ComSIS **Computer Science and Information Systems**

Volume 23, Number 3, June 2026

Contents

ISSN: 2406-1018 (Online)

Computer Science and Information Systems

Vol 23, Number 3, June 2026

ComSIS

Computer Science and Information Systems

Published by ComSIS Consortium

Volume 23, Number 3
June 2026

ComSIS is an international journal published by the ComSIS Consortium

ComSIS Editorial Office:
University of Belgrade, Faculty of Organizational Sciences,
Jove Ilića 154, 11000 Belgrade, Serbia
Phone: +381 11 3950 866; +381 66 8896 155
www.comsis.org; Email: comsis@fon.bg.ac.rs

Volume 23, Number 3, 2026
Belgrade

Computer Science and Information Systems

ISSN: 2406-1018 (Online)

The ComSIS journal is sponsored by:

Ministry of Education, Science and Technological Development of the Republic of Serbia
http://www.mpn.gov.rs/

ComSIS

Computer Science and Information Systems

AIMS AND SCOPE

Computer Science and Information Systems (ComSIS) is an international refereed journal, published in Serbia. The objective of ComSIS is to communicate important research and development results in the areas of computer science, software engineering, and information systems.

We publish original papers of lasting value covering both theoretical foundations of computer science and commercial, industrial, or educational aspects that provide new insights into design and implementation of software and information systems. In addition to wide-scope regular issues, ComSIS also includes special issues covering specific topics in all areas of computer science and information systems.

ComSIS publishes invited and regular papers in English. Papers that pass a strict reviewing procedure are accepted for publishing. ComSIS is published semiannually.

Indexing Information

ComSIS is covered or selected for coverage in the following:

- Science Citation Index (also known as SciSearch) and Journal Citation Reports / Science Edition by Thomson Reuters, with 2024 two-year impact factor 1.8,
- Computer Science Bibliography, University of Trier (DBLP),
- EMBASE (Elsevier),
- Scopus (Elsevier),
- Ex Libris Knowledge Center by Clarivate,
- IET bibliographic database Inspec,
- FIZ Karlsruhe bibliographic database io-port,
- Index of Information Systems Journals (Deakin University, Australia),
- Directory of Open Access Journals (DOAJ),
- Google Scholar,
- Journal Bibliometric Report of the Center for Evaluation in Education and Science (CEON/CEES) in cooperation with the National Library of Serbia, for the Serbian Ministry of Education and Science,
- Serbian Citation Index (SCIndeks),
- doiSerbia.

Information for Contributors

The Editors will be pleased to receive contributions from all parts of the world. An electronic version (LaTeX) prepared as described in "Manuscript Requirements" (which may be downloaded from http://www.comsis.org), along with a cover letter containing the corresponding author's details should be submitted via journal submission system.

Criteria for Acceptance

Criteria for acceptance will be appropriateness to the field of Journal, as described in the Aims and Scope, taking into account the merit of the content and presentation. The number of pages of submitted articles is limited to 30 (using the appropriate LaTeX template).

Manuscripts will be referred in manner customary with scientific journals before being accepted for publication.

Copyright and Use Agreement

Computer Science and Information Systems

Volume 23, Number 3, June 2026

CONTENTS

Computer Science and Information Systems 23(3): i–iii https://doi.org/10.2298/CSIS260300iL

Editorial

Ivan Luković[1], Mirjana Ivanović[2], Miloš Radovanović[2], and Vladimir Kurbalija[2]

[1]University of Belgrade, Faculty of Organizational Sciences
Belgrade, Serbia
ivan.lukovic@fon.bg.ac.rs

[2]University of Novi Sad, Faculty of Sciences
Novi Sad, Serbia
{mira, radacha, kurba}@dmi.uns.ac.rs

Vol. 23, No. 3. issue of ComSIS is in front of you. We are glad to say that we have gained the trust of the authors from several countries all over the world. This issue of ComSIS contains 6 regular research papers.

In the paper: The Soft Factors in Digital Transformation Strategies: A Study of Employees Digital Competencies and Support Measures, František Sudzina, Andreja Pucihar, and Gregor Lenart present a study that examines whether the digital competencies of employees and the way in which management introduces new digital technologies to employees influence the extent of selected strategic areas of digital transformation and the importance of different types of support. The data for the study was collected in collaboration with Slovenian chambers of commerce using an online questionnaire. The authors find that the way in which new digital technologies are introduced has an impact on customer experience, processes and digital solutions to support the business, as well as the digitalization of business models, products and services. In addition, the customer experience was also influenced by digital competencies. The ability to obtain funding for digital transformation was influenced by enterprise size and published local and international examples of best practice in the field of digitalization were influenced by the methods used to introduce new digital technologies.

While Environmental, Social, and Governance (ESG) assessment plays a key role in sustainable finance, data scarcity and noise in emerging economies hinder robust model development. Ivana Marković, Adela Ljajić, Jelena Z. Stanković, Miloš Košprdić, and Jovica Stanković, in their paper Model Parameter-Based Transfer Learning for ESG Score Prediction in Developing Markets, propose a model parameter-based transfer learning with random forest (MPBTL-RF) approach for domain adaptation situations where source data are not available. The proposed model is evaluated using three traditional learning approaches: Random Forest (RF), eXtreme Gradient Boosting (XGB), and Feedforward Neural Networks (FNN). Cross-validation is used to assess model generalizability, while domain adaptation is tested through in-domain and out-of-domain settings. The authors state that proposed MPBTL-RF approach achieves competitive performance compared to traditional baselines in scenarios with limited training data, offering time advantages with predictive efficiency and stability. The authors demonstrate how machine learning pipelines can adapt to data-constrained, real-world domains, fostering the synergy between Artificial Intelligence and business.

The research presented in the paper DICYME: Dynamic Industrial Cyber Risk Modelling Based on Evidence, by Javier García-Ochoa, Jaime Rueda, Rubén R. Fernández, Alberto Fernández-Isabel, Isaac Martín de Diego, Emilio L. Cano, Romy R. Ravines,

Ovidio López Espinosa, and Jaume Puigbó Sanvisens, was also inspired by the digital transformation. The authors advocate that traditional cybersecurity methods are increasingly inadequate for addressing the rapidly evolving threat landscape, emphasizing the critical need for intelligent, adaptive, and proactive defensive strategies. In the study introduced in the paper they present Dynamic Industrial Cyber Risk Modelling Based on Evidence (DICYME), a comprehensive system that integrates diverse analytical techniques to identify patterns and characteristics that reveal emerging threat trends, enabling organizations to proactively defend against potential future attacks. Beyond threat detection, DICYME operates as a pipeline that retrieves data from diverse cyber incident reports, specialized databases, and other relevant sources of cyber-related information, applies specialized techniques for victim identification, indicator computation, threat actor profiling, Common Vulnerability and Exposure (CVE) relationship mapping, and ultimately performs the Cyber Risk Quantification (CRQ).

The Examination Timetabling Problem (ETP) is an NP-hard combinatorial optimization problem that has been extensively studied over the past decades, with methodologies ranging from classical heuristics to contemporary machine learning approaches. In their paper Examination Timetabling with Independent Resource Pools: A Multi-Stage Genetic Algorithm Approach, Khalil Chebil, Mahdi Khemakhem, Abdallah Hammad, and Mohammed Alanazi provide a comprehensive literature review and propose a novel two-stage genetic algorithm-based decomposition combined with a clear linear mathematical formulation. A contribution of this work is in addressing the practical complexity of two-separated examination facilities at PSAU, where male and female students study in separate buildings with independent examination rooms, capacities, and supervisory staff, yet must follow a synchronized examination schedule. This creates a coupled dual-ETP problem where two independent resource allocation problems must share identical time slot assignments, significantly increasing model complexity and constraint interactions.

Filip Prentović, Zoran Budimac, Marjan Heričko, and Gordana Rakić, the authors of the paper Role of Software Metrics in Practical Quality Management, advocate a use of software metrics for continuous insight into products and processes, while software metrics are a critical tool for building reliable software and assessing software quality, especially in complex domains and mission-critical environments. The aim of the research conducted in this paper is to determine the existence of different perspectives regarding the impact of software metrics on software quality between participants in software development. The main contribution is in analyzing the impact that the application of software metrics has on the quality of software and on the overall project success by statistically analyzing data obtained from the survey. The paper contains a comprehensive survey of code analysis techniques, software metrics, and the analysis of the application of software metrics in practice. Also, importance, influence, and the level of usage of software metrics in IT companies have been considered, as well as the estimation of results significance obtained in the process of measuring software.

The paper Artificial Neural Network Modeling for Air Pollution Prediction: LSTM versus the Levenberg-Marquardt Approach, by Goran Keković, Rade Božović, Sonja Ketin, Vladimir Mikić, Miloš Ilić, and Boban Vesin considers the problem of accurate prediction of air pollutant concentrations. It remains a critical challenge for environmental monitoring and public health, demanding robust and adaptive artificial intelligence approaches. The authors investigate the effectiveness of various types of artificial neural

networks (ANNs), including Long Short-Term Memory networks (LSTM) and networks based on the Levenberg-Marquardt algorithm (LM) and its variant with Bayesian regularization (LMBR), in predicting air pollution under different data conditions. Since LSTM networks are based on first derivative loss function algorithms and the LM algorithm is usually superior to this type of algorithm, a comparison of these networks was conducted. This is further supported by the limited coverage of this topic in existing literature. ANNs were tested on two different datasets: the Air Quality dataset, where the target variable was the concentration of benzene (C6H6) and the Beijing PM2.5 dataset, where the target was the concentration of PM2.5 particles.

On behalf of the ComSIS Consortium, we would like to take this opportunity to give our great thanks to the reviewers and all the authors for their high-quality work, great efforts, and remarkable enthusiasm.

Computer Science and Information Systems 23(3):947-968 https://doi.org/10.2298/CSIS250731029S

The Soft Factors in Digital Transformation Strategies: A Study of Employees Digital Competencies and Support Measures

František Sudzina[1], Andreja Pucihar[2], and Gregor Lenart[2]

[1] Department of Systems Analysis, Faculty of Informatics and Statistics Prague University of Economics and Business
W. Churchilla 1938/4, 130 67 Prague, Czech Republic
frantisek.sudzina@vse.cz (corresponding author)
[2] Faculty of Organizational Sciences, University of Maribor
Kidričeva cesta 55a, 4000 Kranj, Slovenia
{andreja.pucihar, gregor.lenart}@um.si

Abstract. The aim of the study is to examine whether the digital competencies of employees and the way in which management introduces new digital technologies to employees influence the extent of selected strategic areas of digital transformation and the importance of different types of support. The data was collected in collaboration with Slovenian chambers of commerce using an online questionnaire. Study finds that the way in which new digital technologies are introduced has an impact on the customer experience, processes and digital solutions to support the business, as well as the digitalization of business models, products and services. In addition, the customer experience was also influenced by digital competencies. The ability to obtain funding for digital transformation was influenced by enterprise size and published local and international examples of best practice in the field of digitalization were influenced by the methods used to introduce new digital technologies.

Keywords: Digital Transformation, Digital Transformation Strategy, Digital Competencies, Support Measures For Digital Transformation.

1. Introduction

The contemporary business landscape is undergoing a profound and accelerating shift driven by the proliferation of digital technologies. This phenomenon, termed Digital Transformation (DT), has evolved from a peripheral IT concern into a central strategic imperative for enterprises across all industries and sizes [14], [30]. The strategic stakes of this transformation are exceptionally high [12]. Firms that successfully navigate this shift can unlock significant competitive advantages, including enhanced operational efficiency, deeper customer engagement, and the creation of novel revenue streams. Conversely, those that fail to adapt, risk obsolescence as new, digitally-native competitors disrupt established markets [30].

In response, a rich body of academic literature has emerged to understand this complex process. Initial research focused on defining the phenomenon and distinguishing it from earlier forms of IT-enabled change [31], [32]. Scholars have established that successful DT is far more than the mere implementation of new technology; it represents a fundamental

rethinking of how an organization uses technology, people, and processes to radically change its business performance and value proposition [17]. It is now widely accepted that organizational factors such as strategy, culture, and skills are critical enablers of this change [5], [15].

However, as the field matures, the focus of inquiry is shifting from what DT is, to how it is successfully executed [31]. While we know that internal organizational factors are important [17], [33], [3], a more granular understanding is needed of the specific mechanisms through which they influence the scope and nature of an enterprise's transformation journey. Several key questions remain at the forefront of this research agenda. First, how do fundamental enterprise characteristics, such as enterprise size, interact with more dynamic, capability-based factors like employee competencies and change management processes to shape an enterprise's strategic priorities [14], [19]? Does an enterprise's internal capacity determine whether it focuses more on transforming its operational core or on enhancing its customer-facing and cultural dimensions [33], [3]?

Second, enterprises do not transform within vacuum-like environment. They operate within a broader environmental context that includes a growing ecosystem of institutional support, such as government programs and Digital Innovation Hubs (DIHs), designed to mitigate the significant risks and costs associated with DT [24]. Yet, there is a need for greater empirical insight into how enterprises perceive and prioritize these external support mechanisms. Is financial assistance that de-risks investment perceived as more critical than knowledge-based support that closes the skills gap? And how do an enterprise's internal characteristics shape its reliance on this external ecosystem?

This study addresses these critical questions by providing an empirical investigation into the determinants, scope, and support mechanisms of digital transformation. Drawing on an exploratory survey-based study, this research analyzes data from 102 Slovenian enterprises, predominantly small and medium-sized enterprises (SMEs), which reflects the structural characteristics of European economies, where approximately 99% of businesses are SMEs [9]. In this study we explore the relationships between key organizational characteristics—namely enterprise size, the digital competencies of employees, and the methods used to introduce new technologies—and the extent of enterprises'DT strategies across a range of operational and people-centric domains. Furthermore, we examine how these organizational factors relate to the perceived importance of different types of financial and knowledge-based support.

