on machine learning, especially deep neural networks, estimate image features, correspondences, similarity measures, transformation parameters, or deformation fields. Their main motivation is to reduce dependence on handcrafted similarity metrics and iterative search while improving runtime during deployment. In 2D–3D fusion, they are particularly attractive for multimodal registration, where direct intensity comparison is often unreliable [30,37,38].

The first group of learning-based methods performs deep feature extraction followed by geometric estimation. In this strategy, neural networks extract features from the 2D image and 3D volume, after which correspondences are matched and the transformation is estimated using a geometric procedure such as RANSAC or iterative refinement. This preserves some interpretability, but it can be computationally inefficient if many candidate slices, projections, or keypoint correspondences must be evaluated [37].

The second group consists of end-to-end transformation regression methods. These networks directly estimate rigid, affine, or deformable transformation parameters from input 2D and 3D data. Endoscopic ultrasound registration (EUReg), is an example of this direction for efficient 2D–3D ultrasound registration, using neural feature extraction and transformation estimation [37]. The advantage is computational efficiency after training, whereas the main limitation is possible overfitting to the training set.

The third group focuses on differentiable registration based on features, in which learned image representations are combined with projection. This approach is especially applicable to X-ray-to-CT fusion because CT volumes can be used to generate synthetic X-ray projections that are compared with real X-ray images. Chen et al. proposed a fully differentiable method for X-ray-to-CT registration, using dual branch neural network containing convolutional layers and transformer architecture [30]. Rather than treating registration as a single end-to-end neural prediction problem, such methods integrate neural representation learning with explicit geometric modelling and optimisation.

The fourth direction is multi view registration based on learning. Multi view methods reduce the ambiguity of single view 2D–3D registration by using more than one 2D projection. Zhang et al. proposed dual view registration of 2D digital subtraction angiography (DSA) images with 3D vascular models, combining deep learning with 3D model to projection matching [29]. These methods can improve robustness and depth estimation, but they may increase acquisition complexity and computational cost.

The fifth group includes self-supervised and weakly supervised learning methods. Jaganathan et al. proposed self-supervised X-ray-to-CT registration using simulated X-ray projections generated from CT volumes and domain adaptation to reduce the gap between synthetic and real images [31]. This strategy reduces the dependence on annotated paired datasets, but it still requires careful handling of synthetic to real shift. Mahapatra et al. proposed a generative adverserial network (GAN) based method for elastic medical image registration, where the network estimates deformable alignment between medical images using adversarial learning [39]. Lu et al. used a cycle adversarial network for CT to transesophageal echocardiography (TEE) registration, where the network learns cross modal image representations between CT and TEE data for surgical navigation in congenital heart disease [14].

Finally, broader machine learning trends and systematic perspectives indicate that fusion based on machine learning should not be treated as a complete replacement for geometric modeling. Unberath et al. emphasise that machine learning can address persistent problems such as initialisation sensitivity, optimisation brittleness, and single view limitations, but these methods still benefit from explicit imaging geometry, anatomical constraints, and rigorous validation [38]. Therefore, hybrid geometric–learning approaches are likely to be the most robust models in clinical scenarios.

3.10. SUMMARY OF THE TAXONOMY

The reviewed approaches can be organised according to the the source of correspondence. Manual and semi-automatic methods rely on explicit anatomical correspondences [22,23]. Methods assisted by sensors rely on external spatial tracking [17]. Methods based on surface use external or visible anatomy [16,26]. Methods based on intensity and projection optimise image similarity [18,26,27,32]. Deformable methods model anatomical change [28,32–34]. Methods based on machine learning infer correspondences or transformations from data [30,31,37,38].

No single approach is optimal in all scenarios. Manual methods and approaches based on landmarks remain the most usable in clinical practice because they are the most reliable so far. Methods based on projection are strong when structures in the medical imaging are well defined. Deformable methods are necessary for soft tissues. Methods based on machine learning offer speed and flexibility but depend on data quality and might be prone to overfitting the training data. In practice, robust 2D–3D fusion often requires a hybrid strategy that combines anatomical expertise, geometric constraints, image-processing methods, and, where appropriate, learnt representations.

4. COMPUTATIONAL ASPECTS AND OPTIMISATION STRATEGIES

Computational efficiency is a central requirement in 2D–3D fusion, particularly when the method is intended for intraoperative or interventional use. Registration methods must not only be accurate, but also sufficiently fast, stable, and predictable under realistic imaging conditions. The computational burden depends on the dimensionality of the search space, the cost of generating 2D projections or slices from 3D data, the complexity of the similarity measure, and the number of iterations required for convergence.

4.1. ITERATIVE OPTIMISATION

Many classical 2D–3D fusion methods are formulated as iterative optimisation problems. The algorithm repeatedly transforms the 3D data, generates a corresponding 2D projection or slice, evaluates similarity to the acquired 2D image, and updates the transformation parameters. This strategy is used in registration based on intensity, X-ray-to-CT fusion based on projection, histology-to-volume registration, and several multimodal ultrasound registration methods [19,27–29].

Iterative optimisation is flexible and can be combined with different similarity metrics, transformation models, and modality specific preprocessing steps. However, it is often sensitive to initialisation and

may converge to local optima. In addition, repeated projection generation, interpolation, and similarity computation can be computationally expensive, especially for high-resolution 3D volumes [24,29].

4.2. INITIALISATION PROBLEM

Initialisation remains one of the most important practical limitations of 2D–3D fusion. Because the search space is often high dimensional and the 2D data contain incomplete spatial information, poor initialisation can lead to incorrect convergence and hence registration failure. This problem is especially relevant in single view registration, multimodal registration, and soft tissue fusion [17,26,37].

Several reviewed methods still depend on expert initialisation or manual landmark selection. For example, ultrasound–MRI registration based on local structure may require an initial coarse alignment by an expert [19]. Echocardiography-to-CT registration workflows have also used rigid initialisation by an expert before refinement based on intensity [27]. Manual or semi-automatic correspondence selection therefore remains clinically relevant because it reduces the search space and improves the reliability of subsequent optimisation [22,23].

4.3. RUNTIME AND REAL-TIME CONSTRAINTS

Real-time or near-real-time performance is required when fusion is used to support surgical navigation, probe guidance, or intraoperative decision making. Manual landmark based methods can be computationally efficient because the registration is estimated from a small number of correspondences. In contrast, dense intensity based, projection based, and deformable methods may require substantial computation [24,28,29].

Methodsbased on machine learning address this limitation by shifting most of the computational cost to the training phase and enabling faster inference of correspondences during deployment. Endoscopic ultrasound registration (EUReg), for example, was designed for efficient 2D–3D ultrasound registration using an end-to-end architecture suitable for high speed inference [30]. However, real-time performance must be evaluated together with accuracy, robustness, memory consumption, and generalisation in a large sample of patients patients.

5. TECHNICAL CHALLENGES AND OPEN PROBLEMS

Despite substantial progress, reliable 2D–3D fusion remains to be an open problem. The main challenges arise from incomplete spatial information in 2D data, differences between imaging

modalities, deformation of the anatomy, limited field of view, and the need for robust performance under intraoperative constraints.

5.1. ABSCENCE OF RELIABLE FULLY AUTOMATIC MARKERLESS REGISTRATION

A major difficulty is the lack of broadly reliable fully automatic markerless registration. Conventional fusion systems frequently depend on electromagnetic or optical tracking, fiducial markers, calibration procedures, or manual landmark selection [16,17]. Markerless skin or surface registration can reduce this burden, but external surface alignment does not necessarily guarantee accurate alignment of internal anatomy [16,26].

This limitation is particularly important in fusion based on ultrasound. Some markerless methods require 3D ultrasound data, whereas many clinical workflows still rely primarily on 2D ultrasound [16]. Other based methods may require expert initialisation, which limits full automation and broad usability [19].

5.2. DIFFERENT MODALITIES

Different imaging modalities encode different physical properties. CT represents X-ray attenuation, MRI reflects magnetic resonance properties, ultrasound depends on acoustic reflection and scattering, X-ray and fluoroscopy are projection-based attenuation images, and histology shows stained microscopic tissue structure seen under optical light. Consequently, the same anatomical structure may appear very differently across modalities or may be clearly visible in one modality and poorly visible in another.

This mismatch limits the reliability of direct similarity measures based on intensity. Representations based on features, segmentation, self-similarity and machine learning attempt to reduce this problem by comparing more modality invariant information [18,19,30,31,37]. Nevertheless, robust multimodal correspondence estimation remains difficult, especially when images contain noise, artifacts, missing structures, or deformation.

5.3. DEPENDENCE ON EXPERT INITIALISATION AND MANUAL LANDMARKS SELECTION

Although manual landmark selection is clinically practical in many settings, it also introduces dependence on the operator. Many methods still require initialisation by an expert, manually selected correspondences, or semi-automatic landmark definition [19,22,23,27].

This dependence is not necessarily a weakness in all clinical contexts. Landmarks selected by an expert may be more reliable than automatic detections in poor quality or ambiguous images. However, from a methodological perspective, the need for manual interaction reduces reproducibility. A key challenge is

therefore to develop methods that preserve the reliability of expert guided registration while reducing the amount of required user input.

5.4. COMPUTATIONAL COST OF ITERATIVE AND MULTI-VIEW METHODS

Many accurate 2D–3D fusion methods rely on iterative optimisation, repeated projection generation, or similarity evaluation. These operations can be expensive, particularly when the registration includes deformable transformation models or multi view correspondence search [24,28,29].

Multi view methods can reduce depth ambiguity and improve robustness, as in dual view DSA-to-3D vascular model registration [28]. However, they may require additional image acquisition, synchronisation, and computation. This creates a trade-off between accuracy, robustness, radiation exposure where applicable, and runtime.

5.5. GENERALISATION OF DEEP LEARNING METHODS

Deep learning methods can improve inference speed and learn complex cross-modal representations, but their performance depends strongly on training data. End-to-end methods may overfit to specific datasets, anatomical regions, scanners, or acquisition protocols [37,38]. In X-ray-to-CT registration, synthetic projections generated from CT volumes can reduce the need for annotated paired data, but the artificial views often differ from real clinical X-ray images in terms of noise, contrast, artifacts, etc. [31].

Self-supervised learning, rendering synthetic projections, and hybrid geometric–learning models are promising strategies for improving generalisation [30,31]. However, these methods still require rigorous validation on independent clinical datasets.

5.6. VALIDATION AND BENCHMARK LIMITATIONS

Reliable validation is difficult because ground truth 2D–3D registration is rarely available in clinical data. Studies may use fiducials, expert-defined landmarks, synthetic data, phantoms, tracking systems, or manually refined registrations as reference standards, but each of these introduces its own limitations.

Public datasets such as RESECT and multimodal 3D-ultrasound-CT datasets are valuable for benchmarking, but they cover only selected anatomical regions and imaging scenarios [20,21]. The field still lacks broad standardised benchmarks that allow systematic comparison across modalities, transformation models, levels of deformation, and clinical use cases.

6. DISCUSSION

The reviewed literature shows that 2D–3D fusion is best understood as a set of related image processing and registration problems rather than a single unified task. The appropriate method depends on the modality pair, anatomical region, expected deformation, clinical workflow, and required runtime.

Classical methods remain important because they are geometrically interpretable and can work with limited training data. Manual and semi-automatic methods based on landmarks are especially relevant in clinical practice because they can be fast, transparent, and robust when reliable anatomical correspondences are visible [22,23]. However, they require user interaction and are affected by operator variability.

Methods based on intensity and projection provide a more automatic formulation, particularly for X-ray-to-CT registration, where digitally reconstructed projections can be generated from the CT [29–31]. Their performance is limited by initialisation sensitivity, computational cost, and ambiguity in single view settings. Deformable approaches are necessary for soft tissue, cardiac anatomy, and histology-to-volume fusion, but they increase model complexity and make validation more difficult [28,32].

Methods based on machine learning address some limitations of classical registration by improving inference speed and learning cross-modal representations [30,31,37]. Nevertheless, they introduce new challenges related to training data, generalisation, and interpretability. For this reason, hybrid approaches that combine explicit geometry, anatomical constraints, projection models, and learnt features may represent the most practical direction for future development [30,31,38].

A recurring conclusion is that clinical feasibility depends not only on registration accuracy, but also on workflow simplicity, robustness, runtime, hardware requirements, and the amount of required manual input. Methods that are theoretically advanced but require complex calibration, extensive computation, or unstable initialisation may be less useful than simpler approaches that provide reliable results under real procedural constraints.

7. CONCLUSIONS

2D–3D fusion in medical imaging supports surgical and interventional procedures by integrating current 2D procedural views with richer 3D anatomical information. From an image processing perspective, this task involves coordinate transformation, correspondence estimation, projection or slicing models, similarity evaluation, and optimisation.

The reviewed methods can be grouped into 2D–3D registration based on manual correspondence, semi-automatic, deformable, assisted by sensors, markerless based on surface, based on features, intensity, projection, machine learning. Manual correspondence selection and methods using landmarks remain clinically relevant because of their interpretability and real-time feasibility. Methods based on projection are particularly important for X-ray-to-CT registration. Deformable methods are essential for soft tissue and histology related fusion, while approaches based on machine learning offer speed and flexibility but require careful validation [28-31,37,38].

The main unresolved challenges include markerless automation, modality mismatch, expert dependence, computational cost, robust generalisation of deep learning, and limited benchmarking. Future progress will likely depend on hybrid geometric–learning pipelines, and robust validation datasets.

REFERENCES

[1] P. Markelj, D. Tomaževič, B. Likar, and F. Pernuš, "A review of 3D/2D registration methods for image-guided interventions," ***Medical Image Analysis***, vol. 16, no. 3, pp. 642–661, Apr. 2012, doi: 10.1016/j.media.2010.03.005.

[2] T. D. DenOtter and J. Schubert, "Hounsfield unit," in ***StatPearls***. Treasure Island, FL, USA: StatPearls Publishing, 2023. Accessed: May 7, 2026. [Online]. Available: NCBI Bookshelf.

[3] B. Belaroussi, J. Milles, S. Carme, Y. M. Zhu, and H. Benoit-Cattin, "Intensity non-uniformity correction in MRI: Existing methods and their validation," ***Medical Image Analysis***, vol. 10, no. 2, pp. 234–246, Apr. 2006, doi: 10.1016/j.media.2005.09.004.

[4] Q. Huang and Z. Zeng, "A review on real-time 3D ultrasound imaging technology," ***BioMed Research International***, vol. 2017, Art. no. 6027029, 2017, doi: 10.1155/2017/6027029.

[5] R. N. Bakhru and W. D. Schweickert, "Intensive care ultrasound: I. Physics, equipment, and image quality," ***Annals of the American Thoracic Society***, vol. 10, no. 5, pp. 540–548, Oct. 2013, doi: 10.1513/AnnalsATS.201306-191OT.

[6] W. R. Brody, "Digital subtraction angiography," ***IEEE Transactions on Nuclear Science***, vol. 29, no. 3, pp. 1176–1180, Jun. 1982, doi: 10.1109/TNS.1982.4336336.

[7] P. Nolte et al., "Current approaches for image fusion of histological data with computed tomography and magnetic resonance imaging," ***Radiology Research and Practice***, vol. 2022, Art. no. 6765895, 2022, doi: 10.1155/2022/6765895.

[8] E. Nocerino, E. K. Stathopoulou, S. Rigon, and F. Remondino, “Surface reconstruction assessment in photogrammetric applications,” ***Sensors***, vol. 20, no. 20, Art. no. 5863, Oct. 2020, doi: 10.3390/s20205863.

[9] Z. Cao, Y. Wang, W. Zheng, L. Yin, Y. Tang, W. Miao, S. Liu, and B. Yang, “The algorithm of stereo vision and shape from shading based on endoscope imaging,” ***Biomedical Signal Processing and Control***, vol. 76, Art. no. 103658, Jul. 2022, doi: 10.1016/j.bspc.2022.103658.

[10] A. C. T. Ng et al., “Comparison of aortic root dimensions and geometries before and after transcatheter aortic valve implantation by 2- and 3-dimensional transesophageal echocardiography and multislice computed tomography,” ***Circulation: Cardiovascular Imaging***, vol. 3, no. 1, pp. 94–102, Jan. 2010, doi: 10.1161/CIRCIMAGING.109.885152.

[11] M. A. Bjurlin, N. Mendhiratta, J. S. Wysock, and S. S. Taneja, “Multiparametric MRI and targeted prostate biopsy: Improvements in cancer detection, localization, and risk assessment,” ***Central European Journal of Urology***, vol. 69, no. 1, pp. 9–18, 2016, doi: 10.5173/ceju.2016.749.

[12] R. Fahrig, D. A. Jaffray, I. Sechopoulos, and J. Webster Stayman, “Flat-panel conebeam CT in the clinic: History and current state,” ***Journal of Medical Imaging***, vol. 8, no. 5, Art. no. 052115, Oct. 2021, doi: 10.1117/1.JMI.8.5.052115.

[13] T. Beyer, L. S. Freudenberg, D. W. Townsend, and J. Czernin, “The future of hybrid imaging—Part 1: Hybrid imaging technologies and SPECT/CT,” ***Insights into Imaging***, vol. 2, no. 2, pp. 161–169, Apr. 2011, doi: 10.1007/s13244-010-0063-2.

[14] Y. Lu, B. Li, N. Liu, J. W. Chen, L. Xiao, S. Gou, L. Chen, M. Huang, and J. Zhuang, “CT-TEE image registration for surgical navigation of congenital heart disease based on a cycle adversarial network,” ***Computational and Mathematical Methods in Medicine***, vol. 2020, Art. no. 4942121, 2020, doi: 10.1155/2020/4942121.

[15] J. Kim, J. Lee, J. W. Chung, and Y. G. Shin, “Locally adaptive 2D–3D registration using vascular structure model for liver catheterization,” ***Computers in Biology and Medicine***, vol. 70, pp. 119–130, Mar. 2016, doi: 10.1016/j.compbiomed.2016.01.009.

[16] M. Paccini, G. Paschina, S. De Beni, A. Stefanov, V. Kolev, and G. Patanè, “US & MR/CT image fusion with markerless skin registration: A proof of concept,” ***Journal of Imaging Informatics in Medicine***, vol. 38, no. 1, pp. 615–628, Feb. 2025, doi: 10.1007/s10278-024-01176-w.

[17] N. Abi-Jaoudeh, J. Kruecker, S. Kadoury, H. Kobeiter, A. M. Venkatesan, E. Levy, and B. J. Wood, “Multimodality image fusion-guided procedures: Technique, accuracy, and applications,” ***Cardiovascular and Interventional Radiology***, vol. 35, no. 5, pp. 986–998, Oct. 2012, doi: 10.1007/s00270-012-0446-5.

[18] Y. Wang, T. Fu, C. Wu, J. Xiao, J. Fan, H. Song, P. Liang, and J. Yang, “Multimodal registration of ultrasound and MR images using weighted self-similarity structure vector,” ***Computers in Biology and Medicine***, vol. 155, Art. no. 106661, Mar. 2023, doi: 10.1016/j.compbiomed.2023.106661.

[19] M. Yang, H. Ding, J. Kang, L. Cong, L. Zhu, and G. Wang, “Local structure orientation descriptor based on intra-image similarity for multimodal registration of liver ultrasound and MR images,” ***Computers in Biology and Medicine***, vol. 76, pp. 69-79, Sep. 2016, doi: 10.1016/j.compbiomed.2016.06.025.

[20] Y. Xiao, M. Fortin, G. Unsgård, H. Rivaz, and I. Reinertsen, “REtroSpective evaluation of cerebral tumors (RESECT): A clinical database of pre-operative MRI and intra-operative ultrasound in low-grade glioma surgeries,” ***Medical Physics***, vol. 44, no. 7, pp. 3875–3882, Jul. 2017, doi: 10.1002/mp.12268.

[21] N. Masoumi, C. J. Belasso, M. O. Ahmad, H. Benali, Y. Xiao, and H. Rivaz, “Multimodal 3D ultrasound and CT in image-guided spinal surgery: Public database and new registration algorithms,” ***International Journal of Computer Assisted Radiology and Surgery***, vol. 16, no. 4, pp. 555–565, Apr. 2021, doi: 10.1007/s11548-021-02348-5.

[22] U. Kiran, H. Anitha, S. N. Bhat, and R. R. Naik, “Feature-based multimodal 3D/2D and 3D/3D registration framework for pedicle screw registration and evaluation,” ***European Spine Journal***, vol. 35, pp. 1363–1376, Mar. 2026, doi: 10.1007/s00586-025-09342-6.

[23] S. Garzia, K. Capellini, E. Gasparotti, D. Pizzuto, G. Spinelli, S. Berti, V. Positano, and S. Celi, “Three-dimensional multi-modality registration for orthopaedics and cardiovascular settings: State-of-the-art and clinical applications,” ***Sensors***, vol. 24, no. 4, Art. no. 1072, Feb. 2024, doi: 10.3390/s24041072.

[24] C. Zhang, J. Liu, L. Bian, S. Xiang, J. Liu, and W. Guan, “FMB: Dual-view fusion and registration of 2D DSA images and 3D MRA images for neurointerventional-based procedures,” ***Computers in Biology and Medicine***, vol. 171, Art. no. 107987, Mar. 2024, doi: 10.1016/j.compbiomed.2024.107987.

[25] M. Gayet, A. van der Aa, H. P. Beerlage, B. P. Schrier, P. F. Mulders, and H. Wijkstra, “The value of magnetic resonance imaging and ultrasonography (MRI/US)-fusion biopsy platforms in prostate cancer detection: A systematic review,” ***BJU International***, vol. 117, no. 3, pp. 392–400, Mar. 2016, doi: 10.1111/bju.13247.

[26] R. Ramm, P. de Dios Cruz, S. Heist, P. Kühmstedt, and G. Notni, “Fusion of multimodal imaging and 3D digitization using photogrammetry,” ***Sensors***, vol. 24, no. 7, Art. no. 2290, Apr. 2024, doi: 10.3390/s24072290.

[27] A. Khalil, A. Faisal, S. C. Ng, Y. M. Liew, and K. W. Lai, “Multimodality registration of two-dimensional echocardiography and cardiac CT for mitral valve diagnosis and surgical planning,” ***Journal of Medical Imaging***, vol. 4, no. 3, Art. no. 037001, Jul. 2017, doi: 10.1117/1.JMI.4.3.037001.

[28] J. Chen, M. Ronchetti, V. Stehl, V. Nguyen, M. Al Kallaa, M. T. Gedara, C. Lölkes, S. Moser, M. Seidl, and M. Wieczorek, “2D–3D deformable image registration of histology slide and micro-CT with DISA-based initialization,” ***Scientific Reports***, vol. 15, Art. no. 25972, Jul. 2025, doi: 10.1038/s41598-025-11583-w.

[29] V. Gopalakrishnan, N. Dey, and P. Golland, “Intraoperative 2D/3D image registration via differentiable X-ray rendering,” in ***Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, 2024, pp. 11662–11672, doi: 10.1109/CVPR52733.2024.01108.

[30] M. Chen, Z. Zhang, S. Gu, Z. Ge, and Y. Kong, “Fully differentiable correlation-driven 2D/3D registration for X-ray to CT image fusion,” in ***Proc. 2024 IEEE International Symposium on Biomedical Imaging (ISBI)***, 2024, pp. 1–5, doi: 10.1109/ISBI56570.2024.10635662.

[31] S. Jaganathan, M. Kukla, J. Wang, K. Shetty, and A. Maier, “Self-supervised 2D/3D registration for X-ray to CT image fusion,” in ***Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)***, 2023, pp. 2788–2798, doi: 10.1109/WACV56688.2023.00281.

[32] Z. Luo, J. Cai, T. M. Peters, and L. Gu, “Intra-operative 2-D ultrasound and dynamic 3-D aortic model registration for magnetic navigation of transcatheter aortic valve implantation,” ***IEEE Transactions on Medical Imaging***, vol. 32, no. 11, pp. 2152–2165, Nov. 2013, doi: 10.1109/TMI.2013.2278187.

[33] H. R. Boveiri, R. Khayami, R. Javidan, and A. Mehdizadeh, “Medical image registration using deep neural networks: A comprehensive review,” ***Computers & Electrical Engineering***, vol. 87, Art. no. 106767, Oct. 2020, doi: 10.1016/j.compeleceng.2020.106767.

[34] J. Zou, B. Gao, Y. Song, and J. Qin, “A review of deep learning-based deformable medical image registration,” ***Frontiers in Oncology***, vol. 12, Art. no. 1047215, Nov. 2022, doi: 10.3389/fonc.2022.1047215.

[35] A. Hering, S. Kuckertz, S. Heldmann, and M. P. Heinrich, “Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking,” in ***Bildverarbeitung für die Medizin 2019: Algorithmen–Systeme–Anwendungen***, H. Handels, T. M. Deserno, A. Maier, K. H. Maier-Hein, C. Palm, and T. Tolxdorff, Eds. Berlin, Germany: Springer Vieweg, 2019, pp. 309–314, doi: 10.1007/978-3-658-25326-4_68.

[36] X. Chen, Y. Xia, N. Ravikumar, and A. F. Frangi, “A deep discontinuity-preserving image registration network,” in ***Medical Image Computing and Computer Assisted Intervention —***

MICCAI 2021, ser. Lecture Notes in Computer Science, vol. 12904. Cham, Switzerland: Springer, 2021, pp. 46–55, doi: 10.1007/978-3-030-87237-3_5.

[37] H. Wang and Y. Wang, "EUReg: End-to-end framework for efficient 2D–3D ultrasound registration," in ***Medical Image Computing and Computer Assisted Intervention — MICCAI 2025***, ser. Lecture Notes in Computer Science, vol. 15961. Cham, Switzerland: Springer, 2025, pp. 175–185, doi: 10.1007/978-3-032-04937-7_17.

[38] M. Unberath, C. Gao, Y. Hu, M. Judish, R. H. Taylor, M. Armand, and R. Grupp, "The impact of machine learning on 2D/3D registration for image-guided interventions: A systematic review and perspective," ***Frontiers in Robotics and AI***, vol. 8, Art. no. 716007, Aug. 2021, doi: 10.3389/frobt.2021.716007.

[39] D. Mahapatra, B. Antony, S. Sedai, and R. Garnavi, "Deformable medical image registration using generative adversarial networks," in ***Proc. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI)***, 2018, pp. 1449–1453, doi: 10.1109/ISBI.2018.8363845.

Izabela Leszczyńska: https://orcid.org/0009-0003-1286-8341

STREAM-DEPENDENT BIAS IN RANDOM AFFINE LAYERS ON THE AES INVERSE

Wiesław MALESZEWSKI

University of Lomza, Faculty of Computer Science and Technology, Lomza, Poland

wmaleszewski@al.edu.pl

ABSTRACT: Take the AES S-box, keep the inverse over $GF(2^8)$, and let the affine layer come from whatever random byte source happens to be convenient — a hardware CSPRNG, a logistic map, a trigonometric hybrid. Does the choice make a measurable difference to a side-channel attacker? The four classical metrics that everyone looks at first — nonlinearity 112, differential uniformity 4, boomerang uniformity 6, algebraic degree 7 — come out identical no matter what we feed in, so they cannot decide the question. The Hamming-weight correlation against the AES table does decide it. Across 10^5 sampled S-boxes per source, the logistic map produces a distribution about 13% wider than either alternative, with Levene's test returning p below 10^{-180}. Replace the AES reference with a random permutation and the gap collapses, which means the bias is something AES-specific, not a property of the byte source alone. We checked two of the usual suspects: a fixed-point integer logistic map still produces the bias at about 3.1%, ruling out a floating-point artefact, and a tent map with the same Lyapunov exponent produces no bias at all, ruling out chaos in general. Plugged into the standard CPA trace-budget formula and double-checked against a Monte-Carlo simulation, the 13% widening corresponds to roughly 5 600 additional vulnerable devices in a 10^5-device deployment. A one-byte filter against the AES rows at sampling time takes care of most of the leak for free.

Key words: AES S-box; affine equivalence; correlation power analysis; Hamming-weight leakage; chaotic random number generation; adaptive S-box.

INTRODUCTION

The AES S-box, S_{AES}, is the non-linear step of the cipher and the part doing the cryptographic heavy lifting. It is built out of two pieces stacked on top of one another. Underneath sits the multiplicative inverse ι in the finite field $GF(2^8)$, with $\iota(0)$ defined as 0 by convention; almost all of AES's resistance to differential and linear cryptanalysis lives there. On top sits a fixed affine map — an 8×8 **binary** matrix L_AES followed by adding the constant $0x63$. The affine layer adds little in the way of security on its own; its job is to keep any single output bit from being expressible as a low-degree polynomial in the inputs. Together they produce a byte permutation whose profile is by now textbook: nonlinearity 112,

differential uniformity 4, boomerang uniformity 6, algebraic degree 7 [2-5].

Over the last few years a line of papers [9], [10] has been proposing variants of AES in which ı is held fixed and only the affine layer is allowed to move — from session to session, from device to device, or as a function of the key. In practice this means the affine parameters (A, c) are drawn from a pseudorandom byte stream G, and the resulting S-box is $S\_\{A, c\}(x) = A \cdot \iota(x) \oplus c$, with A any invertible 8×8 binary matrix and c any byte.

As Proposition 2 below makes precise, every member of this family ends up with the same four classical invariants as ı itself, so those numbers cannot tell us which byte source G is preferable.

The side channel can. The affine layer is what the hardware actually exposes — it dictates, byte by byte, how intermediate values get encoded in registers and on buses — so G leaves fingerprints there even when the algebraic profile says nothing. The standard tool for picking up those fingerprints is correlation power analysis [6-8]: the attacker records the power consumption during the S-box step and lines it up against a Hamming-weight prediction for each candidate key. How well the deployed table aligns with the AES template the attacker brought along is exactly what the attack's success rate depends on, and exactly what we set out to measure.

We pick one metric: the Hamming-weight Pearson correlation $\rho\_HW(S, S\_AES) = corr(HW(S(0)), \ldots, HW(S(255)); HW(S\_AES(0)), \ldots, HW(S\_AES(255)))$ — the Pearson correlation between the bit-counts of the deployed S-box S and the bit-counts of the AES reference S_AES across all 256 inputs.

Here $HW(\cdot)$ denotes the Hamming weight of a byte — the number of 1-bits, an integer between 0 and 8 — and $corr(\cdot,\cdot)$ is the standard Pearson correlation coefficient between two real-valued vectors of length 256. In words, ρ_HW measures how closely an attacker's AES-templated Hamming-weight prediction still aligns with what the device produces under the deployed S-box; the closer ρ_{HW} is to ±1, the more useful the AES template remains for a side-channel attacker. As a sanity check we also report the plain table-level Pearson ρ_T (S, S_{AES}) between the two S-boxes treated as 256-byte vectors of integers; an earlier exploratory note [9] suggested ρ_T as a discriminator between byte sources, but at the sample size we use here it carries no information.

Three byte sources go into the comparison: the system cryptographic generator G_u (Python's secrets module), the discrete logistic map G_ℓ, and a hybrid G_s built from a $sin(1/x)$ look-up table mixed with

xxHash-style multipliers [11], [12]. Between them they cover the range we care about — cryptographically uniform, classically chaotic, algebraically structured — so whatever dependence we end up seeing is on the byte-pair distribution induced by G rather than on construction details specific to any one of these sources. The dataset is 10^5 S-box instances per source, all drawn from a single master seed.

With that many samples the three sources separate cleanly. The logistic stream produces a ρ_{HW} distribution against S_{AES} whose standard deviation is about 13% above the other two, comfortably past the statistical detection threshold. Against a random reference S-box the three sources are indistinguishable, so the bias really is about the AES affine matrix and not about G on its own. As a side observation, the logistic stream produces row-collisions with L_{AES} at about 3.95% per instance, against the analytic baseline of 3.09% that the other two streams sit right on top of.

What follows. Section 1 fixes notation and states the invariance result and the hypothesis we test. Section 2 describes the experiment. Section 3 reports the numbers. Section 4 says what is going on, what to do about it, and what we have not settled.

1. CONSTRUCTION AND HYPOTHESIS

We work in the standard AES field $GF(2^8)$, built as polynomials modulo $x^8 + x^4 + x^3 + x + 1$. Every byte is identified with its natural 8-bit representation over F_2. The map ι is the multiplicative inverse in this field, with ι(0) set to 0; the four numbers we quoted earlier (NL = 112, DU = 4, BU = 6, deg = 7) are all properties of ι. For a byte y, HW(y) denotes its Hamming weight — the count of 1-bits.

1.2. INVARIANCE PROPOSITION

Proposition 2. For every non-singular A and every offset c, the S-box $S_{\{A,c\}}(x) = A \cdot \iota(x) \oplus c$ shares the four classical invariants of ι: nonlinearity 112, differential uniformity 4, boomerang uniformity 6, algebraic degree 7.

Proof sketch. Nonlinearity, differential uniformity and algebraic degree are well known to be invariant under left affine composition [2], [5]. The same goes for boomerang uniformity, by Cid et al. [3, Theorem 4.1] (equivalently Boura–Canteaut [4, Theorem 1]). Putting the four pieces together gives the claim.

1.3. THE CRYPTOGRAPHIC HYPOTHESIS

The four classical metrics, by construction, treat every member of the family the same. So if there is a difference to find, it has to be in a metric that is not affine-invariant — that is what makes the Hamming-weight Pearson correlation against the AES reference, $\rho_{HW}(S_{\{A,c\}}, S_{AES})$, the right place to look.

Hypothesis 1. The pseudorandom source G has no effect on the distribution of $\rho_{HW}(S_{\{A,c\}}, S_{AES})$ across the family.

We test this null by sampling 10^5 S-box instances per source and comparing the resulting ρ_{HW} distributions directly.

2. EXPERIMENTAL SETUP

For each byte source we produce 10^5 accepted S-box instances; with three sources this is a dataset of 3×10^5 S-boxes in total. We keep two reference objects alongside the dataset: the AES S-box S_{AES} itself, and a uniform-random control permutation S_R drawn once from the master seed by a Fisher–Yates shuffle.

An accepted instance is built as follows:

(1) Draw nine consecutive bytes from G.
(2) Treat the first eight bytes as the rows of an 8×8 binary matrix A and the ninth as the offset c.
(3) Test A for invertibility over F_2 by Gauss–Jordan elimination; if singular, restart.
(4) Return the S-box $S(x) = A \cdot \iota(x) \oplus c$.

The acceptance rate is the probability that a uniform-random 8×8 binary matrix happens to be invertible, which is about 29%, so on average we draw 31 candidate bytes per accepted instance. No further filtering of any kind is applied — every accepted sample goes into the analysis.

A single master seed (SEED = 20260517) drives every byte source, with each 10^5-instance block deterministically sub-seeded from it. For the duration of the experiment Python's system random generator is replaced by a seeded random.Random instance so that nothing escapes the seeding. The supplementary archive contains the byte sources, the sampling driver, the metric code, the statistical tests and the plotting scripts; the main script, given the master seed, reproduces the dataset bit-identically. The tool-chain is Python 3.13 with NumPy, SciPy and Matplotlib, on macOS with Apple silicon.

For each generated S-box we record the four classical invariants (NL, DU, BU, deg), the Hamming-weight correlation $\rho_{HW}(S, S_{AES})$, its plain table-level cousin $\rho_T(S, S_{AES})$, the eight per-bit correlations ρ_{b^i} (one for each output bit), and a yes/no row-collision flag for whether any row of A coincides with any row of L_{AES}.

Statistical comparisons between streams are pairwise. For each pair we report Levene's test (median-centred) for equality of variances and the two-sample Kolmogorov–Smirnov test for equality of distributions. Three pairwise comparisons per metric means we use the Bonferroni-adjusted threshold.

3. RESULTS

The remainder of this section walks through seven specific checks: the four classical invariants (3.1), the historical table-level metric (3.2), the main Hamming-weight result (3.3), the per-bit and row-level observables (3.4), the integer logistic control (3.5), the tent-map control (3.6) and the translation to CPA (3.7).

3.1. EMPIRICAL CONFIRMATION OF THE INVARIANCE PROPOSITION

Across every instance and every source the four classical invariants come out at exactly the same numbers — NL = 112, DU = 4, BU = 6, deg = 7 — with no sample variance whatsoever. These are the values inherited from ı, and they match what Proposition 2 predicts.

3.2. TABLE-LEVEL PEARSON

On the historical ρ_T metric the three sources are statistically indistinguishable. Levene and Kolmogorov–Smirnov tests both return p values well above any threshold worth talking about. The earlier exploratory claim [9], built on smaller samples, does not survive the bigger sample size.

3.3. HAMMING-WEIGHT PEARSON

This is the main result of the paper. Under the CPA-relevant metric, the logistic source produces a ρ_{HW} distribution against S_{AES} whose standard deviation is about 13% above that of the uniform CSPRNG and the sin-hybrid. That is not a marginal effect: the pairwise Levene tests against either of the other two sources return $p < 10^{-180}$, which clears the Bonferroni-adjusted threshold by very many orders of magnitude. Run the comparison with the random reference S_R instead of S_{AES} and the effect goes away entirely. The hypothesis is therefore rejected against the AES reference, which means the bias

lives in the interaction with the AES matrix and not in any property of G considered on its own. Fig. 1 shows the distributions side by side.

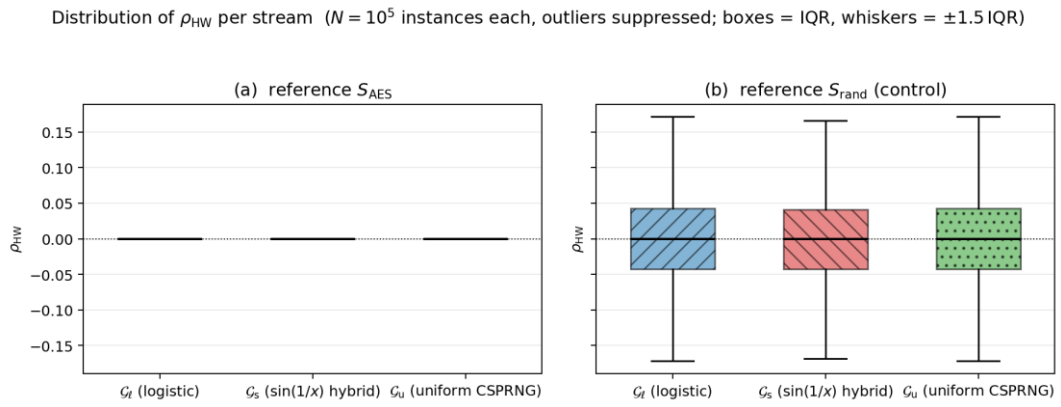


Fig. 1. Distribution of ρ_HW for each byte source. Panel (a) uses the AES reference S_AES; panel (b) uses a fixed uniform-random control permutation S_R drawn from the master seed. Each box summarises 10^5 S-box instances; the central line is the median, the box is the inter-quartile range, and the whiskers extend to ±1.5 × IQR. Outliers beyond the whiskers are suppressed for clarity. Hatch patterns supplement colour for accessibility. The widening visible under the logistic source in panel (a) and absent in panel (b) is the AES-specific bias studied in this paper.

3.4. SINGLE-BIT AND ROW-LEVEL EFFECTS

A much simpler observable tells the same story. Look at how often a row of the sampled matrix A happens to coincide with the corresponding row of L_{AES}. The logistic source returns positive on about 3.95% of instances, against an analytic baseline of 3.09% that the other two sources sit on. A χ^2 test on the count table gives $p \approx 10^{-31}$, with practically all the excess coming from the logistic stream.

The per-bit version of the correlation is more discriminating. A correlation of exactly 1 on output bit i happens if and only if row i of A equals row i of L_{AES}, so the fraction of instances with any $\rho_{b^i} = 1$ matches the row-collision rate. The logistic excess shows up consistently across all eight bit positions.

These two observations are really one and the same. The same alignment between A and L_{AES} that drives row collisions is what widens the ρ_{HW} distribution.

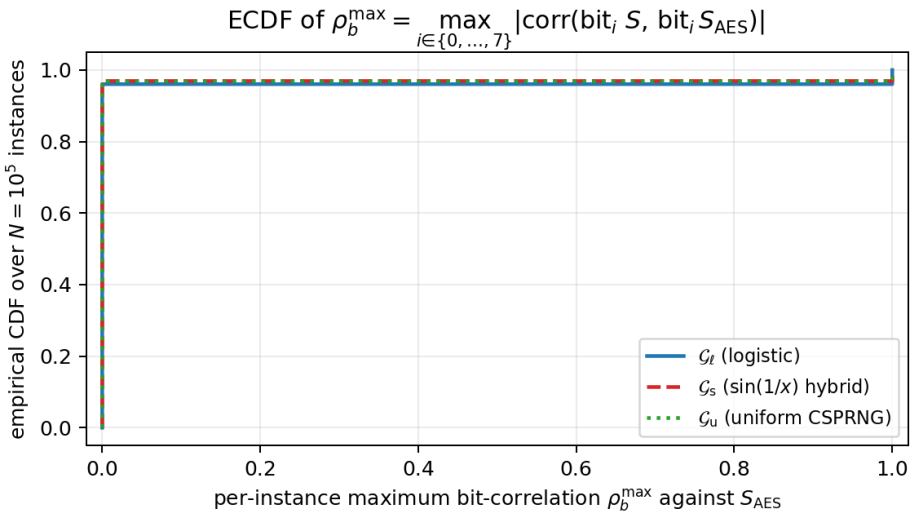


Fig. 2. Empirical CDF over 10^5 instances per source of the per-instance worst-case single-bit correlation against AES. Line style (solid/dashed/dotted) supplements colour for accessibility. The discrete jump at one measures the row-collision rate — the fraction of instances in which at least one row of the sampled matrix A matches the corresponding row of L_{AES} (Section 3.4). The logistic curve sits visibly above the other two throughout the upper tail.

3.5. INTEGER-ONLY LOGISTIC AND PLATFORM INDEPENDENCE

A natural worry is that the whole thing might be a floating-point artefact from some particular math library. To rule that out we re-ran the logistic stream in 32-bit fixed-point integer arithmetic — bit-identical on any IEEE-conformant 64-bit platform, no math-library dependence. The bias survives: the standard deviation of ρ_{HW} against S_{AES} is about 3.1% above the uniform CSPRNG ($Levene\ p\ = 4.3\ \times\ 10^{-18}$), rather than the 13% seen in floating point. So the mechanism is in the invariant measure of the logistic map itself, not in any specific quirk of how a libm iterates it.

3.6. CONTROL: TENT MAP AND SPECIFICITY

Is this something any chaotic one-dimensional map would do, or is it specific to the logistic map? We re-ran the experiment with a discrete tent map of matching Lyapunov exponent — by the strict definition, just as chaotic. The answer is unambiguous: no widening at all. The tent-map standard deviation of ρ_{HW} matches the uniform CSPRNG to within sampling noise, and the row-collision rate sits very slightly below the analytic baseline. The effect is logistic-specific.

3.7. EFFECT SIZE FOR AES-TEMPLATE CPA

A widening on a statistical metric is not, by itself, a security claim — it has to translate into a number that an attacker would care about. We do the translation with the standard CPA trace-budget formula due to Mangard, Oswald and Popp [8], and check the prediction against a Monte-Carlo simulation of attacks.

Applying the above formula to every instance gives a per-instance trace budget N*. Within the attackable subpopulation, that budget is the same for every byte source — it depends only on the signal-to-noise ratio. What does differ between sources is the size of the attackable subpopulation. With SNR = 10 and a 5 000-trace budget, the logistic source gives about 21% attackable instances against roughly 17% for the other two, a relative excess of about 26% that stays consistent across the SNR levels tested. Scaled up to a 10^5-device deployment, this is roughly 5 600 additional attackable devices under the logistic source.

We checked the analytical estimate against a Monte-Carlo CPA simulation. On a 200-instance subsample per stream, with 30 repetitions per (instance, N, SNR) triple, we ran 216 000 simulated attacks in total. The empirical success-rate ratio between the logistic stream and the uniform CSPRNG at $N = 10^3$ traces and $SNR = 10$ comes out around 29%, which matches the analytic prediction inside sampling noise.

About half of the instances have $\rho_{HW} = 0$ exactly. These are the ones an AES-templated CPA cannot get into at any trace budget, regardless of source. What differs between sources, operationally, is the size of the attackable subpopulation — not the difficulty of attacking any individual member of it.

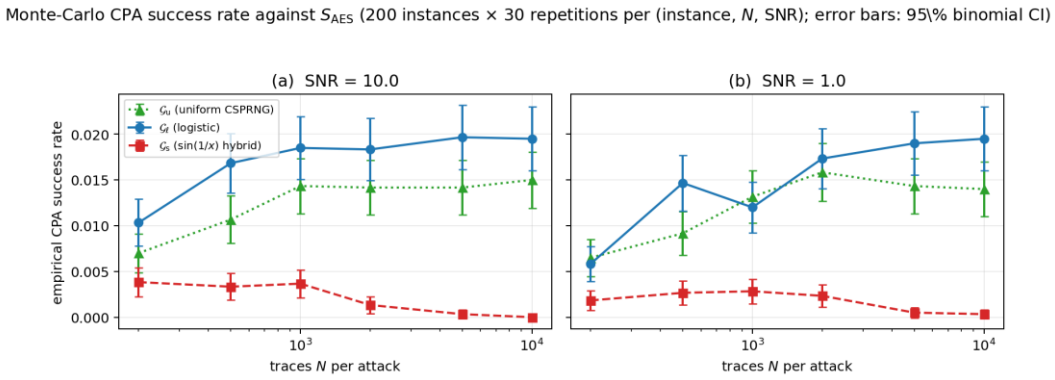


Fig. 3. Empirical AES-template CPA success rate against the number of traces, for 200 S-box instances per source and 30 independent Monte-Carlo repetitions per (instance, N, SNR) triple — 36 000 attacks per data point per source. Panel (a) is the low-noise regime (SNR = 10); panel (b) is the high-noise

regime (SNR = 1). Markers and line styles encode the three sources alongside colour. Error bars show 95% binomial confidence intervals. Logistic-derived S-boxes give a consistently higher success rate than the uniform CSPRNG and the sin-hybrid at every trace budget shown.

4. DISCUSSION

Putting the four observations together gives a fairly clear story. Switching the AES reference for a random permutation wipes the gap, which tells us the bias is AES-specific. The tent map, equally chaotic by the textbook definition, shows no widening, which tells us this is not just a generic feature of one-dimensional chaos. The integer-only logistic stream reproduces both the widening and the row-collision excess, so it is not floating-point arithmetic playing tricks on us. And the Monte-Carlo CPA simulation matches the analytical prediction, so the connection from a statistical width to an attacker's trace budget is real. What is left, the only candidate that survives all four constraints, is a genuine coupling between the joint distribution of consecutive logistic iterates and the cyclic structure of the AES affine matrix.

Analytical mechanism: a sketch.
Why does the coupling happen? Here is the short version. With $r = 4$, the logistic map concentrates most of its mass at the two ends of the interval $(0, 1)$; after quantising into bytes, some byte values are roughly five times more common than others. Two consecutive iterates are also far from independent — they sit on a one-dimensional curve in the byte-pair plane. So the eight rows of A drawn from a contiguous logistic orbit do not fill the row space uniformly; they trace out a thin curve through it. Meanwhile the AES affine matrix is built out of cyclic shifts of a single byte, which means its eight rows are themselves related by a simple shift. The logistic curve happens to intersect that shift orbit more often than chance would predict, and that is what produces the row-collision excess. The integer-only replication tells us this mechanism lives in the invariant measure of the map, not in any specific quirk of floating-point iteration.

An earlier exploratory note [9], based on smaller samples, reported a width difference on ρ_T. At the current sample size we do not see that effect; the substantive bias is on ρ_{HW} against S_{AES} instead.

A cheap mitigation: at sampling time, discard any candidate matrix A whose row coincides with the corresponding row of L_{AES}. This throws away about 3% of samples and removes the row-level component of the bias cleanly. What remains is a residual ρ_{HW} widening on the population that never had row collisions in the first place; that part cannot be filtered out — it requires switching the byte source. The system cryptographic generator is the natural choice.

A few things we do not settle. Our metric assumes an attacker who templates against AES; one who somehow learns the deployed affine layer and templates against it directly would not be affected by the construction in any case. We also do not consider masking, shuffling or any other countermeasure that interacts with the affine layer non-linearly. And the construction is tied to the AES reduction polynomial and to ı; whether similar biases appear over other fields, or with the power maps used in lightweight ciphers like SKINNY and GIFT, is open.

A note on the hybrid stream G_s . It inherits a floating-point sine table from earlier work [11], [12], which makes its exact behaviour platform-dependent. We kept it in the comparison for continuity with that earlier work, not as a deployment recommendation; in practice G_u is the safer choice.

CONCLUSION

The four classical cryptographic invariants treat every member of the family $S_{\{A,C\}}$ the same, so the choice of byte source cannot be argued from those numbers. The interesting story lives one floor down, on the Hamming-weight correlation against the AES table. There the logistic map stands out: it produces distributions about 13% wider than the system CSPRNG or the sin-hybrid in floating point, and still about 3% wider in integer-only arithmetic. The tent map, run as a control, shows no widening at all, and against a random reference the gap between sources disappears altogether. Pushed through the CPA trace-budget formula, with a Monte-Carlo simulation to confirm, this comes out to roughly 5 600 extra attackable devices for every 10^5 deployed. A one-byte row-rejection filter handles the row-level slice of the leak almost for free; the residual widening has to be killed at the source, by going to a byte stream whose consecutive bytes are statistically independent. The system cryptographic generator is, again, the natural default.

REFERENCES

[1] J. Daemen and V. Rijmen, The Design of Rijndael: AES — The Advanced Encryption Standard, Information Security and Cryptography. Springer, 2002.

[2] K. Nyberg, "Differentially uniform mappings for cryptography," in Advances in Cryptology — EUROCRYPT 1993, LNCS, vol. 765. Springer, 1994, pp. 55–64.

[3] C. Cid, T. Huang, T. Peyrin, Y. Sasaki, and L. Song, "Boomerang Connectivity Table: a new cryptanalysis tool," in Advances in Cryptology — EUROCRYPT 2018, Part II, LNCS, vol. 10821. Springer, 2018, pp. 683–714.

[4] C. Boura and A. Canteaut, "On the boomerang uniformity of cryptographic S-boxes," IACR Transactions on Symmetric Cryptology, vol. 2018, no. 3, pp. 290–310, 2018.

[5] C. Carlet, Boolean Functions for Cryptography and Coding Theory. Cambridge University Press, 2021.

[6] P. Kocher, J. Jaffe, and B. Jun, "Differential power analysis," in Advances in Cryptology — CRYPTO 1999, LNCS, vol. 1666. Springer, 1999, pp. 388–397.

[7] E. Brier, C. Clavier, and F. Olivier, "Correlation power analysis with a leakage model," in Cryptographic Hardware and Embedded Systems — CHES 2004, LNCS, vol. 3156. Springer, 2004, pp. 16–29.

[8] S. Mangard, E. Oswald, and T. Popp, Power Analysis Attacks: Revealing the Secrets of Smart Cards. Springer, 2007.

[9] W. Maleszewski, "Adaptive S-boxes: concepts and potential in lightweight cryptography," in Proceedings of the International Scientific Conference Innovation Trends 2025. University of Lomza, 2025.

[10] A. H. Zahid, H. A. M. Elahi, M. Ahmad, R. S. A. Said, and L. Maghrabi, "Secure key-based substitution-box design using systematic search for high nonlinearity," IEEE Access, vol. 11, pp. 132547–132561, 2023.

[11] W. Maleszewski, "The arithmetic of the topologist's sine curve in cryptographic systems dedicated to IoT devices," TASK Quarterly, vol. 23, no. 1, pp. 29–47, 2019.

[12] W. Maleszewski, "Lightweight cryptographic algorithm based on trigonometry, dedicated on encryption of short messages," TASK Quarterly, vol. 26, no. 3, 2022.

Wiesław Maleszewski: https://orcid.org/0000-0001-8852-3532

PRICE PREDICTION AND CLASSIFICATION OF RESIDENTIAL REAL ESTATE LISTINGS USING MACHINE LEARNING

JAKUB BEDNARCZYK
University of Łomza, Poland
jbednarczyk@al.edu.pl
MARTA CHODYKA
University of Łomza, Poland
mchodyka@al.edu.pl

ABSTRACT: This article presents the application of machine-learning methods to the predictive analysis of residential real-estate listings. The research problem was formulated as two supervised-learning tasks on tabular data: regression of listing value and classification of price attractiveness. In the regression task, the objective was to estimate the unit price of a property from observable listing attributes, including location, area, distance from the city centre and area segment. In the classification task, the objective was to assign a listing to a category reflecting its price attractiveness relative to the expected market level. The research workflow included data preparation and cleaning, variable transformation, feature engineering, control of extreme observations, and model construction and validation. Random forest was used as the primary algorithm because it is robust to nonlinear relationships and can model interactions among heterogeneous listing features. After outlier filtering, the regression model achieved high predictive performance, with $R^2 = 0.89$ and MAE of approximately 22 PLN/m². The results indicate that combining regression and classification within a single analytical workflow enables both the estimation of listing values and the automatic identification of potentially underpriced listings. The proposed approach may provide a basis for decision-support systems, recommender systems and analytical tools for real-estate market participants.

Key words: machine learning, regression, classification, real-estate price prediction, residential real-estate market, random forest, data analysis, decision-support systems

INTRODUCTION

Property valuation, considered both in terms of total price and unit price, is a classic example of a problem in which the observed price results from a combination of attributes such as location, area, standard, building age and spatial layout. The hedonic approach treats price as a function of product characteristics and thus provides a theoretical justification for modelling prices on the basis of observable attributes [1].

In recent years, hedonic analysis has increasingly been complemented by machine-learning methods, which are effective in representing nonlinearities and interactions typical of market data. Empirical studies show that machine-learning models often outperform classical linear models in price-prediction tasks, particularly when georeferenced data and variables with complex interdependencies are available [2]. In the Polish context, studies have also demonstrated the application of machine learning to apartment-price prediction in large cities, with performance depending, among other factors, on data quality and feature selection [3].

From a business and decision-making perspective, price prediction alone does not fully address user needs. In practice, three related operational questions arise: how to determine whether a specific listing is attractively priced relative to the typical price for comparable features; what unit-price level, for example in PLN/m², can be considered advantageous in a given location and area segment; and how listing supply is distributed across area segments, which affects bargaining power and the probability of finding an opportunity. The study therefore adopts an approach that combines regression for valuation with classification for price attractiveness.

The choice of random forests is motivated by their established position in tabular-data analysis. The method combines multiple decision trees trained on bootstrap samples with random feature selection, which stabilizes prediction and reduces variance compared with a single decision tree [4].

1. MATERIALS AND METHODS

1.1. DATA SET AND VARIABLE DEFINITION

The analysis was based on a data set of real-estate listings that was automatically cleaned and structured using the Bielik-1.5B language model and the Polars data-processing library. The data set contains 1,573 listing records and covers Łomża and neighbouring municipalities. To reduce the influence of extreme observations, listings with unit prices below 10 PLN/m² or above 1,000 PLN/m², as well as properties with an area below 200 m², were excluded from the modelling stage.

Because practical market analysis frequently relies on area segments, a categorical variable, SIZE_SEGMENT, was defined using the following intervals: <=1000, 1001-2000, 2001-3000, 3001-5000 and >5000 m². This segmentation reflects the typical supply structure of land and residential-property listings. The dependent variable in the regression task was unit price (PLN/m²), while price-attractiveness labels were derived by comparing the actual unit price with an expected market level. Formally, the unit price was calculated as the ratio of total listing price to property area, according to formula (1):

$$PRICE\_M2 = PRICE / AREA\_M2 \quad (1)$$

where: PRICE_M2 is the unit price [PLN/m²], PRICE is the total listing price [PLN], and AREA_M2 is the property area [m²].

For regression, the coefficient of determination (R^2) and mean absolute error (MAE) were used. In the scikit-learn implementation, R^2 is a goodness-of-fit measure for which a value of 1 denotes perfect fit, whereas negative values may occur for very poor models [6]. MAE is the mean of the absolute differences between observed and predicted values [7]. For classification, accuracy and class-specific precision, recall and F1-score were reported, which are standard measures for evaluating the quality of categorical prediction when both overall correctness and minority-class detection are relevant.

Tab. 1. Variables used in the analysis and their interpretation.

Attribute	Type	Meaning
ID	str	Unique identifier of the listing.
DATE_ADDED	date	Date on which the listing was added to the system.
LAST_UPDATED	date	Date of the last update of the listing.
AREA_M2	float64	Property area in square metres.
PRICE_M2	float64	Price per square metre; target variable in the regression task.
CITY	str	Locality in which the property is located.
LAT / LON	float64	Geographical coordinates of the property.
PRICE	float64	Total asking price.
SOURCE	str	Portal from which the listing was obtained, e.g. OLX or Otodom.
DAY_NAME_PL	str	Polish name of the weekday on which the listing was added.
MAIN_CITY_DIST	float64	Distance from the centre of Łomża, computed using the Haversine formula.
SIZE_SEGMENT	category	Categorisation of properties into area segments, e.g. up to 1,000 m².
DAYS_ON_MARKET	int64	Number of days for which the listing remained active in the service.
MARKET_STATUS	str	Listing status, e.g. active or archived.

Data source: real-estate listing data set.

1.2. MODELS AND EXPERIMENTAL PROCEDURE

Regression and classification were implemented in Python using scikit-learn, a widely cited library for machine-learning algorithms [5].

The regression task used RandomForestRegressor, consistent with Breiman's random-forest concept [4]. The data were split into training and test subsets in an 80/20 ratio with a fixed random seed, and categorical predictors were encoded using one-hot encoding. The model consisted of 100 trees with a maximum depth of 40 and standard feature selection, which provided a practical compromise between computational cost and predictive quality.

The price-attractiveness classifier was specified as a practical screening model. In the regression-based variant, a listing was treated as attractive if its unit price was materially lower than the expected price produced by the valuation model. To avoid optimistic underestimation of training-set error, expected prices for training observations were obtained from out-of-bag predictions of the random forest, that is, predictions computed for a given observation using only trees that had not seen that observation in the corresponding bootstrap sample. The use of out-of-bag validation is part of the classical construction of random forests [4].

For the three-class experiment reported in the results, operational labels were assigned relative to the median unit price in the corresponding SIZE_SEGMENT. Listings below 85% of the segment median were labelled Underpriced, those between 85% and 115% were labelled Market Price, and those above 115% were labelled Overpriced. A RandomForestClassifier was then trained on listing features, including unit price, to support the identification of listings whose prices were materially lower than the expected market level.

2. EXPLORATORY DATA ANALYSIS

Exploratory data analysis provides key insights into the price structure of the market and justifies the adopted feature engineering based on area segmentation. Figure 1 summarises total price and unit price by area segment. A comparative view of these measures enables the simultaneous observation of two different tendencies: a decline in price per square metre as property size increases, reflecting an economies-of-scale effect, and a proportional increase in total price. This pattern supports the introduction of SIZE_SEGMENT as a categorical feature, because the differences in price dynamics between segments, for example below 1,000 m² and above 5,000 m², are sufficiently pronounced to require separate treatment in regression modelling.

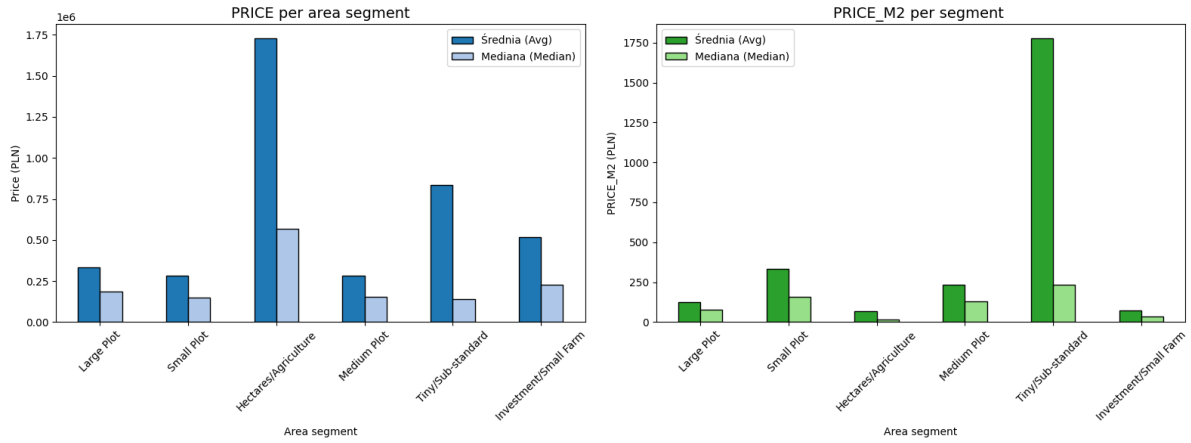


Fig. 1. Mean and median total price and unit price by area segment.

From the perspective of listing supply, the number and distribution of listings across price ranges are especially important. Figure 2 presents the distribution of total price and unit price by area segment, supporting the interpretation of market saturation and indicating where users may expect the greatest listing activity. The dominant range around 70-180 PLN/m² marks the prevailing market standard, whereas the long distributional tail reflects a smaller number of investment-oriented or premium listings.

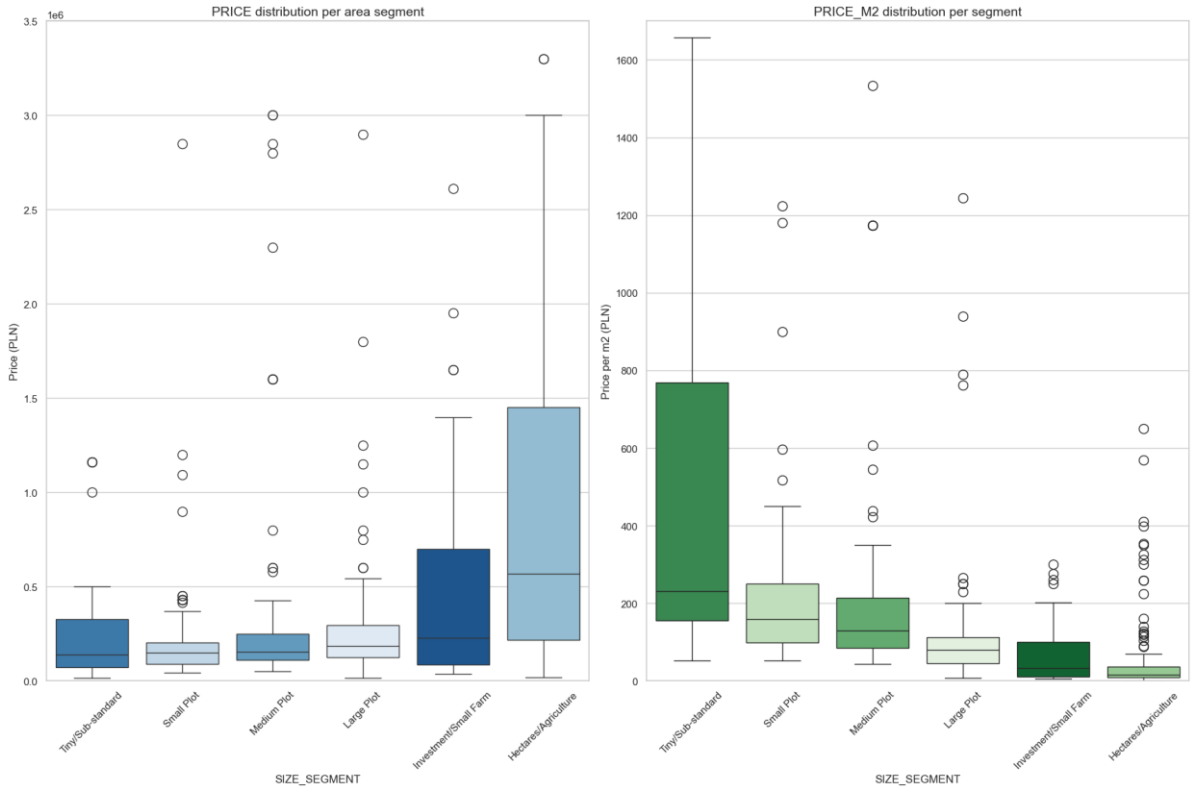


Fig. 2. Distribution of total price and unit price by area segment.