The findings of this research offer several important contributions. By analyzing the interplay between static enterprise attributes and dynamic capabilities, this study provides a nuanced understanding of the true drivers of digital transformation. It offers empirical evidence on the distinct dimensions of DT strategy and identifies which organizational capabilities are most critical for each. Finally, it provides actionable insights for managers on how to prioritize their transformation efforts and for policymakers on how to design more effective support interventions.

The remainder of this paper is structured as follows. Section 2 develops a comprehensive theoretical framework, drawing on the Technology-Organization-Environment (TOE) model and the dynamic capabilities perspective to ground our research questions. Section 3 details the research methodology, including the sample, data collection, and analytical

techniques. Section 4 presents the core empirical results of our statistical analysis. Section 5 provides a detailed discussion of these findings, interpreting their significance and linking them back to the theoretical framework. Finally, Section 6 concludes the paper by summarizing its key contributions and outlining avenues for future research.

2. Theoretical Framework

DT has evolved from a peripheral IT concern into a central strategic imperative for enterprises across all industries and sizes [14], [30]. Literature has moved beyond simple definitions to offer a more nuanced understanding of this process. Verhoef et al. [30] usefully distinguish between three stages of digital change: digitization, the conversion of analog information to digital format; digitalization, the use of digital technologies to alter existing business processes; and digital transformation, a more fundamental, enterprise-wide change that leads to the development of new business models. This study focuses on the latter, more pervasive form of change, which represents holistic strategic reorientation rather than a series of discrete technological upgrades.

In this context, DT represents far more than the mere implementation of new technology; it is a fundamental rethinking of how an organization uses technology, people, and processes to radically change its business performance and value proposition [17]. Vial [31] offers a comprehensive process-oriented definition, framing DT as "a process that aims to improve an entity by triggering significant changes to its properties through combinations of information, computing, communication, and connectivity technologies." This process compels organizations to re-evaluate not only their internal operations but also their value creation paths, customer relationships, and even their core business models, with a key outcome being enhanced product and service innovation [5].

The strategic stakes of this transformation are exceptionally high. Firms that successfully navigate this shift can unlock significant competitive advantages, including enhanced operational efficiency, deeper customer engagement, and the creation of novel revenue streams. Conversely, those that fail to adapt risk obsolescence as new, digitally-native competitors disrupt established markets [30]. The central challenge for incumbent enterprises, therefore, is not if they should transform, but how they should transform to overcome challenges and exploit opportunities from DT. This requires a coherent strategy that aligns technological capabilities with organizational goals, a challenge that is particularly acute for enterprises of all sizes. This study seeks to explore the determinants, scope, and support mechanisms of digital transformation within the specific context of Slovenian enterprises, contributing to a deeper understanding of how enterprises can successfully navigate this complex journey.

To understand the multifaceted nature of DT, a holistic framework is required. The adapted Technology-Organization-Environment (TOE) framework, first articulated by Tornatzky and Fleischer [29] and later widely used in information systems research, provides a robust and widely accepted theoretical lens for this study. Its enduring relevance stems from its comprehensive approach, positing that three primary, interacting contexts influence an enterprise's decision to adopt and implement new technology:

- The Technological Context (T): As originally conceived, this context refers to the pool of technologies available to an enterprise, both those already in use and those available externally. It encompasses not just the technologies themselves but also the enterprise's perception of their characteristics—their potential benefits, their complexity, and their compatibility with existing systems and processes [29]. In the realm of DT, this context is exceptionally rich, including a vast array of digital tools (e.g., Cloud, Analytics, Cybersecurity, Industry 4.0) that can be combined in novel ways.

- The Organizational Context (O): This context encompasses the internal characteristics and resources of the enterprise. Tornatzky and Fleischer [29] highlighted several key factors, including the enterprise's size and scope, the complexity of its managerial structure, its internal communication processes, and the availability of slack resources. In modern DT research, this context has expanded to include the enterprise's human capital, its strategic orientation, its culture, and its overall readiness for change. It is the internal environment that shapes the enterprise's ability to recognize the need for change and its capacity to enact it.

- The Environmental Context (E): This context includes the broader arena in which the enterprise operates. It encompasses factors such as the industry's structure, the presence and actions of competitors, the regulatory landscape, and the enterprise's relationships with its network of suppliers, customers, and other institutions [29]. For DT, this context is particularly dynamic, encompassing not only competitive pressures but also the institutional support systems available, such as government grants or partnerships with entities like Digital Innovation Hubs (DIHs).

This study operationalizes the TOE framework by examining specific variables within each context. The organizational context is investigated through the independent variables of enterprise size, the existence of a formal DT strategy, the level of employee digital competencies, and the method of introducing new technology. The technological context is explored through the dependent variables which measure the extent to which an enterprise's strategy covers various DT domains, representing the scope and nature of the technological application. Finally, the environmental context is addressed by analyzing the perceived importance of various external support mechanisms. This framework allows for a structured investigation into how organizational factors influence enterprise's technological choices and its interaction with the external support environment, providing a holistic view of the transformation process.

While technology provides the tools for transformation, a strong consensus in the literature emphasizes that organizational factors are the primary drivers of success. As the influential work by Kane et al. [17] argues, "strategy, not technology, drives digital transformation." This perspective suggests that the most digitally mature organizations are not necessarily those with the most advanced technology, but those with clear, coherent strategies and the internal capabilities to execute them. This people-centric point of view is evident also in recent literature review [23] that systematically categorizes the "soft factors" that are critical for success of digital transformation projects. Their findings point out, that for digitalization projects to succeed, the "people dimension" requires explicit attention. Furthermore, their research identifies four core organizational enablers—Leadership, Culture, Support, and Learning—that are crucial for both the implementation and adoption phases

of DT. In this light, this study's variables, "digital competencies" and "IT introduction methods," can be seen as tangible indicators of these broader soft factors, particularly organizational learning and support.

The role of enterprise size in DT is a subject of ongoing debate in literature. A theoretical dichotomy is often drawn between Small and Medium-sized Enterprises (SMEs) and large enterprises, as they are perceived to face fundamentally different sets of challenges and opportunities.

Large enterprises typically possess greater financial resources and specialized skilled employees to invest in large-scale transformation projects. They can leverage their scale to absorb the high costs of new systems and dedicated R&D. However, they are also frequently encumbered by complex legacy systems, rigid organizational structures, and a bureaucratic culture that can slow down decision-making and create significant resistance to change [30]. This "liability of bigness" can stifle the very agility needed to respond to fast-moving digital disruptions.

In contrast, SMEs are often characterized by greater agility, organizational flexibility, and faster, more centralized decision-making, which could theoretically facilitate quicker adaptation to market shifts [4]. However, they face a significant "liability of smallness," with well-documented constraints in terms of capital, access to specialized human resources, and market information [10]. These limitations often manifest as critical barriers to DT, such as prohibitive investment costs and a lack of employees with the requisite digital skills [20].

However, this dichotomy is not absolute. Recent research suggests that while resource constraints are real, they do not predetermine failure. Müller, Buliga, and Voigt [21] argue that SMEs can be highly successful innovators in the Industry 4.0 context when they possess a "prepared mind"—a proactive strategic orientation that allows them to leverage their agility effectively. Similarly, Li, et al [19] demonstrate from a capability perspective that the digital transformation success of SME entrepreneurs is critically determined by the development and deployment of digital competencies and dynamic capabilities, not merely by the enterprise's size or resource endowment. This shifts the focus from a deterministic view based on size to a more nuanced perspective based on capability. By including small, mid-sized, and large enterprises, this study is uniquely positioned to test this theoretical tension empirically. It allows for an investigation into whether the strategic imperative to transform is felt equally across enterprises of all sizes and how size might influence the nature, scope, and support needs of their respective transformation journeys.

A formal DT strategy is the cornerstone of a successful transformation, providing a roadmap that aligns technology investments with business objectives [14]. However, a strategy is only as good as the organization's ability to execute it. This is where the concept of dynamic capabilities becomes essential. Coined by Teece, Pisano, and Shuen [28], dynamic capabilities are defined as “the enterprise’s ability to integrate, build, and reconfigure internal and external competences to address rapidly changing environments” (p. 516). This framework is particularly relevant to DT, as it focuses on how enterprises manage change in turbulent conditions.

To understand how these capabilities operate, recent research has focused on their micro-foundations, the underlying individual and group-level factors that enable enterprise-level change. A critical contribution in this area comes from Helfat and Peteraf [13], who argue that the managerial cognitive capabilities of individual leaders fundamentally underpin enterprise-level dynamic capabilities. They define this as the capacity of a manager to perform the mental activities—such as perception, attention, reasoning, and problem-solving—that comprise cognition. This cognitive perspective provides a powerful theoretical lens for interpreting the variables in this study. The Teece [27] framework of sensing, seizing, and transforming can be understood through the cognitive abilities of the enterprise's managers and employees:

- Sensing: The ability to identify and shape new opportunities and threats. Helfat and Peteraf [13] link this to the cognitive capabilities of perception and attention—the ability to recognize emerging patterns and focus on relevant stimuli in a noisy environment.
- Seizing: The ability to mobilize resources to capture value. This is linked to cognitive capabilities of problem-solving and reasoning—the ability to design new business models and make sound investment decisions under uncertainty.
- Transforming: The ability to continuously renew the organization. This is linked to cognitive capabilities of language, communication, and social cognition, which are essential for persuading others, overcoming resistance, and orchestrating change.

In the context of this research, the variable “digital competencies of employees” can be understood as an empirical proxy for these underlying cognitive capabilities. A workforce rated as having ‘advanced knowledge and skills that they are constantly upgrading’ is not merely technically proficient; it reflects an organizational-level aggregation of the cognitive capacity to sense new technological opportunities, reason through their implications, and solve the complex problems associated with their implementation.

Building on this, recent research has sought to operationalize these concepts into a measurable, multidimensional construct. Cao, Duan, and Edwards [5] introduce the concept of a "digital transforming capability," which they define as an integration of three key pillars: digital strategy, digital skills, and digital technology use. This capability, they argue, is the direct mechanism through which enterprises achieve DT and drive outcomes like product innovation. Furthermore, they demonstrate that this capability is significantly influenced by organizational culture. Specifically, they find that an adhocracy culture (valuing innovation and agility) has the strongest positive influence, followed by clan, market, and hierarchy cultures. This provides a powerful theoretical link, suggesting that the organizational context, particularly its cultural norms, is a critical antecedent to the development of the very capabilities needed to execute a digital transformation.

Therefore, this study conceptualizes "digital competencies" and their "introduction methods" as measurable manifestations of an enterprise's underlying dynamic and transforming capabilities, which are shaped by the enterprise's organizational culture.

Digital transformation is not a monolithic event but a complex phenomenon that unfolds across various organizational domains. Recent critical reviews of the field argue that DT should be understood as a "network of interrelated processes" rather than a single, linear

progression [34]. This perspective challenges researchers to move beyond simplistic models and examine the interplay between different transformation activities.

Literature provides several frameworks for deconstructing this network. Westerman, Bonnet, and McAfee [33] identify three core pillars of transformation: transforming customer experience, operational processes, and business models. Furthermore, they refined their framework to incorporate employee experience and digital platforms [3]. This aligns with other models that distinguish between different dimensions of transformation (e.g., [25]. Drawing on this, the present study investigates two broad categories of transformation activities:

- Operational and Technological Transformation: This dimension concerns the technical core of the business. It includes initiatives such as formalizing a data management strategy, optimizing processes and digital solutions to support business, enhancing cybersecurity, and implementing Industry 4.0 technologies. These efforts are primarily aimed at improving internal efficiency, resilience, and operational excellence.
- Customer and People-Centric Transformation: This dimension focuses on the enterprise's external interface and its human capital. It encompasses strategies related to enhancing customer experience, developing digital-skilled employees and digital workplaces, and cultivating a digital culture. These initiatives are focused on creating superior market value, fostering an adaptive workforce, and building an organizational culture that can sustain innovation.

By conceptualizing DT as a network of processes across these domains, this study investigates whether the organizational determinants discussed above have a differential impact on these "hard" versus "soft" areas of transformation, providing a more nuanced understanding of the DT process and answering the call for more holistic research [34].

For many enterprises, particularly SMEs, the barriers to digital transformation can be prohibitive. The literature consistently identifies two primary categories of obstacles: financial constraints (the high cost of investment in new technology and expertise) and a knowledge gap (a lack of in-house skills and access to reliable information) [10], [20]. The external environment, specifically the institutional support ecosystem, is therefore critical for helping enterprises overcome these challenges.

Digital Innovation Hubs (DIHs) such as DIH Slovenia, which were supported by European Commission, have emerged as key institutional actors designed to address these specific barriers [24]. The services offered by such hubs and other support programs can be theoretically categorized based on the barriers they aim to solve:

- Financial Support Mechanisms: These are designed to alleviate capital constraints. They include the possibility of obtaining grants, co-financing for projects, and access to dedicated returnable funds. These mechanisms de-risk investment and enable projects that might otherwise be unaffordable.
- Knowledge-Based Support Mechanisms: These are designed to close the information and skills gap. They include access to a directory of credible experts, published examples of good practices, events and workshops on digitalization, and voucher-based consulting in the field of digitalization by the Digital Innovation Hub (DIH Slovenia)

registered experts. These services facilitate knowledge transfer, skills development, and ecosystem networking.

This theoretical division provides a clear framework for investigating the perceived importance of different support types, allowing this study to provide crucial empirical feedback to policymakers and support institutions on how to best design and prioritize their interventions to effectively foster digital transformation.