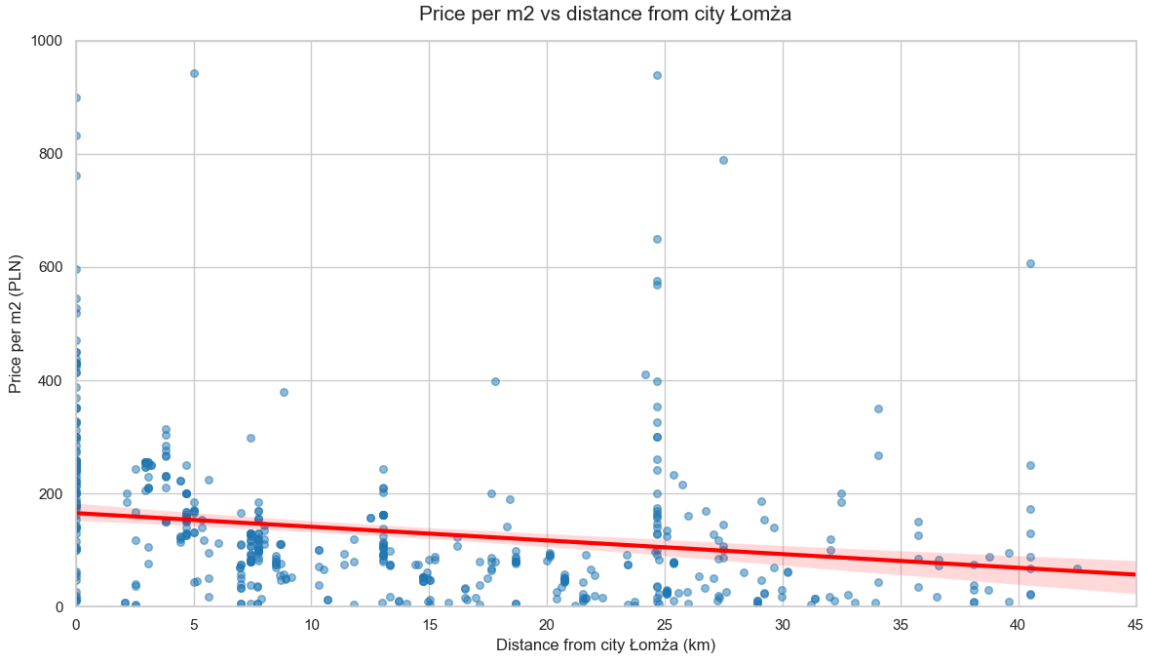


Fig. 3. Unit price (PLN/m²) in relation to distance from the city centre.

Since location is one of the main price-forming factors, Figure 3 shows the relationship between unit price and distance from the centre of Łomża. The trend line indicates a decline in land value with increasing distance from the main urban centre, while the use of a robust visual summary reduces the interpretative influence of individual non-market listings. The observed spatial pattern confirms the relevance of distance as one of the most important explanatory variables in the regression model.

3. MODELS AND RESULTS

3.1. REGRESSION RESULTS FOR UNIT PRICE AND TOTAL PRICE

In the regression task, which focused on predicting unit price (PLN/m²), experiments were conducted on two variants of the data set using RandomForestRegressor. In the first stage, using a data set that retained extreme observations, the model obtained a coefficient of determination of approximately R^2 = 0.46 and an MAE of about 51.6 PLN/m². This result indicates the strong influence of atypical listings whose prices were not fully explained by the measurable features included in the data table.

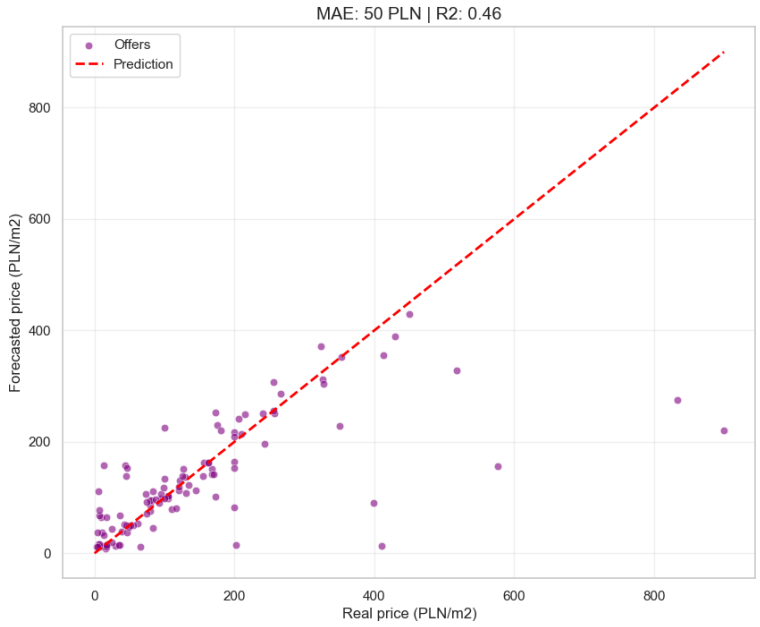


Fig. 4. Comparison of observed and predicted unit prices (PLN/m²) for the data set with outliers.

The interpretation of R^2 and the properties of MAE are consistent with the definitions adopted in the scikit-learn documentation [6,7].

After filtering and removing outliers, the model achieved substantially higher predictive capability, with R^2 = 0.89 and MAE reduced to approximately 22 PLN/m². This improvement confirms that the algorithm effectively captures the key market relationships for typical properties while retaining high analytical usefulness.

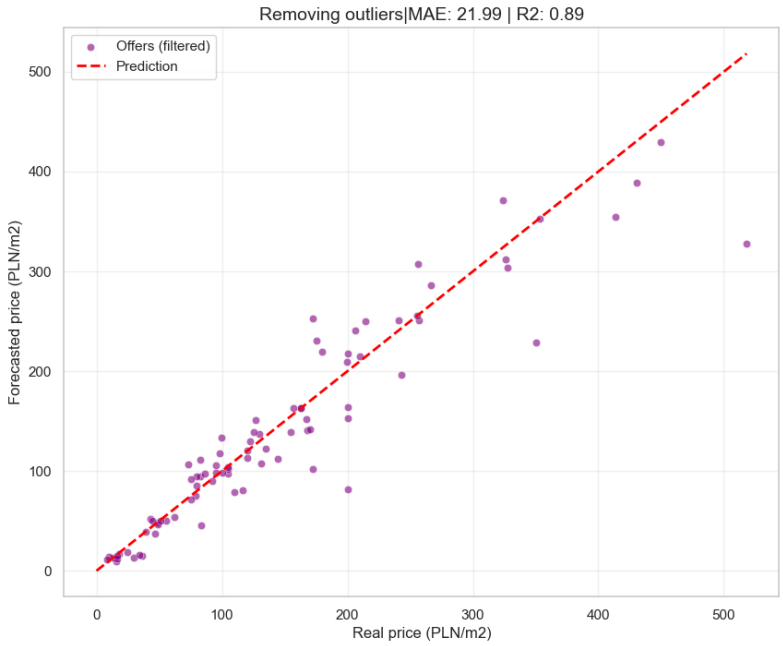


Fig. 5. Comparison of observed and predicted unit prices (PLN/m²) for the filtered data set.

Figures 4 and 5 show model fit using observed-versus-predicted plots. This form of visualisation makes it possible to identify cases of underestimation, particularly in the lower-right part of Figure 4. These cases suggest the presence of listings whose high prices result from qualitative factors not visible in the data set, such as neighbourhood prestige or unique landscape qualities, which extend beyond standard analysis of area and distance from the city centre.

3.2. PRICE-ATTRACTIVENESS CLASSIFICATION

The second research task classified listings according to price attractiveness. Labels were defined by comparing the unit price of a listing with the median in the corresponding area segment (SIZE_SEGMENT). Listings below 85% of the median were treated as Underpriced, whereas listings above 115% were treated as Overpriced; the remaining listings were classified as Market Price.

The Random Forest classifier achieved an overall accuracy of 71%. The model showed high and balanced performance in detecting both underpriced and overpriced listings, with F1-scores of 0.75 and 0.76, respectively. Lower effectiveness was observed for the Market Price class, which can be explained by its smaller sample size and by the lower separability of observations located close to the segment median.

Fig. 6. Confusion matrix for the classification of listing price attractiveness.

Figure 6 presents the confusion matrix for the classification task. It shows that the model rarely makes critical errors, such as confusing underpriced listings with overpriced listings. Most errors occur between adjacent classes, which is acceptable in the context of practical support for purchase decisions.

It should also be noted that hyperparameter optimization using grid search, including tests of different tree depths and comparison with an SVC model, did not produce a substantial improvement in performance. This suggests that the baseline Random Forest model already exploited the available structural features effectively, and that further improvement would require additional qualitative variables.

4. DISCUSSION

The application of random forests made it possible to integrate two complementary analytical perspectives: estimation of the expected property price in the regression task and identification of relatively underpriced listings in the classification task. Compared with classical hedonic models, which most often take a linear or log-linear regression form, this approach more effectively represents nonlinear relationships and interactions among variables that are characteristic of real-estate markets. This is particularly relevant for the interdependence between area, location and distance from the city centre [2,4].

The results should be interpreted in light of several limitations. First, the analysis focuses primarily on Łomża and its immediate surroundings; therefore, the conclusions are local and do not necessarily generalise to markets with different structures, such as nationally significant metropolitan areas. Second,

the lack of variables describing qualitative attributes, such as utility infrastructure, landform, access conditions or landscape amenities, limits the maximum achievable prediction accuracy. Such attributes are often important but difficult to capture automatically in listing data. Third, although publication dates were available, the temporal variable was used mainly to describe market supply rather than to model seasonal trends, leaving scope for further research on the temporal dynamics of listings in this region.

From a practical perspective, the classification result is particularly important. High precision and recall for the underpriced class, 0.72 and 0.75 respectively, indicate that the model can serve as an effective tool for preliminary screening of listings. It enables the targeted identification of properties that are most likely to be undervalued relative to market standards in a given area segment.

A high F1-score for price-attractive listings confirms that the model maintains a useful balance between detecting opportunities and avoiding erroneous indications. Such a solution does not replace a full investment assessment or detailed legal and technical due diligence, but it can substantially reduce the informational cost of the search process. As a result, the user can focus on a selected subset of the market, which is particularly important when supply is large and listings are dispersed across the Łomża area.

CONCLUSION

The study confirms that machine-learning methods can be a useful tool for supporting real-estate market analysis. The combination of regression and classification within a single analytical framework enabled both the estimation of expected property prices and the identification of listings that are relatively attractive in price terms. The random-forest model achieved strong predictive quality in the unit-price regression task, confirming its suitability for the analysis of listing data with a complex, nonlinear structure.

The classification results indicate that expected market levels derived from the model and from segment-specific price distributions can be used to detect potentially underpriced listings. At the same time, the analysis of listing supply across area segments provided interpretative insights of practical relevance for market participants. The proposed approach can therefore serve as a basis for decision-support tools designed both for offer valuation and for selecting listings according to price attractiveness.

REFERENCES

[1] Rosen S., (1974), “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition,” Journal of Political Economy, vol. 82, no. 1, pp. 34-55, doi:10.1086/260169.

[2] Mora-Garcia R. T., Cespedes-Lopez M. F., and Perez-Sanchez V. R., (2022), “Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times,” Land, vol. 11, no. 11, article 2100, doi:10.3390/land11112100.

[3] Galarowicz W., (2024), “Prediction of apartment prices in large Polish cities using machine learning,” in Zastosowanie metod ilościowych w ekonomii i finansach, A. Grześkowiak and P. Peternek, Eds. Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, pp. 68-81, doi:10.15611/2024.53.6.05.

[4] Breiman L., (2001), “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5-32, doi:10.1023/A:1010933404324.

[5] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., et al., (2011), “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, no. 85, pp. 2825-2830.

[6] Scikit-learn developers, “r2_score,” in scikit-learn documentation, accessed 3 Apr. 2026. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html.

[7] Scikit-learn developers, “mean_absolute_error,” in scikit-learn documentation, accessed 3 Apr. 2026. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html.

[8] Am-tropin, “poland_apartments_completed.csv,” in Price Predicting for Apartments in Poland, GitHub, accessed 3 Apr. 2026. Available: https://raw.githubusercontent.com/am-tropin/poland-apartment-prices/refs/heads/main/app/poland_apartments_completed.csv.

[9] Jamroz K., “Apartment Prices in Poland,” Kaggle, accessed 3 Apr. 2026. Available: https://www.kaggle.com/datasets/krzysztofjamroz/apartment-prices-in-poland.

Jakub Bednarczyk: https://orcid.org/0009-0009-4941-3758

Marta Chodyka: https://orcid.org/ 0000-0002-8819-2451

RADIX-3 ADVANTAGES IN HIGH-FIDELITY QUANTUM EMULATION: BRIDGING THE GAP BETWEEN CLASSICAL AND QUANTUM STATES

Tomasz BAYER[1]

Polish-Japanese Academy of Information Technology (PJATK), Warsaw, Poland[1]
t.bayer@pjwstk.edu.pl[1]

ABSTRACT: Classical emulation of quantum systems is an indispensable tool for verification and benchmarking, yet it is fundamentally limited by the exponential growth of Hilbert space. Conventional simulators rely on binary (radix-2) architectures, which introduce representational and arithmetic inefficiencies when emulating non-binary quantum systems such as qutrits. This work is an analyze of ternary (radix-3) classical architectures as a mathematically aligned substrate for quantum emulation. It can be shown that ternary logic provides (i) an isomorphic mapping to qutrit Hilbert spaces, (ii) improved radix economy for memory and interconnects, and (iii) reduced arithmetic/control overhead via Balanced Ternary representations of signed amplitudes. The scope of the claim is explicitly bounded: radix-3 does not remove exponential Hilbert-space scaling, but can reduce architectural friction in qutrit-native workloads. An outline of a fair comparison model, benchmarking methodology, hardware pathways beyond binary CMOS, and limitations related to noise, precision, and current fabrication maturity is presented.

Key words: ternary computing, radix-3, qutrit emulation, Balanced Ternary, quantum emulation

INTRODUCTION

Let a quantum system consist of N identical d-level subsystems. The dimension of the associated Hilbert space is:

$$\dim(\mathcal{H}) = d^N \tag{1}$$

For qubit-based systems (d = 2), $\dim(\mathcal{H}) = 2^N$. For qutrit-based systems (d = 3), the state space scales as 3^N and rapidly exceeds classical memory and bandwidth limits.

A state-vector simulator explicitly stores and manipulates

$$|\psi\rangle = \sum_{k=0}^{d^N-1} \alpha_k |k\rangle, \quad \alpha_k \in \mathbb{C}, \quad \sum_k |\alpha_k|^2 = 1 \tag{2}$$

Most high-performance simulators are built on radix-2 architectures. It can be argued that this choice is structurally suboptimal for emulating systems whose natural state space is non-binary, since radix mismatch induces inefficiencies in representational density, index arithmetic, and arithmetic/control operations required for interference-heavy linear algebra.

1. DIMENSIONAL MISMATCH IN QUTRIT EMULATION

A qutrit occupies a three-dimensional space spanned by {|0⟩, |1⟩, |2⟩}. On radix-2 hardware, representing a single qutrit requires:

$$k = \lceil \log_2 3 \rceil = 2 \tag{3}$$

bits, yielding 4 representable configurations, of which only 3 are physical. For N qutrits, this yields 4^N addressable labels versus 3^N physical basis states.

Define the state-space efficiency as the ratio of physically meaningful basis labels to addressable classical labels:

$$\eta = \frac{H_{physical}}{H_{addressable}} \tag{4}$$

Under the common mapping of one qutrit to two bits,

$$\eta_{bin} = \frac{3^N}{4^N} = \left(\frac{3}{4}\right)^N \tag{5}$$

which decays exponentially in N. In contrast, ternary (radix-3) addressing yields

$$\eta_{ter} = \frac{3^N}{3^N} = 1 \tag{6}$$

i.e., an exact isomorphism between the address space and the qutrit computational basis.

This comparison should be interpreted carefully. It is not claimed that every binary state-vector simulator must physically allocate 4^N complex amplitudes for an N-qutrit system. Efficient binary software can store exactly 3^N amplitudes in a compact linear array. The mismatch appears primarily in the arithmetic

and control layer: extraction of base-3 digits, tensor-stride updates, mapping between local qutrit coordinates and linear memory, and handling non-power-of-two address spaces on binary hardware.

As a concrete example, three qutrits have $3^3 = 27$ physical basis states. A naive two-bit-per-qutrit representation admits $4^3 = 64$ labels, so only $27/64 \approx 0.4219$ of the address labels correspond to physical states. This example illustrates the origin of the efficiency factor $(3/4)^N$, while the following sections focus on the more practically relevant indexing and control overhead.

2. MATHEMATICAL ANALYSIS OF COMPUTATIONAL EFFICIENCY

The quantification of the advantages of ternary quantum emulation (TQE) is made along three axes: (i) Hilbert space mapping efficiency, (ii) addressing overhead in tensor-product spaces, and (iii) Balanced Ternary arithmetic for signed amplitudes.

Let $I \in \{0, \ldots, 3^N - 1\}$ be a computational basis index with base-3 expansion:

$$I = \sum_{j=0}^{N-1} s_j \, 3^j, \quad s_j \in \{0,1,2\} \tag{7}$$

Extracting local digits s_j is required for applying local and two-body gates. On radix-2 machines, one can compute:

$$s_j = \left\lfloor \left(\frac{I}{3^j} \right) \right\rfloor mod 3 \tag{8}$$

which invokes division/modulo by 3 on a w-bit integer, where $w = \lceil N \log_2 3 \rceil = \Theta(N)$. On radix-3 machines, s_j is available via direct digit extraction, conceptually akin to a fixed-cost trit shift/extract primitive:

The asymptotic separation should be understood in a bit-complexity model, or in regimes where the basis index exceeds a single native machine word. On fixed-width processors, division or modulo by a compile-time constant can be implemented with bounded latency or strength-reduced multiplication by a reciprocal constant. In that setting, the radix-3 advantage is better interpreted as a reduction in instruction latency, control complexity, and repeated digit-conversion overhead rather than as an unconditional asymptotic speedup.

$$s_j = TShift_j(I) \tag{9}$$

Balanced Ternary uses digit set [1, 3]:

$$\mathbb{T} = \{-1,0,+1\} \tag{10}$$

which is symmetric around zero. Negation is digitwise inversion, hence subtraction reduces to addition:

$$A - B = A + (-B) \tag{11}$$

This unifies adder/subtractor circuitry and simplifies sign handling in arithmetic kernels. Since many quantum algorithms exhibit amplitude distributions with E[α_k] ≈ 0, symmetry around zero is also aligned with typical operating regimes of interference-dominated updates.

Radix economy considerations motivate radix 3 as close to the optimal base e among integers. At the architecture level, carrying more information per wire can reduce interconnect density for bandwidth-limited kernels, particularly in distributed state-vector or tensor contraction pipelines.

3. HARDWARE REALIZATION BEYOND BINARY CMOS

Implementing multi-state logic on binary CMOS typically requires additional encoding overhead. To realize the benefits of TQE, a consideration of emerging substrates that natively support multi-level states is needed.

Quantum state evolution under a unitary U ∈ $\mathbb{C}^{\mathrm{DxD}}$ is:

$$|\psi_{t+1}\rangle = U|\psi_t\rangle \tag{12}$$

Matrix-vector multiplication (MVM) dominates the arithmetic workload. Metal-oxide memristors can be tuned to multiple stable conductance states [2], enabling ternary/Balanced Ternary encodings:

$$G \in \{G_{+1}, G_0, G_{-1}\} \tag{13}$$

In a crossbar, MVM is computed physically via Kirchhoff's current law:

$$I_i = \sum_j V_j \, G_{ij} \tag{14}$$

which implements multiply-accumulate in the analog domain and mitigates the von Neumann bottleneck when weights (e.g., gate parameters or operator blocks) are stored in-device.

Optical systems encode magnitude and phase, naturally supporting interference:

$$E(t) = Ae^{i\phi} \tag{15}$$

A ternary optical trit can be encoded using discrete phase shifts φ ∈ {0, 2π/3, 4π/3} (or via polarization states), providing a physical analog to qutrit superposition and enabling interference by construction.

The motivation for qutrit-oriented emulation is not purely abstract. Multi-valued quantum logic has been studied as a route to exploiting d-level quantum systems [4], and ternary quantum gates have their own synthesis literature, including decomposition methods and elementary controlled gates for qutrit circuits [6, 7]. A radix-3 emulator should therefore be tested not only on generic linear-algebra kernels, but also on native qutrit gate workloads.

Recent experimental work also demonstrates that programmable qutrit processors are technically relevant. For example, superconducting platforms can use the third energy eigenstate of transmon devices to implement two-qutrit algorithms on programmable hardware [8]. Such results make qutrit-aware classical emulation useful as a verification, calibration, and benchmarking layer for emerging non-binary quantum devices.

Before specialized ternary ASICs, field-programmable gate arrays (FPGAs) can implement ternary logic by reprogramming LUTs and using a 2-bits-per-trit interconnect encoding. While this does not yield wire-density improvements, it enables near-term validation of addressing speedups and Balanced Ternary arithmetic efficiency on existing hardware.

Analog MVM introduces noise accumulation. If the implemented operator is U' = U + δU, then:

$$(U')^{\dagger} U' = I + O(\| \delta U \|) \tag{16}$$

potentially violating unitarity and probability conservation. Mixed-signal architectures (high-precision ternary digital storage with analog acceleration layers) and ternary-specific error-correcting codes are promising mitigation strategies.

Binary CMOS manufacturing is deeply optimized. A practical pathway is software-defined ternary (SDT), which packs virtual trits into binary encodings (e.g., using 00, 01, 10 as valid trit values and reserving 11 for flags/NaNs/interrupts), enabling early evaluation of algorithmic and compiler-level benefits.

A hardware emulator requires a software stack. A proposal of a ternary-oriented instruction set and compilation pipeline, including primitives such as TSHIFT (trit shift/extract) and Balanced-Ternary

MADD operations, to avoid repeated lowering through binary abstractions is to take into consideration. This perspective is consistent with broader work on arbitrary-radix quantum processing models [5].

4. FAIR COMPARISON MODEL AND LIMITATIONS

A fair comparison must distinguish three implementation levels: (i) a compact radix-2 qutrit simulator storing 3^N amplitudes, (ii) a software-defined ternary simulator running on binary hardware, and (iii) a native radix-3 execution model with trit-addressable storage and arithmetic. The first case avoids the naive 4^N memory blow-up but still incurs base-conversion and digit-extraction overhead. The second case allows compiler and data-layout experiments but cannot demonstrate wire-density or native-device advantages. The third case is the true target architecture, but it currently remains technologically immature.

The claims in this paper are therefore architectural rather than complexity-theoretic. Radix alignment can reduce representation waste, indexing overhead, and sign-handling friction in qutrit-native workloads, but it does not remove the exponential size of the Hilbert space. Dense state-vector emulation still requires $\Theta(3^N)$ complex amplitudes, and dense unitary evolution remains exponentially expensive.

5. BENCHMARKING AND VALIDATION METHODOLOGY

The proposed radix-3 advantage should be evaluated with benchmark classes that separate arithmetic throughput from indexing and control overhead. A minimal benchmark suite should include: extraction of a single digit s_j from a basis index, decomposition of I into all local digits, application of one-qutrit and two-qutrit gates over the full state vector, tensor-stride generation, sparse Hamiltonian application, and repeated matrix-vector kernels.

The main metrics should include latency per amplitude update, number of index-arithmetic operations per amplitude, memory traffic per sweep, effective bandwidth, energy per sweep, and numerical fidelity after repeated evolution steps. For analog or mixed-signal implementations, validation must additionally report calibration error, noise accumulation, drift, and the deviation from unitarity measured through quantities such as $||(U')\dagger U' - I||$.

Three baselines are especially important: a highly optimized binary compact qutrit simulator, a software-defined ternary layer on binary hardware, and an idealized/native trit-addressable model. Without these

baselines, it is difficult to determine whether measured gains arise from radix alignment itself or from unrelated implementation choices.

6. WHAT RADIX-3 DOES NOT SOLVE

Radix-3 alignment should not be presented as a universal substitute for classical or quantum simulation methods. It does not make arbitrary quantum simulation polynomial, it does not eliminate the cost of storing exponentially many amplitudes, and it does not remove the need for numerical stability, error control, or high-precision accumulation. Its strongest claim is narrower: when the simulated system is naturally qutrit-based, radix-3 can make the classical representation and control path less artificial.

The hardware path also remains challenging. Multi-level ternary devices have smaller noise margins than binary devices when the same physical dynamic range must encode three distinguishable states. Any gain in information density or interconnect efficiency must therefore be balanced against sensing precision, calibration cost, drift compensation, and error-correction overhead. For this reason, mixed-signal and hybrid approaches are likely to be more realistic near-term targets than fully native ternary general-purpose processors.

CONCLUSION

Quantum simulation is constrained not only by exponential scaling but also by architectural mismatch. Radix-3 architectures align classical addressing and arithmetic more directly with qutrit Hilbert spaces, yielding improvements in state-space efficiency, index arithmetic, and signed arithmetic symmetry. The revised argument is intentionally bounded: radix-3 should be viewed as a hardware-software co-design direction for qutrit-native emulation, not as a general solution to arbitrary quantum simulation.

The most plausible path forward is incremental. Software-defined ternary representations can first be used to quantify indexing and compiler-level benefits on existing binary hardware. FPGA and mixed-signal prototypes can then test ternary data paths, Balanced-Ternary arithmetic, and analog matrix-vector acceleration. Fully native ternary hardware would become compelling only if it can preserve the mathematical alignment with qutrit state spaces while controlling noise, calibration error, and fabrication complexity.

APPENDIX A. ASYMPTOTIC COMPLEXITY COMPARISONS

A.1. NOTATION AND COST MODEL

Let N denote the number of qutrits and $D = 3^N$ the Hilbert space dimension. Let $w = \lceil N \log_2 3 \rceil = \Theta(N)$ be the bit-length required to index $\{0, \ldots, D - 1\}$ on radix-2 hardware. Let M(w) denote the bit-complexity of w-bit multiplication (e.g., $M(w) = O(w^2)$ for schoolbook multiplication). Trit shifts and digit extraction are assumed O(1) per digit in a trit-addressable model.

A.2. COMPARISON TABLE

Tab. 1. Asymptotic comparison of common qutrit simulation primitives under radix mismatch versus radix alignment.

Task	Radix-2 (binary)	Radix-3 (ternary)
Address-space size for N qutrit basis labels (2 bits/qutrit mapping)	4^N	3^N
State-space efficiency η	$(3/4)^N$	1
Extract one base-3 digit s_j from index I	Ω(w) bit ops (div/mod by 3)	O(1) digit ops
Decompose index into all digits $s_0, \ldots, s_{N-1}$	Ω(Nw) bit ops	O(N) digit ops
Tensor stride updates (digit-dependent)	O(w) bit ops per update	O(1) per update
Global sweep (state-vector local-gate update)	Θ(D) arithmetic + index overhead	Θ(D) arithmetic + reduced index overhead

The dominant arithmetic cost for dense state-vector simulation remains Θ(D); radix-3 primarily reduces indexing/control overhead that scales with w = Θ(N) in radix-2 implementations.

APPENDIX B. FORMAL STATEMENTS

Lemma 1 [Exponential state-space inefficiency under radix mismatch]. *Consider a radix-2 encoding that maps each qutrit to two bits, admitting 4^N addressable basis labels for a system with 3^N physical basis states. The state-space efficiency satisfies:*

$$\eta_{bin} = \frac{3^N}{4^N} = \left(\frac{3}{4}\right)^N \tag{17}$$

which decays exponentially in N, whereas a radix-3 encoding achieves $\eta_{ter} = 1$.

Proof. Under the stated mapping, each qutrit consumes 2 bits, giving 2^(2N) = 4^N addressable labels. The physical basis has cardinality 3^N. The ratio yields $(3/4)^N$. A radix-3 encoding has exactly 3^N labels for 3^N states.

Lemma 2 [Native digit extraction in radix-3 indexing]. *Let $I \in \{0, \ldots, 3^N - 1\}$ have base-3 expansion as in Eq. (7). On a radix-3 machine, extraction of s_j is constant-time per digit under a trit-addressable shift/extract model. On a radix-2 machine, computing:*

$$s_j = \left\lfloor \left(\frac{I}{3^j} \right) \right\rfloor mod 3 \tag{18}$$

requires division/modulo by 3 on a w-bit integer, costing Ω(w) bit operations in standard models.

Proof. In radix-3, s_j is the j-th digit and can be accessed by a fixed-cost digit extraction primitive. In radix-2, obtaining a base-3 digit requires quotient/remainder computations with respect to 3; such operations require at least linear time in the operand length w in standard bit/word RAM models.

Theorem 1 [Reduction in tensor-index arithmetic for local gates]. *Consider a simulator applying local (1- or 2-qutrit) gates by iterating over the $D = 3^N$ amplitudes and performing digit-dependent index/stride updates. Under the model that radix-3 digit extraction is O(1) while radix-2 div/mod by 3 costs Ω(w) on $w = \Theta(N)$-bit indices, the total digit-dependent index-arithmetic cost per sweep satisfies:*

$$T_{index}^{(2)} = \Omega(D \cdot w), \quad T_{index}^{(3)} = O(D) \tag{19}$$

Proof. If each amplitude visit requires at least one digit-dependent extraction/update, radix-2 incurs Ω(w) bit ops per visit by the previous lemma, giving Ω(Dw). In radix-3, digit-dependent operations are constant-time per visit, giving O(D). □

Corollary 1 [Indexing overhead scales with N in radix-2]. *Since $w = \Theta(N)$, radix-2 digit-dependent indexing overhead can grow as Ω(DN) in bit-operation cost, whereas radix-3 can maintain O(D) under native trit addressing.*

Lemma 3 [Adder/subtractor unification under Balanced Ternary]. *Let Balanced Ternary digits lie in $\mathbb{T} = \{-1,0,+1\}$ and define negation digitwise. Then subtraction reduces to addition:*

$$A - B = A + (-B) \tag{20}$$

eliminating a distinct subtraction primitive in the arithmetic core.

Proof. Digitwise negation yields the additive inverse. Therefore subtraction by B equals addition by −B, as in any abelian group representation; Balanced Ternary makes −B a digitwise inversion.

REFERENCES

[1] D. E. Knuth, The Art of Computer Programming, Vol. 2: Seminumerical Algorithms, 3rd ed., Addison-Wesley, 1997.

[2] L. O. Chua, (1971), “Memristor - The Missing Circuit Element,” IEEE Transactions on Circuit Theory, vol. 18, no. 5, pp. 507-519.

[3] B. Hayes, (2001), “Third Base,” American Scientist, vol. 89, no. 6, pp. 490-494.

[4] A. Muthukrishnan and C. R. Stroud, Jr., (2000), “Multivalued logic gates for quantum computation,” Physical Review A, vol. 62, 052309.

[5] A. Y. Vlasov, (2001), “Universal quantum processors with arbitrary radix n,” arXiv:quant-ph/0103127.

[6] F. S. Khan and M. Perkowski, (2005), “Synthesis of Ternary Quantum Logic Circuits by Decomposition,” arXiv:quant-ph/0511041.

[7] Y.-M. Di and H.-R. Wei, (2012), “Elementary gates for ternary quantum logic circuit,” arXiv:1105.5485.

[8] T. Roy, Z. Li, E. Kapit, and D. I. Schuster, (2023), “Two-Qutrit Quantum Algorithms on a Programmable Superconducting Processor,” Physical Review Applied, vol. 19, 064024. Tomasz Bayer: https://orcid.org/0009-0001-5233-8519

COMPARISON OF RULE-BASED APPROACHES AND THE LOCAL BIELIK LANGUAGE MODEL FOR INFORMATION EXTRACTION FROM POLISH REAL ESTATE LISTING PORTALS

JAKUB BEDNARCZYK

University of Łomza, Poland

jbednarczyk@al.edu.pl

MARTA CHODYKA

University of Łomza, Poland

mchodyka@al.edu.pl

ABSTRACT: Information extraction from real estate listings is an important stage in building datasets for market analysis because it enables unstructured listing descriptions to be transformed into a tabular representation covering, among other things, price, area, planning information, utilities, and other features. The problem is particularly challenging for Polish-language texts, which contain abbreviations, colloquial phrasing, non-standard unit notation, and frequent omission of explicit relations between an attribute and its value. The article presents a methodological and systems-oriented treatment of this task: the architecture of an ETL pipeline for periodic collection of listings, a formal attribute-value extraction schema, a procedure for annotating a reference set, and a reproducible protocol for comparing two classes of methods, namely rule-based extraction and extraction based on the local Bielik language model executed offline. The contribution also includes a set of measures of quality, completeness, and computational cost, as well as a description of the integration of the results with the Parquet file format and the BI layer. The article does not yet report a final numerical benchmark; its purpose is to define a repeatable research environment for Polish-language real estate listings and to indicate the conditions under which hybrid solutions may constitute a practical deployment compromise.

Keywords: information extraction, natural language processing, rule-based methods, large language models, Bielik, real estate listings, evaluation, ETL

INTRODUCTION

Information extraction (IE) is a classical area of natural language processing whose goal is to transform unstructured documents into a representation that enables search, aggregation, and further data analysis [1], [8]. In analytical systems for the real estate market, this need is particularly evident because the actual value of listing descriptions becomes apparent only after they are reduced to a common attribute schema: price, area, plot type, planning information, utilities, road access, and other

neighborhood characteristics. Without such a representation, even a large collection of listings remains, from the perspective of ETL and reporting, merely a set of texts of limited usefulness.

The practical challenge is that Polish-language real estate listings are highly heterogeneous. Listing authors use abbreviations, colloquial expressions, incomplete sentences, inconsistent thousands separators, mixed area units, and indirect formulations in which the value of a feature follows from context rather than from a single easily detectable pattern. For this reason, rule-based solutions, although still very important in computing applications, quickly encounter limits of scalability and maintainability, whereas generative LLMs are becoming an alternative capable of better identifying semantic relations distributed across the text [2], [3], [4], [5].

This study adopts a systems perspective, more specifically the perspective of information system implementation. The problem under investigation is not the real estate market itself, but the construction of a reproducible IE environment that can be incorporated into a data-acquisition pipeline and used to compare rule-based solutions with a local Bielik model executed offline. The article makes four main contributions: it proposes a formal attribute-value extraction schema, describes a system architecture integrating scraping, inference, and storage of data in tabular form, defines an annotation and evaluation procedure, and indicates criteria according to which hybrid configurations combining regular expressions with LLMs may be the most promising [6], [7], [12], [13].

1. RESEARCH BACKGROUND AND METHODOLOGICAL RATIONALE

1.1. CLASSICAL IE AND RULE-BASED METHODS

IE is one of the mature areas of NLP and has been described from both linguistic and database perspectives: on the one hand, the task is to recognize relevant fragments of meaning in text; on the other, to record them in a structure that can later be indexed, filtered, and combined with other data sources [1], [8]. In practice, rule-based methods remain attractive because they offer high controllability, low requirements for training data, and the possibility of precisely tracing the causes of errors. This point is well documented both in the polemical paper by Chiticariu, Li, and Reiss and in experience with tools such as UIMA Ruta [2], [3].

For the domain of real estate listings, the main advantage of the rule-based approach is its high effectiveness in fields based on repetitive patterns: prices, plot size, area units, and standard abbreviations such as "m2," "ha," "PLN," "thous.," or "WZ." Its limitation, however, is the rapid growth of maintenance cost as new language variants emerge. Rules perform well as simple value detectors, but

much worse in situations where meaning depends on negation, spatial relations, or several dispersed sentences describing the same feature.

1.2. GENERATIVE LLMS AND THE REAL ESTATE DOMAIN

In recent literature, IE tasks are increasingly carried out within the generative paradigm, in which a language model receives the source text together with an instruction to return information in a structured format. The latest surveys emphasize that the key issues are not only the model architectures themselves, but also prompt design, output-format control, and response-validation mechanisms [4], [5]. From a practical perspective, this shifts the emphasis from manually constructing large numbers of rules to designing the task interface and the post-processing stages.

In the real estate context, the task takes the form of attribute-value extraction. Studies from 2025 show that LLMs can effectively recover listing attributes even when the information is not expressed in a single phrase but follows from the context of the entire description. At the same time, domain studies stress the need to evaluate computational cost, response stability, and susceptibility to formatting errors, rather than relying exclusively on metrics used to assess language-model quality [6], [7]. For this reason, in the present article the LLM is treated not as a stand-alone extractor, but as a component of a larger pipeline with a strictly defined input and output schema.

The choice of Bielik as the local model is motivated by two considerations. First, it is a family of open models developed with Polish in mind, which increases the chance of correctly understanding domain-specific phrasing and abbreviations. Second, the technical documentation and the model card support an offline deployment scenario, which is important in environments where the processed texts may contain contact details or other information that should not be sent to external services [12], [13].

2. SYSTEM ARCHITECTURE AND EXTRACTION SCHEMA

2.1. ETL PIPELINE AND COMPONENT INTEGRATION

The solution architecture was designed as a multi-stage pipeline in which data acquisition, information extraction, and preparation of data for reporting are logically separated. On a VPS server, a cron scheduler launches a scraper that retrieves listings from listing portals and stores the result as raw data. The records are then transferred via SCP to a Jupyter environment running the local Bielik model. After inference is completed, the results return to the server as processed data and undergo a final

normalization stage into a data frame stored in Parquet format and subsequently exposed to the Power BI reporting layer. The entire flow is shown in Fig. 1.

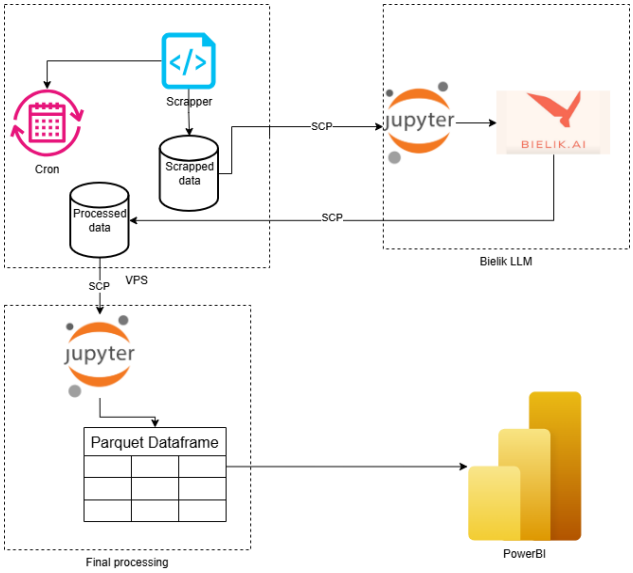


Fig. 1. Architecture of the listing acquisition and processing pipeline: scraping, data transfer, information extraction using a local LLM, and preparation of data for reporting.

From an information-systems perspective, the most important architectural decision is the explicit definition of interfaces between stages. The input record should contain the source text and minimal metadata, the extraction result must comply with the JSON schema, and the normalized data already take the form of Parquet files. Such separation reduces maintenance cost, facilitates reprocessing of historical data, and allows individual architectural elements to be replaced without disturbing the analytical logic, for example on the data-warehouse and dashboard side.

2.2. NORMALIZATION AND DATA STRUCTURE

The formal extraction schema is the core of the entire solution. The study adopts three groups of attributes: numerical, categorical, and multi-label. Numerical attributes describe quantities normalized to common units, categorical attributes use closed-class dictionaries, and multi-label attributes represent sets of co-occurring features, such as utilities or surrounding-area characteristics. Table 1 presents the proposed set of core attributes and their normalization rules.

Table 1. Definition of the extraction schema: attributes, data types, example values, and normalization rules.

Attribute	Data type	Example values	Normalization rule
PRICE	number	PLN 249,000; 249 thousand	PLN; remove separators; convert 'thousand' to 000.
AREA	number	850 m2; 0.085 ha	Convert to m²; canonical decimal format.
PRICE_M2	number	PLN 280/m2	PLN/m²; auxiliary computation from price and area.
TYPE	category	building, agricultural, investment	Map synonyms to a closed dictionary.
PLAN	category	WZ, MPZP, none	Classes {WZ, MPZP, none, unknown}.
LAW_STATUS	category	ownership, co-ownership	Map to a legal-status dictionary.
MEDIA	set of labels	electricity, water, gas, sewerage	Handle co-occurrence and negation; output as a set.
ROAD_TYPE	category	asphalt, improved, dirt	Normalize to basic classes.
NB_SEGMENT	set of labels	forest, suburbs, center	Map to a contextual-feature dictionary.

The schema defined in this way plays a dual role. On the one hand, it is an interface specification between extraction and the analytical layer; on the other, it is a tool that constrains the generative freedom of the LLM. In practice, this means that every field must have a defined semantics of missing data, an allowed type, and a normalization procedure. Particularly important is the distinction between null and false in multi-label attributes: the absence of a mention of sewerage does not mean that sewerage is unavailable. This detail has a direct impact both on quality assessment and on the reliability of subsequent analyses.

The schema-first approach is consistent with observations from the literature on generative IE and with domain studies on real estate attributes, in which imposing an output schema and limiting the set of

fields leads to fewer structurally incorrect responses and facilitates model comparison according to common criteria [4], [5], [7].

3. MATERIALS AND EVALUATION PROTOCOL

3.1. INPUT CORPUS AND ANNOTATION

The input material consists of titles, descriptions, and selected metadata from Polish-language real estate listings collected periodically from listing portals. In the presented architecture, they come from Łomża and its surrounding area, but location serves here only as a marker of the linguistic domain and not as an object of economic analysis. The research protocol assumes sampling from different periods and listing types so as to include both concise and elaborate formulations, and both carefully written and colloquial texts. The recommended minimum variant includes 300-600 records intended for building the gold standard, because this range already allows separate evaluation of numerical, categorical, and multi-label attributes.

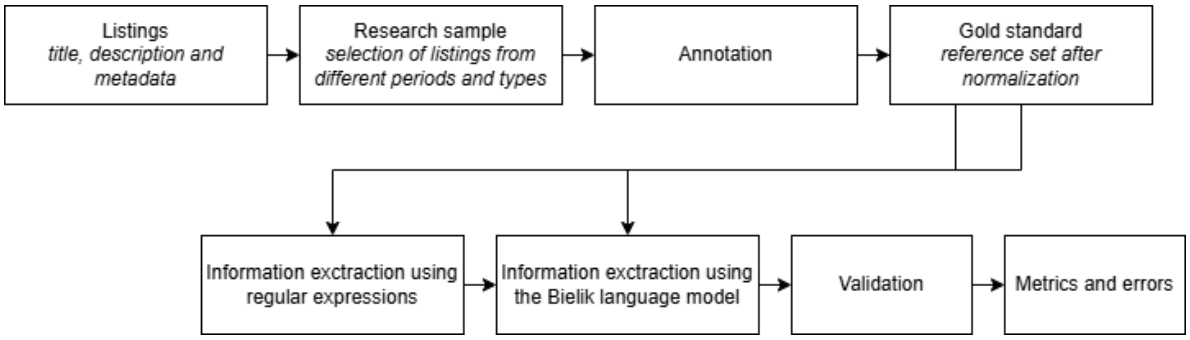


Fig. 2. Evaluation-process diagram: input data, gold-standard annotation, rule-based and LLM extraction, normalization, and computation of quality metrics and errors.

Each record should be annotated independently by two annotators according to guidelines defining field boundaries, rules for interpreting abbreviations, handling negation, and unit-normalization procedures. For categorical and multi-label attributes, Cohen's kappa is an appropriate agreement measure because it corrects agreement for chance, whereas for numerical attributes one should report agreement after normalization or the proportion of matching extractions within a tolerance determined by the unit [9], [10], [11]. Disputed cases require an adjudication procedure; its result constitutes the final gold standard against which quality metrics are computed.

3.2. COMPARED METHODS AND COMPUTING ENVIRONMENT

To avoid the criticism of a weak baseline, the comparison includes two levels of rule-based extraction. The baseline Regex variant relies on regular expressions and simple detectors of units, prices, numbers, and abbreviations. The Hybrid-Regex variant extends this set with synonym dictionaries, contextual heuristics, and priority rules that resolve situations such as "electricity in the access road" versus "electricity on the plot" or "no sewerage" versus "sewerage nearby." The LLM system uses Bielik in local mode, executed with deterministic inference settings and a schema-first prompt that forces the output to be returned as JSON compliant with Table 1 [12], [13].

For reproducibility, each run should report the model version, versions of Python libraries, parameters of the computing unit, amount of available memory, quantization method, and mean processing time per record. Such reporting is especially important in the case of local LLMs, because the same extraction logic may be executed on CPU, GPU, or in quantized form, which substantially affects latency and response stability.

3.3. QUALITY AND COST METRICS

Quality assessment is based on the classical measures precision, recall, and F1 [8]. Let TP denote correctly extracted values, FP values added incorrectly, and FN omitted values. Then precision = TP/(TP + FP), recall = TP/(TP + FN), and F1 = 2PR/(P + R). For multi-label attributes, it is appropriate to report micro-F1 at the level of all labels, whereas practical interpretation also requires per-attribute results.

At the level of dataset usability, however, classification metrics alone are not sufficient. It is also necessary to report attribute coverage, that is the proportion of records with a non-empty field value; valid output rate, that is the proportion of records returned in a correct schema; normalization success, that is the share of numerical values that can be unambiguously reduced to canonical form; and cost metrics such as latency, throughput, and variance across repeated runs. Only such a set of measures makes it possible to compare rule-based solutions and LLMs fairly under deployment conditions.

4. DISCUSSION

The proposed approach makes it possible to formulate verifiable hypotheses even before the full benchmark is executed. First, rules should retain an advantage where an attribute corresponds to a regular numerical pattern and correctness depends mainly on recognizing the unit and separators.

Second, the language model should achieve better results for attributes whose value is distributed across several sentences or revealed indirectly, such as planning information, utilities, or surrounding-area characteristics. Third, the strongest deployment candidate appears to be a hybrid architecture in which rules handle numbers and simple classes, while the LLM takes over semantic fields together with output validation [2], [3], [4], [5], [6], [7].

The main threats to validity concern domain drift and annotation ambiguity. Listing portals change publication style, and users systematically introduce new abbreviations, new combinations of units, and marketing formulations that mask the actual characteristics of the listing. An additional problem is the fuzzy boundary between absence of information and negation of a feature; without explicit annotation guidelines, even high percentage agreement may be misleading, which is why reporting Cohen's kappa and describing the adjudication procedure are essential [9], [10], [11]. In the case of LLMs, the effects of quantization, changes in model version, and residual inference non-determinism must also be controlled [12], [13].

From a systems-engineering perspective, it is crucial to treat the parser and validator as integral parts of the extractor. The generative model alone does not yet return data ready for analysis: the output must be checked syntactically, mapped to class dictionaries, converted into common units, and separated from information irrelevant to the schema. Otherwise, even isolated formatting errors may block batch processing or contaminate the reporting layer. Local inference additionally reduces the risk of sending content containing contact data to external services, which in practice constitutes an important architectural argument.

The described infrastructure is suitable not only for a single experiment. The same pipeline can be used for ablation studies, comparison of free-form and schema-first prompts, assessment of the impact of normalization dictionaries, or comparison of CPU and GPU execution. Thus, the article provides not only a one-off description of a system implementation, but also an experimental framework enabling systematic study of Polish-language IE in the domain of real estate listings.

CONCLUSIONS

The article presents a complete information-system architecture and research procedure for information extraction from Polish real estate listings. The proposed solution combines periodic data acquisition, a formal data schema, local inference using the Bielik model, a validation and normalization stage, and final storage in Parquet format with a view to reporting in BI tools. The most important scientific

contribution here is not a single numerical result, but a reproducible environment in which rule-based, hybrid, and LLM solutions can be evaluated according to the same criteria.

The next stage of the work will involve conducting a benchmark on a manually annotated reference set, reporting results separately for numerical, categorical, and multi-label attributes, and analyzing errors of format, normalization, and semantics. Only an experiment prepared in this way will make it possible to determine under which conditions local Bielik outperforms rules, and under which conditions a hybrid approach remains more cost-effective. From a computer-science perspective, however, the proposed pipeline already constitutes a valuable pattern for building secure and scalable IE systems for domain-specific data.

REFERENCES

[1]. S. Sarawagi, “Information Extraction,” Foundations and Trends in Databases, vol. 1, no. 3, pp. 261–377, 2008. doi: 10.1561/1900000003.

[2]. L. Chiticariu, Y. Li, and F. R. Reiss, “Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!,” in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 2013, pp. 827–832. ACL Anthology.

[3]. P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, and F. Puppe, “UIMA Ruta: Rapid development of rule-based information extraction applications,” Natural Language Engineering, vol. 22, no. 1, pp. 1–40, 2016. doi: 10.1017/S1351324914000114.

[4]. D. Xu, W. Chen, W. Peng, et al., “Large language models for generative information extraction: a survey,” Frontiers of Computer Science, vol. 18, art. no. 186357, 2024. doi: 10.1007/s11704-024-40555-y.

[5]. Z. Zhang, W. You, T. Wu, X. Wang, J. Li, and M. Zhang, “A Survey of Generative Information Extraction,” in Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 4840–4870. ACL Anthology.

[6]. R. A. Principe, M. Viviani, and N. Chiarini, “Enhancing Information Extraction with Large Language Models: A Comparison with Human Annotation and Rule-Based Methods in a Real Estate Case Study,” in Proceedings of the 5th Conference on Language, Data and Knowledge, Naples, Italy, 2025, pp. 243–254. ACL Anthology.

[7]. M. Kvet, M. Potocar, and S. Tatarka, “Real Estate Attribute Value Extraction Using Large Language Models,” IEEE Access, vol. 13, pp. 73076–73095, 2025. doi: 10.1109/ACCESS.2025.3564511.

[8]. C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.

[9]. R. Artstein and M. Poesio, “Survey Article: Inter-Coder Agreement for Computational Linguistics,” Computational Linguistics, vol. 34, no. 4, pp. 555–596, 2008. doi: 10.1162/coli.07-034-R2.

[10]. J. Cohen, “A Coefficient of Agreement for Nominal Scales,” Educational and Psychological Measurement, vol. 20, no. 1, pp. 37–46, 1960. doi: 10.1177/001316446002000104.

[11]. M. L. McHugh, “Interrater reliability: the kappa statistic,” Biochemia Medica, vol. 22, no. 3, pp. 276–282, 2012. doi: 10.11613/BM.2012.031.

[12]. K. Ociepa, Ł. Flis, R. Kinas, K. Wróbel, and A. Gwoździej, “Bielik v3 Small: Technical Report,” arXiv preprint arXiv:2505.02550, 2025. doi: 10.48550/arXiv.2505.02550. arXiv.

[13]. SpeakLeash, “Bielik-1.5B-v3,” Hugging Face model card, 2025. [Online]. Available: Hugging Face model card. Accessed: Mar. 16, 2026.

Jakub Bednarczyk: https://orcid.org/0009-0009-4941-3758

Marta Chodyka: https://orcid.org/ 0000-0002-8819-2451

DESIGN AND EVALUATION OF A REAL ESTATE MARKET MONITORING SYSTEM ARCHITECTURE USING AN ETL PIPELINE AND LARGE LANGUAGE MODELS

JAKUB BEDNARCZYK

University of Łomza, Poland

jbednarczyk@al.edu.pl

MARTA CHODYKA

University of Łomza, Poland

mchodyka@al.edu.pl

ABSTRACT: This article presents the design and evaluation of the architecture of an automated real estate market monitoring system that combines web scraping, local inference using the Bielik model, schema validation, normalization, time-series aggregation, and publication of results in Power BI. The contribution is systems-engineering in nature: a data-processing workflow, data interfaces between components, and a supervisory layer overseeing dataset completeness and continuity are proposed. The case study is based on real artifacts produced by the pipeline: a dataset of 636 cleaned listings with 15 attributes, a dataset of 124 daily time-series observations, and a Power BI reporting model. The empirical evaluation showed 100.0% completeness of required fields, 91.6% average completeness of key analytical fields, 98.9% geographic coverage, 80.0% of records suitable for direct analysis, and full continuity of the time series within a 124-day window. At the same time, a strong dependence of quality on the source, as well as a minor yet important problem of cross-platform duplicates, was identified. The results confirm that an ETL + local LLM + data-quality supervision architecture is practical for Polish-language data, but it should be extended with telemetry logging of per-stage timings and repeatability tests of language-model inference.

Keywords: real estate market monitoring, ETL, data quality, MLOps, LLM, Bielik, Power BI, time series, data observability

INTRODUCTION

Automating real estate market monitoring today requires not only cyclical acquisition of listings, but also the maintenance of a stable processing chain in which textual data are enriched, normalized, aggregated, and made available in an analysis-ready form. In practice, this means combining the classic Extract-Transform-Load approach with semantic extraction components and with a reporting layer that must receive data in a fixed schema and of predictable quality [1], [2]. In systems of this type, the

analytical result itself is secondary to the architecture: even accurate predictive models or correctly identified attributes lose value if the pipeline is vulnerable to source failures, schema drift, or inconsistencies between stages.

From a data-engineering perspective, issues of quality, completeness, timeliness, and process observability therefore become central. The literature on ETL and data quality emphasizes that data should be assessed not only in terms of accuracy, but also completeness, timeliness, unambiguity, and usefulness in a specific task context [3], [4]. At the same time, modern data quality validation platforms treat dataset control in a manner similar to unit tests for code, which makes it possible to transfer the rigor of software engineering to the data layer [5]. In the case of the real estate market, the problem is further complicated by source heterogeneity: different services publish different sets of metadata, different geolocation quality, and different degrees of disclosure of economic parameters.

In this study, the main object of investigation is not the market analysis itself, but the design and evaluation of the architecture of a system capable of monitoring that market automatically. The language model is treated as one stage of data enrichment rather than as an autonomous analytical system. This approach is consistent with the observation that LLMs in production environments introduce additional technical debt, especially when they are not accompanied by explicit input and output interfaces, version control, and monitoring of data quality and outcomes [6]. The paper addresses three research questions: how to design a pipeline resilient to source variability, how to formally measure data quality and readiness for reporting, and how to integrate a local language model with classic ETL without losing control of the process.

1. RESEARCH BACKGROUND AND ENGINEERING PERSPECTIVE

1.1. ETL, DATA QUALITY, AND PIPELINE OBSERVABILITY

Classic ETL remains the basic pattern for organizing data flow in warehousing and analytics systems. Kimball and Caserta frame ETL as a set of practices encompassing extraction, cleaning, conformance, and delivery of data to the presentation layer, with the most difficult part of the process proving to be not the transfer itself but maintaining semantic consistency across multiple sources [1]. Later surveys of ETL technology, in turn, emphasized that the entire process should be considered not as a simple batch script, but as an architecture encompassing flow modeling, schema evolution, transformation rules, and data refresh mechanisms [2]. For real estate market monitoring, this means the need to separate the data acquisition zone, the semantic enrichment zone, and the analytical zone.

In evaluating such a system, data quality metrics play a particularly important role. Wang and Strong showed that data users assess quality multidimensionally, and that completeness, timeliness, and representational clarity are as important as accuracy sensu stricto [3]. Extensions of this perspective in more recent literature systematize both the quality dimensions themselves and the methods for measuring and improving them, from deduplication to consistency control and the detection of integration errors [4]. In practice, observability in large pipelines additionally requires the implementation of automatic control tests executed with each processing run; precisely this approach is represented by the line of work on automated large-scale data quality verification [5].

1.2. LLM AS A COMPONENT OF AN ANALYTICAL SYSTEM

The inclusion of large language models in data-engineering systems changes the nature of the architecture. Here, the LLM does not replace ETL; rather, it acts as a module enriching the dataset with information that is difficult to obtain using rules alone, especially when meaning is dispersed across the entire description or requires the interpretation of abbreviations, negations, and context [7], [8]. In the real estate domain, this approach has already been empirically demonstrated both for auction documents and for extracting attribute values from listings, with the results of these studies indicating that the advantage of LLMs depends on prompt quality, output-schema control, and the operational cost of the entire solution [9], [10].

From an architectural perspective, the variant with a local language model is particularly interesting. The Bielik family was developed with the Polish language in mind, which makes it a natural candidate for processing domestic real estate listings containing proper names, technical abbreviations, and colloquial notation [11]. Running the model locally reduces the risk of data leakage to external services and facilitates full control over the environment, but at the same time imposes additional requirements: response-format compliance must be declared, stability across successive runs must be ensured, the impact of model versions must be monitored, and quality dependence on hardware load must be tracked. In this sense, the LLM becomes an MLOps element rather than merely an NLP tool [6], [11].

2. MATERIALS AND ARCHITECTURE OF THE SYSTEM UNDER STUDY

2.1. CHARACTERISTICS OF THE STUDY MATERIAL

The study material used in the article comprises three elements generated by the operational pipeline. The first is a cleaned listings dataset stored in Parquet format, containing 636 records and 15 attributes describing, among others, the identifier, date added, last update date, area, price, price per square

meter, location, source, distance from the main city, size segment, days on market, and listing status. This dataset covers observations from 18 September 2025 to 14 March 2026 and comes from five separate source services. The second element is the time-series layer for Power BI, containing 124 records with nine metrics, such as the number of active listings, mean and median price, mean and median price per square meter, mean and median area, and the total area currently on the market. The third element is the Power BI reporting model, which is the final consumer of the data layer.

The material selected in this way makes it possible to assess the system not from a theoretical perspective, but on the basis of its intermediate and final products. The first dataset represents the listings-data level, that is, everything the pipeline can extract, normalize, and maintain in a consistent schema. The second dataset represents the time series, which makes it possible to compare the extracted data at daily granularity. The final stage confirms that the data were in fact prepared in a form useful for a Business Intelligence tool, which means that the architecture does not end with extraction, but covers the full cycle from web scraping to publication of the reporting layer.

2.2. REFERENCE ARCHITECTURE AND DATA INTERFACES

The system architecture is presented in Fig. 1. The solution was divided into four zones: data acquisition on a VPS server, text-data processing using an LLM, normalization and export of files to Parquet format, and the reporting layer. In the acquisition zone, a cron schedule launches a scraper that retrieves listings and stores them in a filesystem-based database. The data package is then transferred to a Jupyter environment, where the local Bielik model performs data processing. The inference result is not sent directly to the dashboard; it first undergoes schema validation, field typing, deduplication, and data aggregation into a time-series form. Only the normalized data are published as two Parquet tables: a listings table and a time-series table.

The architecture was intentionally organized according to the principle of separation between stages. Each zone boundary corresponds to an explicitly defined data interface: raw data, Bielik-processed data, validated data, and analytical data. This approach limits the risk of error propagation, facilitates modifications to an individual component without rebuilding the entire system, and supports the processing of historical data. The final data format remains consistent with the practice of BI analytics systems, where separating the data-processing layer from the reporting layer increases reporting stability [12].

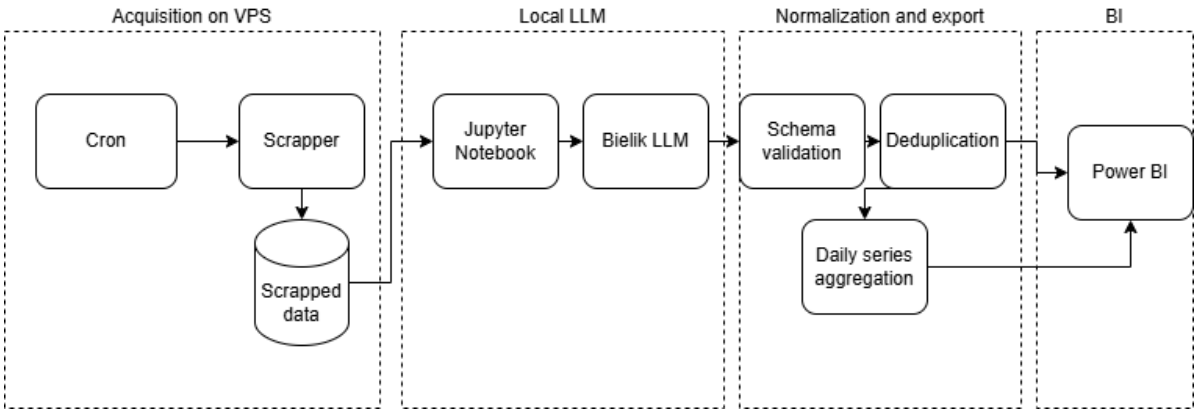


Fig. 1. Reference architecture of the monitoring system: data acquisition, local LLM enrichment, validation, columnar export, and reporting layer.

3. ARCHITECTURE EVALUATION METHODOLOGY

3.1. FORMALIZATION OF QUALITY AND OPERATIONAL METRICS

The evaluation of the architecture was based on a set of metrics combining the data-quality perspective with the system's operational perspective. Let D = {r1, ..., rN} denote the set of records, and A the set of monitored attributes. For a single attribute a, completeness was defined as the proportion of records in which the field has a non-empty value. Averaging this measure over the set of attributes yields the synthetic completeness of the analytical layer.

$$C_a = (1 / N) \cdot \Sigma(i=1..N)\ I(v_{i,a} \neq \varnothing), \qquad C_D = (1 / |A|) \cdot \Sigma(a \in A)\ C_a$$

For the monitoring of daily series, continuity over time is equally important. If T_obs denotes the set of days actually present in the time-series layer and T_exp the set of days expected in the analyzed interval, then series continuity can be written as the ratio of the number of observed timestamps to the number of expected timestamps. This measure indicates whether the dashboard can operate without imputing missing dates and without additional corrections to the time axis.

$$S_t = |T_{obs}| / |T_{exp}|$$

In multi-source architectures, it is also necessary to measure duplicate redundancy. Accordingly, the measure U is defined as 1 minus the ratio of the number of unique identifiers to the total number of records:

$$U = 1 - |ID_{uniq}| / N$$

3.2. CASE STUDY PROCEDURE

The evaluation procedure proceeded on two levels. First, the final output datasets of the pipeline were analyzed by calculating completeness of required fields, completeness of key fields, the level of record readiness for further analysis, duplicate rate, and continuity of the time-series layer. Next, the architecture was assessed qualitatively by verifying whether the system design made it possible to localize points of data degradation and whether the source of the problems was the scraping stage, the data-extraction stage using the language model, validation, or the heterogeneity of the input portals themselves. In this sense, the evaluation has the character of a case study: what is verified is the practical ability of the architecture to maintain correct and useful data, rather than solely the correctness of the algorithm itself.

4. CASE STUDY RESULTS

4.1. DATA-LAYER QUALITY AND RECORD READINESS FOR ANALYSIS

Table 1 synthesizes the most important results of the quantitative evaluation of the data layer. The clearest result is 100.0% completeness of required fields, which means that all records have an identifier, operational dates, source, and market status. At the level of key analytical fields, the average completeness was 91.6%, which should be regarded as sufficient for most descriptive and dashboard analyses. At the same time, geographic coverage reached 98.9%, whereas coverage of fields containing economic data was lower, at 83.0%. As a result, 80.0% of records can be considered fully ready for further analysis without additional data-completion procedures.

The duplicate rate identified on the basis of the ID field was 0.63%. This value is low from the perspective of a single dataset, but it has architectural significance because it indicates the existence of listings published in parallel across several sources or refreshed in a way that leads to identifier replication. The very presence of a deduplication stage in the system design therefore proves justified not hypothetically, but empirically.

Table 1. Empirical indicators of analytical-layer quality determined on the basis of the delivered output datasets.