The literature provides a robust understanding of DT as a strategic imperative, best analyzed through the TOE framework. It highlights that success is driven by organizational factors, particularly dynamic capabilities rooted in managerial cognition and culture, and that this transformation unfolds across a network of operational and customer-facing processes. However, as the field matures, there is a growing need for greater specificity and empirical validation of these complex relationships.

This study is positioned to contribute to this developing conversation in several ways. While the link between capabilities and DT is established, there is an opportunity to empirically test how specific competencies and change management practices influence prioritization across different transformation domains. Although the role of support ecosystems is understood, there is a need for empirical evidence on which types of support are most valued by enterprises on their transformation journey, and how these needs relate to enterprise-specific characteristics like size. By empirically investigating these relationships in the Slovenian context, this research will provide a nuanced and integrated view of the digital transformation process, directly responding to recent calls for more holistic and process-oriented research in the field.

3. Methodology

Data from Slovenian enterprises were collected in collaboration with Chamber of Commerce and Digital Innovation Hub Slovenia (DIHS) using an online questionnaire using 1ka.arnes.si on-line survey platform. The effective sample size is 102. This study adopts an organizational capability perspective by focusing on structural, managerial, and competency-related factors, rather than on financial performance or revenue outcomes. Therefore we did not require from respondents to report revenue and balance sheet total [9] and have consequently use only number of employees as simplified criteria for enterprise size.

Collected were data on the enterprise size, measured only by the employee headcount (coded as 1 – small enterprise (less than 50 employees), 2 – mid-size enterprise (50 to 249 employees), 3 – large enterprise (250+ employees)), digital transformation strategy (0 – no, 1 – yes), rating of digital competencies of employees (1 - lack of sufficient digital competencies, 2 - basic knowledge and skills, 3 - advanced knowledge and skills that they are constantly upgrading), introduction of new digital technologies to employees by management (1 - employees are informed about the new features after the introduction, 2 - employees are informed about the introduction process, 3 - employees are involved in the planning and introduction of the new features), rating of the coverage extent of digital transformation strategy areas, namely Customer experience, Data Management Strategy, Processes and Digital Solutions to Support Business, Digitalization of Business

Models, Products and Services, Development of Digital Human Resources and Digital Workplaces, Development of Digital Culture, Cybersecurity, and Industry 4.0 (a 1-6 Likert scale, where 1 meant not at all, and 6 meant to the largest extent), and the importance of different types of support, namely possibility of obtaining grants for digital transformation, possibility of obtaining co-financing for digital transformation projects, possibility of obtaining dedicated returnable funds for digital transformation, directory of credible information about experts in the field of digitalization, published examples of domestic and international good practices in the field of digitalization, events on the topic of "How to approach digitalization", voucher based consulting in the field of digitalization by the Digital Innovation Hub (DIH Slovenia) registered experts (0 – no, 1 – yes).

Bivariate testing was used due to the sample size. Chi-2 test was used to test the relationship between the enterprise size and having a digital transformation strategy. Pearson's correlation coefficients were used to assess the relationships between the enterprise size, digital competencies of employees, and ways of introduction of new digital technologies to employees by management on one side, and the extent of selected digital transformation strategy areas and the importance of different types of support on the other side. Since only Pearson's correlation coefficients is used, "Pearson" is omitted in the rest of the article. Principal component analysis with Varimax rotation was used to create components from the digital transformation strategy areas and from the importance of different types of support. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy and Bartlett's Test of Sphericity were conducted prior to the actual principal component analysis. To contrast solution in rotated component matrix, we have set cutoff value of 0.40 in order to have single variable representing each principal component [7]. All statistical analyses were conducted in IBM SPSS.

4. Results

In the sample at hand, approximately a fifth of enterprises had a digital transformation strategy. Table 1 provides a crosstab for the enterprise size and having a digital transformation strategy.

Table 1. Relationship between the Company Size and Digital Transformation Strategy

	No DT Strategy	DT Strategy	Total
Small Companies	67	17	84
Mid-sized Companies	9	2	11
Large Companies	4	3	7
Total	80	22	102

To investigate whether the adoption of a digital transformation strategy is associated with enterprise size, a Chi-squared test of independence was conducted. The null hypothesis (H_0) stated that there is no relationship between an enterprise's size (small, mid-sized, large) and its likelihood of having a formal digital transformation strategy. The analysis of the observed frequencies in Table 1 against the expected frequencies under the null

hypothesis yielded a non-significant result ($\chi^2(2) = 2.038$, p = 0.361). As the p-value is greater than 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant evidence in our sample to suggest an association between enterprise size and the presence of a digital transformation strategy.

These 22 enterprises were additionally asked which areas do their digital transformation strategy cover and to what extent (a 1-6 Likert scale, where 6 meant high). Table 2 provides descriptive statistics for the extent of digital transformation strategy areas for the enterprises that have a digital transformation strategy.

Table 2. Descriptive statistics for extent of digital transformation strategy areas

	Mean	Std. Deviation
Customer Experience	4.546	1.845
Data Management Strategy	4.909	1.151
Processes and Digital Solutions to Support Business	5.455	1.101
Digitalization of Business Models, Products and Services	5.455	1.101
Development of Digital Human Resources and Digital Workplaces	4.143	1.276
Development of Digital Culture	4.500	1.472
Cybersecurity	4.524	1.632
Industry 4.0	3.905	1.895

Averages are obviously in the upper half of the scale, ranging from 3.9 to 5.45, the highest being for processes and digital solutions to support business, and for digitalization of business models, products and services.

To test whether enterprise size, digital competencies of employees, and the methods of introducing new digital technologies to employees by management influence the extent of selected digital transformation strategy areas, correlation coefficients were calculated, and the results are provided in Table 3.

Table 3. Cross-correlations for extent of digital transformation strategy areas

	Company Size	Digital Competencies	Introduction of new Digital Technologies
Customer Experience	-0.013	0.608**	0.558*
Data Management Strategy	0.155	0.293	0.358
Processes and Digital Solutions to Support Business	0.081	0.264	0.449*
Digitalization of Business Models, Products and Services	-0.038	0.358	0.519*
Development of Digital Human Resources and Digital Workplaces	0.009	0.423†	0.365
Development of Digital Culture	0.089	0.374	0.247
Cybersecurity	0.231	0.141	0.181
Industry 4.0	-0.059	-0.114	-0.121

Note: † significant at 0.1, * significant at 0.05, ** significant at 0.01.

From the selected digital transformation strategy areas, customer experience was influenced the strongest by competencies and ways of introducing new digital technologies. Additionally, ways of introducing new digital technologies impact processes and digital solutions to support business, and digitalization of business models, products and services. All at the 0.05 significance level. Company size was not correlated significantly with any of the selected digital transformation strategy areas; all significances were above 0.1. To ensure the robustness of this finding, additional parametric and non-parametric correlations were computed using the raw number of employees (continuous variable) rather than categorical size groups; however, the results remained statistically insignificant.

Principal component analysis (PCA) provides an insight into how the areas group together. Kaiser-Meyer-Olkin test of sampling adequacy is 0.75, i.e. middling [16], and Bartlett's test of sphericity Chi-2 is 98.678 resulting into significance below 0.001, therefore, it makes acceptable to use principal component analysis.

The eigenvalue-one criterion is method for determining how many components to retain in a PCA. An eigenvalue less than one indicates that the component explains less variance that a variable would and hence shouldn't be retained [6]. PCA revealed two components that have eigenvalues greater than one and which explained 53.54% of variance for component 1 and 16.66% of the total variance for component 2 (Table 4). Components 1 and 2 explained 70.20% of total variance of digital transformation strategy areas. Obtained eigenvalues indicated that the study should proceed with two-component solution.

Table 4. Digital transformation strategy areas total variance explained*

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative total %	Total	% of Variance	Cumulative total %	Total	% of Variance	Cumulative total %
1	4.283	53.544	53.544	4.283	53.544	53.544	3.170	39.622	39.622
2	1.333	16.663	70.207	1.333	16.663	70.207	2.447	30.585	70.207
3	0.970	12.124	82.331						
4	0.723	9.041	91.371						
5	0.247	3.088	94.460						
6	0.216	2.696	97.155						
7	0.147	1.838	98.993						
8	0.081	1.007	100.000						

*Extraction Method: Principal Component Analysis; eigenvalue-one criterion

Principal component analysis of digital transformation strategy areas was performed with Varimax rotation and resulted in two components solution. The rotated component matrix is provided in Table 5.

Table 5. Rotated Component Matrix for extent of digital transformation strategy areas*

	Component 1	Component 2
Customer Experience	-0.006	*0.917*
Data Management Strategy	*0.845*	0.346
Processes and Digital Solutions to Support Business	*0.902*	0.286
Digitalization of Business Models, Products and Services	*0.808*	0.360
Development of Digital Human Resources and Digital Workplaces	0.294	*0.872*
Development of Digital Culture	0.399	*0.713*
Cybersecurity	*0.723*	0.036
Industry 4.0	*0.470*	0.055

*Extraction Method: Principal Component Analysis – Varimax

Component 2 consists of softer factors, such as customer experience, development of digital human resources and digital workplaces, development of digital culture. And component 1 consists of harder factors, such as data management strategy, processes and digital solutions to support business, digitalization of business models, products and services, cybersecurity, Industry 4.0.

Table 6 provides descriptive statistics for the importance of different types of support measures that could help enterprises that have a digital transformation strategy.

Table 6. Descriptive statistics for the importance of support measures for digital transformation

	Mean	Std. Deviation
Possibility of obtaining grants for digital transformation	0.859	0.350
Possibility of obtaining co-financing for digital transformation projects	0.549	0.501
Possibility of obtaining dedicated returnable funds for digital transformation	0.183	0.390
Directory of credible information about experts in the field of digitalization	0.225	0.421
Published examples of good domestic and foreign practices in the field of digitalization	0.423	0.497
Events on the topic of "How to approach digitalization"	0.338	0.476
Voucher based consulting in the field of digitalization by the Digital Innovation Hub (DIH Slovenia) registered experts	0.338	0.476

Importance ranges from 0.18 to 0.86, the highest being possibility of obtaining grants for digital transformation, the lowest being possibility of obtaining dedicated returnable funds for digital transformation.

To test whether enterprise size, digital competencies of employees, and ways of introduction of new digital technologies to employees by management influence the importance of different types of support, correlation coefficients were calculated, and they are provided in Table 7.

Table 7. Cross-correlations for the importance of support measures for digital transformation

	Company Size	Digital Competencies	Introduction of new Digital Technologies
Possibility of obtaining grants for digital transformation	-0.255*	0.058	-0.100
Possibility of obtaining co-financing for digital transformation projects	-0.149	0.146	0.044
Possibility of obtaining dedicated returnable funds for digital transformation	0.113	0.190	-0.167
Directory of credible information about experts in the field of digitalization	0.058	0.135	0.081
Published examples of domestic and international good practices in the field of digitalization	0.020	0.176	0.319**
Events on the topic of "How to approach digitalization"	-0.005	-0.017	0.107
Voucher based consulting in the field of digitalization by the Digital Innovation Hub (DIH Slovenia) registered experts	-0.113	0.061	0.107

Note: * significant at 0.05, ** significant at 0.01

With regards to support measures, Kaiser-Meyer-Olkin test of sampling adequacy is 0.66, i.e. mediocre [16], and Bartlett's test of sphericity Chi-2 is 84.592 resulting into significance below 0.001, therefore, it is acceptable to use PCA. PCA revealed two components that have eigenvalues greater than one and which explained 33.05% of variance for component 1 and 21.38% of the total variance for component 2. Components 1 and 2 explained 54.43% of total variance of importance for support of digital transformation (Table 8). Obtained eigenvalues indicated that the study should proceed with two-component solution.

Table 8. Support measures for digital transformation total variance explained*

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative total %	Total	% of Variance	Cumulative total %	Total	% of Variance	Cumulative total %
1	2.314	33.051	33.051	2.314	33.051	33.051	2.221	31.735	31.735
2	1.497	21.380	54.432	1.497	21.380	54.432	1.589	22.697	54.432
3	.955	13.643	68.075						
4	.770	11.001	79.076						
5	.573	8.188	87.263						
6	.520	7.426	94.689						
7	.372	5.311	100.000						

*Extraction Method: Principal Component Analysis; eigenvalue-one criterion

Principal component analysis of support of digital transformation factors was performed with Varimax rotation method and resulted in two components solution. The rotated component matrix is presented in Table 9.

Table 9. Rotated Component Matrix for the importance of support measures for digital transformation*

	Component 1	Component 2
Possibility of obtaining grants for digital transformation	-0.071	*0.715*
Possibility of obtaining co-financing for digital transformation projects	0.387	*0.686*
Possibility of obtaining dedicated returnable funds for digital transformation	0.023	*0.622*
Directory of credible information about experts in the field of digitalization	*0.498*	-0.387
Published examples of domestic and international good practices in the field of digitalization	*0.754*	0.080
Events on the topic of “How to approach digitalization”	*0.820*	-0.034
Voucher based consulting in the field of digitalization by the Digital Innovation Hub (DIH Slovenia) registered experts	*0.759*	0.249

*Extraction Method: Principal Component Analysis – Varimax

Component 2 consists of funding-related support measures, such as possibility of obtaining grants for digital transformation, the possibility of obtaining co-financing for digital transformation projects, and the possibility of obtaining dedicated returnable funds for digital transformation. And component 1 consists of knowledge-related support, such as a directory of credible information about experts in the field of digitalization, published examples of domestic and international good practices in the field of digitalization, events on the topic of "How to approach digitalization", and voucher based consulting in the field of digitalization by the Digital Innovation Hub (DIH Slovenia) registered experts.

5. Discussion

A central finding of this research study is the impact of digital competencies and introduction of new digital technologies on the extent of digital transformation. The strong, significant correlations between the digital competencies of employees and/or the ways of introduction of new digital technologies with nearly all measured DT strategy areas provide empirical support for the arguments of the dynamic capabilities framework, defined as an enterprise's ability to integrate, build, and reconfigure competences to address rapidly changing environments [28].

This study suggests that these two variables are more than just operational choices; they are manifestations of an enterprise's higher-order capabilities. Advanced digital competencies are not merely technical skills but represent the organization's capacity to sense new technological opportunities and seize them by understanding their strategic implications - the core of the sensing and seizing capabilities described by Teece [27]. This aligns with the microfoundational view of Helfat and Peteraf [13] that enterprise-level capabilities are rooted in the cognitive abilities of its managers and employees to perceive, reason, and problem-solve in complex environments.

Conversely, a lack of these capabilities can be framed as a form of Organizational Debt (OD), a phenomenon described as the accumulation of "suboptimal decisions, outdated procedures, misaligned structures and cultural barriers that limit an organization's ability to adapt and innovate quickly" [1]. Our findings empirically support this concept from a digital transformation perspective. When an organization fails to cultivate employee digital competencies or foster collaborative methods for introducing new technology, it is effectively incurring OD. This debt manifests in the softer DT areas we identified, such as a diminished customer experience, which [1] link to consequences like user frustration and reduced quality. Therefore, the soft factors we analyze are not merely enablers of success but are critical mechanisms for preventing the accumulation of organizational debt that mortgages the company's future agility and competitiveness.

Statistically significant correlations between digital transformation strategy areas, digital competencies and method for introduction of new digital technologies reported in Table 3, represent empirical evidence for the mechanisms of transformation competency at work. The data results suggest that a digital transformation strategy is correlated with cognitive capacity to envision the future (high employee competencies) and the organizational capacity to execute change collaboratively (participatory introduction methods). This aligns with the framework proposed by Blanka et al. [2], where an employee's ability to drive digital transformation is composed of two intertwined elements: generic digital

skills and, critically, intrapreneurial competencies. This frames the paper's contribution from simply identifying "important factors" to providing empirical validation for a specific, theoretically grounded model of human-centric digital transformation.

Similarly, the method of introducing technology—specifically, the involvement of employees in the planning and introduction process—is a clear indicator of an enterprise's transforming capability, the crucial third pillar in Teece's [27] framework that focuses on continuous organizational renewal. It reflects an organization's ability to manage change, foster a collaborative culture, and ensure that technological adoption is aligned with human processes. The findings can be further enriched by applying the integrated framework of soft factors proposed by Ngereja et al [23]. This framework shows that organizational enablers (like culture and support) translate into success through specific project-level actions and individual characteristics. Our finding that employee competencies (an individual characteristic) strongly correlate with "softer" DT areas like customer experience aligns with this model. It suggests that success in people-centric transformation requires not only skilled individuals but also project-level actions such as end-user involvement and organizational-level support, to foster a learning culture, which are among the critical soft factors identified.

In the context of the two components of digital transformation strategy areas from the principal component analysis results, it is possible to see a trend that softer areas (component 2) are positively correlated with higher digital competencies. In case of customer experience, significance is 0.004; for development of digital human resources and digital workplaces, significance is 0.063, and for development of digital culture, significance is 0.104. Harder component 1 includes Data Management Strategy, Processes and Digital Solutions, Cybersecurity, and Industry 4.0 corresponds to those identified by Sousa and Rocha [26], such as the Internet of Things (IoT), Robotization, and Artificial Intelligence, which directly enable process optimization and new data-driven operations. While harder areas from component 1 do not significantly correlate with digital competencies at the 0.05 level. These two components have a structure similar to the foundational pillars of business models, operational processes, and customer experience identified by Westerman et al. [33], our component 1 containing business models and a part of operational processes, and our component 2 containing customer experience and a part of operational processes.

With regards to the way new digital technology is introduced, it positively correlates to customer experience, processes and digital solutions to support business, and digitalization of business models, products and services, all significant at 0.05. These three areas are more at the operational level. While the remaining five areas, such as data management strategy, development of digital human resources and digital workplaces, development of digital culture, cybersecurity, and Industry 4.0 are more at the strategic level.

With regards to the two components of types of DT support from the principal component analysis, it is possible to see that financial related support on average (even when the possibility of obtaining dedicated returnable funds for digital transformation is ranked the lowest) is perceived as more important than knowledge-related support. This demonstrates a clear and pragmatic risk aversion. Enterprises prioritize support that directly mitigates the financial downside of transformation. This finding empirically confirms that the financial constraints identified as primary obstacles for SMEs by scholars like Ghobakhloo & Ching

[10] and Mittal et al. [20] are perceived by enterprises themselves as the most critical barrier to overcome. While knowledge-based support (consulting, events, best practices) is valued, it is perceived as secondary to the immediate need of financial support to initiate projects. This finding has significant implications for policymakers and support institutions like DIHs. It suggests that while knowledge-sharing initiatives are beneficial, the most impactful intervention, especially for encouraging SMEs to engage in DT, is the provision of direct, non-refundable financial support that lowers the initial barrier to investment. However, to ensure this investment leads to sustainable transformation, policymakers should adopt a two-tiered approach. Financial mechanisms (grants, co-financing) should serve as the primary catalyst to de-risk entry, while knowledge-based initiatives (expert directories, best-practice workshops) should be integrated as complementary capability builders. A practical application of this balance would be hybrid support models, where financial funding is paired with mandatory voucher-based consulting or training. This ensures that the necessary capital injection is matched by the development of internal competencies, thereby addressing the resource and skills gaps simultaneously.

From a managerial perspective, the findings underline that digital competencies should be treated as a core strategic asset. Results of our study suggest that investments in skills development directly enhance an organization's sensing and seizing capabilities. Managers should therefore prioritize continuous upskilling, reskilling, and learning mechanisms that develop not only technical digital skills but also employees' ability to fully understand and exploit opportunities of digital technologies in operational and strategic context. This aligns with findings of Neumann et al [22], which outline importance of learning culture practices that encourages experimentation and treats mistakes as growth opportunities rather than failures to fully exploit agile transformation of enterprises.

The results of our study also show that the ways in which new digital technologies are introduced in an enterprise, more precisely the degree of employee involvement in planning and implementation of digital technologies, act as a key indicator of an enterprise's transforming capability. Participatory introduction and implementation approaches should become integral managerial mechanisms for effective execution of digital transformation.

The distinction between softer and harder digital transformation strategy areas has important implications for managerial prioritization. Different approaches are needed to support overall enterprise-wide development of digital culture on one side and execute targeted technological investments and ensure specialized knowledge and expertise for harder (more technologically orientated knowledge) domains.

The results of our study indicated that financial and knowledge-based support are needed for successful digital transformation. Prior research has shown that SMEs continue to face considerable difficulties in keeping pace with ongoing digital transformation, particularly when compared to larger enterprises [35], [18].

These challenges are not only a recent phenomenon. Similar difficulties and limitations have been observed for several decades, dating back to the period when information technology first became a strategic organizational resource and a source of competitive advantage. Earlier research consistently reports that SMEs encountered substantial barriers to information technology adoption, including limited availability of digital skills and

expertise, as well as constraints related to managerial capacity and financial resources for ICT-related investments [8], [11].

For managers, this implies that digital transformation initiatives are more likely to gain traction when financial uncertainty is explicitly addressed. However, it is widely known that financial resources alone are insufficient to sustain transformation. Managers should therefore combine financial investments with building up digital capabilities, especially digital culture, including supporting learning, experimentation, and knowledge-sharing practices to ensure that funded initiatives translate into lasting organizational capabilities.

Finally, our findings also emphasize the central role of managerial cognition and leadership in enabling human-centric digital transformation. Managers should actively shape shared understanding, strategic vision, and collective engagement with digital transformation, meaning building up digital culture in an enterprise-wide context. This requires managers to act not only as decision-makers but also as sense-givers who articulate why digital transformation matters and how employees as the most strategic asset in the enterprise valuably contribute to it.

6. Conclusion

This study embarked on an investigation into the determinants and nature of digital transformation (DT) within the Slovenian business landscape, a phenomenon of critical importance in the contemporary global economy. By grounding the research in the adapted Technology-Organization-Environment (TOE) framework and leveraging the theoretical lens of dynamic capabilities, this article has moved beyond a simple assessment of technology adoption to provide a nuanced, multi-dimensional view of the transformation process. The empirical findings offer a clear and compelling narrative: successful digital transformation is fundamentally a matter of organizational capability, not merely technological acquisition or enterprise size.

Our research confirms that the journey to digital maturity is driven by what an enterprise can do, rather than what it is. The initial Chi-2 analysis revealed that the strategic decision to pursue DT is not confined to large, resource-rich corporations; it is an imperative felt across enterprises of all sizes. This finding suggests that the environmental pressures of the digital age act as a great leveler, making transformation a near-universal strategic concern. However, the subsequent analysis demonstrated that enterprise's internal dynamic capabilities profoundly shape the scope and depth of this transformation. The strong, positive correlations between employee digital competencies and collaborative change management processes with the extent of DT initiatives underscore a main findings of this study: it is the investment in human capital and organizational processes - the microfoundations of sensing, seizing, and transforming which truly enables an enterprise to navigate the complexities of the digital transformation.

The principal component analysis provided a clear structure to this transformation journey, revealing two distinct but complementary fronts. The first, an operational and technical core, comprises the foundational investments in data, processes, and security necessary for efficiency and resilience. The second, a people and customer-centric front, encompasses the initiatives in culture, human resources, and customer experience that drive market

value and long-term adaptability. The finding that digital competencies are most strongly correlated with this second, "softer" dimension is particularly insightful. It suggests that while the operational core can be engineered, the value-creating front must be cultivated through a skilled, engaged, and empowered workforce.

Finally, this study sheds light on the pragmatic realities faced by enterprises undertaking this journey. The clear preference for direct financial support, such as grants, over knowledge-based assistance or refundable funds, highlights the critical role of de-risking the initial investment. For policymakers and support institutions like DIH Slovenia, this sends an unambiguous message: lowering the financial barrier to entry is the most potent catalyst for encouraging enterprises, especially SMEs, to embark on their digital transformation.

With these results, this study provides a significant contribution to both theory and practice. It provides robust empirical support for the application of dynamic capabilities theory to the DT phenomenon and challenges the deterministic role often ascribed to enterprise size. For managers, this study offers specific practical guidance beyond general investment in human capital. First, the strong correlation between introduction methods and transformation scope indicates that managers must abandon top-down implementation in favor of participatory approaches. Involving employees in the planning phase of new technology adoption acts as a practical mechanism for seizing opportunities and reducing resistance. Second, managers should prioritize the development of soft transformation areas - specifically customer experience and digital workplace culture - as these were found to be the primary drivers of employee digital competencies. Rather than viewing digital skills training as a separate activity, it should be integrated into the redesign of customer-facing processes, where the immediate value of the technology is most visible to the workforce. For policymakers, it offers clear guidance on how to structure support mechanisms for maximum impact.

Although this study provides valuable insights into how human factors - such as employee digital competencies and managerial approaches to supporting technology adoption - influence enterprise digital transformation, several limitations should be acknowledged. First, the study is based on cross-sectional data collected through an online survey, which captures organizational practices and perceptions at a single point in time. Given the dynamic characteristics of digital transformation, longitudinal research designs would be more suitable for examining causal relationships and changes over time.

Second, the sample is limited in size and comprises enterprises from a single country - Slovenia. In addition, the sampling was not random since the survey was promoted by Digital Innovation Hub of Slovenia and Chamber of Commerce. It is possible that respondents are those enterprises who are more aware or even advanced in the digital transformation journey. Therefore, the generalizability of results is limited and should be interpreted with awareness of these limitations.

Nevertheless, this context is highly relevant, as SMEs enterprises dominate not only the Slovenian economy but also in other European countries. To enhance transparency and interpretability, we have addressed these limitations by providing detailed descriptive statistics on enterprise demographics.

Additionally, the study categorized enterprise size solely based on the number of employees. Future research should aim to incorporate financial metrics, such as annual turnover and balance sheet totals, to fully align with the EU definition and capture potential nuances related to financial resources.

While this study provides a detailed snapshot of the Slovenian context, the path forward for research is clear. Longitudinal studies are needed to trace the causal pathways between capabilities and performance over time, and cross-country comparisons would enrich our understanding of how different institutional environments shape the DT journey. Ultimately, this research study affirms that digital transformation is a deeply human endeavor. Technology may provide tools, but it is the strategic vision, the cognitive capabilities, and the collaborative spirit of an organization's people that will determine its success in the digital age.

Acknowledgments. University of Maribor, Faculty of Organizational Sciences and the authors acknowledge the financial support from the Slovenian Research and Innovation Agency (research core funding P5-0018).