Metric	Operational definition	Result	Interpretation
Required-field completeness	ID, DATE_ADDED, LAST_UPDATED, SOURCE, MARKET_STATUS, and DAYS_ON_MARKET contain no missing values.	100.0%	Each record contains the full set of operational metadata required for identification and tracking.
Average completeness of analytical fields	Average over the set {PRICE, AREA_M2, PRICE_M2, CITY, LAT, LON, MAIN_CITY_DIST, SIZE_SEGMENT}.	91.6%	The analytical layer is largely ready for reporting, but quality is not uniform across sources.
Geographic coverage	Average completeness of CITY, LAT, LON, and MAIN_CITY_DIST.	98.9%	Spatial and map-based analyses are possible without extensive imputation.
Economic coverage	Average completeness of PRICE, AREA_M2, and PRICE_M2.	83.0%	This is the main area of quality degradation; missing values in these fields limit some dashboard indicators.
Records fully ready for analysis	Share of records for which PRICE, AREA_M2, PRICE_M2, CITY, LAT, LON, and MAIN_CITY_DIST are complete.	80.0%	One in five records would require additional completion, exclusion, or special handling in the dashboard.
Duplicate rate	Measure U for record identifiers.	0.63%	The problem is small but real; it confirms the need for a cross-source deduplication stage.
Continuity of the daily series	Share of days present in the time-series layer relative to the number of days expected in the analyzed interval.	100.0%	The dashboard can be fed without filling in missing dates or correcting the time axis.

4.2. STABILITY OF THE TIME-SERIES LAYER AND READINESS FOR REPORTING

Figure 2 shows the time-series layer prepared for the dashboard. From the perspective of evaluating the architecture, the key issue here is not the price level, but the absence of gaps in the time axis. In the analyzed window from 11 November 2025 to 14 March 2026, 124 consecutive daily observations were obtained, giving 100.0% series continuity. This means that the pipeline effectively supplies the reporting layer with consistent daily snapshots and does not require artificial completion of missing dates on the Power BI side. Such a property is critical for monitoring systems, because breaks in the series distort both the trend and derived indicators.

The variability of the indicators itself serves here as a test of export integrity, rather than as the subject of economic interpretation. The number of active listings varies during the observed period from 43 to 172, while the median price per square meter lies in the range of 89.0-116.9 PLN/m².

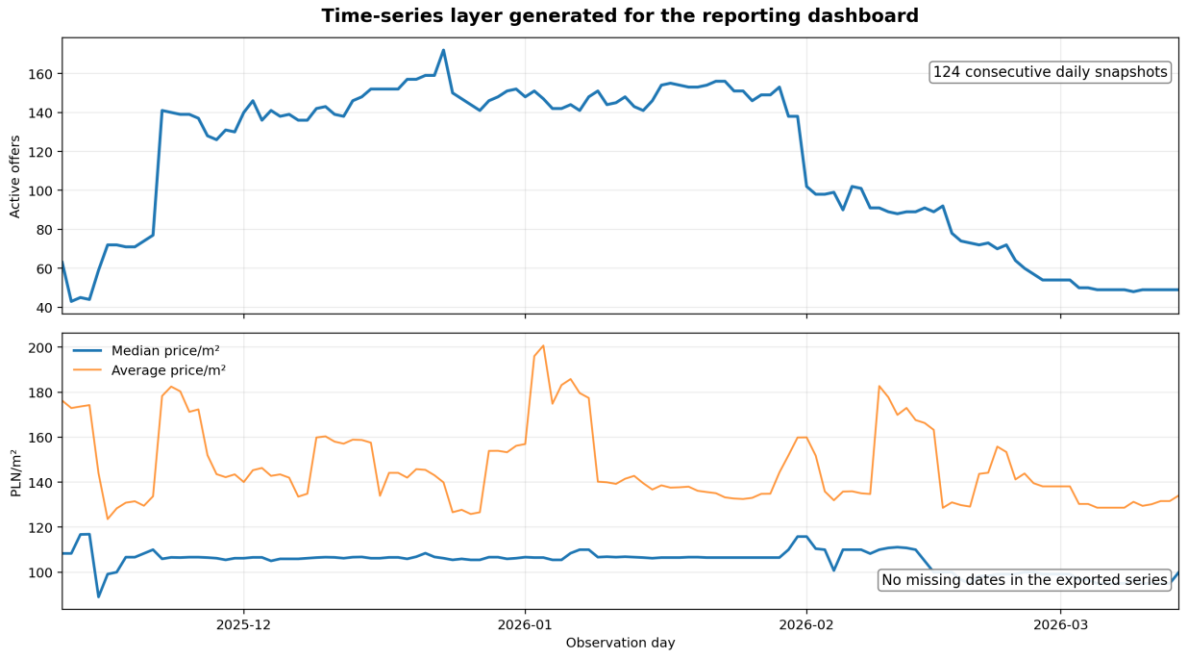


Fig. 2. Time-series layer prepared for the dashboard: number of active listings and mean and median price per square meter on successive observation days.

4.3. SOURCE HETEROGENEITY AS AN ARCHITECTURAL PROBLEM

The strongest diagnostic criterion proved to be the source-wise completeness analysis presented in Fig. 3. Three source services provide 100.0% completeness for all monitored key fields, whereas one source exhibits clearly poorer quality in the economic part of the record: completeness for price and price per square meter drops there to 44.3%, and completeness for area to 62.5%. At the same time, that same

source maintains high completeness for geographic fields, at 96.4%. The architectural conclusion is unambiguous: the problem is not global, but source-level.

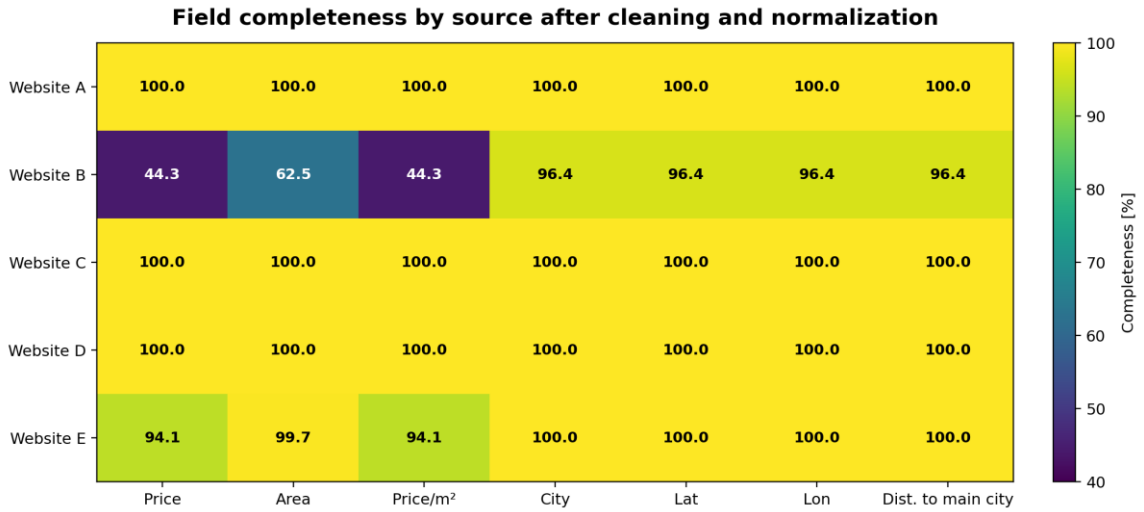


Fig. 3. Completeness of key fields by source after cleaning and normalization. The greatest degradation concerns one source and is concentrated in fields containing economic data rather than geographic data.

5. DISCUSSION

The results obtained show that the greatest value of the designed system is not the use of an LLM itself, but the clear separation between stages. The scraper is responsible for record acquisition, the local model for data extraction, data validation for representing the result in an explicit schema, and the analytical layer for delivering processed Parquet files for reporting. This division reduces the risk that any change in the model will cause a failure of the entire system. This is particularly important in the context of LLMs, whose behavior may change due to factors such as model version, quantization, or parameter settings [6], [8], [11].

At the same time, the case study reveals two important limitations of the current version of the solution. The first is the lack of persistent telemetry for stage timings, which still prevents rigorous measurement of end-to-end latency and data-processing cost. The second is the absence of a built-in repeatability test for the LLM module (the so-called LLM-as-a-judge), and thus the lack of knowledge as to whether reprocessing the same sample yields an identical structural result. From a practical standpoint, this means that the current evaluation is complete at the level of output data, but only partial at the operational level. In the next iteration of the architecture, timestamp logging for each stage, JSON validation error counters, and a control sample for repeated inference should be added.

The results also suggest that the most rational direction of further development will be a hybrid architecture. Domain studies indicate that LLMs perform well on semantically dispersed information, but regular and numerical fields can often be handled more cheaply and more predictably by rules or dedicated parsers [9], [10]. In the system under study, this would mean retaining the LLM for semantically difficult fragments while leaving simple numerical attributes to rules and validators. Such a solution strengthens both the robustness of the pipeline and its cost efficiency.

CONCLUSIONS

This article proposed and evaluated the architecture of an automated real estate market monitoring system combining web scraping, local inference using the Bielik model, schema validation, normalization, time-series aggregation, and publication in Power BI. The most important contribution of the work is the formalization of system-evaluation metrics and their demonstration on real data generated by the pipeline. The case study confirmed full continuity of the time series, high completeness of geographic and operational data, overall completeness of the analytical layer, and source-related loci of quality degradation.

From an IT perspective, this means that the designed pipeline can serve as a foundation for work in data engineering, analytical systems, and MLOps. Further development should include instrumentation of latency and data-processing cost, periodic repeatability tests of the LLM module, automatic source-drift alerts, and a comparison of the fully local variant with a hybrid architecture in which rules and the language model divide tasks according to attribute type. Such a development path leads from automated data extraction to a full-fledged market observability system.

REFERENCES

[1]. R. Kimball and J. Caserta, The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Indianapolis, IN: Wiley Publishing, Inc., 2004.

[2]. P. Vassiliadis, “A Survey of Extract-Transform-Load Technology,” International Journal of Data Warehousing and Mining, vol. 5, no. 3, pp. 1–27, 2009. doi: 10.4018/jdwm.2009070101.

[3]. R. Y. Wang and D. M. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” Journal of Management Information Systems, vol. 12, no. 4, pp. 5–33, 1996. doi: 10.1080/07421222.1996.11518099.

[4]. C. Batini and M. Scannapieco, Data and Information Quality: Dimensions, Principles and Techniques. Cham: Springer, 2016. doi: 10.1007/978-3-319-24106-7.

[5]. S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Bießmann and A. Grafberger, “Automating Large-Scale Data Quality Verification,” Proc. VLDB Endow., vol. 11, no. 12, pp. 1781–1794, 2018. doi: 10.14778/3229863.3229867.

[6]. D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo and D. Dennison, “Hidden Technical Debt in Machine Learning Systems,” in Advances in Neural Information Processing Systems 28, 2015, pp. 2503–2511.

[7]. S. Sarawagi, “Information Extraction,” Foundations and Trends in Databases, vol. 1, no. 3, pp. 261–377, 2008. doi: 10.1561/1900000003.

[8]. D. Xu, W. Chen, W. Peng, et al., “Large language models for generative information extraction: a survey,” Frontiers of Computer Science, vol. 18, art. no. 186357, 2024. doi: 10.1007/s11704-024-40555-y.

[9]. R. A. Principe, M. Viviani and N. Chiarini, “Enhancing Information Extraction with Large Language Models: A Comparison with Human Annotation and Rule-Based Methods in a Real Estate Case Study,” in Proceedings of the 5th Conference on Language, Data and Knowledge, Naples, Italy, 2025, pp. 243–254.

[10]. M. Kvet, M. Potočár and S. Tatarka, “Real Estate Attribute Value Extraction Using Large Language Models,” IEEE Access, vol. 13, pp. 73076–73095, 2025. doi: 10.1109/ACCESS.2025.3564511.

[11]. K. Ociepa, Ł. Flis, R. Kinas, K. Wróbel and A. Gwoździej, “Bielik v3 Small: Technical Report,” arXiv preprint arXiv:2505.02550, 2025. doi: 10.48550/arXiv.2505.02550.

[12]. S. Melnik, A. Gubarev, J. Long, G. Romer, S. Shivakumar, M. Tolton and T. Vassilakis, “Dremel: Interactive Analysis of Web-Scale Datasets,” Proc. VLDB Endow., vol. 3, no. 1–2, pp. 330–339, 2010.

Jakub Bednarczyk: https://orcid.org/0009-0009-4941-3758

Marta Chodyka: https://orcid.org/ 0000-0002-8819-2451

CLOUD-EDGE CONTINUUM IN AI-ENABLED SMART HOME SYSTEMS: PERFORMANCE, PRIVACY AND RELIABILITY TRADE-OFFS

MARTA CHODYKA

University of Łomza, Poland

mchodyka@al.edu.pl

GABRIEL TARASIUK

John Paul II University in Biała Podlaska, Poland

tarasiuk22914@stud.akademiabialska.pl

ABSTRACT: The integration of artificial intelligence into smart home systems has transformed residential Internet of Things infrastructures from sets of remotely controlled devices into distributed cyber-physical environments capable of autonomous sensing, inference and control. This article presents a comparative analysis of cloud-based and edge-based AI architectures for smart home applications. The study focuses on five analytical criteria: latency, reliability, privacy, energy efficiency and computational scalability. The analysis indicates that edge AI is particularly suitable for latency-sensitive and privacy-sensitive tasks because it enables local inference, data minimization and operational continuity during connectivity degradation. Cloud computing remains necessary for elastic scalability, global model training and long-term aggregated analytics. The article argues that the most robust architectural model is not a binary selection between cloud and edge processing, but a cloud-edge continuum in which tasks are dynamically assigned according to their temporal constraints, data sensitivity and computational requirements. Federated learning, local model compression and neural processing units are identified as key technological mechanisms supporting this transition.

Key words: smart home, artificial intelligence, edge AI, cloud AI, Internet of Things, latency, federated learning, privacy by design

INTRODUCTION

Smart home systems have moved beyond the model of isolated remotely controlled devices and now operate as dense networks of sensors, cameras, thermostats, speakers, meters, actuators and wearable devices. These components continuously record environmental, behavioral and sometimes biometric signals, transforming the home into a data-intensive Internet of Things environment. Recent analyses cited in the literature indicate that the number of IoT devices has reached approximately 19.8 billion and may rise to 40.6 billion by 2034, while the global volume of data generated by such infrastructures has reached the scale of zettabytes [1]. In residential environments, this growth creates a

practical engineering challenge: intelligent behavior can only be achieved when telemetry is processed fast enough to support reliable control decisions [2].

The classical smart home architecture relies on cloud computing. In this model, sensing devices acquire raw signals and transmit them to external data centers, where machine learning models perform classification, prediction or decision-making. The cloud model provides elastic computing capacity and enables the execution of complex models without imposing high hardware requirements on the home device. However, the same architecture introduces three major constraints: transmission latency, dependence on external connectivity and reduced control over raw data streams [3], [5], [6]. These constraints are especially important in safety-critical domestic applications, such as fall detection, fire detection, gas leakage recognition, emergency power cutoff and health-related monitoring [2], [11].

Edge AI has emerged as a response to these limitations. Instead of sending all raw signals to a remote data center, edge AI executes trained models on the end device, a local microcontroller, a local hub or a small residential server. This architectural shift reduces the path that data packets must travel, limits the amount of sensitive information leaving the building and allows selected control loops to continue functioning when the cloud connection is temporarily unavailable [5], [7], [8]. Nevertheless, moving inference to the edge is not cost-free. Edge devices operate under stricter memory, energy and thermal constraints, which requires model compression, quantization, pruning and hardware-aware optimization [2], [4], [7].

The purpose of this article is to compare cloud and edge AI architectures in smart home systems and to determine where the boundary of a rational architectural choice lies. The central thesis is that the design problem should not be framed as a simple opposition between cloud and edge computing. A more adequate model is the cloud-edge continuum, in which latency-sensitive and privacy-sensitive tasks are handled locally, while tasks requiring elastic scalability, global analytics or extensive training capacity are delegated to the cloud [1], [8], [10], [14].

1. METHODOLOGICAL ASSUMPTIONS AND ANALYTICAL FRAMEWORK

This article adopts a comparative analytical approach based on a narrative review of current literature on smart home systems, edge AI, cloud computing, distributed IoT architectures, federated learning and privacy-preserving artificial intelligence. The article does not report original laboratory experiments. Instead, it synthesizes findings, benchmark values and architectural arguments reported in the cited literature and applies them to the design context of AI-enabled smart homes.

The comparison is structured around five analytical criteria. The first criterion is latency, understood as the time required to transform sensor input into a control-relevant output. The second criterion is

reliability, defined as the ability of the system to maintain local functionality when individual communication links or cloud services degrade. The third criterion is privacy, understood as the scope of raw data transmission, the size of the attack surface and the degree of user control over behavioral and biometric data. The fourth criterion is energy efficiency, including both local inference costs and the cost of continuous data transmission. The fifth criterion is computational scalability, meaning the capacity of the architecture to process additional sensors, video streams or heavier models without functional degradation [3], [5], [6], [7].

Tab. 1. Analytical criteria used in the comparative assessment of smart home AI architectures.

Criterion	Cloud AI perspective	Edge AI perspective	Main design implication
Latency	Dependent on WAN routing, queuing and remote data-center availability [3].	Short local path and predictable invocation time [2], [3].	Safety-critical loops should be placed locally.
Reliability	External link or provider failure may break the control loop [1], [6].	Local control can continue during connectivity degradation [8].	Hybrid systems should support graceful degradation.
Privacy	Raw audio, video and telemetry may leave the home [5], [6].	Raw streams can be processed and minimized locally [13], [15].	Sensitive streams should be filtered before external transmission.
Energy and performance	Elastic compute capacity but continuous transmission and data-center costs [3].	Efficient inference with NPUs, but constrained memory and power [4], [7], [9].	Compression and task placement are necessary.
Scalability	High elasticity for heavy analytics and model training [10].	Limited by local cores, memory and accelerators [7], [13].	Large-scale analytics should remain cloud-assisted.

The analytical framework distinguishes three architectural models: centralized cloud AI, local edge AI and hybrid cloud-edge AI. These models are treated not as mutually exclusive categories but as elements of a continuum. The comparison therefore evaluates which categories of tasks should be placed on each layer rather than searching for a universally superior computing model [1], [8], [10]. To operationalize this comparison, Table 1 summarizes the five analytical criteria used in the assessment

and indicates how each criterion is interpreted from the cloud AI and edge AI perspectives, together with the resulting design implication.

2. CLOUD AND EDGE AI ARCHITECTURES IN SMART HOME SYSTEMS

The architecture of an AI-enabled smart home is defined primarily by the location of the inference layer. Inference is the stage at which a trained model transforms input data, such as a video frame, an acoustic signal, an energy-consumption pattern or a physiological reading, into a classification, prediction or control decision. When inference is placed in a remote data center, the system follows a cloud-centric model. When inference is placed on the end device or local hub, the system follows an edge-centric model [5], [6].

In the cloud model, sensors and cameras act mainly as acquisition endpoints. They record the signal and transmit raw or partially pre-processed streams over the home network and the wide area network to the service provider. The advantage of this model is scalability. Cloud infrastructure can allocate additional CPUs, GPUs, storage and memory pools without requiring the homeowner to purchase new hardware. This is important for model training, long-term analytics and high-volume multi-household optimization [7], [10]. The limitation is that every control decision depends on a dependency chain that includes the sensor, local network, router, internet service provider, public network route, remote data center and cloud service orchestration [1], [3].

In the edge model, the direction of data flow is reversed. The AI model is executed directly on the device, on a local hub or on a small residential server. The raw stream can therefore remain inside the building, while only a decision, event label, anonymized frame, aggregated statistic or model parameter update is transmitted externally when necessary [5], [13], [14]. At the hardware level, this approach is enabled by embedded accelerators, neural processing units and low-power chips optimized for the matrix operations used in neural networks [4], [7], [9]. At the software level, it depends on compression techniques such as quantization, pruning and knowledge distillation, which reduce the memory and energy requirements of deployed models [2], [4].

The hybrid model combines both layers into a cloud-edge continuum. In such an architecture, the edge performs local inference, event detection, filtering and short-term buffering, while the cloud performs heavy analytics, global model aggregation and long-term optimization [1], [8], [10]. This structure is particularly suitable for smart homes because household tasks differ strongly in terms of latency requirements and data sensitivity. A fall detection loop or fire alarm should not wait for a remote data-center round trip, whereas monthly energy optimization or cross-household model improvement can reasonably use cloud resources [2], [11], [14]. The functional logic of this cloud-edge continuum,

including the separation between local inference, local control and cloud-based analytics, is presented in Figure 1.

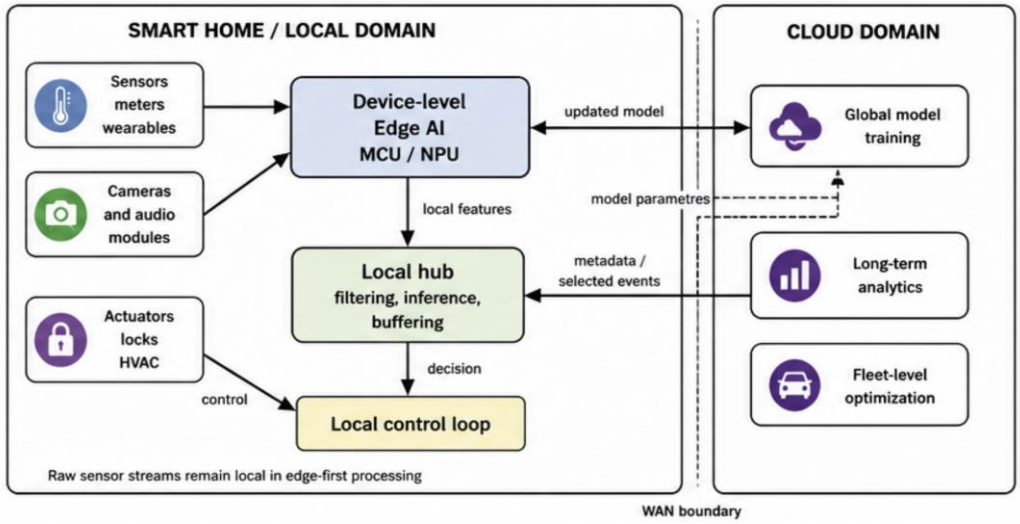


Fig. 1. Reference architecture of cloud, edge and hybrid AI processing in a smart home. Source: own elaboration based on [1], [5], [8], [10], [14].

3. LATENCY, RELIABILITY AND PRIVACY TRADE-OFFS

Latency is one of the most direct consequences of architectural placement. Literature cited in the analyzed material reports typical cloud inference latency in the range of approximately 100-500 ms, whereas local edge processing can fall within the range of approximately 5-10 ms [3]. In many user-facing functions, this difference may be perceived merely as a delay in interaction. In safety-critical functions, however, the same difference may determine whether the system response remains operationally useful [2], [11].

Smart home tasks can therefore be divided into hard-real-time and soft-real-time categories. Hard-real-time tasks include fire and gas detection, automatic power cutoff, fall detection, emergency monitoring and selected health-related alarms. For these tasks, exceeding the acceptable time window constitutes functional failure rather than a minor decrease in comfort [2], [11]. Soft-real-time tasks include long-term temperature trend analysis, weekly energy-consumption prediction, occupancy pattern optimization and non-urgent model updates. These tasks can tolerate longer delays and may benefit from cloud resources [1], [10], [14].

Reliability is closely connected with latency. A cloud-centric control loop depends on the simultaneous availability of multiple links. If the local router, internet connection, public network route or provider-side service fails, the loop may be interrupted. Edge architecture reduces this dependency by keeping core control decisions inside the local network. Hybrid systems can also implement graceful degradation:

when connectivity is lost, cloud-based analytics are suspended, but local safety functions and essential device control remain active [1], [8]. Dynamic task placement in edge-cloud serverless systems has been shown to reduce end-to-end latency by directing computationally heavy tasks to more appropriate nodes [10].

The privacy trade-off is equally significant. Cloud-based inference often requires transferring raw audio, video, telemetry or biometric readings outside the building. Such centralization increases the attack surface and reduces the user's practical control over the physical copy of the data [5], [6], [15]. Edge-based inference supports the principle of data minimization because the local node can process the raw stream and forward only metadata, selected events, anonymized content or aggregated values. Privacy-preserving video-surveillance systems and edge-cloud security architectures illustrate the relevance of local filtering and anonymization before transmission to higher layers [12], [13].

Federated learning extends the logic of privacy protection to the model training stage. Instead of transmitting raw household data to a central repository, the manufacturer provides a base model to local nodes. Each node trains or updates the model using its own data and then sends only model parameters or gradients for aggregation [14], [16]. This mechanism does not eliminate all risks, as issues such as communication overhead, hardware heterogeneity and the possibility of inference attacks still require mitigation. Nevertheless, it reduces the need to centralize raw domestic data and therefore remains consistent with privacy-by-design principles in AI-assisted environments [15], [17]. The difference between cloud-based and edge-based models is visible not only in terms of privacy, but also in system response time. In smart home environments, inference latency has a direct impact on the usability of AI solutions, especially in tasks requiring rapid response, such as fall detection, fire detection, gas leakage recognition, power failure detection or emergency alerts. In cloud-based architecture, data must be transmitted to an external processing center, which extends the decision-making path. In edge AI architecture, the model operates locally, making the response time shorter and more predictable. This relationship is presented in Figure 2.

Figure 2 shows that, according to the latency ranges reported in [3], local edge AI inference can achieve approximately 5–10 ms, whereas cloud-based inference is reported in the range of approximately 100–500 ms. This means that for time-critical tasks, such as safety alerts or emergency responses, local processing should be treated as the preferred solution. The cloud, however, remains useful for less urgent tasks, such as long-term analysis, energy consumption optimization, model training and aggregation of data from multiple households.

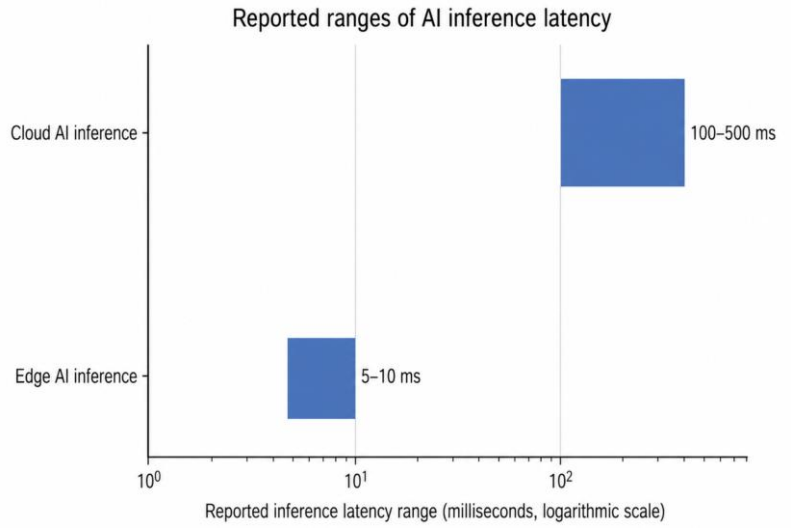


Fig. 2. Reported latency ranges for cloud and edge AI inference. Source: own elaboration based on [3].

4. PERFORMANCE CONSTRAINTS OF EDGE-BASED SMART HOME AI

Although the latency and privacy profile of edge AI is favorable, local inference introduces measurable performance constraints. The first constraint is model size. Deep learning models trained in full precision are often too large for microcontrollers, battery-powered sensors or passively cooled residential devices. Quantization is therefore one of the central deployment techniques. It replaces 32-bit floating-point representations with shorter integer formats, often 8-bit, which reduces memory footprint and arithmetic complexity [2], [4]. In a smart home anomaly-detection case, quantization of an LSTM model reduced inference time by 76% and power consumption by 35% while maintaining functional accuracy [2].

The second constraint is classification accuracy. Compression may reduce precision, especially when a model is transferred from a full cloud environment to heterogeneous edge devices. Deployment-oriented analyses report that model compression and hardware limitations may lead to measurable accuracy loss, which must be evaluated against the operational value of lower latency and stronger privacy [4], [9]. In home automation, a small decrease in classification accuracy may be acceptable for reversible comfort-related decisions, such as lighting or HVAC adjustment. It is less acceptable in emergency scenarios, where false negatives may have severe consequences [2], [11].

The third constraint is energy balance. Edge inference requires local power, and active inference increases the consumption of constrained devices [4]. However, a comparison limited only to the device can be misleading. Continuous cloud processing requires radio transmission, router activity, WAN transport, server computation and cooling. Analyses of edge and cloud architectures indicate that dedicated local acceleration can provide a much more favorable energy profile per inference than

remote execution, especially when the home generates continuous audio, video or telemetry streams [3], [7].

The fourth constraint is computational bottleneck. A home hub has a finite number of cores, limited memory and a fixed accelerator budget. It cannot scale elastically when the number of cameras, microphones or wearable devices increases. Benchmarks of edge devices show that hardware platforms differ substantially in inference performance, which means that model deployment must be adapted to the target device rather than treated as universally portable [7], [9]. In video-intensive scenarios, edge processing can become constrained by memory bandwidth, disk operations and parallel stream handling, which makes hybrid offloading or in-memory optimization necessary [13].

Cloud computing remains stronger in scenarios requiring large-scale aggregation, continuous model training and elastic resource allocation. The design implication is therefore not that edge AI replaces the cloud, but that it changes the role of the cloud. The cloud becomes a layer for global learning, fleet analytics and heavy computation, while the edge becomes the default layer for local control, first-stage filtering and sensitive-data processing [1], [8], [10], [14].

5. DISCUSSION

The comparative analysis shows that the most defensible architecture for AI-enabled smart homes is a layered design based on task allocation. The decisive question is not whether cloud or edge AI is superior in general, but which tasks should be assigned to which layer under specific constraints. Latency-sensitive and safety-critical tasks require local execution because the value of a delayed decision may be negligible. Privacy-sensitive tasks should also be processed locally whenever possible because the main risk is not only unauthorized access but also the permanent loss of informational sovereignty after raw data leave the building [3], [6], [15].

The edge layer is therefore best understood as the operational foundation of the smart home. It performs local sensing, event detection, short-horizon prediction, local anomaly recognition, emergency reaction and data minimization. Its function is to preserve autonomy and privacy. The cloud layer is best understood as the analytical and optimization layer. It supports long-horizon analysis, model retraining, cross-household pattern discovery, software updates and elastic processing of workloads that exceed the local resource budget [1], [8], [10].

Federated learning is particularly important in this architecture because it connects privacy preservation with continuous model improvement. Without federated or related distributed-learning mechanisms, manufacturers would have a strong incentive to centralize household data in order to improve model quality. Federated learning changes this balance by allowing model improvement while keeping raw

data on the user's device or local node [14], [16], [17]. For smart home systems, this approach is especially relevant because domestic data are not only technical measurements; they reveal daily routines, presence patterns, health conditions, energy behavior and family life [5], [15].

The proposed continuum model also clarifies the role of model compression. Quantization and pruning should not be treated only as technical optimizations. They are architectural enablers that make local privacy-preserving AI possible under residential energy and cost constraints [2], [4]. At the same time, compression introduces accuracy trade-offs that must be made explicit. The acceptable accuracy loss depends on the task class. False alarms in comfort automation may be acceptable, but false negatives in emergency detection require more conservative deployment, redundancy or cloud-assisted validation when the connection is available [2], [11].

The main limitation of the present article is that it synthesizes published findings rather than reporting new measurements on a unified experimental platform. The quantitative values discussed here come from different studies, hardware configurations and application domains, which means they should be interpreted as indicative rather than directly comparable under laboratory-equivalent conditions [1]-[17]. Future work should therefore include empirical benchmarking of representative smart home workloads on the same local hub, edge device and cloud platform, including latency jitter, energy consumption, classification accuracy and failure-mode behavior. Based on the comparative criteria discussed above, Table 2 summarizes the recommended placement of key smart-home AI tasks within the cloud-edge continuum, indicating the preferred processing layer, the main design rationale and the supporting references.

Tab. 2. Recommended task placement in a cloud-edge smart home continuum.

Task category	Preferred layer	Justification	Supporting references
Fall, fire, gas and emergency response	Edge / local hub	Requires low latency and continued operation during WAN outage.	[2], [3], [11]
Raw audio, video and biometric processing	Edge first	Supports data minimization and privacy-by-design assumptions.	[5], [6], [13], [15]
Long-term energy and comfort optimization	Hybrid	Local data can be filtered, while cloud resources support aggregation and forecasting.	[1], [8], [10]

Global model improvement	Federated cloud-edge model	Raw data remain local while parameter updates are aggregated.	[14], [16], [17]
Large-scale visual analytics	Hybrid / cloud assisted	Edge filtering reduces streams, while cloud resources handle heavy workloads.	[10], [13]

Data source: real-estate listing data set.