References

1. Ahmad, M., Al-Baik, O., Hussein, A., Abu-Alhaija, M.: Unraveling the organisational debt phenomenon in software companies. Computer Science and Information Systems 22(1), 369–399 (2025)
2. Blanka, C., Krumay, B., Rueckel, D.: The interplay of digital transformation and employee competency: A design science approach. Technological Forecasting and Social Change 178, 121575 (May 2022)
3. Bonnet, D., Westerman, G.: The New Elements of Digital Transformation. MIT Sloan Management Review 62(2), 83–89 (Nov 2020)
4. Bouwman, H., Nikou, S., De Reuver, M.: Digitalization, business models, and SMEs: How do business model innovation practices improve performance of digitalizing SMEs? Telecommunications Policy 43(9), 101828 (Oct 2019)
5. Cao, G., Duan, Y., Edwards, J.S.: Organizational culture, digital transformation, and product innovation. Information & Management 62(4), 104135 (Jun 2025)
6. Cattell, R.B.: The Scree Test For The Number Of Factors. Multivariate Behavioral Research 1(2), 245–276 (Apr 1966)
7. Dunteman, G., H.: Principal Components Analysis. No. 07-069 in Quantitative Applications in the Social Sciences, Sage Publications, Inc. (1989)
8. Eller, R., Alford, P., Kallmünzer, A., Peters, M.: Antecedents, consequences, and challenges of small and medium-sized enterprise digitalization. Journal of Business Research 112, 119–127 (May 2020)
9. European Commision: SME definition. https://single-market-economy.ec.europa.eu/smes/sme-fundamentals/sme-definition_en (2025)
10. Ghobakhloo, M., Ching, N.T.: Adoption of digital technologies of smart manufacturing in SMEs. Journal of Industrial Information Integration 16, 100107 (Dec 2019)
11. Ghobakhloo, M., Iranmanesh, M.: Digital transformation success under Industry 4.0: A strategic guideline for manufacturing SMEs. Journal of Manufacturing Technology Management 32(8), 1533–1556 (Oct 2021)
12. Hanelt, A., Bohnsack, R., Marz, D., Antunes Marante, C.: A Systematic Review of the Literature on Digital Transformation: Insights and Implications for Strategy and Organizational Change. Journal of Management Studies 58(5), 1159–1197 (Jul 2021)

13. Helfat, C.E., Peteraf, M.A.: Managerial cognitive capabilities and the microfoundations of dynamic capabilities. Strategic Management Journal 36(6), 831–850 (Jun 2015)
14. Hess, T., Matt, C., Benlian, A., Wiesböck, F.: Options for Formulating a Digital Transformation Strategy. MIS Quarterly Executive 15(2), 151–173 (Jun 2016)
15. Jeansson, J., Bredmar, K.: Digital Transformation of SMEs: Capturing Complexity. In: BLED 2019 Proceedings. vol. 33. AIS eLibrary, Bled, Slovenia (Jan 2019)
16. Kaiser, H.F., Rice, J.: Little Jiffy, Mark Iv. Educational and Psychological Measurement 34(1), 111–117 (Apr 1974)
17. Kane, G.C., Palmer, D., Phillips, A.N., Kiron, D., Buckley, N.: Strategy, not Technology, Drives Digital Transformation. MIT Sloan Management Review (Jul 2015)
18. Kljajić Borštnar, M., Pucihar, A.: Multi-Attribute Assessment of Digital Maturity of SMEs. Electronics 10(8), 885 (Apr 2021)
19. Li, L., Su, F., Zhang, W., Mao, J.Y.: Digital transformation by SME entrepreneurs: A capability perspective. Information Systems Journal 28(6), 1129–1157 (Nov 2018)
20. Mittal, S., Khan, M.A., Romero, D., Wuest, T.: A critical review of smart manufacturing & Industry 4.0 maturity models: Implications for small and medium-sized enterprises (SMEs). Journal of Manufacturing Systems 49, 194–214 (Oct 2018)
21. Müller, J.M., Buliga, O., Voigt, K.I.: Fortune favors the prepared: How SMEs approach business model innovations in Industry 4.0. Technological Forecasting and Social Change 132, 2–17 (Jul 2018)
22. Neumann, M., Kuchel, T., Diebold, P., Schön, E.M.: Agile culture clash: Unveiling challenges in cultivating an agile mindset in organizations. Computer Science and Information Systems 21(3), 1013–1031 (2024)
23. Ngereja, B.J., Hussein, B., Wolff, C.: A comparison of soft factors in the implementation and adoption of digitalization projects: A systematic literature review. International Journal of Information Systems and Project Management 12(2), 70–86 (Apr 2024)
24. Sassanelli, C., Terzi, S., Panetto, H., Doumeingts, G.: Digital Innovation Hubs supporting SMEs digital transformation. In: 2021 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC). pp. 1–8. IEEE, Cardiff, United Kingdom (Jun 2021)
25. Schallmo, D., Williams, C.A., Boardman, L.: Digital Transformation of Business Models — Best Practice, Enablers, and Roadmap. International Journal of Innovation Management 21(08), 1740014 (Dec 2017)
26. Sousa, M.J., Rocha, Á.: Digital learning: Developing skills for digital transformation of organizations. Future Generation Computer Systems 91, 327–334 (Feb 2019)
27. Teece, D.J.: Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance. Strategic Management Journal 28(13), 1319–1350 (Dec 2007)
28. Teece, D.J., Pisano, G., Shuen, A.: Dynamic capabilities and strategic management. Strategic Management Journal 18(7), 509–533 (Aug 1997)
29. Tornatzky, L.G., Fleischer, M., Chakrabarti, A.K.: The Processes of Technological Innovation. Issues in Organization and Management Series, Lexington Books (1990)
30. Verhoef, P.C., Broekhuizen, T., Bart, Y., Bhattacharya, A., Qi Dong, J., Fabian, N., Haenlein, M.: Digital transformation: A multidisciplinary reflection and research agenda. Journal of Business Research 122, 889–901 (Jan 2021)
31. Vial, G.: Understanding digital transformation: A review and a research agenda. The Journal of Strategic Information Systems 28(2), 118–144 (2019)
32. Warner, K.S., Wäger, M.: Building dynamic capabilities for digital transformation: An ongoing process of strategic renewal. Long Range Planning 52(3), 326–349 (Jun 2019)
33. Westerman, G., Bonnet, D., Mcafee, A.: The Nine Elements of Digital Transformation. MIT Sloan Management Review 55(3) (Jan 2014)
34. Wiener, M., Strahringer, S., Kotlarsky, J.: Where are the processes in IS research on digital transformation? A critical literature review and future research directions. The Journal of Strategic Information Systems 34(2), 101900 (Jun 2025)

35. Zare, L., Ali, M.B., Rauch, E., Matt, D.T.: Navigating challenges of small and medium-sized enterprises in the Era of Industry 5.0. Results in Engineering 27, 106457 (Sep 2025)

František Sudzina is an Assistant Professor at Prague University of Economics and Business, Faculty of Informatics and Statistics, Department of Systems Analysis. His research interests include information systems, consumer behavior, and psychology. He teaches undergraduate courses in Information for Business and Philosophy of Science, and a graduate course in Gamification. František Sudzina is a member of the Academy of Management, the Association for Information Systems, the Operational Research Society, and the Strategic Management Society.

Andreja Pucihar is a Full Professor of Information Systems at University of Maribor, Faculty of Organizational Sciences. Her research focuses on digital innovation, digital transformation, digital business models, and data-driven decision-support systems. She has over 25 years of experience in industrial and European projects related to SME digitalization. She leads the research program Decision support systems in digital business and directs the eCenter digital business lab. Dr. Pucihar is chair of the Bled eConference on digital business and serves on editorial boards of international journals such as Electronic Markets and Journal of Theoretical and Applied Electronic Commerce Research.

Gregor Lenart is an Assistant Professor at University of Maribor, Faculty of Organizational Sciences, Information Systems Department. His primary research areas are information systems, enterprise architecture, business process management, e-business, digital transformation, and business model innovation. As part of his research, he has collaborated in several European research projects, which were focused on innovation and digitalization of small and medium-sized enterprises.

Received: July 31, 2025; Accepted: March 2, 2026.

Computer Science and Information Systems 23(3):969–1000 https://doi.org/10.2298/CSIS250616028M

Model Parameter-Based Transfer Learning for ESG Score Prediction in Developing Markets

Ivana Marković[1], Adela Ljajić[2], Jelena Z. Stanković[1], Miloš Košprdić[2], and Jovica Stanković[1]

[1] Faculty of Economics, University of Niš
Trg kralja Aleksandra Ujedinitelja 11, 18000 Niš
ivana.markovic@eknfak.ni.ac.rs (corresponding author),
jelenas@eknfak.ni.ac.rs,jovica.stankovic@eknfak.ni.ac.rs
[2] The Institute for Artificial Intelligence Research and Development of Serbia
Fruškogorska 1, 21000 Novi Sad
adela.ljajic@ivi.ac.rs, milos.kosprdic@ivi.ac.rs

Abstract. While ESG (Environmental, Social, and Governance) assessment plays a key role in sustainable finance, data scarcity and noise in emerging economies hinder robust model development. To address this, we propose a model parameter-based transfer learning with random forest (MPBTL-RF) approach for domain adaptation situations where source data are not available. The proposed model is evaluated using three traditional learning approaches: Random Forest (RF), eXtreme Gradient Boosting (XGB), and Feedforward Neural Networks (FNN). Cross-validation is used to assess model generalizability, and domain adaptation is tested through in-domain and out-of-domain settings. The proposed MPBTL-RF approach achieves competitive performance compared to traditional baselines in scenarios with limited training data, offering time advantages with predictive efficiency and stability. This work demonstrates how machine learning pipelines can adapt to data-constrained, real-world domains, fostering the synergy between AI (Artificial Intelligence) and business.

Keywords: Machine Learning, Domain Adaptation, Transfer Learning, Small Datasets, ESG Score, Developing Markets, Corporate Financial Performance.

1. Introduction

Small data samples are challenging for machine learning across various domains, including finance, where databases are frequently limited by numerous factors such as the number of available firms, reporting restrictions, or institutional privacy policies. As noted by Kokol et al. [11], "using machine learning on small-sized datasets presents a problem because, in general, the 'power' of machine learning in recognizing patterns is proportional to the size of the dataset; the smaller the dataset, the less powerful and accurate the machine learning algorithms".

Transfer learning has emerged as a crucial paradigm in machine learning, enabling knowledge learned in a source domain to improve predictive performance in a related target domain under data-scarce conditions [15]. This approach addresses the critical challenge of data scarcity by allowing models to improve performance on target tasks or domains through knowledge transfer from related domains with larger labeled datasets.

Recently, source-free unsupervised domain adaptation, also known as unsupervised model adaptation, has attracted significant attention, enabling effective generalization of pretrained models to target domains without labeled data. This setting is particularly relevant in real-world applications, where organizations often share only trained models rather than raw data due to privacy, security, or scalability constraints [8].

ESG scores, widely used as indicators of corporate sustainability, benefit from advances in artificial intelligence (AI), which plays a crucial role in finance and is an influential factor in ESG investing [12]. The existing work on this topic can be divided into eight categories: Trading and Investment, ESG Disclosure, Measurement and Governance, Firm Governance, Financial Markets and Instruments, Risk Management, Forecasting and Valuation, Data, and Responsible Use of AI [14]. Distinct AI and machine learning techniques are employed across these categories. Among these, the categories of Data and Forecasting include either predicting the ESG score or exploring novel approaches for measuring ESG performance [14].

However, most existing studies focus on developed economies, while ESG score prediction in developing markets remains particularly challenging due to data scarcity, heterogeneous reporting standards, and incomplete coverage. The annual publication cycle of ESG scores complicates the data collection necessary for model training. Furthermore, rigid privacy policies increase challenges in data sharing within economics, especially in the insurance and finance sectors. To the best of our knowledge, no previous studies have addressed ESG score prediction in developing markets.

The research presented here fills this gap by proposing and empirically validating the model parameter-based domain adaptation approach to facilitate transfer learning based on RF (MPBTL-RF). We compare our approach with strong baselines (RF, XGB and FNN) and evaluate its performance in both in-domain and out-of-domain settings, including a real-world case study on the Belgrade Stock Exchange (BELEX). Additionally, in contrast to most research focused on the transfer learning approach within classification tasks [23, 24, 10], this study explores regression problems.

The contributions of this paper are threefold:

- We design a computationally efficient and practical algorithm and formalize a model parameter-based transfer learning pipeline for RF (MPBTL-RF).
- We conduct a comprehensive experimental evaluation comparing MPBTL-RF against state-of-the-art baselines (RF, XGB, FNN), analyzing both in-domain and out-of-domain scenarios, together with a statistical validation of the obtained results.
- We discuss robustness and out-of-domain transferability, positioning ESG score prediction as a benchmark task that highlights broader machine learning challenges in small-data and heterogeneous environments.

The remainder of the paper is structured as follows: Section 2 reviews related work, including machine learning for ESG prediction and transfer learning. Section 3 describes the MPBTL-RF model and baseline models. Section 4 presents in-domain and out-of-domain results with comparative analysis. Section 5 applies the model to the BELEX dataset. Section 6 provides insights into the practical application of the proposed method, while Section 7 defines its limitations. Finally, Section 8 concludes with key findings and directions for future research.

2. Background

This section summarizes key research contributions in this field, highlighting the methodologies and datasets utilized for ESG score prediction in developed markets. It concludes with a literature review that motivates the proposed research.

2.1. Related work

A significant amount of literature now employs machine learning techniques for ESG score estimation, demonstrating the potential of these approaches to provide valuable insights into corporate sustainability performance.

D'Amato et al. [5] explained the determinants of the ESG score by performing the RF algorithm. They used financial statement information, such as profitability indicators, liquidity, and solvency ratios, from a subset of 109 companies listed in the STOXX Europe 600 Index between 2014 and 2018. The study found that financial statement items are powerful tools for explaining and predicting ESG scores, with performance measured by Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2) metrics.

Krappel et al. [12] proposed a heterogeneous ensemble model to predict ESG ratings using fundamental data. The model combines FNN, CatBoost, and XGB ensemble members. Their dataset included 7,413 companies with annual observations between 2002 and 2019, amounting to 57,310 observations. They utilized 475 numerical and 44 categorical features and measured performance using MAE and R^2 metrics. The study highlighted the importance of a comprehensive dataset and advanced machine learning techniques for accurate ESG score prediction.

Garcia et al. [7] developed a rough set model to relate ESG scores to key corporate financial performance measures from the investor's perspective. Their dataset comprised 1,688 observations from 2013 to 2018, with seven numerical features. They suggested that the industry sector and financial variables reveal significant differences across firms regarding ESG, but the model's significance diminishes when examining small differences in ESG performance.

Sokolov et al. [20] proposed an approach to automatically convert unstructured text data into ESG scores using deep learning for Natural Language Processing (NLP). They incorporated Bidirectional Encoder Representations from Transformers (BERT) to improve the accuracy of assessing the relevance and content of documents in an ESG context using social media data. This approach emphasizes the potential of automating ESG scoring and constructing ESG portfolios through advanced NLP techniques.

D'Amato et al. [6] predicted ESG scores using data from 401 companies that are constituents of the STOXX Europe 600 index. They used seven numerical features and employed the RF algorithm. The performance was measured using RMSE and MAPE metrics. Their findings indicate that the RF model can better grasp the nonlinear aspects of ESG score prediction than a classical generalized linear model.

Chowdhury et al. [4] applied six machine learning algorithms, including Artificial Neural Networks Classifier (ANNC), Bagging Classifier (BGC), k-Nearest Neighbors Classifier (KNNC), Naive Bayes Classifier (NBC), Random Forest Classifier (RFC), and Support Vector Machines Classifier (SVMC) on a global data sample of 6,166 firms in 73 countries from 2005 to 2019. They found that the RFC provided the highest accuracy

(78.50%) among the six algorithms. The study revealed that the lagged ESG score had the highest contribution to the model, with performance assessed using accuracy, Kappa, the area under the curve, receiver operating characteristic, and logLoss metrics.

The available literature analyzes the problem of predicting ESG scores from the point of view of developed markets, but no literature refers to the same problem in developing markets.

2.2. Motivations for the work

We found motivation for this research in the methods proposed in Tan et al. [22] transitive transfer learning, as well as in the Pinto et al. [16], Jin et al. [9] and active transfer learning model in Shim et al. [19].

Inspired by the human ability for transitive inference, where seemingly unrelated concepts can be connected through intermediate bridges using auxiliary knowledge, [22] introduced a novel learning paradigm called Transitive Transfer Learning (TTL). They proposed a framework that emulates this human-like learning process through two main components: intermediate domain selection and knowledge transfer. Extensive empirical evidence shows that the framework yields state-of-the-art classification accuracies on several classification datasets.

In [16] the authors suggested that model parameters or hyperparameters generated for similar tasks would also be similar, and that the information collected from the source task could be sent to another task in the form of shared model weights. According to their experiments, where different neural networks share the same feature and label space, they investigated a homogeneous inductive problem using model-based transfer learning. The authors conducted experiments that leveraged 250 data-driven models based on a synthetic dataset of a building archetype and studied, among other things, the influence of data availability for the transfer process of thermal dynamics.

Authors in [19] predicted reaction conditions from limited data through active transfer learning. They showed that specifically tuned machine learning models based on RF classifiers improved the applicability of Pd-catalyzed cross-coupling reactions to nucleophile types unknown to the model. They stated that in active transfer learning, simple models that are composed of a small number of decision trees with limited depths are key to ensuring generalizability, interpretability, and performance.

According to Jin et al. [9], model-based transfer learning is particularly efficient because it leverages the source domain model to directly transfer high-level knowledge, eliminating the need to reprocess training data or engage in complex relational reasoning, thereby enhancing its ability to generalize insights from the source domain to the target domain.

Let D_S represent the source domain and D_T the target domain with their corresponding feature spaces as X_S and X_T and let *P(X)* defines the probability distribution of the feature space. Denote the learning task τ_T and τ_S, and target the predictive learning function as $f_T(.)$. According to [16], transfer learning can be classified based on label availability, domain and task similarity, and the technique used for knowledge transfer.

According to label availability, there is inductive transfer learning, where both D_S and D_T have labeled data but $\tau_T \neq \tau_S$; transductive transfer learning where $\tau_T = \tau_S$, but $D_S \neq D_T$ while there are labeled data only in D_S; and unsupervised transfer learning where $D_S = D_T$ and $\tau_T \neq \tau_S$ but the tasks are related and there are no labeled data in

both domains. Within transductive transfer learning, we differentiate between cases where $X_S \neq X_T$ and the second case where $X_S = X_T$, but $P(X_S) \neq P(X_T)$ [16].

Regarding domain and task similarity, there is homogeneous transfer learning, where both the feature and labeled spaces in both source and target domains are the same, and heterogeneous transfer learning, where either the feature or label space is different.

Relative to a strategy that is adopted to share knowledge, according to [15], there are four different transfer learning approaches: instance-based methods, feature-based approaches, model-based transfer methods, and relational knowledge transfer. Instance-based techniques are often employed when X_T=X_S, allowing certain source data to be reweighted and utilized as training data in the target domain. Feature-based methods focus on uncovering a hidden feature space from the source domain to enhance performance in the target domain. Model-based transfer methods, on the other hand, transfer knowledge by sharing parameters or prior distributions of model hyperparameters.

This approach assumes that similar tasks should have similar models or hyperparameters, enabling knowledge transfer from the source task to the target task through shared parameters. Lastly, relational knowledge transfer relies on the assumption that the source and target domain data exhibit similar relationships, enabling this group of methods to transfer relational structures between the two domains.

Also, a common way of transferring is with a setting where X_S=X_T in the source and target domain, which minimizes the differences in the data distribution [21]. A traditional machine learning model relies solely on data from the target building, whereas a transfer learning model reuses knowledge from a source building to lower implementation costs, accelerate training, and improve performance [16].

However, because in this work, the source and target tasks share the same feature space and there are no labeled data in the target domain, a hybrid method, as a homogeneous transductive approach using model parameter-based transfer learning, was explored. This setting is also, according to [15], related to domain adaptation.

In the next section, we define how the proposed approach can expand the applicability of ESG score prediction.

3. Methodology

This section outlines two approaches to machine learning: model parameter-based transfer learning and traditional machine learning.

We address scenarios in which a large transitive dataset with target variables is not available. We applied parameter transfer from a source domain to perform model parameter-based transfer learning with RF, inspired by D'Amato et al. [5]. This approach leverages model parameters from the source domain to adapt to the target domain, despite limited target data in the transitive dataset.

A limitation of the proposed workflow concerns the availability of transitive domain data. In domains with the same data restriction problems but with multiple candidate transitive domains or transitive datasets, a selection algorithm according to the domain and dataset properties should be provided. The presented model tackles problems where transitive domain data samples are limited to provide source learning, so that knowledge transfer on a hyperparameter-based approach from a model trained on larger data is provided for training a model on a transitive dataset. The proposed approach is not specific to

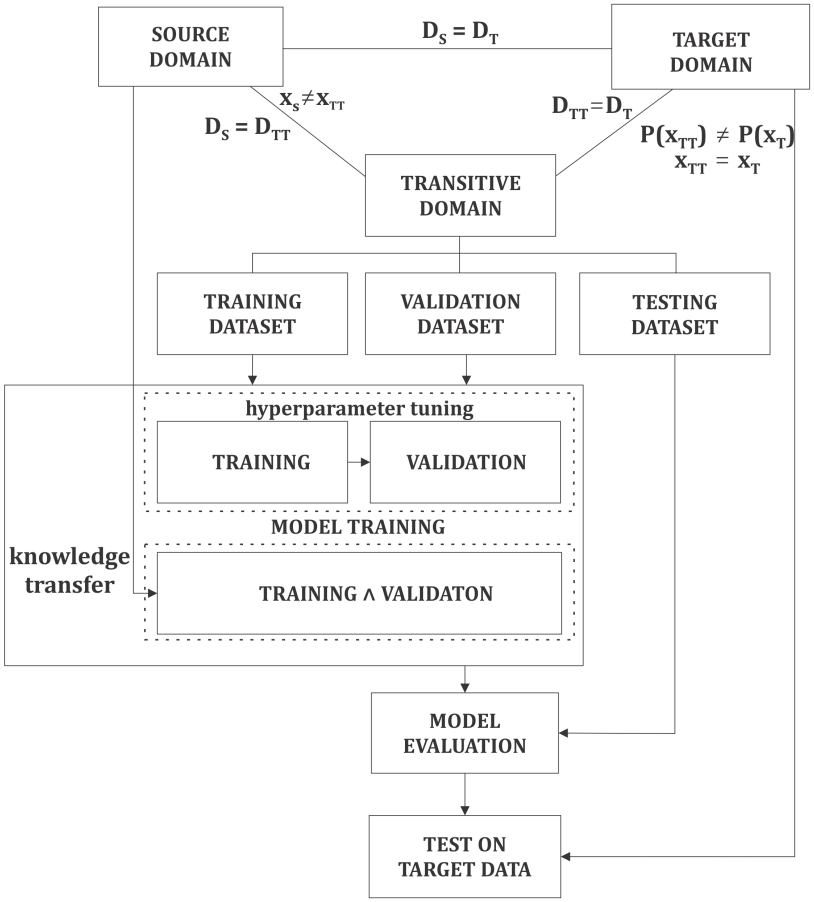


Fig 1. Proposed transfer learning workflow

any traditional machine learning algorithm, although the RF is used as a prediction model in this paper.

We compared our MPBTL-RF with traditional machine learning approaches, using the transitive dataset for training. We experimented with RF, XGB, and FNN for traditional model training and analyzed how our model performs relative to them under varying data availability conditions in the transitive dataset. Figure 1 presents the complete flow for all model trainings.

3.1. Datasets

We define two datasets: a transitive dataset and a target dataset. In the application scenario, there is no available labeled data in the target dataset, while the transitive dataset has a limited number of labeled data. The approach assumes that the learning tasks across the source, transitive, and target domains are the same. While the feature space between the

transitive and target domains is the same, X_{TT}=X_T, there is a difference in data distribution $P(X_{TT}) \neq P(X_T)$.

Transitive dataset. As a transitive dataset in this work, we used the Euronext Tech Leaders (TECLP) index, which covers the high-growth and leading companies in Europe from the Technology super-sector according to the Industry Classification Benchmark (ICB) industry classification.

The technology sector has exhibited slower ESG score development compared to other industries [10], due in part to challenges such as energy consumption and evolving regulatory expectations. However, recent trends show an increasing integration of ESG practices driven by regulatory pressures, stakeholder demands, and the sector's pivotal role in innovation. This makes the technology sector a relevant focus for studying ESG adoption and prediction in developing markets, where transparency and sustainability efforts are still emerging.

The available TELCP dataset consists of 94 companies with annual observations between 2018 and 2022, with some missing data for certain years. In our experiments, we used financial statements and ESG Refinitiv scores from this dataset for the companies included in the index.

We partitioned the transitive dataset into training, validation, and test subsets with a 60%-20%-20% split to prepare it for model training. Each model was assessed using five different fold splits to ensure robust evaluation, as detailed in the results section.

Considering the panel nature of the data, similar to [12], which shows a high correlation between scores for the same company across years, the train-validation-test split (60%-20%-20%) was done in a way that included each company in only one part of the split. This approach ensures the independence of the splits, preventing any sample from being included in more than one subset. This approach is crucial for maintaining the integrity of our proposed methodology, as it prevents data leakage between the training and validation sets. This precaution ensures that the machine learning model can generalize effectively to unseen data, thereby providing reliable predictions on the target dataset. Furthermore, consistent with the proposed transfer learning workflow, the feature space remains the same for both transitive and target data, X_{TT}=X_T. The test part of the transitive dataset is subsequently utilized for in-domain model evaluation to compare the performance of the machine learning and transfer learning aproaches.

Target datasets. As part of the model evaluation process, we created two target datasets. The first target dataset (holdout) in this study is based on available real data and consists of 14 blue-chip stocks traded in European markets classified as frontier according to the Financial Times Stock Exchange (FTSE) equity country classification[3]. According to the FTSE classification scheme from September 2024, there are 11 countries whose equity markets are classified as frontier markets, namely Bulgaria, Croatia, Cyprus, Estonia, Latvia, Lithuania, Malta, the Republic of North Macedonia, Serbia, the Slovak Republic, and Slovenia.

We examined the main stock market indices from these markets and selected the stocks according to the availability of the companies' ESG scores. Observed stocks are

[3] https://www.lseg.com/en/ftse-russell//equity-country-classification

constituents of the leading stock exchange indices - 8 stocks from the Bucharest Exchange Trading (BET) index, 2 from the Ljubljana Stock Exchange index (SBITOP), 2 from the Bulgarian Stock Exchange index (SOFIX), 1 from the Slovak share index (SAX), and 1 from the Cyprus Stock Exchange (CSE) index. The stocks come from different industries: Financials (35.7%), Industrials (28.6%), Technology (21.4%), Utilities (7.14%), and Energy (7.14%). The dataset, representing diverse sectors from developing markets, is used as an out-of-domain evaluation set relative to the technology-focused Euronext dataset, allowing the assessment of model generalization across markets with differing maturity and sectoral composition. Since this dataset is, in effect, cross-domain relative to the technology-focused transitive dataset, it provides a suitable setting for evaluating model generalization.

The second target dataset (synthetic) was created by expanding the first target dataset with additional artificially generated samples produced using the Gaussian Copula Synthesizer from the Python Synthetic Data Vault (SDV) library. This augmentation preserves the statistical characteristics of the original data while increasing its size, allowing for a more comprehensive evaluation of model performance in cross-domain scenarios.

Case study dataset. The case study dataset contains real-world data derived from the indices of the regulated market, the Belgrade Stock Exchange. In this experiment, the case study dataset consists of 14 companies listed on the BELEX, whose stocks were continuously traded on the regulated market and served as constituents of the BELEXline Index from 2018 to 2022 – a total of seventy data instances. Through this case study, we contribute to bridging the gap between theoretical advancements and practical applications in ESG score prediction, fostering the synergy between AI and business.