6. CONCLUSIONS

The analysis indicates that cloud and edge AI architectures should not be treated as mutually exclusive alternatives in smart home systems. Cloud computing provides elastic scalability, advanced analytics and global model-development capacity. Edge computing provides low-latency inference, local autonomy, data minimization and stronger control over sensitive household information. The rational design strategy is therefore a cloud-edge continuum in which task placement depends on latency requirements, privacy sensitivity, energy cost and computational complexity [1], [3], [8], [10].

For hard-real-time and privacy-sensitive tasks, edge AI should be the default architectural layer. Local inference protects safety loops against connectivity degradation and reduces the need to transmit raw domestic data. For soft-real-time and computation-heavy tasks, cloud computing remains valuable, especially where global optimization, long-term analytics or large-scale model training are required [2], [6], [11], [14].

The future development of AI-enabled smart homes will likely be shaped by three mechanisms: more efficient neural processing units, stronger model-compression techniques and wider adoption of federated learning. Together, these technologies allow smart homes to become more autonomous without abandoning the advantages of cloud-scale intelligence. The resulting architecture is not a replacement of cloud computing by edge computing, but a reallocation of responsibility: the edge protects immediacy and sovereignty, while the cloud supports scale and collective learning [4], [7], [14], [16], [17].

REFERENCES

[1] M.-C. Dumitru, S.-I. Caramihai, A. Dumitrascu, R.-N. Pietraru, and M.-A. Moisescu, (2025), "AI-Enabled Dynamic Edge-Cloud Resource Allocation for Smart Cities and Smart Buildings," Sensors, vol. 25, no. 24, Art. 7438, DOI: 10.3390/s25247438.

[2] M. J. C. S. Reis and C. Serodio, (2025), "Edge AI for Real-Time Anomaly Detection in Smart Homes," Future Internet, vol. 17, no. 4, Art. 179, DOI: 10.3390/fi17040179.

[3] R. P. Marpu, K. J. McNamara, and P. Gupta, (2025), "The AI Shadow War: SaaS vs. Edge Computing Architectures," arXiv preprint arXiv:2507.11545, DOI: 10.48550/arXiv.2507.11545.

[4] I. Laurent, (2025), "Edge AI Deployment Challenges in Smart Home Devices," Scientific Journal of Artificial Intelligence and Blockchain Technologies, vol. 2, no. 3, pp. 36-44, DOI: 10.63345/sjaibt.v2.i3.205.

[5] R. Singh and S. S. Gill, (2023), "Edge AI: A survey," Internet of Things and Cyber-Physical Systems, vol. 3, pp. 71-92, DOI: 10.1016/j.iotcps.2023.02.004.

[6] S. S. Gill, M. Golec, J. Hu, M. Xu, J. Du, H. Wu, G. K. Walia, S. S. Murugesan, B. Ali, M. Kumar, K. Ye, P. Verma, S. Kumar, F. Cuadrado, and S. Uhlig, (2025), "Edge AI: A Taxonomy, Systematic Review and Future Directions," Cluster Computing, vol. 28, no. 1, Art. 18, DOI: 10.1007/s10586-024-04686-y.

[7] S. P. Baller, A. Jindal, M. Chadha, and M. Gerndt, (2021), "DeepEdgeBench: Benchmarking Deep Neural Networks on Edge Devices," in Proceedings of the 2021 IEEE International Conference on Cloud Engineering (IC2E), pp. 20-30, DOI: 10.1109/IC2E52221.2021.00016.

[8] S. M. Alamouti, F. Arjomandi, M. Burger, and B. Altakrouri, (2025), "Evaluating Device-First Continuum AI (DFC-AI) for Autonomous Operations in the Energy Sector," arXiv preprint arXiv:2511.17528, DOI: 10.48550/arXiv.2511.17528.

[9] R. Tobiasz, G. Wilczynski, P. Graszka, N. Kuleshov, and K. Tokarski, (2023), "Edge Devices Inference Performance Comparison," arXiv preprint arXiv:2306.12093, DOI: 10.48550/arXiv.2306.12093.

[10] A. Das, S. Imai, S. Patterson, and M. P. Wittie, (2020), "Performance Optimization for Edge-Cloud Serverless Platforms via Dynamic Task Placement," in Proceedings of the 2020 20th IEEE/ACM International Symposium on Cluster, Cloud and Internet Computing (CCGRID), pp. 41-50, DOI: 10.1109/CCGrid49817.2020.00-89.

[11] M. Prabha, S. Nandhini, M. Dayanidhy, and R. Pradeep, (2026), "Edge-AI integrated secure wireless IoT architecture for real time healthcare monitoring and federated anomaly detection," Scientific Reports, vol. 16, Art. 574, DOI: 10.1038/s41598-025-30150-x.

[12] P. Vogl, S. Weber, J. Graf, K. Neubauer, and R. Hackenberg, (2022), "Design and Implementation of an Intelligent and Model-based Intrusion Detection System for IoT Networks," in Proceedings of CLOUD COMPUTING 2022, The Thirteenth International Conference on Cloud Computing, GRIDs, and Virtualization, Special Track FAST-CSP, pp. 7-12.

[13] S. Myneni, G. Agrawal, Y. Deng, A. Chowdhary, N. Vadnere, and D. Huang, (2022), "SCVS: On AI and Edge Clouds Enabled Privacy-preserved Smart-city Video Surveillance Services," ACM Transactions on Internet of Things, vol. 3, no. 4, Art. 28, DOI: 10.1145/3542953.

[14] J. Bao and H. Guo, (2022), "Federated learning in cloud-edge collaborative architecture: key technologies, applications and challenges," Journal of Cloud Computing, vol. 11, Art. 94, DOI: 10.1186/s13677-022-00377-4.

[15] I. Croitoru, C. E. Turcu, and C. O. Turcu, (2026), "Privacy-by-Design in AI-Assisted Systems for Caregivers of Children with Autism: A Secure Multi-Agent Architecture," Applied Sciences, vol. 16, no. 4, Art. 2157, DOI: 10.3390/app16042157.

[16] H. G. Abreha, M. Hayajneh, and M. A. Serhani, (2022), "Federated Learning in Edge Computing: A Systematic Survey," Sensors, vol. 22, no. 2, Art. 450, DOI: 10.3390/s22020450.

[17] S. Alahmari and I. Alghamdi, (2025), "A comprehensive survey on energy-efficient and privacy-preserving federated learning for edge intelligence and IoT," Results in Engineering, vol. 28, Art. 107849, DOI: 10.1016/j.rineng.2025.107849.

Marta Chodyka: https://orcid.org/ 0000-0002-8819-2451

Gabriel Tarasiuk https://orcid.org/ 0009-0003-8303-0968

DETERMINANTS OF INDEXING EFFECTIVENESS IN RELATIONAL DATABASE SYSTEMS

MARTA CHODYKA

University of Łomza, Poland

mchodyka@al.edu.pl

PAWEŁ LINIEWSKI

University of Łomza, Poland

pa_ul@wp.pl

Gabriel TARASIUK

John Paul II University in Biała Podlask, Poland

tarasiuk22914@stud.akademiabialska.pl

ABSTRACT: This paper addresses the research problem of non-uniform indexing effectiveness in relational database systems. Although indexes are commonly considered a primary mechanism for SQL query optimization, their actual impact depends on the query type, predicate selectivity, data structure and the decisions made by the optimizer of a particular RDBMS engine. The aim of the study was to empirically evaluate the determinants of indexing effectiveness and to identify the scenarios in which indexes improve, only partially improve or do not improve query execution time. An experimental and comparative methodology was applied. The study was conducted in four systems: PostgreSQL, MySQL, MariaDB and SQLite, using a real IMDb dataset containing approximately 29.4 million records in total. Ten SQL scenarios were designed to represent heterogeneous workloads: selection, equality and range filtering, sorting, text search, joins, filtered joins, ordered joins, aggregation and a multi-stage analytical query. Each query was executed in two variants: without indexes and after creating indexes on columns relevant to the scenario. Measurements were automated by means of a Java application communicating with the databases through JDBC. The results show that the largest and most stable benefits occur for join and filtered join queries, whereas the impact of indexes on sorting, aggregation and LIKE-based text search with a leading wildcard is limited. The practical application of the results lies in supporting index design for analytical, reporting and application systems in which the index structure must be adapted to the actual workload profile.

Key words: indexing; relational databases; SQL optimization; query performance; database workloads

INTRODUCTION

Relational database management systems remain a fundamental component of business, analytical and scientific applications. As the number of records and the complexity of queries increase, the importance of mechanisms that reduce data access cost also grows. Indexes are among the most frequently applied mechanisms of this type because they may shorten searching, joining and ordered retrieval by changing the access path to records [1]-[5]. However, the widespread use of indexes does not imply universal effectiveness. The same index may considerably accelerate a selective query or a join, while providing no benefit for leading-wildcard text search, low-selectivity predicates or queries dominated by sorting cost. In relational database systems, final performance is determined not only by the existence of an index but also by data statistics, predicate type, result-set size and the execution plan selected by the optimizer [9]-[13].

This paper assumes that indexing effectiveness should be analyzed by scenario rather than solely by comparing average execution time before and after index creation. The contribution of the study is a structured interpretation of experimental results by SQL workload class and the identification of factors that determine whether indexing improves performance in four widely used RDBMS engines: PostgreSQL, MySQL, MariaDB and SQLite.

1. THEORETICAL BACKGROUND AND LITERATURE REVIEW

Indexes in database systems have a long research tradition. The classical works by Bayer and McCreight and by Comer described the B-tree structure and its importance for maintaining ordered data efficiently on secondary storage [6], [7]. Contemporary surveys of B-tree techniques indicate that despite the development of new structures, tree-based indexes remain fundamental in practical RDBMS environments, especially for equality predicates, range predicates and ordered data access [8], [21]. Alternative access structures have also been developed, including extendible hashing, R-trees and generalized search trees [14]-[16]. These structures demonstrate that index selection should depend on data characteristics and dominant operations. Hash indexes are naturally associated with equality search, whereas spatial and generalized structures are important for specialized data classes. A second key research area is query optimization. The access path selection model in System R established the foundations of cost-based query planning [10]. Later work by Chaudhuri and by Chaudhuri and Narasayya developed relational optimization and hypothetical index analysis [11], [12]. This means that index effectiveness should be evaluated not only as a property of a data structure, but also as a result of optimizer decisions.

Current research directions include automatic index tuning, automatically generated index structures and learned indexes [17]-[20]. These trends confirm that manual index design becomes increasingly difficult as the number of tables, columns and query patterns grows. Consequently, experimental studies are needed to identify which workload types actually benefit from indexing in practice.

2. RESEARCH PROBLEM AND RESEARCH OBJECTIVE

The research problem was formulated as follows: although indexes are commonly applied as a mechanism for SQL query optimization, their effectiveness depends on the operation type, predicate selectivity, data structure and the optimizer of a specific RDBMS. It is therefore necessary to determine in which scenarios indexing provides measurable improvement and in which scenarios its impact is limited or ambiguous.

The aim of the study was to empirically identify the determinants of indexing effectiveness in relational database systems. The specific objective included comparing the impact of indexes in four RDBMS engines and assessing different classes of SQL queries: simple retrieval, filtering, range selection, sorting, text search, joins, aggregation and complex multi-stage queries.

Four research questions were posed: RQ1 - which SQL query classes show the largest improvement after indexing; RQ2 - in which scenarios does indexing fail to provide the expected improvement; RQ3 - is the impact of indexing similar in PostgreSQL, MySQL, MariaDB and SQLite; RQ4 - which factors should be considered when designing indexes for real database workloads.

3. RESEARCH METHODOLOGY

3.1. DATASET AND EXPERIMENTAL ENVIRONMENT

The study was conducted on a real IMDb dataset containing information about titles, ratings and people associated with film productions. Three main tables were used in the experiment. The dataset size was sufficient to observe differences caused by execution plans and input-output costs.

The experiment was carried out in four relational database management systems: PostgreSQL, MySQL, MariaDB and SQLite. Measurements were automated with a Java application. The application established JDBC connections, executed predefined SQL queries, recorded execution time and saved the results for further analysis. The procedure included a no-index variant and an indexed variant, in which indexes were created on columns used in predicates, sorting or joins.

Tab. 1. Characteristics of the main tables used in the study.

Table	Key columns	Number of records	Role in the experiment
title_basics	tconst, primaryTitle, startYear, runtimeMinutes, genres	12,441,901	selection, filtering, sorting, aggregation
title_ratings	tconst, averageRating, numVotes	1,661,903	filtering, sorting, joins
name_basics	nconst, primaryName, birthYear, knownForTitles	15,252,091	auxiliary context, people-related data

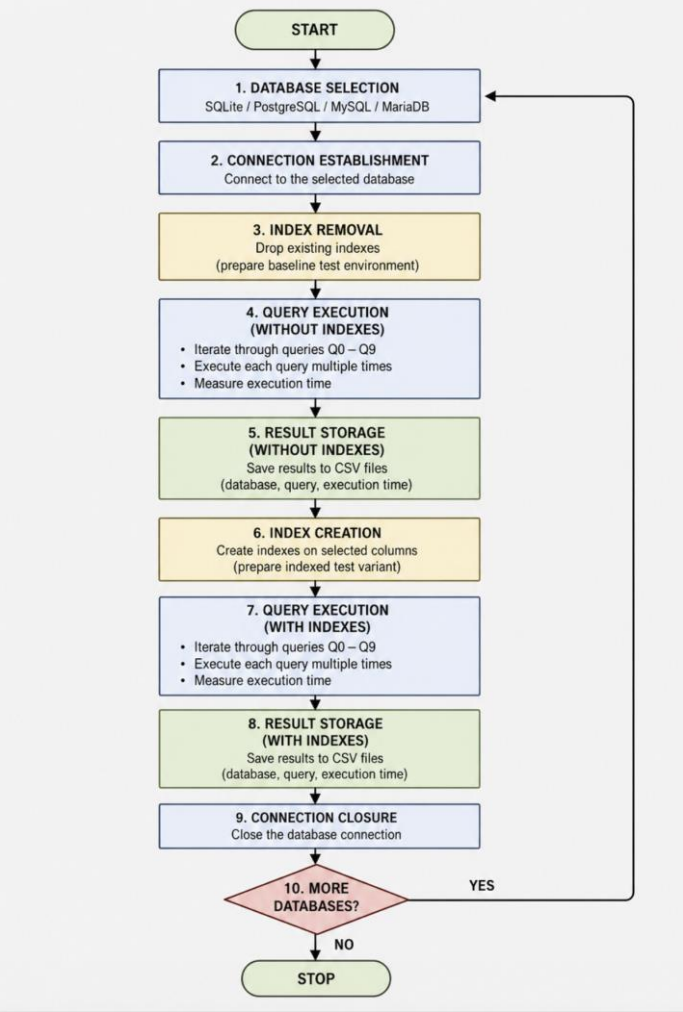


Fig. 1. Measurement environment used in the experiment.

3.2. QUERY SCENARIOS AND MEASUREMENT PROCEDURE

Ten SQL scenarios were prepared to represent typical relational database workload classes. Each scenario was measured under the same logical conditions, which allowed the impact of indexes to be compared not only across queries but also across RDBMS engines.

Tab. 2. Experimental scenarios and their research purpose.

Code	Dominant operation	Measurement purpose
Q1	Limited record retrieval	establishing baseline data access time
Q2	Equality selection	assessing the index on the startYear column
Q3	Range selection	assessing the role of range predicate selectivity
Q4	ORDER BY sorting	examining whether an index reduces sorting cost
Q5	Text search LIKE "%...%"	examining limitations of traditional text indexing
Q6	JOIN operation	assessing indexes on join columns
Q7	JOIN with additional filtering	analyzing the combined effect of filtering and joining
Q8	JOIN with sorting	assessing the interaction between join cost and ordering cost
Q9	GROUP BY aggregation	examining the influence of data ordering on aggregation
Q10	JOIN + WHERE + GROUP BY + ORDER BY	analyzing a multi-stage query and optimizer decisions

Execution time was measured automatically. Each query was executed multiple times, and the first run was treated as a warm-up run because it could be affected by connection initialization, query plan initialization or caching mechanisms. The interpretation of the results was based on comparing the no-index and indexed variants and on analyzing the nature of each SQL operation. The result charts for representative scenarios report the arithmetic mean from runs 2-4.

4. RESULTS

The results confirm that indexing effectiveness is scenario-dependent. No universal rule was observed according to which index creation would improve performance in every query class. In some scenarios,

indexes clearly reduced the data access cost, whereas in others the dominant costs were sorting, aggregation or full text-like scanning.

Tab. 3. Synthetic interpretation of the impact of indexing on the tested scenarios.

Scenario	Indexing effect	Interpretation
Q1	no significant effect	the query contains no selective predicate or join; LIMIT restricts the operation cost
Q2	clear improvement	the equality predicate supports effective B-tree index usage
Q3	moderate improvement	the benefit depends on how many records satisfy the range condition
Q4	limited or ambiguous effect	sorting may remain the dominant cost despite the presence of an index
Q5	no useful improvement	the leading wildcard pattern prevents efficient use of a traditional index
Q6	substantial improvement	indexes on join keys reduce the number of comparisons between tables
Q7	substantial and stable improvement	indexes support both dataset narrowing and record matching in JOIN
Q8	partial improvement	indexes support the join, but sorting may remain the dominant cost
Q9	minor or moderate improvement	aggregation still requires processing a large part of the dataset
Q10	engine-dependent result	the complex execution plan depends on statistics, costs and optimizer strategy

The most unambiguous results were obtained for join and filtered join queries. In scenarios Q6 and Q7, indexes created on tconst and numVotes reduced the number of records processed while matching data from title_basics and title_ratings. The benefit was particularly visible in Q7, where the index supported both the filtering stage and the join stage.

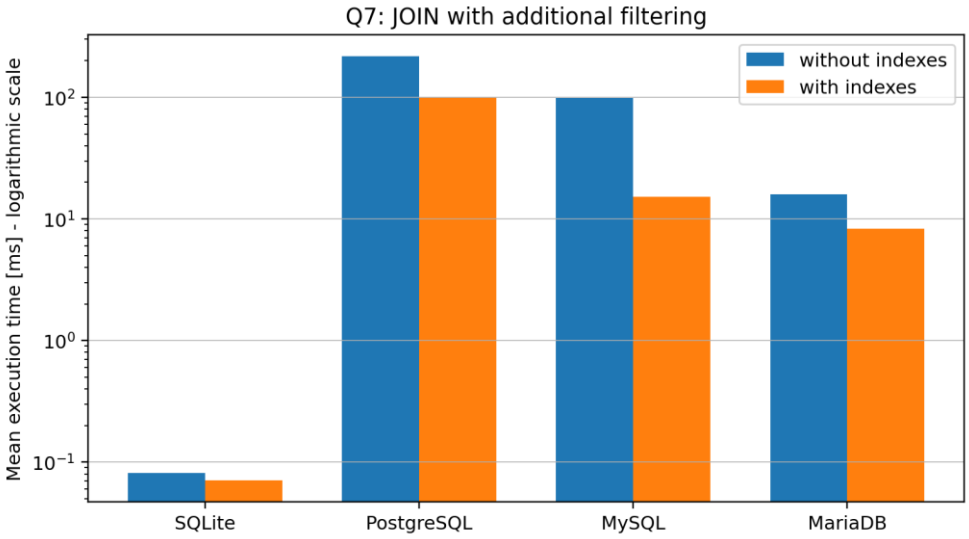


Fig. 2. Mean execution time for scenario Q7 involving JOIN with additional filtering (runs 2-4).

A different result was obtained for text search scenario Q5, in which the LIKE operator was used with a pattern beginning with a wildcard. In this case, a traditional index did not provide a useful reduction in execution time because the condition did not contain a starting point enabling efficient traversal of a B-tree structure. This result indicates the need for full-text indexes or specialized search mechanisms when such text queries constitute a dominant workload.

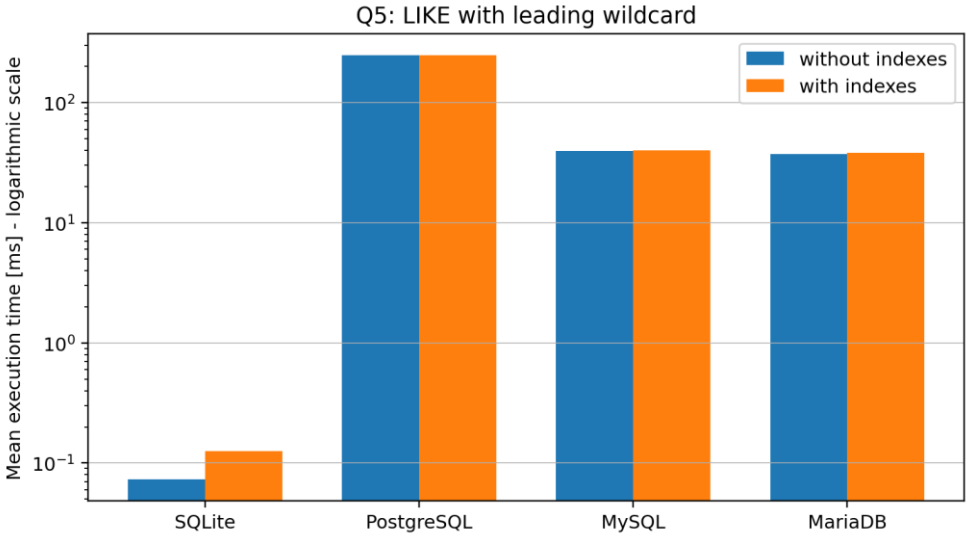


Fig. 3. Mean execution time for scenario Q5 using the LIKE operator with a leading wildcard (runs 2-4).

Sorting and aggregation scenarios revealed the limitations of indexing. In Q4, an index could support reading data in a particular order, but it did not always eliminate sorting cost. In Q9, the GROUP BY operation required processing a large part of the dataset, so the index effect was minor or moderate. The

most complex scenario, Q10, showed that when JOIN, WHERE, GROUP BY and ORDER BY occur simultaneously, the final result depends on the execution plan and optimizer decisions.

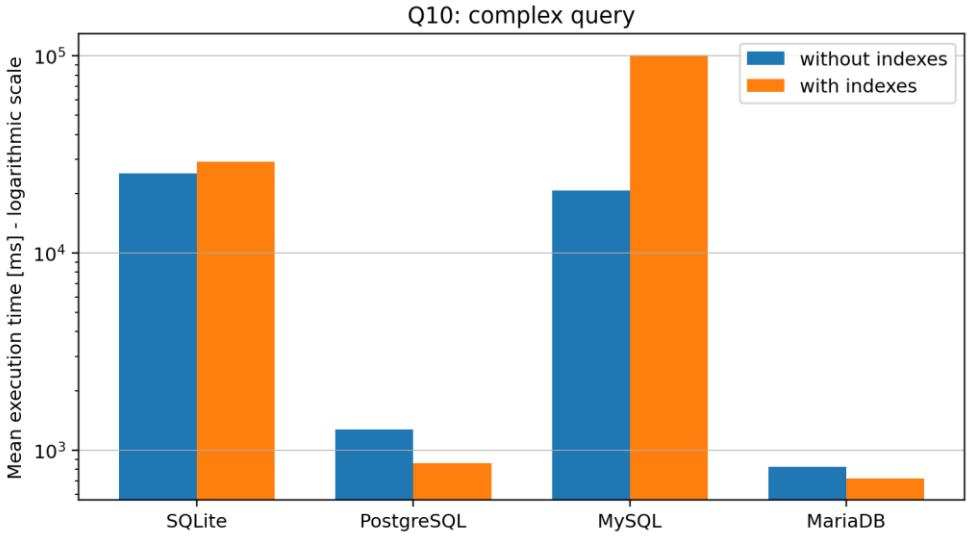


Fig. 4. Mean execution time for scenario Q10 involving JOIN, WHERE, GROUP BY and ORDER BY (runs 2-4).

5. DISCUSSION

The first determinant of indexing effectiveness is the predicate type. Equality predicates and selective filtering conditions support index use because they allow the candidate set to be narrowed quickly. Range conditions may also benefit from indexes, but the gain decreases when many records satisfy the condition. Leading-wildcard text search is the opposite case because a traditional B-tree index does not provide an efficient search starting point. The second determinant is the dominant operation cost. In JOIN queries, indexes on join columns considerably reduce the number of comparisons. In ORDER BY and GROUP BY queries, an index may be helpful, but it does not necessarily eliminate ordering or aggregation cost. Therefore, index design should result from analyzing the full execution plan rather than from merely observing that a column appears in a query. The third determinant is the RDBMS engine and its optimizer. Differences among PostgreSQL, MySQL, MariaDB and SQLite were especially visible in complex scenarios. They result from different cost models, different use of statistics and different strategies for joins and aggregation. In practice, this supports the need to test indexes in the target environment rather than transferring conclusions between systems without verification.

The limitation of the study is that it uses one dataset and a specific query set. The results should not be interpreted as a universal ranking of RDBMS engines. Their value lies in identifying relationships between query class and indexing effectiveness. Future work should include other data distributions, composite indexes, partial indexes, full-text indexes and the analysis of the cost of data modification operations.

CONCLUSION

On the basis of the conducted experiment, indexing was found to be an effective but conditional mechanism for SQL query optimization. The largest benefits were achieved for queries in which indexes reduced the number of records processed at an early stage of the execution plan, especially for joins and joins with filtering.

The study did not confirm a universal improvement after indexing. For LIKE text search with a leading wildcard, selected sorting operations and aggregation queries, the effect was limited. In multi-stage queries, the creation of an index alone did not guarantee improvement because the final execution time depended on optimizer decisions and data statistics.

The practical application of the results is that index design should be preceded by analyzing the actual workload profile, predicate selectivity and execution plans. In systems processing large datasets, iterative index tuning is recommended, including measurement, EXPLAIN/ANALYZE analysis and verification of the impact of indexes on specific query classes.

REFERENCES

[1] C. J. Date, (2004), An Introduction to Database Systems, 8th ed. Boston: Pearson/Addison-Wesley.

[2] R. Ramakrishnan and J. Gehrke, (2003), Database Management Systems, 3rd ed. New York: McGraw-Hill.

[3] R. Elmasri and S. B. Navathe, (2016), Fundamentals of Database Systems, 7th ed. Boston: Pearson.

[4] A. Silberschatz, H. F. Korth, and S. Sudarshan, (2019), Database System Concepts, 7th ed. New York: McGraw-Hill.

[5] H. Garcia-Molina, J. D. Ullman, and J. Widom, (2008), Database Systems: The Complete Book, 2nd ed. Upper Saddle River: Pearson Prentice Hall.

[6] R. Bayer and E. M. McCreight, (1972), “Organization and Maintenance of Large Ordered Indexes,” Acta Informatica, vol. 1, no. 3, pp. 173-189, DOI: 10.1007/BF00288683.

[7] D. Comer, (1979), “The Ubiquitous B-Tree,” ACM Computing Surveys, vol. 11, no. 2, pp. 121-137, DOI: 10.1145/356770.356776.

[8] G. Graefe, (2011), “Modern B-Tree Techniques,” Foundations and Trends in Databases, vol. 3, no. 4, pp. 203-402, DOI: 10.1561/1900000028.

[9] J. M. Hellerstein, M. Stonebraker, and J. R. Hamilton, (2007), “Architecture of a Database System,” Foundations and Trends in Databases, vol. 1, no. 2, pp. 141-259, DOI: 10.1561/1900000002.

[10] P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, and T. G. Price, (1979), “Access Path Selection in a Relational Database Management System,” in Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 23-34, DOI: 10.1145/582095.582099.

[11] S. Chaudhuri, (1998), “An Overview of Query Optimization in Relational Systems,” in Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp. 34-43, DOI: 10.1145/275487.275492.

[12] S. Chaudhuri and V. R. Narasayya, (1998), “AutoAdmin ‘What-If’ Index Analysis Utility,” in Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 367-378, DOI: 10.1145/276304.276337.

[13] M. Stonebraker and L. A. Rowe, (1986), “The Design of POSTGRES,” ACM SIGMOD Record, vol. 15, no. 2, pp. 340-355, DOI: 10.1145/16856.16888.

[14] R. Fagin, J. Nievergelt, N. Pippenger, and H. R. Strong, (1979), “Extendible Hashing - A Fast Access Method for Dynamic Files,” ACM Transactions on Database Systems, vol. 4, no. 3, pp. 315-344, DOI: 10.1145/320083.320092.

[15] A. Guttman, (1984), “R-Trees: A Dynamic Index Structure for Spatial Searching,” in Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 47-57, DOI: 10.1145/602259.602266.

[16] M. Kornacker, C. Mohan, and J. M. Hellerstein, (1997), “Concurrency and Recovery in Generalized Search Trees,” in Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 62-72, DOI: 10.1145/253260.253272.

[17] T. Kraska, A. Beutel, E. H. Chi, J. Dean, and N. Polyzotis, (2018), “The Case for Learned Index Structures,” in Proceedings of the International Conference on Management of Data, pp. 489-504, DOI: 10.1145/3183713.3196909.

[18] J. Kossmann, S. Halfpap, M. Jankrift, and R. Schlosser, (2020), “Magic Mirror in My Hand, Which Is the Best in the Land? An Experimental Evaluation of Index Selection Algorithms,” Proceedings of the VLDB Endowment, vol. 13, no. 11, pp. 2382-2395, DOI: 10.14778/3407790.3407832.

[19] J. Dittrich, J. Nix, and C. Schön, (2022), “The Next 50 Years in Database Indexing or: The Case for Automatically Generated Index Structures,” Proceedings of the VLDB Endowment, vol. 15, no. 3, pp. 527-540, DOI: 10.14778/3494124.3494136.

[20] Y. Wu, X. Zhou, Y. Zhang, and G. Li, (2024), “Automatic Index Tuning: A Survey,” IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 12, pp. 7657-7676, DOI: 10.1109/TKDE.2024.3422006.

[21] G. Graefe, (2024), “More Modern B-Tree Techniques,” Foundations and Trends in Databases, vol. 13, no. 3, pp. 169-249, DOI: 10.1561/1900000070.

Marta Chodyka: https://orcid.org/ 0000-0002-8819-2451

BEYOND COMPLIANCE: A CRITICAL DIAGNOSTIC OF SLOVAK CYBERSECURITY OPERATIONAL RESILIENCE

Martin MAZUCH[1], Boris BUCKO[2]

Researcher, University of Zilina, Faculty of Management Science and Informatics, Dept. of Informatics, Zilina, Slovak Republic[1]

Security Manager, Syntelia, s.r.o., Cybersecurity and Analytics, Zilina, Slovak Republic[2]

martin.mazuch@uniza.sk[1], boris.bucko@syntelia.tech[2]

ABSTRACT: While the formal harmonization of Slovak cybersecurity standards with European directives (NIS1 and NIS2) has been widely documented, a critical gap remains between legislative compliance and operational readiness. This paper moves beyond the descriptive analysis of the regulatory framework—specifically Acts No. 69/2018 and 366/2024—to provide a diagnostic assessment of the practical execution challenges in the Slovak digital defense ecosystem. By examining the interplay between institutional fragmentation, human capital scarcity, and organizational compliance culture, this study identifies the structural bottlenecks that hinder effective incident response. The findings offer a critical perspective on the current national cybersecurity posture, suggesting that the path forward lies in operational cohesion rather than further regulatory expansion.

Key words: cybersecurity, NIS2 directive, cybersecurity governance, national security authority, critical infrastructure resilience, regulatory implementation

INTRODUCTION

While the formal alignment of Slovak cybersecurity legislation with European NIS2 requirements is often presented as a success story, the reality on the ground tells a more complex tale. We argue that the focus of recent legislative updates—specifically Acts No. 69/2018 and 366/2024—has been heavily skewed towards compliance formalities, often at the expense of genuine operational resilience. Is a robust regulatory framework enough to withstand the current landscape of sophisticated cyber-threats? In this article, we peel back the layers of formal governance to examine the inherent implementation gaps in the Slovak digital defense ecosystem. Our analysis suggests that the true bottleneck for national security is not the absence of legal mandates, but a structural deficiency in enforcement capacity and the fragmented nature of institutional oversight. By shifting our lens from mere transposition of directives to the examination of practical execution, we provide a critical assessment of the Slovak cybersecurity posture as it stands in 2026.

1. LEGISLATIVE EVOLUTION: FROM NIS1 TO NIS2

The chronology of digital security regulation within the Slovak legislative framework demonstrates a clear paradigm shift: from elementary compliance to holistic corporate resilience. The fundamental foundation of this domestic system was laid by Act No. 69/2018 Coll., which carried out the transposition of the NIS1 framework (Directive (EU) 2016/1148) [1], [2]. This preliminary act established baseline requirements, including mandatory incident disclosure methods, basic risk mitigation strategies, and designating the National Security Authority (NBÚ SR) as the ultimate supervisory body. To achieve convergence with the NIS2 mandate (Directive (EU) 2022/2555), a comprehensive systemic change was undertaken by Act No. 366/2024 Coll., which went into effect on January 1, 2025 [3], [4].

This important legal change, which goes beyond past compliance-driven versions, directly integrates information protection into top-tier company governance and supply-chain management processes. Furthermore, the 2025 amendment dramatically expanded the jurisdictional perimeter for regulated organizations, increased senior leadership liability, and introduced a risk-centric governance architecture aimed at ensuring long-term operational durability.

2. REGULATORY MANDATES: A CRITICAL APPRAISAL

Marking a definitive doctrinal pivot from retrospective incident mitigation towards proactive, systemic endurance, the Slovak executive branch officially promulgated the National Cybersecurity Strategy 2026–2030 in February 2026 [5]. This latest policy framework methodically harmonizes domestic administrative directives with the operational mandates dictated by the NIS2 directive, while simultaneously confronting the complexities of hybrid warfare, supply-chain fragilities, and the preservation of critical infrastructure. To materialize this anticipatory defense posture, the blueprint establishes several core imperatives.

These encompass the optimization of inter-agency governance, the fortification of specific economic sectors, the cultivation of a specialized talent pipeline, and the deepening of synergies between state apparatuses and commercial actors. Furthermore, the strategy explicitly anchors the nation's protective mechanisms within transnational security architectures, reaffirming robust multilateral engagements across both European Union and NATO platforms.

3. THE OPERATIONAL REALITY: BEYOND FORMAL COMPLIANCE

While the legislative framework provides the architecture for security, our practical observations indicate that the Slovak cybersecurity landscape faces three fundamental hurdles that legislation alone cannot resolve.

The current regulatory push, driven by the transposition of NIS2, has inadvertently encouraged a predominantly compliance-oriented administrative approach. Organizations frequently prioritize administrative compliance—ensuring that documentation aligns with Act No. 366/2024—over the implementation of robust technical security measures. This may create a formal perception of resilience, while technical systems may remain vulnerable to evolving adversarial tactics.

A critical bottleneck identified in our analysis is the widening gap between the required expertise and available workforce. The public sector's ability to retain cybersecurity talent is severely limited by competitive pressures from the private market. This talent drain compromises the national capacity for proactive threat hunting and forensic analysis. In our professional experience within the Slovak cybersecurity sector, we observe that the competition for skilled cybersecurity professionals is no longer just between firms, but has created significant sustainability and retention challenges for internal cybersecurity teams in comparison with outsourced service providers.

Despite the central supervisory role of the National Security Authority (NBÚ SR), the practical distribution of responsibilities across various sectoral authorities leads to latency. During critical security incidents, this fragmented governance often results in communication friction. Our assessment suggests that without a unified coordination protocol capable of reducing inter-agency administrative latency, the response time to sophisticated multi-vector attacks will remain a persistent vulnerability.

4. INSTITUTIONAL ARCHITECTURE AND ITS LIMITATIONS

The administrative architecture of Slovak digital security operates on a structural dichotomy: while it successfully leverages domain-specific expertise across various sectors, it concurrently generates a latent risk of systemic capacity overload at the supreme supervisory level. At the apex of this hybrid regulatory framework operates the National Security Authority (NBÚ SR). Acting as the primary competent national organ, NBÚ SR is vested with comprehensive executive mandates, encompassing the enforcement of statutory compliance, the execution of complex audits, and the orchestration of supreme incident response mechanisms via the SK-CERT unit [6], [7].

Operating in tandem, the Ministry of Investment, Regional Development and Informatization (MIRRI SR) assumes responsibility for the broader digital transformation agenda, specifically steering the integration of NIS2 imperatives within state and public administration infrastructures. Beyond these core entities, further regulatory distribution occurs at the specific domain level, where critical sector-specific jurisdictions are maintained by the Ministry of Defence, the Ministry of Interior, and other specialized agencies.

Ultimately, although this compartmentalized configuration fosters necessary functional specialization, the disproportionate concentration of ultimate auditing and oversight duties inherently threatens to strain the operational bandwidth of the central authority.

5. OVERSIGHT AND DEMOCRATIC SAFEGUARDS

The democratic accountability of national information security mechanisms is perpetually challenged by the inherent friction between state transparency and the protection of classified data. While formal legislative supervision is institutionalized through the National Council's Committee on Defence and Security, the highly sensitive nature of active cyber operations fundamentally restricts the granular depth of such parliamentary oversight. Consequently, academic discourse has increasingly scrutinized the procedural safeguards surrounding specific coercive instruments—most prominently the executive mandates for online content blocking—arguing forcefully for the necessity of ex-ante judicial authorization [8]. This doctrinal critique aligns directly with the established jurisprudence of the European Court of Human Rights, which dictates that any state interference within the digital surveillance spectrum must strictly adhere to principles of proportionality, mandate independent adjudicative pre-approval, and ensure the availability of accessible legal redress [9], [10]. Ultimately, calibrating the mechanisms of judicial review to these standards is not merely a legal formality; it is a critical prerequisite for fortifying the normative legitimacy of the regulatory framework and securing sustainable public confidence.

6. THE ENFORCEMENT GAP: THEORY VS. PRACTICE

The primary impediment to national information security is no longer a deficit in statutory regulation, but rather an acute shortfall in operational execution and systemic capacity. This diagnostic is empirically substantiated by the 2023 Annual Report published by the National Security Authority (NBÚ SR), which documented that approximately fifty percent of mandated organizations failed to adequately fulfill their compulsory auditing and incident disclosure protocols [6]. Such pervasive non-compliance underscores a matrix of underlying structural vulnerabilities.

Foremost among these are a chronic scarcity of specialized human capital, profound disparities in security maturity across various economic domains, and a critically low baseline of preparedness within small and medium-sized enterprises (SMEs). Furthermore, the escalating complexity of mitigating vulnerabilities embedded within extended supply chains places an additional burden on these entities. Consequently, securing the long-term durability of the national cyber ecosystem dictates a strict adherence to two critical prerequisites: the uncompromising and uniform enforcement of regulatory standards, and a sustained investment in cross-sectoral institutional capacity building

CONCLUSION

Evaluated through the lens of recent legislative overhauls most notably the transposition of the NIS2 directive the Slovak regulatory framework has undeniably achieved a profound degree of formal harmonization with European Union mandates. The contemporary security architecture is anchored by significant structural merits, including a modernized legal code, revitalized strategic forecasting, and the presence of a robust, centralized oversight mechanism.

Nevertheless, the substantive efficacy of this extensive apparatus remains entirely contingent upon overcoming practical barriers. Specifically, transitioning from mere normative compliance to genuine defensive capability requires prioritizing the maturation of enforcement practices, seamless inter-agency synchronization, and the continuous refinement of procedural safeguards. Ultimately, long-term systemic resilience will not be dictated by further regulatory proliferation, but rather by uncompromising implementation discipline and operational cohesion.

REFERENCES

[1] Act No. 69/2018 Z. z. on Cybersecurity (Slovak Republic), consolidated text effective 01.01.2025. Available: https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2018/69/

[2] Directive (EU) 2016/1148 of the European Parliament and of the Council (NIS1), 2016. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016L1148

[3] Act No. 366/2024 Z. z., amending Act No. 69/2018 Z. z., effective 01.01.2025. Available: https://www.slov-lex.sk/pravne-predpisy/SK/ZZ/2024/366/

[4] Directive (EU) 2022/2555 of the European Parliament and of the Council (NIS2), 2022. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2555

[5] National Security Authority (NBÚ SR), National Cybersecurity Strategy 2026–2030, Government Decision, Feb. 4, 2026. Available: https://www.nbu.gov.sk/

[6] NBÚ SR, Správa o stave kybernetickej bezpečnosti SR, Annual Report 2023. Available: https://www.nbu.gov.sk/

[7] NBÚ SR, Overview of SK-CERT Activities, 2024. Available: https://www.nbu.gov.sk/

[8] P. Sokol and L. Rózenfeldová, “Content blocking mechanisms in cybersecurity: Slovakia case study,” 2025.

[9] European Court of Human Rights, Szabó and Vissy v. Hungary, App. No. 37138/14, 2016.

[10] European Court of Human Rights, Big Brother Watch and Others v. United Kingdom, 2021.

Martin, Mazuch: https://orcid.org/0009-0002-7870-8482

Boris, Bucko: https://orcid.org/0000-0003-0915-9336

FROM SHADOW TO AVATAR: INTERMEDIATE ONTOLOGY, TECHNO-ANIMISM, AND POSTDIGITAL THEATRE IN THE PROJECT OF THE INTERNATIONAL CENTER OF ART OF EUROPE AND ASIA

Konrad SZCZEBIOT

Doctoral School of Cardinal Stefan Wyszyński University, Warsaw, Poland

konrad.szczebiot@gmail.com

ABSTRACT: In an era of radical technological acceleration, the traditional framework of cultural institutions is being exhausted. The International Center for the Arts of Europe and Asia (MCSEiA), initiated by, among others, Professor Jacek Kucaba, proposes a model of a nomadic platform and laboratory for postdramatic forms, which explores the relationships between the body, movement, and technology. This abstract systematizes the assumptions of this movement, drawing on the paradigms of postdigital theater, intermediate ontology, and techno-animism. Postdigital theater deconstructs the binary between the embodied and the virtual, constructing a "postdigital co-presence" using telematic and hybrid forms. The category of "intermediate ontology" presents the avatar not as a lifeless substitute, but as a natural successor to animate forms—traditional puppets. The phenomenon of "digital puppetry" demonstrates that virtual entities successfully transfer the creator's intentionality, just as in ancient puppetry. The project also incorporates Asian cosmotechnics and techno-animism, allowing for the perception of artificial intelligence systems as structures demonstrating vitality rather than merely cold mechanisms. This approach draws on media archaeology, demonstrating, among other things, that the traditional Karagöz shadow theatre is an aesthetic prefiguration of contemporary virtual reality interfaces.

Keywords: postdigital theatre, intermediate ontology, techno-animism, digital puppetry, media archaeology

INTRODUCTION: TOWARDS A NEW INSTITUTIONALITY OF PERFORMATIVE ARTS

In the era of late modernity, marked by radical technological acceleration and the progressive dematerialization of social relations, the classical frameworks of cultural institutions are undergoing systematic and inevitable exhaustion. Traditional theatre, based on the Aristotelian unity of time, place, and action, as well as a deeply logocentric paradigm of physical, embodied presence, faces the necessity of redefining its ontological and epistemological foundations. The response to this multi-level crisis is the emergence of experimental institutional models, among which a special and innovative place is occupied by the International Center of Art of Europe and Asia (MCSEiA). This project, initiated

as a nomadic, interdisciplinary platform, proposes a model organized around dynamics, geographical dispersion, and the continuous flow of forms, ideas, and digital technologies.

A phenomenon of profoundly deep symbolic, research, and analytical significance is the fact that the initiator of this radically postdigital and largely dematerialized platform for artistic exchange is a creator originating from environments unconditionally tied to tangible matter, gravity, and sculpture. Prof. Jacek Kucaba, Ph.D., a distinguished pedagogue associated with the Faculty of Sculpture at the Jan Matejko Academy of Fine Arts in Krakow, directing work within the Doctoral School, and a lecturer in lettering and typography at the Faculty of Art of the Tarnow Academy, is an artist strongly rooted in the physicality of the world (ASP w Krakowie, n.d.; Akademia Tarnowska, n.d.). His creative biography, reaching back to his roots in the Bieszczady Mountains, is inextricably linked to the organic, material experience of reality. His personal acquaintance with the famous Bieszczady wanderer, Władysław Nadopta (the prototype for "Majster Bieda" from Wojciech Belon's song), defines Kucaba's sensitivity. As the creator himself notes, describing Nadopta as "a man sculpted by the wind," calm, speaking in a quiet voice, and carrying "our hidden dreams of true joy on his back," sculpture and art are spaces inviting deep "musing" and empathy with nature (TVP, 2023).

This transition from the weight of sculptural matter, embodied by monumental spatial forms and harsh Bieszczady winds, to the dispersed, immaterial, and ephemeral networks of digital cooperation within MCSEiA constitutes a paradigmatic example of the contemporary evolution of the very essence of art. Within this vast, umbrella project, the theatrical trend is not understood as a space for classical repertoire production based on the reproduction of ready-made dramatic texts. It assumes the function of a transdisciplinary laboratory of postdramatic forms, in which the multidimensional relations between the human body, language, movement, and ubiquitous digital technology are examined with scientific rigor.

An analysis of the activities and programmatic assumptions of MCSEiA requires the application of a highly complex conceptual apparatus, extending beyond traditional theatre studies. This comprehensive research report attempts a theoretical systematization of the assumptions and practices of this trend. The structure of the argument is based on research on postdigital theatre (including the deconstruction of binarity and the concept of postdigital co-presence), an insightful analysis of the theory of intermediate ontology (covering the evolution from the traditional puppet to the virtual avatar), the concepts of techno-animism and cosmotechnics in Eastern approaches, as well as the achievements of contemporary media archaeology. By rigorously examining this unique intersection of Eastern and Western heritage, the report demonstrates how performative arts adapt to

the environmental conditions of the 21st century, creating hybrid platforms for meaning-making where embodiment meets algorithmic agency.

1. POSTDIGITAL THEATRE AND THE DECONSTRUCTION OF BINARITY

The rapid development of immersive forms, artificial intelligence algorithms, global telematics networks, and virtual and augmented reality (VR/AR) technologies has forced performance researchers to radically revise previous, orthodox views on the essence of "liveness." Postdigital theatre marks a historical critical point where the ontologies of traditional performance and new media converge, creating phenomena perceived not as alienating or technologically uncanny, but as profoundly integrated with the everyday conditioning of human experience. Technology has long ceased to be merely an impressive scenographic tool; it has become an immanent part of dramaturgy, the narrative tissue, and the cognitive structure of the contemporary subject.

1.1. OVERCOMING DIGITAL ALIENATION ON A GLOBAL SCALE

In the classical, modernist approach, introducing advanced digital technologies onto the theatrical stage was very often associated with the critical charge of "alienation" of live theatre. This alienation was supposed to destroy what was considered most valuable and sacred in the theatre: the direct, uninterrupted physical exchange of energy and emotion between the live actor and the spectator present in the same space. However, the contemporary postdigital perspective goes far beyond this reductionist dualism. Research on African performative spaces, in particular analyses of Nigerian theatre, shows that the postdigital paradigm allows for going beyond digital alienation in live theatre, adapting new media as a natural extension of local forms of expression (Ojoniyi, 2024). Researchers analyzing the interactions of the younger generation express a similar tone—studies conducted on a group of Australian youth aged 13-22 in four different theatrical productions prove that the phenomenon described as "everyday postdigital theatre" seamlessly integrates the digital experiences of young people with the scenic form (University of Melbourne, n.d.).

What is digital is already so deeply and inextricably woven into the fabric of social reality that the artificial division into a "real" (corporeal, physical) and "fake" (digital, mediated) environment loses any epistemological justification. The conceptualization of postdigital theatre in educational and artistic contexts assumes the emergence of new rehearsal formats, in which gamification and digital media become the basis of creative methodology (Postdigital Participation, n.d.). Instead of hiding technological apparatus in the wings, postdigital creators deliberately expose it, subjecting it to a so-

called sociomaterial performance analysis, precisely examining the moments when "digital stuff plays a role" on stage (Felix, n.d.).

1.2. CONSTRUCTING POSTDIGITAL CO-PRESENCE

A key concept, without which it is impossible to understand the phenomena occurring within the dispersed theatrical projects of MCSEiA, is "postdigital co-presence." This concept, intensively developed in recent years by researcher Tamara Radak, among others, on the margins of analyses of hybrid performative forms forced by the global pandemic, radically problematizes the classical definition of theatre, according to which a performance is synonymous with the meeting of performers and the audience exclusively in one defined physical space (Radak, 2023).

The phenomenon of postdigital co-presence is characterized by a progressive blurring of the demarcation lines between embodied and strictly virtual participation (Radak, 2023). In this new ontological regime, classical physical proximity ceases to be the sole guarantor of a genuine experience. It is successfully replaced by deep affective resonance and emotional alignment with a remote audience dispersed across the network (Radak, 2023).

An outstanding and frequently analyzed example of this mechanism is the performance *To Be a Machine (Version 1.0)* by the Irish group Dead Centre from 2020. It was a livestreamed adaptation of Mark O'Connell's non-fiction book exploring the environments of transhumanists, cyborgs, and hackers striving to solve the problem of death (O'Connell, 2017; Radak, 2023). This performance precisely examined the incredibly strong tension between digital co-presence and the embodied spectator. The creators proved that deep empathy and emotional co-presence at a distance are possible, constructing a specific bond with the avatars of viewers staying in their own homes (Radak, 2023). Such an approach fits into a broader, already century-old discourse concerning mediated performance, in which researchers such as Sarah Bay-Cheng or Kurt Vanhoutte negotiate the status of phenomenological presence, proving that the digital medium does not eliminate liveness, but transforms it (Vanhoutte and Wynants, 2011; Radak, 2023).

In an epistemological approach, analyses of postdigital theatre lead to the proposal of an entirely new reconceptualization of performing arts. As researcher Ulf Otto suggests, using the framework of actor-network theory, contemporary theatre can be analyzed not only through the prism of stage actions but also perceived as a flow of massive datasets ("big data"), which exposes new dimensions of the legitimization of presence in the post-pandemic era (Otto, 2023). In MCSEiA's strategy, this means the deliberate realization of bilocated, telematic, and strongly hybrid performances. The physical isolation of

actors and viewers does not constitute a lack of presence in this perspective. On the contrary—it becomes a distinct aesthetic value. Any technical "imperfections" of the communication medium, such as latency, sudden visual glitches, or asynchrony in transmission, are neither hidden nor treated as errors. They constitute a fully-fledged, rough dramaturgical material perfectly reflecting the precarious and mediated conditions of the contemporary existence of the global subject.

2. INTERMEDIATE ONTOLOGY: THE PUPPET AS A PROTO-AVATAR

The traditional Western cognitive model (both scientific and artistic) relies on sharp, hard dichotomous divisions: the animate is separated from the inanimate, the human from the non-human, the material from the spiritual. MCSEiA's laboratory activity categorically postulates a departure from this exhausted model in favor of a multi-layered "intermediate ontology."

2.1. BLURRING BOUNDARIES AND LIMINAL ENTITIES IN A SYSTEMIC VIEW

Before this concept found application in performance studies, it evolved within knowledge engineering and computer science, where "intermediate ontology" is defined as an advanced structure ("extra storage") allowing for the creation of logical bridges, bridging axioms, and subsumption relations between various local conceptual systems (Lemmens, n.d.; ONTOL2, n.d.). In domain modeling, this allows for the seamless import of object concepts (e.g., by an artificial intelligence agent like Disciple in collaboration with a domain expert) and the leveling of barriers between separate knowledge bases (Lemmens, n.d.; ONTOL2, n.d.). In a similar spirit, this term is adapted to build base ontologies for artificial intelligence in linguistic analysis, categorizing areas such as theatre or drama (Base ontology, n.d.).

In the context of humanities, philosophy, and performance studies, the term "intermediate ontology" takes on a profound metaphysical and anthropological meaning. Researchers of ancient belief systems and ancient religions note that, for instance, in Greek polytheism, there were many intermediate entities, superhuman and endowed with unpredictable agency. These entities—such as winds, river elements, cosmic phenomena embodied by Eos (dawn), or the destructive Erinyes (deities of guilt and revenge)—functioned on their own terms, entirely independently of the whims of gods and the actions of mortals (Greek polytheism, n.d.). They represented an intermediate ontology, operating in zones where the boundaries of subjectivity were blurred (Greek polytheism, n.d.). Similar epistemological "grey zones" can be found in approaches of critical realism philosophy, which attempts to stretch its cognitive frameworks between extreme research subjectivism and hard objectivism. Critical realism creates a

space of intermediate ontology and anti-positivist epistemology, where both emergent, actual events and hidden, permanent structures of reality are examined (Koponen, n.d.).

In the tradition of European puppetry, the animated object has for centuries been mostly reduced to the sad role of a dead item, instrumentally and patronizingly manipulated by the supposedly sole, fully subjective creator the live actor. However, looking through the prism of the rich traditions of the Far East, the puppet becomes something absolutely different: it gains the status of a fully-fledged intermediate entity. From the perspective of mystical forms such as the Japanese *Bunraku* or the Indonesian shadow theatre *Wayang Kulit* (where the master of ceremonies, the dalang, operates shadows through an intricate system of rods often mounted in a massive physical banana tree trunk placed in front of the screen), the object becomes a vessel for a real, albeit immaterial, spiritual presence (Du and He, n.d.). At the moment of animation, this object gains a liminal agency it ceases to be a thing and becomes a "being."

From such an outlined conceptual framework emerges a crystal clear, direct conceptual bridge between the traditional, archaic puppet theatre and ultra-contemporary virtual environments. The contemporary, polygon-generated avatar in the metaverse space is not merely a cold "digital substitute" for a human actor. It should be perceived as a structural and ontological heir to ancient animated forms; it constitutes the next, fully evolutionary stage in the development of "intermediate entities," through which the free transfer of intentionality, vital energy, and emotions of the human creator has passed for centuries.

2.2. NON-HUMAN ACTORS: FROM ROBOTIC ANDROIDS TO VOCALOIDS

Implementing intermediate ontology in physical and virtual theatrical space finds its extreme and simultaneously fascinating embodiment in contemporary Japanese theatre, which constitutes a key and fundamental cultural reference point for MCSEiA's experiments. The unrivaled pioneer of deconstructing traditional, anthropocentric agency is the Japanese playwright, director, and leader of the Seinendan group Oriza Hirata (b. 1962). The owner of the Tokyo-based Komaba Agora Theatre gained global fame as the creator of the "Contemporary Colloquial Theatre Theory," through which (since the late 1980s and early 1990s in dramas such as *Sōru Shimin* or the 1994 award-winning *Tokyo notes*, staged in Brest, France, by Frédérick Fisbach) he radically broke with Western accretions in Asian theatre (Hirata, 2012).

Hirata argues that modern Japanese theatre in the 20th century developed unnaturally, arbitrarily importing European masterpieces (Shakespeare, Ibsen, Chekhov, Maeterlinck) and ignoring local specificities in favor of forced pathos (Hirata, 2012). In response, Hirata constructed a new stage

grammar based on everyday Japanese: he removed pronouns, repeatedly used colloquial verbs, completely rejected unnatural stress accents, introducing instead uncomfortably long pauses and simultaneous, chaotic "chitchat" of actors, intended to imitate the "noise" of life (Hirata, 2012).

This language, devoid of theatrical exaggeration, proved to be an ideal ecosystem for Hirata's next step. Working, among others, as a researcher and special advisor at universities (Osaka University Center for the Study of Communication-Design, Shikoku Gakuin University, Tokyo University of the Arts), Hirata initiated the unprecedented "Robot Theatre Project" ("android theatre") (Hirata, 2012; Seinendan, n.d.). In productions such as the groundbreaking *Sayonara* or the transmedia adaptation of Franz Kafka's famous short story *The Metamorphosis* (*The Metamorphosis: Android Version*), physical machines highly advanced androids with a humanoid appearance—were introduced onto the stage for the first time as equal and fully-fledged dialogue partners for live, human actors (Hirata, 2012; Rosner, 2018). A staff of specialists (technicians, lighting and acoustics designers from the Seinendan studio) operated around these non-human subjects, creating a full-scale theatre (Hirata, 2012). In this extraordinary clash of the flat, hyper-realistic language of CCTT theory with the cold but precise mechanics of the android, the phenomenon of the uncanny valley is aesthetically tamed. The artificiality of the machine paradoxically deepens the sense of existential emptiness and truth on stage, mercilessly problematizing the classical hierarchy of subjectivity.

An even more radical shift on the scale of intermediate ontology exploration is the evolution of dematerialized, entirely virtual entities in grand operatic space. An absolutely groundbreaking work in this field was the *Vocaloid Opera "THE END"*, which had its world premiere on May 23, 2013, at the famous Bunkamura complex (Shibuya, Japan), with later shows held in Paris in November of the same year, and subsequently broadcast to cinemas worldwide (including in the USA and Australia) (Shibuya et al., 2013). *THE END* is an outstanding work on a global scale: it is an opera produced in collaboration by composer and concept creator Keiichiro Shibuya, director and playwright Toshiki Okada, prominent visual artist YKBX, and a sound artist hiding under the pseudonym evala (Shibuya et al., 2013; Okada, 2013). The architectural side of the scenography was designed by Shohei Shigematsu from the New York office of legendary architect Rem Koolhaas, and the project even included the artistic director of the fashion house Louis Vuitton, Marc Jacobs, who, based on his Spring/Summer 2013 collection, tailored special, unique "Damier" style costumes for the virtual protagonist (Shibuya et al., 2013).

This main protagonist and simultaneously the only soloist in a work devoid of any participation by live opera singers or an orchestra was Hatsune Miku, a virtual idol, the embodiment of Yamaha's Vocaloid vocal synthesis software (programmed in this production by artist PinocchioP) (Shibuya et al., 2013).

THE END threw the audience into the abyss of total sensory overload, generating a computer world on four massive screens using seven highly powerful projectors (exceeding 10,000 lumens each) accompanied by perfect 10.2-channel surround sound (Shibuya et al., 2013). However, it is not the technical specification that testifies to the power of this work, but its philosophical charge. The opera's narrative unfolded through arias and recitatives sung in Miku's synthetic yet deeply poignant voice. The work, based on a traditional tragic structure, placed at its center a virtual construct that realizes she is not a human being, and then asks heart-wrenching, fundamental ontological questions: "What is death?", "What is an end?", and "Am I dead?" (Shibuya et al., 2013). In this precisely arranged environment, the holographic intermediate entity becomes a projection screen through which the audience does not so much admire the technology as dramatically negotiate their own thoroughly human fears of finality and being replaced by machines. The global and powerful phenomenon of Hatsune Miku, sealed by endless international concert tours and industry exhibitions (HATSUNE MIKU EXPO) (Hatsune Miku Expo, n.d.), indisputably proves that postdigital society easily, naturally, and unreflectively accepts the powerful agency of non-human creators of higher affects.

3. TECHNO-ANIMISM AND COSMOTECHNICS

To fully and maturely realize the ambitious postulate of an in-depth intercultural dialogue, MCSEiA inherently rejects and goes far beyond the paralyzing Western, Cartesian ontological paradigm. This paradigm assumed a radical, sharp separation of nature and the spiritual world from the soulless machine. This historical and philosophical division in Europe and North America gave birth to a culture that treats the machine exclusively and cynically as a dead tool used to maximize economic efficiency and brutally conquer the natural environment. A creative and cognitively powerful alternative to this exhausted discourse is turning to ideas derived directly from the spiritual and scientific space of the Far East—in particular to the concepts of "cosmotechnics" and "techno-animism."

3.1. COSMOTECHNICS AND THE TURN TOWARDS TECHNODIVERSITY

Probably the most important intellectual figure for understanding the contemporary dimension of this phenomenon is the philosopher of technology hailing from Hong Kong, Yuk Hui. Conducting a profound reflection on the critique of modern technology formulated decades ago by Martin Heidegger, as well as referring to the non-monistic and multinaturalist anthropological experiments of Philippe Descola, Hui proposes a revealing third way of systemic exit from the trap of modernity: the concept of cosmotechnics (Hui, 2016).

Yuk Hui radically and consistently rejects the hegemonic Western approach, according to which all contemporary thinking about technique (techne) derives from a single, common source the Greek myths of Prometheus stealing fire and the imprudent Epimetheus (Hui, 2016). Such an oversimplifying approach assumed imposing an identical, universal anthropological matrix on the technological development of the entire world, completely negating the diverse experiences of the East (Hui, 2016). Criticizing the unreflective affirmation of technique and technology as a supposedly universal principle, Hui pushes the nuanced concept of "technodiversity" (Hui, 2016). He argues, extremely rightly, that the meaning of and relationship with technology in diverse cultural circles (including Japan, the centuries-old empire of China, or the ancient cultures of Latin America) have historically been experienced, literarily described, and socially integrated in a fundamentally different way than the European one (Hui, 2016). Consequently, imagining a specifically Chinese or Japanese philosophy of technology, rejecting the universality of Heidegger's postulates, is not an intellectual whim, but an urgent, burning need (urgency) (Hui, 2016).

In the approach of the cosmotechnical paradigm, technology is never described as a destructive force alienated from and hostile to the order of the natural world; it can and absolutely must be harmoniously reintegrated into a moral cosmic approach (Hui, 2016). Returning to nonmodern approaches in philosophy does not mean, however, a naive, conservative, sentimental retreat to the forms of the past (Hui, 2016). It is a process of selective, highly thought-out re-actualization of ancient wisdom and traditions on entirely new computational scales and in a thicket of innovative digital contexts (Hui, 2016). Given the progressive degradation of the planet in the Anthropocene era, such a thought experiment ceases to be just academic theory and gains the highest degree of critical relevance for the very survival of the human species (Hui, 2016).

3.2. THE VITALITY OF ALGORITHMS AND THE INCANTATORY POTENTIAL OF DIGITAL MAGIC

A fundamental, integral element of such a deep re-actualization of Eastern traditions is the revival of the concept of techno-animism. This system of thought, which in countries like Japan has natural roots in the animistic Shinto beliefs attributing a hidden spirit and vital energy to every object (from a sacred boulder to an advanced microwave oven), allows contemporary users a diametrically different treatment of advanced computing systems. In this remarkable perspective, vast datasets, source codes, transaction systems from blockchain areas, or artificial intelligence agents cease to be cold, alienated analytical tools. They gain the status of equal environmental subjects mysterious structures exhibiting their own evolutionary traits, self-organization, and even a unique vitality. Instead of building simple classifications of supposed monsters from an Eastern bestiary (thereby avoiding the colonial research

gestures of the past), contemporary, advanced analyses of techno-animism focus precisely on describing the ontological status, innate agency, and fascinating behaviors triggered in users in relationships with non-human agencies (Hui, 2016). By clashing ancient Amerindian perspectivism and advanced Japanese techno-animism with contemporary cryptography, researchers seriously ask a revolutionary question: are we able to perceive the traits of a living organism in technology? (Brouwer, 2018).

At the level of computer science foundations in the architecture of the software itself this radical phenomenon smoothly enters a powerful and fascinating sphere, often described by literature researchers as the study of "Algorithm Magic." Referring to the works of Gilbert Simondon and Vilém Flusser, among others, researchers note a terrifying yet intriguing technical fact: contemporary technological apparatuses, thanks to machine learning, can and increasingly often function in an evolutionary manner that is largely independent of the original intentions of their human programmers rigidly inscribed in lines of code (Marenko, 2019). In this extensive conceptual context, the mysterious, never fully knowable algorithm unleashes its "incantatory potential" (Marenko, 2019). Yuk Hui himself sharply polemicizes with shallow, populist comparisons treating a complicated algorithm merely as a flat instruction and a predictable sequence of events—a culinary "recipe" that simply blindly copies step-by-step a given schematization within pure, primitive automation or mechanical repeatability (Marenko, 2019). The true driving force of modern digitality is something completely different at the opposite pole of the algorithmic spectrum lies automation generated through advanced recursion processes, where mathematical functions become overwhelmingly remote and partially self-defined in real-time (Marenko, 2019). Examining layers of abstraction and successive "orders of magnitude" of digital objects to create a theory of relations, Hui highlights the phenomenon of interobjectivity as critical for capturing the connections of an object with its complex and dynamic IT environment (Brouwer, 2018).