3.2. Model parameter-based transfer learning with random forest (MPBTL-RF)

Among machine learning models, RF is frequently utilized in transfer learning studies due to its ease of use and its well-documented high performance across a wide range of tasks [21, 19, 18]. Following the approach in [19], we hypothesize that overly complex models tend to overfit and struggle to generalize to dissimilar data and that simplifying model architecture can enhance predictive accuracy in the target domain. According to [17], the implementation of the RF requires setting two main parameters: the number of trees and the number of randomly selected predictor variables. In our experiment, we utilized the RF regression model for parameter transfer, as detailed by [5], recognizing that these parameters were tuned on a dataset from the STOXX® Europe Index.

Source model. Small transitive datasets make it challenging to train a model effectively, and parameters from a source domain model can be inherited for training. Conversely, when the transitive dataset is large enough, it serves as the source model itself. In both cases, the source model trained on a larger and well-sampled dataset transfers its parameters to facilitate learning on the target dataset.

Briefly explained, this model is in our work referred to as the source model and is used only for the transfer of model parameters, given that the available transitive data are from the Euronext Tech Leaders Index. The experiments in [5] utilized financial statement items and Bloomberg ESG scores collected for the constituents of the STOXX Europe 600

Index, representing large, mid, and small-capitalization companies across 17 countries in the European region. They selected a sample of 109 companies listed in the STOXX Europe 600 Index between 2014 and 2018, representing 21% of the entire set of companies included in the index. The selected companies belong to the Communications, Energy, Technology and Utilities industry sectors [5].

We applied a RF regressor with transferred parameters from the source model by setting: minimum samples per leaf to 1, maximum features per split to 5, and the number of trees in the forest to 500 to optimize predictive performance on the target dataset. Setting minimum samples per leaf to 1 allows each leaf node to capture finer patterns, which is beneficial for smaller datasets. Setting the maximum features per split parameter to the value of 5 limits the features considered at each split, balancing variance and interpretability, while the number of trees in the forest is set to 500 to create a large forest of trees to enhance stability and accuracy. These parameters should enable the model to leverage source knowledge effectively, yielding substantial generalization and performance in our target domain.

Feature selection. According to the proposed workflow in Figure 1, our study also employs the same financial features except the categorical sector feature presented in the source domain dataset, as there are no sector differences in our transitive dataset. The indicators used, listed in Table 1, represent liquidity, solvency, profitability, operating efficiency, and the company's market value. Variables used in the formulas are defined as follows: Net sales – the total revenue generated by the company from the sale of goods or services during the year, after deducting sales returns, allowances, and discounts; Total assets – the sum of current and long-term assets owned by the company; EBIT – a company's operating profit, calculated as earnings before deducting interest expenses and income taxes; Annual dividend per share – the total dividends paid divided by the number of outstanding shares over the year; Share price – the closing price of a company's share on the last trading day of the year; Net income – the company's total profit after deducting all expenses and taxes from total revenue; Equity – the residual owner's interest in a company's assets after all liabilities have been deducted; Total liabilities – the sum of all current and long-term financial obligations owed by the company; Cash – cash and cash equivalents readily available to the company; Cash flow – operating cash flow calculated by adjusting net income for non-cash items like amortization and changes in working capital; Total current assets – assets expected to be sold, consumed, or converted to cash within one year or the company's operating cycle; Total current liabilities – financial obligations and debts the company is expected to settle within one year or its normal operating cycle; Total debt – the sum of all interest-bearing debt, including current borrowings and long-term liabilities; Earnings per share – net income divided by the average number of shares outstanding. All financial values are based on the company's financial statements for the relevant fiscal year.

3.2.1. Parameter transfer procedure and overfitting control

In this study, MPBTL-RF was implemented through *structural parameter inheritance* between domains using the RF regressor. Model parameters optimized on the STOXX

Table 1. Financial Indicators

Label	Variable Name	Formula
Sales_to_Assets	Asset Turnover Ratio	Net sales / Total assets
EBIT_to_Sales	Operating Margin	EBIT / Net sales
DY	Dividend Yield	Annual dividend per share / Share price
NI_to_Sales	Net Profit Margin	Net income / Net sales
Rating	Kralicek QuickTest[1]	0.25 × (Equity/Total liabilities)
		+0.25 × ((Total liabilities - Cash)/Cash flow)
		+0.25 × (EBIT/Total assets)
		+0.25 × (Cash flow/Net sales)
LR	Current Liquidity Ratio	Total current assets / Total current liabilities
SR	Solvency Ratio	Total debt / Total assets
P/E	Price to Earnings Ratio	Share price / Earnings per share

[1] Results expressed in points from 1 to 5.

Europe 600 dataset [5] were transferred to the Euronext Tech Leaders dataset and subsequently applied for prediction on the target dataset. The transferred parameters included: (i) the number of trees ($n\_estimators = 500$), (ii) the number of features considered at each split ($max\_features = 5$), and (iii) the minimum number of samples per leaf ($min\_samples\_leaf = 1$).

This approach follows the *homogeneous transductive* parameter-transfer paradigm [15, 16], in which the model structure–rather than learned weights–is reused to regularize learning across domains. By constraining model complexity, the inherited parameters function as a form of inductive bias, stabilizing learning in the transitive and target domains where labeled data are limited. The procedure is summarized on Figure 2.

```
# Source: STOXX Europe 600
source_model = RandomForestRegressor(
    n_estimators=500, max_features=5, min_samples_leaf=1
)
source_model.fit(X_source, y_source)

# Transitive: Euronext Tech Leaders
transitive_model = RandomForestRegressor(
    n_estimators=source_model.n_estimators,
    max_features=source_model.max_features,
    min_samples_leaf=source_model.min_samples_leaf
)
transitive_model.fit(X_transitive, y_transitive)

# Target: BELEX (Frontier Markets)
y_pred = transitive_model.predict(X_target)
```

Fig 2. Transitive Model Training and Prediction

To prevent overfitting and maintain generalization, the model was trained using a 80%–20% train–test split, ensuring that each company appears in only one subset, thereby eliminating data leakage. Additionally, a 5-fold cross-validation procedure was employed to assess model stability. Fixing the structural parameters from the source model serves as a regularization mechanism, reducing the risk of overfitting and enhancing robustness in small-sample target domains.

3.3. Traditional ML with parameter-tuning

To evaluate traditional machine learning models, we employed a systematic hyperparameter-tuning methodology using 5-fold cross-validation. This involved splitting the data into five folds and training a separate model on each fold, using a 60%-20%-20% split for training, validation, and testing within each fold to ensure robust performance comparison. Consistency was maintained by using the same split as in the previous experiment with model parameter-based transfer learning in Section 3.2 for the test set, but reallocated 25% from each training split (previously 80%) to create a validation set for each fold, achieving a final 60%-20%-20% split.

A comprehensive hyperparameter grid search was conducted for each model. This process entailed evaluating the models across the five validation sets to identify the optimal parameter configurations for predicting ESG, E, S, and G targets. The best-performing parameters were then applied to the test set for final evaluation across all targets (ESG, E, S, G).

For this evaluation, we employed three different machine learning approaches:

- **RF**: An ensemble learning method that builds multiple decision trees and outputs the mode of the classes (classification) or the mean prediction (regression).
- **XGB**: Unlike RF, XGB builds trees sequentially, with each tree focusing on correcting the errors of the previous one. This iterative error-correction process results in a highly accurate model by gradually reducing bias. This approach is especially valuable with small datasets, where limited information can benefit from stepwise refinement [3].
- **FNN**: Although typically requiring large datasets, FNN can still perform well on smaller datasets if structured carefully to avoid overfitting. Their flexibility and capacity to model non-linear relationships make them effective even with limited data and features, as in this study. Additionally, FNN offers advantages in terms of training ease and interpretability due to the relatively small number of parameters. This efficiency, combined with their ability to model complex relationships, justifies the use of FNN for predicting ESG scores on the available dataset.

These approaches were selected for their established effectiveness in predictive modeling and their ability to handle the challenges posed by the small dataset and the complexity of ESG score prediction. Additionally, these networks enable efficient hyperparameter tuning, which was performed with various combinations, as detailed in Table 2.

3.4. Baseline method

In assessing our models, we also benchmarked our results against values from [12]. In their study, three individual models were used – FNN, CatBoost (CB), and XGB, along

Table 2. Grid Search Parameter Ranges for Evaluated Models

Model	Parameter	Value Combinations
RF	Number of Estimators	100, 500, 1000
	Max Features	auto, sqrt, log2
	Max Depth	10, 30, 50, None
	Min Samples Split	2, 5, 10
	Min Samples Leaf	1, 2, 4
	Bootstrap	True, False
XGB	Early Stopping Rounds	None, 10
	Learning Rate	10^{-2}, 10^{-3}, 10^{-4}
	Max Depth	3, 5, 10, 15
	Min Child Weight	0.5, 0.7, 1
	Number of Estimators	100, 200, 300
FNN	Number of Layers	5, 6, 7
	Number of Neurons	1000, 600, 300, 200
	Activation	sigmoid, relu, leaky_relu
	Dropout	0, 0.3, 0.5, 0.7
	Learning Rate	10^{-3}, 10^{-4}, 10^{-5}
	Batch Normalization	True, False
	Epochs	100, 200, 300
	Early Stopping	True, False

with two types of ensemble models – NN Ensemble (NNE) and Heterogeneous Ensemble (HE) to predict ESG and pillar scores using data from companies in the S&P 500 index. Their results, along with their simple mean baseline model (BL), are presented in Table 3.

Table 3. Baselines for different models from [12]

	HE	BL	FNN	CB	XGB	NNE
ESG	11.2	16.5	12.1	11.3	11.4	12.1
E	14.9	22.7	16.3	15.0	15.2	16.2
S	13.5	19.0	14.6	13.5	13.6	14.5
G	16.7	18.7	17.4	16.7	16.8	17.4

3.5. Models training, validation, and evaluation parameters

The training subset was used to train all traditional machine learning models. To ensure that the experiments are not influenced by random chance, we implemented a 5-fold cross-validation strategy to optimize model performance and reduce overfitting. We performed hyperparameter tuning across different fold splits of our transfer dataset for each of the pillars: ESG, E, S, and G. To ensure that the model treated each feature equally and to

facilitate faster convergence during training, we standardized the dataset by scaling the features so that they had a mean of 0 and a standard deviation of 1.

Fine-tuning parameters for FNN were performed on the National Platform for AI of Serbia, utilizing a single NVIDIA A100 GPU with 40GB using PyTorch. After evaluating 5,184 models per fold, we identified the optimal parameters, totaling 25,920 models across all folds. The process took up to 50 hours for all five folds. The training was performed in batches of 32. All other fine-tuning, model training and inference were done on local computers.

Validation was performed for the traditional ML models with parameter-tuning in Section 3.3. The validation was performed through cross-validation and cross-sample evaluation (test subset) across multiple folds. A test subset is considered for in-domain evaluation. At the same time, both target datasets were utilized for out-of-domain evaluation to assess the generalization of the best-performing models across different contexts not seen during training.

Furthermore, to ensure methodological coherence with the workflow illustrated in Figure 1, we conducted statistical analyses to validate the suitability of the selected transitive dataset and to assess feature distribution differences across all datasets used in the study. The results of the Kruskal-Wallis test presented in Table 4 show no statistically significant differences in the distribution of ESG scores – both overall and across its individual Environmental (E), Social (S), and Governance (G) components–among the observed datasets. This suggests that ESG adoption levels do not significantly vary among the companies from the technology sector in the transitive dataset and those in the holdout and synthetic datasets.

Table 4. Kruskal-Wallis test results for TECLP, holdout, and synthetic datasets

Indicator	Test Statistic	p-value
ESG	0.8367	0.6581
E	0.3176	0.8532
S	4.6039	0.1001
G	1.2305	0.5405

Note: Significance level (α) set at 0.05.

On the other hand, the Kruskal-Wallis test results followed by the Dunn-Bonferroni post hoc pairwise comparison among TECLP, holdout and BELEX datasets yielded mixed results as presented in Table A1 in the Appendix. Several variables, including Sales_to_Assest, Rating and LR, did not show statistically significant differences after adjustment, indicating insufficient evidence to reject the null hypothesis of equal distributions among these groups. Statistically significant differences (adjusted $p < 0.05$) were observed for EBIT_to_Sales and NI_to_Sales in all group comparisons. Conversely, variables such as DY, P/E, and SR showed significant differences only in selected group comparisons, indicating heterogeneous effects depending on the specific group pairing.

The obtained results correspond to the experimental framework shown in Figure 1.

Additionally, for the purpose of a comprehensive analysis of model robustness and properties, traditional machine learning models were implemented in conjunction with the BORUTA feature selection algorithm.

During cross-validation, we used grid search (Table 2) to find the best combination of hyperparameters on the training dataset, aiming to minimize the value of the scorer function, which calculates the mean squared error (MSE) between predicted and actual values. The best model was used to predict outcomes on the validation and test subsets. The best model was also used to predict values for the out-of-domain evaluation, holdout and synthetic dataset. Similarly, in models with BORUTA feature selection implementation, feature selection was performed for each fold separately.

The evaluation was performed in similar regression tasks as in [12], using MAE, as defined in Equation 1.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| \tag{1}$$

Finally, it is worth noting that the proposed approach is versatile and not limited to any specific traditional machine learning algorithm – it can be applied across various algorithms and data domains.