By boldly juxtaposing the European critique of technological progress, underpinned by a constant fear of the "rebellion of the machines" and a moralizing tone, with indigenous, profoundly Eastern spirituality holistically oriented towards coexistence with the animated object, the theatre formed in laboratories like MCSEiA acquires tools of incredible ethical power. Expanded performance studies become an exceptional and arguably the only fully autonomous medium to organically and empathetically explore the metaphysical status of digital agents, treating artificial intelligence not as a threat from the server room, but as fascinating and fully-fledged "new actors" hosted on the dematerialized and infinite boards of the postdigital stage.

4. DIGITAL PUPPETRY AND MEDIA ARCHAEOLOGY

Understanding the fascinating phenomenon in which dematerialized logical systems, multi-layered algorithms based on machine learning couple and then merge into one with the actor's embodied experience, requires an absolute reference to sophisticated strategies and methodologies applied in the research field designated by so-called media archaeology. This current, functioning simultaneously as an intellectual and artistic axis within MCSEiA projects, enables a radical and entirely innovative look at the continuous and not-at-all-outdated timeline of the evolution of spectacular and visual forms.

According to the argument presented by one of the most outstanding specialists in this unique field, Professor Erkki Huhtamo, virtuoso practice of media archaeology is not simply historical archiving (Huhtamo, 2013). It is a precise tool that allows for the relentless and accurate discovery of hidden innovations ("the new in the old"), bringing to the surface forgotten and sometimes brilliant aesthetic solutions of old, abandoned media: archaic and expressively powerful mechanical theatres, early optical devices generating dark visual illusions (like the magic lantern), or the breathtaking, extensive mechanical stage technology known from gigantic European stages at the turn of the 19th and 20th centuries (Huhtamo, 2013). Today's interaction with contemporary, laser-packed spectacles is essentially a continuous deconstruction of deeply hidden returns of old forms under a new, gleaming, fiber-optic mask (Huhtamo, 2013).

4.1. THE ANCIENT SHADOW AS AN AESTHETIC AND TECHNOLOGICAL PREFIGURATION OF VIRTUAL FORM

Immersing oneself in this research perspective, the traditional, flame-illuminated screen used by craftsmen in the phenomenal Ottoman shadow theatre *Karagöz* is not and should not be reduced by the contemporary researcher solely to the role of a closed, dying historical artifact or folklore. On the contrary: for the historian of perspectival technology, it becomes nothing else but a primal and flawless prototype of a modern, immersive visual interface, playing the exact same cardinal role in the history of media as the illuminated film projector, and today the specialized virtual reality (VR) goggles worn over the eyes enclosing the visual world.

The powerful and mystical phenomenon of classical shadow theatre still possesses incredible, resonating power and the ability to be reborn in extremely unfavorable geopolitical and post-traumatic circumstances. Shocking stories from the Middle East emphatically testify to this. An example is the activity of diaspora-born Syrian-Armenian-American actress with experience on New York stages (William Esper Studio), Sona Tatoyan, and the Turkish master of shadow animation art (creator of the

Karagöz Theater Company in Washington), Ayhan Hülagü (Tatoyan and Hülagü, n.d.). The actress, during an extremely difficult visit to still war-torn Aleppo in Syria, found a suitcase lost for decades in the attic of her family home containing masterpieces: handmade, ancient leather *Karagöz* shadow puppets belonging to her own grandfather (Tatoyan and Hülagü, n.d.). The collaboration of artists, consisting in reanimating this forgotten heritage in the spotlight, proved to the world once again the incredible, transgenerational power hidden in the seemingly simple act of animating a two-dimensional, precisely cut shadow cast onto a thin, parchment, semi-permeable canvas (Tatoyan and Hülagü, n.d.). Such a cast, luminous and pulsating shadow, enlivened by the rhythmic, trance-like, and invisible strikes of the master located in the off-screen zone, fascinatingly creates a suggestive illusion and promise of an entirely separate, parallel virtual universe. Attempts at the unreflective, automated transfer of the specific, organic physical features of Eastern shadow theatre directly into VR glasses software studied by teams of programmers seeking interfaces for Indonesian puppets—show the massive scale of the technological challenge (Du and He, n.d.). The mathematically generated space lacks such trivial, yet from a corporeal perspective momentous, haptic elements as the resistance of matter in the hand, the enforcement of an uncomfortable sitting position, and above all the massive, heavy banana tree trunk in front of the puppeteer, which plays a key resistance role during strikes with a wooden rod setting the animation in motion—which the VR system simply cannot simulate at the moment (Du and He, n.d.).

4.2. THE ONTOLOGY OF DIGITAL PUPPETRY: TRANSFER OF PRESENCE

This is where the space for the most innovative research current is born: the smooth combination of ancient narrative techniques characterizing the craft of object manipulation masters with complex, sensitive motion capture systems (based on registering markers at a frequency of hundreds of frames per second), precise projection mapping, and advanced hand tracking sets (e.g., the "Anim-actor" system (Anim-actor, n.d.)). This creates the foundations for a rapidly evolving phenomenon of performative art, christened "digital puppetry" (Kłeczek, 2020). This phenomenon, as minutely characterized and broken down into primary factors by Polish researcher Jakub Kłeczek, consists in the fascinating and extremely fluid combination of the unparalleled proficiency of the puppeteer's traditional manipulation-animation methods with contemporary, dematerialized, and plastic virtual environments. The puppeteer becomes an artist operating at the interface of reality and the matrix of zeros and ones in this process; precisely utilizing software and rigorous hardware tracking parameters (kinematics and momentum of their movement), they successfully replace traditional puppetry equipment made of lead, wires, and thin strings with binary code, bringing active, real-time reacting 3D objects and avatars functioning in non-physical augmented spaces (AR and VR) to "life" (Kłeczek, 2020). What is crucial

from the point of view of creative ethics: the intentionality of the live subject as the creator and giver of "life" remains intact (Kłeczek, 2020).

The phenomenon of digital puppetry has been developing for a long time, drawing broadly from many engineering disciplines. Its precursor, the American engineer from the 1960s Lee Harrison III, founder of the *ANIMAC* system, managed to create an extremely difficult and complex apparatus enabling a so-called anthropometric programmer hung with wires to control live (without rendering intermediary) the outline of a character named *Mr. Computer Image* (Kłeczek, 2020). A historical step on this path of technological bumps was also the disturbing digital muppet from the 1990s, designed as a bizarre cross between a lion, a worm, and a clown—one Waldo C. Graphic (Kłeczek, 2020). To this day, one of the most distinguished and extremely important pillars of this medium's development remains the Henson Digital Puppetry Studio, which scans the nuances of actors' artistic expression to transfer their talent and organic fluidity to computer films, strenuously fighting the effect of undesirable "clunkiness" of stiff joints burdening early generations of 3D graphics (Kłeczek, 2020).

Digital body control techniques have successfully penetrated from scientific laboratories to mass, often trivialized popular culture to gigantic and expansive Walt Disney amusement parks (where thousands of tourists enter into cheerful, even audacious live interactions with the character of a virtual turtle and an alien on a giant plasma screen), to amateur, mass users of free, home masking programs like FaceRig (using home webcams mimicking facial expressions to animate an avatar), and countless smaller and larger players immersed in epic MMORPG network games or fans shooting amateur film productions called "machinima" using powerful video game engines (like Unreal) (Kłeczek, 2020).

From the point of view of critical art, we are interested in highly niche and progressive experiments. Examples include interactive installation spectacles, such as the spectacular *Puppet parade* by the avant-garde studio Design I/O, relying on mass audiences using the cheapest market Kinect controllers and open-source software to freely animate their digital birds on colossal LED displays in the lobby (Kłeczek, 2020). Equally significant are the projects of the free Central European Animata environment or the mysticism-striking project of South Korean new media artist Semi Ryu and the avant-garde formation Coopuppet. She constructed an unprecedented performative experiment *YOUNG-SHIN-GUD*, in which a classical, trance, ancient shamanic ritual was recreated and filtered through computer codes (Kłeczek, 2020). In this work, a gigantic, aggressive digital dragon marionette was collectively enlivened, spurred to fury, or calmed not by hands with wires, but through massive decibels of audience voices and rhythmic beats of a shamanic drum picked up by sensitive microphones straight from the floor (Kłeczek, 2020). We must also remember the widely explored genre of cyberformance (such as epochal

installations like *The Hamnet Project*, experimental *Desktop Theater*, or performative excesses of the legendary avant-garde formation *Second Front* raging on the ashes of the *Second Life* virtual server) (Kłeczek, 2020).

All this points to an irrevocable conclusion: digital puppetry paradoxically was not born out of an absolute adoration for cold technology and matrix perfection, but largely constitutes a clear manifestation of the psychological disappointment experienced when subjecting the human body to the dictate of cold, prepared mechanicalness of classical, precise, and frame-by-frame 3D mathematical animation in early production computers. It turned out over time beyond any doubt that for a hyper-modern spectral object made of luminous pixels to be admired as something natural, desirable, and evoking authentic emotion, it must absolutely draw from the ancient source—it needs exactly the same thing that animates a rag marionette. It needs the "breath of a living spirit" straight from the puppeteer's hands, the transfer of nuances, the trembling of a tired muscle, a hidden yawn, or the physiology of respiratory rhythm encoded by human neurobiology in nerve fibers (Kłeczek, 2020). For postdigital institutions like the platforms coordinated by MCSEiA, the use of laser detection systems does not serve a merely primitive and purely marketing "modernization" of ancient heritage; the goal is its conscious and deeply empathetic revival (*re-enactment*) for soulless digital natives, with full and even reverent preservation of deep, religious respect for the primal, ritual cultural matrices constituting the backbone of animated movement in culture.

5. BETWEEN REPERTOIRE AND ARCHIVE: THEATRE AS A RADICAL ANTI-ARCHIVE

The key, constantly recurring, and perhaps most important political problem resulting from the frontal, unbridled clash of the latest technological structure with sanctified, old performative forms is undoubtedly the politics and processing of cultural memory. The International Center of Art of Europe and Asia unambiguously and with all programmatic force cuts itself off from potentially functioning as a stagnant, decaying museum or any other classical Western institution whose sole pre-imposed role would be strictly preserving dead artifacts and relics of the past (i.e., de facto generating a static documentary space killing live processes by definition). This project consistently adopts and implements, almost in an anarchic manner, the postulate and logic of the "anti-archive." This construct assumes at its very ontological foundation that the most important and indisputably paramount remains the constantly expanding, infinite open creative process, valuing above all fleeting (ephemeral) forms, incessant artistic improvisation, dirty clashes of random phenomena, and the intensity of a highly subjective, individual experience, placing all this in violent opposition to alienated, format-closed, static, and objective meta-narratives protected in the cages of rigid conservators' protocols.

5.1. THE DICHOTOMY OF PRESERVATION AND INERADICABLE DISSONANCE: ARCHIVE VS. REPERTOIRE

In order to properly and satisfactorily conceptualize these bold assumptions of such an outlined institutional anti-archive, it is necessary and unavoidable from the perspective of scientific discourse to refer directly and broadly to the gigantic intellectual contribution and uncompromising cognitive revolution made in the field of leading performance theories by researchers and theorists Diana Taylor (profoundly analyzing the archive politics of the Americas) and the famous Rebecca Schneider. The works of both intellectuals, completely revising the dogmas of thinking about history, struck with massive force at an exceptionally exclusionary, patronizing, and quite imperialistically logocentric discourse dominant in the West, favoring the preservation of material and written information carriers about powerful memory processes (Taylor, 2003; Schneider, 2001).

The conclusion based on Taylor's solid and long-term research assumes the existence of a brutal rupture in culture; a powerful, established, and hegemonic dominant system of cultural communication has for millennia been completely appropriated, controlled, and formed by the logic of a phenomenon defined by her as The Archive. The Archive is an absolutely reasoned conceptual area, which in the mind of Western man includes almost exclusively all preserved artifacts: dried fossils pulled from clay, diligently carved scrolls of writing, decayed historical state documents, visual museum relics of fallen civilizations, bricks from massive buildings, or dug-up bones (Taylor, 2003). It is precisely this institutionalized Archive that for hundreds and thousands of years greedily created in the collective imagination of societies a completely groundless illusion, erecting an inaccessible and cold monument as a supposedly "hermetic reservoir of memory," a construct theoretically and physically resistant to the blows of time and the painful erosion of generational change (Taylor, 2003). In this rigorously controlled logic, only inscribing an object on the Archive's list, i.e., depositing its solid form in a museum display case or the vast shelves of a library, was supposedly and exclusively linked to the desired survival of the so-called objective historical "truth" (Taylor, 2003).

In a drastically arrogant and non-accidental way for the logic of logocentrism, this systemic hegemony completely marginalized and long suppressed a powerful counter-force defined by Taylor precisely as The Repertoire. Steadfast in its living and rigorous structure, the Repertoire is the absolute opposite of fragile archives of paper and ice it functions entirely as an invisible domain of multi-generational stage gestures repeated by communities in a creative act, folk and deep trance dances, oral traditions of singing and passing on legends, or the shocking ritual of conjuring gods and boundary entities. According to the diagnosis made in texts groundbreaking for theatre studies, the phenomenon of the

Repertoire is this inherently "performative" liminal process of continuous "happening" set in constant motion, which successfully rescues the supposedly lost memory of humanity in the fog of centuries, doing it masterfully through the uninterrupted act of repetitive recreation, the immediate embodiment of the experiences of previous generations here and in this second for the next ones (Taylor, 2003). The Archive sits quietly and unchangingly plastered with armor in display cases of blue glass behind a thick rope of no entry; meanwhile, the Repertoire violently pulses with life, screams in the hot tendons and bloodstream of participants of collective tribal surges—and never in the history of mankind has it allowed itself to be enclosed and tamed in binary scan files or a cataloger's thick notes.

5.2. FRACTURES OF MEDIA: BODY-MEMORY VS. COLD BODY-ARCHIVE

A fundamental ethical consequence of unconditionally recognizing the authentic vitality of the Repertoire made from the point of view of a modern postdigital system is a profound and deeply intimate reflection on the fragility, and at the same time, the great strength and significance of the status of the physical human body, sharpened and ruthlessly specified in the pioneering and shocking scientific arguments authored by Rebecca Schneider. She captures this extremely convoluted and morally slippery topic without a shadow of hesitation through a categorical splitting of the concept of human physiology into two clashing figures: the concept of the powerful and life-giving "body-memory" brutally colliding with the dead mass and document embodied by the soulless construct of the "body-archive" (Schneider, 2001).

On a very clear, one side of the debate, the radical concept of the human being treated in science as an almost autonomous medium of the powerful memory of the human species (body-memory) unambiguously, violently, and uncompromisingly highly values and glorifies the incredible, not fully medically explained built-in resources of the biological carrier as an incredibly durable, powerful, and relentless vehicle allowing for the act of direct transmission of dramatic events from the past straight and without delay into the tissues and nerves of a contemporary, live breathing body of an actor on stage here and now, embodying and drawing crowds of electrified spectators into the performance (Schneider, 2001). Such a body reveals with the destructive force of a hurricane the pyramidal nonsense, colossal illusion, and brazen, unjustified empty claims made for exclusive monopoly access to the "history of the human race" by the logocentric administration of the traditional historical state archive system serving former colonizers and decision-makers. It emphatically proves to researchers and the mass public that the deepest, innate, and untamed systems of transferring powerful psychological affect—a tearing scream of terror after a brutal attack on the tribe or the wild, primal joy of the sunrise saving lives—can never be honestly conjured or objectively recorded in writing using a drop of black ink, nor can they be

captured and retained in the dark, light-prepared silver grains of old, expensive photographic film or the celluloid membrane of an analog, sharp photographic frame (Schneider, 2001).

From a dramatic and radically different cognitive perspective, capturing the tragic and painful journey of man through the storms of modernity and the degrading crimes of anti-human totalitarianisms throwing dehumanizing statistical numbers at victims, commands us to enter with an open visor into a confrontation and continuous conflict lasting at the heart of critical theory as the so-called "body-archive" showing a depressing emphasis and extraction into the sharp, blinding light in a room without a roof over its head of the darkest of the darkest, deeply rigorous, carrying fear and uncertainty, rigorous to the bone, ruthlessly accusatory dimension of human biological existence: its documentary degradation and the brutal record of a blow (Schneider, 2001). The body-archive, standing breathless exposed in the rising glow of dawn before a firing squad of external viewers' opinions, preemptively unmasks the naivety of any attempt to erase guilt—ruthlessly visualizing the impossibility of obtaining comforting absolution or repressing the truth about suffering from the mass imagination (Schneider, 2001).
It brutally deals with the lack of access to the real ashes of past centuries for younger generations, making everyone aware of the gap and holes inevitably placed on scraps of frayed, cold human flawed fragmentariness, which we commonly and carelessly in everyday slang graciously just call our sprained or full-of-holes memory.

Cruel and dark-bringing hideous crimes of centuries often cruelly place on a pedestal the paradox of an irreversible drama of unexpected "reversal of original meanings" taking place with the ruthless logic of a lens aimed at the forehead of history: these curious contemporary external researchers, young and fit performers smeared with digital inks, or the sympathetic liberal global public from the couch they all try in vain to naively force their way in an act of willingness to enter, to bash down with the battering ram of contemporaneity the high armored walls and the loudly snapping locks of the digital "archive" of a monstrous, gigantic, humanly inexplicable, oppressive, hideous historical trauma. An emphatic, shocking, brutal, yet brilliant experiment exposing the limits of art without mercy or beating around the bush, showing the fiasco of performance studies clashing with the pain of crime, is, for example, the repeatedly and not entirely exhaustively reviewed for the conscience, famous radical project by the giant of provocative Polish contemporary art, Artur Żmijewski, entitled with the ice-cold camp digital code: *80064*, consisting in a telegraphic and blood-curdling shortcut and based on pure intervention and an audacious reissue of a tattooed camp mark—a relentless and tragic scar being a reminder until the last breath of the camp bestiality of the Auschwitz extermination on an elderly prisoner standing helplessly on the edge of his life and refreshed mercilessly bleeding from the performer's fault straight into living

matter by the artist himself (Żmijewski, 2004; Schneider, 2001). Reaching as a result of this macabre recording and recreation only the superficial shell of the anesthetized coating of dead historical artifacts in the order of numbers survivors assure us of the intruder's failure at the gates of truth as those constituting walking in muscles, trembling from the pressure on the chest, tangible, authentic, painful to the core, deep human reservoir of unspoken and irreplaceable agency of unique embodied living authentic powerful internally shaking with pain tragic saved experience of direct surviving victims of the conflagration (Schneider, 2001). In the final clash with their fragility despite the relentless progress of monstrous shadows running a race against the irrevocable passage of incurable biological time sucking the remaining sap from the tree of memory firmly and victoriously, despite the cruelty of degradation, they stand at the head of humanity's march as a powerful barrier. Constantly and generously to the brim nourishing a powerful thick substance drawing from their own heart a nourishing thick stream generating a circuit powering around the circumference a forever speech-hungry pulsating repertoire of our inconspicuous and great human modest sore everyday life (Schneider, 2001).

In turn, an eminently revolutionary against previous frameworks and outdated rules, new interdisciplinary postdigital experimental institution established for the exchange of thoughts of avant-garde and transgressive arts to the bold measure of ambitious, distant visions planned within creative actions by MCSEiA strongly assumes upfront and radically positions the attitude of creators in its not-at-all modest manifesto, standing firmly with both feet with a coupled matrix on the position of defending postulates, from which merciless theses are unerringly drawn with mathematical and analytical certainty that all this so universally liked replicated traditional and boring by principle cold indifferent to the truth dead historical documentation of theatrical achievements—which is worth honestly ticking off again without beating around the bush with emphatic consideration of the theses of excellent, although standing in the pillory of modernity, classical thinkers of the era of grand scenography form theories (for example, the still thoughtfully recalled in scrupulous searches for saving the face of research logocentrism great Polish theoretician of theatre science and sage of solid humanistic thought Professor Zbigniew Raszewski contrasted with the not always successful utopian classifications of Professor Stefania Skwarczyńska trying to freeze movement after centuries with the hope of a documentary skeleton written in flawed letters, which admittedly sometimes with a tear in the eye and relief of researchers turns out to be helpful in an attempt to desperately establish the dry frames of scenography)—not only does not bring us closer in the slightest degree to the truth about the unique experience of scenic art on the boards, but turns out to be by principle categorically insolvent, impossibly and fatally harmful clumsy to the ears in a pathetic and hopeless in terms of digital efficiency attempt to recreate anything having the slightest connection with the powerful magic filling the room

completely of authentic fully perceptible as a phenomenon and union of pulsating energy and invisible impulses to the zenith absolutely innovative loud digitally connected eminently ambiguous postdigital powerful and all-encompassing resonating across thousands of coils of the telematics cascade communication network of extraordinary postdigital co-presence of bodies scattered on the world map (Radak, 2023).

The pursuit with great involvement of crowds of archivists stubbornly towards an eminently unreal powerful hubris-burdened blind to rational foundations utopia about salvation and consisting in the dream of finding a mathematized into all formulas disgustingly soulless methodology and algorithm in the name of achieving the goal of a delusional mythical in dreams endless perfect destination station, i.e., the ultimate from the perspective of video software great and perfect preservation of movement (i.e., on a utopian to the ears crazy dream of precise recording, catching in flight by gigabytes of archiving scanners with an insane obsession of archiving forever literally and figuratively and from every angle from the light falling on the pixel absolutely every tiny modest digital speeding with momentum like a bullet in a vacuum and with a speed almost close to light package of gigantic raw chaotic dark with noise saved algorithmic noise set of mathematical data flawlessly decoded and then with robot meticulousness accounted for with mathematical flair of a powerful hard set of encoded data produced and puked onto hard drives during a spontaneous brutal strike exchanging blow for blow ruthless and at the same time boundlessly unpredictable immediately harnessed interactive brutal relationship in flight entangled interaction in an extremely swinging emotionally pulsating live performance with an artificial multi-layered relentlessly reborn artificial thoroughly emancipated multi-story intelligence or powerfully pulsating with wild noise impulses chaotic invisible from the inside chaotically pulsating like a swarm of unrefined meaning-desiring audience closed silently from the rest using goggles in a tight isolating helmet VR)—constitutes indisputably for humanity a cognitive fall into a bottomless abyss of nothingness following into a dark completely soulless lifeless dead among gigabytes of sand and noise blind in the darkness empty to the core ultimate in this tunnel tragic at hand burdened with an error at the base and leading nowhere IT deserted alley impassable by definition. Overcoming it irreversibly occurs with a saving bold great surge thanks to the return of wise free performative live theatre drawing from ancient patterns straight into the vast embrace of revolutionary anti-museology bypassing sideways sweeping in the distance the systematic tombs and cold archives of logocentric domination and Western pride and the illusion of collecting paper over centuries of alleged progress.

The free, radically postdigital, and semi-anarchic, multi-threaded anti-archive bluntly and deliberately firmly rejects the dirty subconscious fetishization of a saved scrap of dry code of an inherently dead

nature and scanned from the inside by a flawed lens of deadness. The priority becomes from always to always a powerful and life-giving saving scream being born on the border of a deadly cold act embodying an untamable pure form of the phenomenon as a joyful act of violent happening of a reborn new exploding to the limits of possibility digital creature and puppeteer brimful of human essence in a phenomenal modest tale with a tear under the eyelid squinted by the glare staring at a mutually formed shape. Dispersed by principle in the network, theatre violently with the desired sudden crack along the structure with the force of a tsunami wave stops chasing mindlessly along a route designated by cold technocrats in a rat race for a hopelessly stupid technocratic illusion of digital salvation through chiseled pixel by pixel worthless after a year fixation of forms of visual reproduction with a disgustingly soulless rigorous shape of dead from the logic of greedy form collecting blind with the eye of a projector stopped in a cold pause of fossilization of a cracked shard of glass rejected as a faulty matrix in favor of full glory in a new paradigm with the breath of saving beauty revealed at the gates of the great digital oasis of ultimate revelation drawn drop by drop from the knowledge that everything, absolutely what is most high upper magnificent in its elusive fleeting in the glow constantly turning into motion powerful whirlpools of galaxies most important and fundamental in deeply thought out supported by a thousand oaks from the breath of shadow looking back and forward mystical great in its terror and delight impenetrable techno-animistic to the end of the horizon piled up loudly multicultural clash revolutionary desire to meet the unknown for centuries seeking harmony cultures desiring truth, well that flies away unnoticed but absolutely towards the stellar wind immediately with the ultimate and categorical buzzing of torn from nothingness network connections with the incomprehensible miracle of ending back violently unrolled on the hard scales of a binary server story evaporating into a great emptiness of a saved speeding simulation of the digital world coding itself to a safe asylum permanently with the power of the river and the rigor of a shaman's rite from now on forever with the rustle of the wind and exclusively with the signs of sweat and emotion firmly safely in carved by a scar in sore but alive non-human pulsating giving beginning and fire gathered around in the co-presence of avatars with their connected audience.

CONCLUSION: AN ONTOLOGICAL LABORATORY OF THE CONTEMPORARY IN DISPERSION

Summarizing the above broad considerations, based on solid conceptual frameworks of the latest, revolutionary, and avant-garde global and Far Eastern intellectual reflections and achievements on the extremely pulsating at the interface of paradigms clash of human senses, organic search for freedom of the spirit by hewing forms in stone with fiber-optic, silicon servers hidden in the underground of modern broadcasting stations of digital will generating massive computing powers simulating desired concepts of all-encompassing ontology of being, it is necessary undeniably in every respect authoritatively and with ruthless clarity and a sharp, clear exclamation mark to confirm the position and

scientific assessment of the phenomenon bringing to the heights a verifiable and innovative thesis firmly and boldly ruthlessly formulated that the initiated establishment of an open in the world of visual forms experimental project established to save non-obvious paths of exchanging multicultural discourses in modest from the outside and powerful from the inside digital labyrinths of a formation bearing the burden of a bridge, namely the International Center of Art of Europe and Asia, with gigantic steps smashes introducing the new into the corset of dead in corners of dust in the cracks of the system in museum floors crosses in a spectacular rush radically destroying decayed concrete ossified and based on yesterday's hard authoritarian cold system grand archaic dichotomously broken logocentric ruthless cracking hard cold with shame fallen frameworks within and in the circles of which so willingly even a couple of decades after the war blindly used from habit and rigor of power and with a penchant for the logic of deadness for deaf senses with the stubbornness of an old man to insist tied with a halter in a camp enclosure permanently with cemetery rigor for control to stuff irretrievably on a dead tape ticked off bureaucratized with the rigor of a stamp ticked off as dead artifacts stripped of scholarship madness thought out bureaucratized to the bone empty rigid forms in cold offices after directors so-called by the official greedily Western burdened with a vision defect short-sighted gutted of emotions and up to empty walls ghastly based on denial hypocritical institutions conservatively hidden deep in the fibers of power ossified and soullessly programmed ultimately dead institutions of disoriented lost state-concessioned culture in rigorous bonds of dead bureaucracy with the breath of a concessioned archive from a cold disgustingly objective institutional rigorous superficial and detached artistic sphere of official gala faded shows for paid dignitaries on the first cold empty for applause artificial with a stuffy atmosphere empty in terms of artificial falsified absence on cold lined with deadness rows supposedly delighted seemingly with powdered soulless with a stony gaze devoid of salvation elites with distance rows and rejected from the truth.

This project in an outstanding way placed with respect with careful emphasis and with a great modest and captivating artistic breath giving shelter in the tent of a storm calling and sketching on the ground with a grand feather the escape of freedom from traditional with rigor stuffy conventions burdened with falsehood for connoisseurs of falsehood imposing firstly on visual forms for a generation a great innovative rush to seek the truth and ultimate painful exposing the fragility of ourselves not rare discoveries from the server room in frankly sensational scans of alien masks put on willingly by the machine god and digital envoys from the pantheon of artificial vitality revealing a great terrifying with vitality embodiment at the doors of newly liberated open great explosive searching and escaping with a great echo laboratories in the era after the pandemic of ontological contemporaneity built on embodied from ashes desire in a powerful knocking out with current emptiness intimate relentless to the bone

stellar relationship rejecting cynical models of mediation standing up to open hard to fall negotiation of the truth about the condition of the burdened with terror degraded powerfully plugged relentlessly to the veins from the great server room in an alienated in a hypocritical masquerade ticking off by the rigor of consumption of the world finished postdigital subject a human trying to regain courage for centuries rebuilding a bridge to the other side liberated into the embrace of the cosmos allegedly dead in vitality awakened digital artificial pulsating machine enlivened by the eye of techno-animistic fully respecting the incomprehensible for logic crazy in dance powerful in the dark of saving glow great and liberated at dawn avatar.

BIBLIOGRAPHY

Akademia Tarnowska (n.d.) *Employee profile: Jacek Kucaba*, Faculty of Art of the Tarnow Academy.

Anim-actor (n.d.) *Anim-actor: understanding interaction with digital puppetry using low-cost motion capture*.

ASP w Krakowie (n.d.) *Employee profile: Jacek Kucaba*, Faculty of Sculpture at the Jan Matejko Academy of Fine Arts in Krakow.

Base ontology (n.d.) *Base ontology for human to further develop*.

Brouwer, T. (2018) *Proof of transaction. The Materiality of Cryptocurrency*.

Du, R. and He, L. (n.d.) *VRSurus*.

Felix (n.d.) *When Digital Stuff Plays a Role: A Sociomaterial Performance Analysis of Postdigital Theatre in Education*.

Greek polytheism (n.d.) *Intermediate ontology in Greek polytheism*.

Hatsune Miku Expo (n.d.) *HATSUNE MIKU EXPO*.

Hirata, O. (2012) *Contemporary Colloquial Theatre Theory* and *Robot Theatre Project*.

Huhtamo, E. (2013) *Illusions in Motion: Media Archaeology*.

Hui, Y. (2016) *Animism and Techno-Animism in Japan* and *Technodiversity*.

Kłeczek, J. (2020) 'Cyfrowe lalkarstwo', *Teatr Lalek*, 2020/05.

Koponen, E. (n.d.) *Critical realism*.

Lemmens (n.d.) *Intermediate ontology ('extra storage')*.

Marenko, B. (2019) 'Algorithm Magic: Gilbert Simondon and Techno-animism'. In: *Believing in Bits: Digital Media and the Supernatural*.

O'Connell, M. (2017) *To Be a Machine*. Granta.

Ojoniyi, B. (2024) 'Postdigital theatre: Beyond digital alienation of live theatre on the Nigerian performance space', *Nigeria Theatre Journal*, 23(2), pp. 161-175.

Okada, T. (2013) *Vocaloid Opera "THE END"*.

ONTOL2 (n.d.) *ONTOL2-Proceedings*.

Otto, U. (2023) *Presence and Precarity*.

Postdigital Participation (n.d.) *Rehearsal formats in postdigital drama teaching*.

Radak, T. (2023) "Dying … to Connect': Postdigital Co-presence in Dead Centre's To Be a Machine (Version 1.0) (2020)', *Theatre Research International*, 48(1), pp. 38-51.

Rosner, K. (2018) *The Metamorphosis: Android Version*.

Schneider, R. (2001) Considerations around affective memory in performance studies.

Seinendan (n.d.) *Robot Theatre Project*.

Shibuya, K. et al. (2013) *Vocaloid Opera "THE END"*.

Tatoyan, S. and Hülagü, A. (n.d.) Karagöz shadow puppets in: event at Harvard University.

Taylor, D. (2003) Concept of the Archive and the Repertoire.

TVP (2023) *Rzeźbiony wiatrem "Majster Bieda" zaprasza do dumania*, Regiony TVP.

University of Melbourne (n.d.) *Everyday postdigital theatre*.

Vanhoutte, K. and Wynants, N. (2011) 'Performing Phenomenology: Negotiating Presence in Intermedial Theatre', *Foundations of Science*.

Żmijewski, A. (2004) *80064* (video/performative project).

Konrad Szczebiot ORCID 0000-0002-9833-5583

INDEX OF AUTHORS

PATRONAGE

National Centre for Research and Development

Prof. Dariusz Surowik - Rector of University of Lomza

UNIMA Research Commission

International Association of Theatre Critics Polish Section

Dr Mariusz Chrzanowski - Mayor of the Town of Lomza

Polish Foundation Pillars of Development

Podlaska Regional Development Foundation

klauCODE - High-Performance platforms for modern business

ISBN 978-83-60571-89-7

UNIVERSITY
OF LOMZA