4. Results

This section introduces the results obtained through empirical analysis of different approaches used in this study. We conducted experiments for ESG score prediction and for the Environmental, Social, and Governance pillars separately. Thus, the obtained results provide comprehensive insight into the possibility of transfer learning in developing markets. As previously described in section 3.1 in order to further verify the robustness of the proposed model, a synthetic dataset was created and an out-of-domain evaluation was performed on it, as well as on the holdout dataset. Figure 3 illustrates the experimental setup.

4.1. In-domain evaluation

We first present the evaluation results on the test subset of the transitive dataset (in-domain), as defined in Section 3.1. The analysis includes the proposed MPBTL-RF approach and traditional machine learning approaches (RF, XGB), with and without feature selection.

MPBTL-RF in-domain.

The results obtained using the proposed MPBTL-RF are shown in Table 5. Based on the average values displayed, it can be observed that the G pillar score is the most challenging to predict, as it has the highest error. In contrast, the ESG component is predicted with the lowest error. Analyzing the results across folds, the difference between the highest and lowest errors is smallest for the ESG score (2.48), while the largest difference is observed in the E component (6.39 absolute error difference). The variation in error levels suggests stability in the predictions obtained with this model, considering the dataset split defined in Section 3.1.

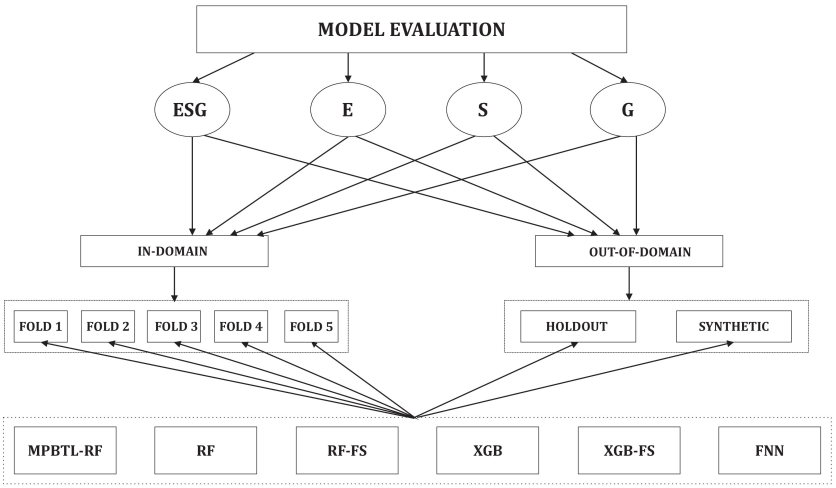


Fig 3. Experimental settings

Table 5. MPBTL-RF in-domain evaluation

MAE	**ESG**	**E**	**S**	**G**
Fold 1	14.61	17.08	18.12	21.40
Fold 2	15.41	18.78	15.94	22.26
Fold 3	13.05	15.03	15.01	17.71
Fold 4	15.53	21.42	15.60	21.13
Fold 5	15.01	20.48	15.44	19.24
Avg MAE	14.72	18.56	16.02	20.34

RF in-domain. Table 6 summarizes the in-domain performance of the RF model. The results exhibit a consistent pattern across folds, with relatively stable MAE values for all ESG components. The ESG score is associated with the lowest prediction error, whereas the G pillar remains the most difficult to predict, yielding the highest MAE. In most cases, validation errors are slightly lower than those on the test set, which is expected given that model tuning is performed on the validation data. The limited discrepancy between validation and test performance indicates that the model does not suffer from pronounced overfitting.

When feature selection is introduced (Table 7), the RF-FS model attains very similar MAE values across all targets. This indicates that restricting the feature space does not adversely affect predictive accuracy. The results suggest that the selected subset of features retains the essential information required for prediction, while less informative variables can be excluded without loss of performance. Overall, the RF model preserves stable behavior under feature reduction in the in-domain setting.

XGB In-domain. Table 8 presents the in-domain results for the XGB model. The model shows consistent performance across folds, with relatively stable MAE values for

Table 6. Traditional RF in-domain evaluation

	Validation Set				Test Set			
MAE	ESG	E	S	G	ESG	E	S	G
Fold 1	13.14	16.57	16.27	19.94	14.45	17.15	20.96	20.66
Fold 2	14.38	20.33	15.22	18.51	16.43	19.87	18.00	22.12
Fold 3	16.13	18.90	16.67	19.31	13.01	14.76	14.32	18.32
Fold 4	15.62	17.98	16.32	19.62	14.78	21.13	15.67	20.31
Fold 5	14.66	16.30	16.75	21.55	15.70	21.21	15.02	17.91
Avg MAE	14.78	18.02	16.25	19.79	14.87	18.83	16.07	19.86

Table 7. Traditional RF -FS in-domain evaluation

	Validation Set				Test Set			
MAE	ESG	E	S	G	ESG	E	S	G
Fold 1	13.45	17.00	16.90	19.89	14.01	17.15	17.77	19.93
Fold 2	13.62	20.21	14.97	17.96	16.59	19.34	17.65	21.86
Fold 3	16.68	19.50	16.92	20.07	12.57	14.83	15.73	19.31
Fold 4	14.63	18.38	15.48	19.44	15.01	20.45	15.78	19.85
Fold 5	15.21	16.68	17.21	21.38	15.44	21.20	16.33	16.22
Avg MAE	14.72	18.35	16.30	19.05	12.72	18.59	16.65	19.43

all ESG components. The ESG score is predicted with the lowest error, while the G pillar remains the most challenging target, exhibiting the highest MAE. Across most components, the average MAE is lower on the validation set than on the test set, which is expected given that hyperparameter tuning is performed on the validation data. The relatively small differences between validation and test errors suggest that the model does not exhibit severe overfitting. However, for the S component, a slight increase in test error (difference of 0.647) indicates mild overfitting.

When feature selection is applied (Table 9), the XGB-FS model achieves comparable MAE values across all targets, indicating that reducing the feature space does not negatively impact predictive performance. This suggests that the selected subset captures the most relevant information, while redundant features can be removed without loss of accuracy. The results further indicate that XGB remains robust under feature reduction, although minor overfitting effects may still occur for certain components in small-sample settings.

Compared to the results in Table 3, where XGB was trained on a much larger dataset and reported MAE errors of 11.4, 15.2, 13.6, and 16.8 for the ESG score and the E, S, and G components, respectively, the higher MAE errors observed with XGB (15.19, 19.51, 16.29, and 19.81) highlight the significant influence of training set size on prediction accuracy.

FNN In-domain. The results in Table 10 show that, although the FNN achieves the lowest MAE on the validation set, its performance on the test set is substantially worse and exhibits the largest variance across folds. This behavior indicates strong overfitting,

Table 8. XGB in-domain evaluation

	Validation Set				Test Set			
MAE	ESG	E	S	G	ESG	E	S	G
Fold 1	14.16	17.44	16.97	19.86	14.61	18.74	18.04	18.67
Fold 2	14.66	20.41	16.43	17.17	17.13	20.74	18.29	22.21
Fold 3	15.80	18.45	17.07	18.06	12.90	15.02	15.77	18.18
Fold 4	15.43	18.12	16.46	19.33	15.32	21.61	13.92	20.77
Fold 5	15.79	17.05	17.75	21.22	15.99	21.48	15.42	19.23
Avg MAE	15.17	18.29	16.94	19.13	15.19	19.51	16.29	19.81

Table 9. XGB-FS in-domain evaluation

	Validation Set				Test Set			
MAE	ESG	E	S	G	ESG	E	S	G
Fold 1	14.14	18.24	18.28	19.40	14.63	18.71	17.76	17.75
Fold 2	14.96	20.54	15.28	17.25	17.65	20.75	17.87	21.47
Fold 3	15.54	18.45	16.65	17.94	13.02	15.02	15.98	18.27
Fold 4	15.43	18.30	16.43	19.00	15.29	21.18	13.91	19.45
Fold 5	16.03	18.52	16.67	20.92	16.68	23.08	15.05	18.76
Avg MAE	15.22	18.81	16.66	18.90	15.45	19.75	16.11	19.14

which is expected when applying high-capacity neural networks to small and heterogeneous datasets. Due to their expressive nature, FNNs are sensitive to variations in training-validation splits, often converging to different local minima and producing unstable predictions across folds.

In contrast, ensemble tree-based models (RF, XGB) benefit from averaging mechanisms that provide more stable and consistent predictions. The observed behavior of the FNN therefore reflects the trade-off between flexibility and stability in small-sample deep learning settings [11]. Although the model captures non-linear relationships between financial indicators and ESG scores, its lack of generalization limits its practical applicability.

This discrepancy between validation and test performance is not observed in the other models, and data splitting procedures ensured strict separation across folds, indicating that the observed overfitting is inherent to the model rather than a consequence of data leakage or methodological bias. Consequently, the FNN model is excluded from further comparative analysis.

In-domain comparative analysis. The comparison of the results for all previously described models is presented in Figure 4.

ESG score. The proposed MPBTL-RF achieves an average MAE of 14.72, which is comparable to the MAE value of 11.3 presented in Table 3. At the same time, the proposed transfer learning approach has the smallest standard deviation value (1.00), compared to

Table 10. FNN In-domain evaluation

	Validation Set				Test Set			
MAE	ESG	E	S	G	ESG	E	S	G
Fold 1	12.01	13.89	13.41	17.08	24.69	24.46	26.69	23.00
Fold 2	12.97	17.91	11.47	16.93	100.08	86.60	102.75	106.75
Fold 3	13.29	16.04	14.89	17.56	62.02	51.96	77.92	68.13
Fold 4	11.11	15.18	12.08	16.51	21.58	26.53	25.56	24.65
Fold 5	12.77	14.21	12.49	20.16	26.53	28.59	29.99	24.51
Avg MAE	12.43	15.45	12.868	17.65	46.98	43.63	52.58	49.41

the value of 1.30 obtained with RF, 1.65 obtained with baseline models, and 1.58 obtained with the XGB algorithm.

The MPBTL-RF model exhibits the smallest variation in MAE values across folds, with a difference of just 2.48 between the highest and lowest values. This suggests that the model is the most robust and least affected by changes in data splits.

In relation to the models with feature selection in term of ESG predictions, the proposed MPBTL-RF approach exhibit somewhat weaker but still comparative results.

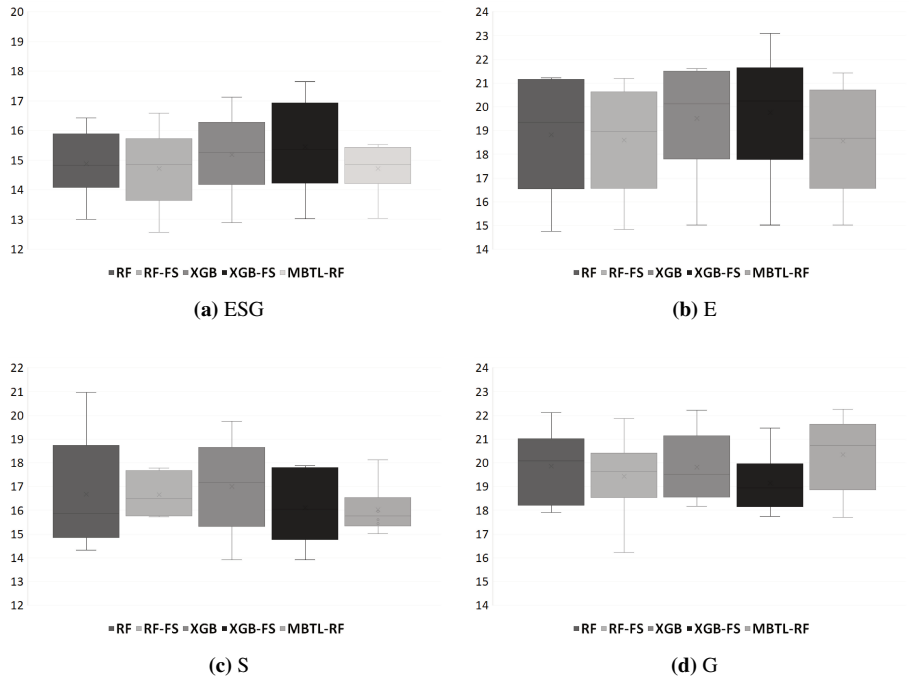
E pillar score. For the E pillar score, the proposed MPBTL-RF results in average MAE to 18.56 which is comparable to the MAE value of 14.9 obtained in Table 3. Additionally, in the case of the E pillar, the proposed transfer learning approach has the smallest standard deviation. The lowest MAE value obtained across folds was in Fold 3 using the RF model. The smallest difference between the highest and lowest MAE values is observed with the MPBTL-RF model, amounting to 6.39, suggesting stability in this model across different splits (train, validation, and test). In relation to the RF and XGB models with feature selection in term of average MAE for E pillar predictions, the proposed MPBTL-RF approach outperformed feature selection approaches.

S pillar score. The proposed MPBTL-RF results in average MAE to 16.02, which is comparable to the MAE value of 13.5 obtained in Table 3. Additionally, for the S pillar, the proposed transfer learning approach exhibits the smallest standard deviation. The smallest difference between the highest and lowest MAE values across folds is observed with the MPBTL-RF model, amounting to 3.11.

For S pillar predictions, the proposed MPBTL-RF approach again has the best average MAE values in comparison to RF and XGB with feature selection.

G pillar score. For the G pillar score, the proposed MPBTL-RF brings the MAE to 20.34, which is comparable to the MAE value of 16.1 obtained in Table 3. Additionally, in Fold 3, the MPBTL-RF approach achieves the lowest MAE among the models studied, indicating a stronger generalization capability. However, the smallest range between the highest and lowest MAE values across models is observed with the traditional RF model.

Regarding standard deviation, the XGB model shows the smallest variation for the G component, which, along with the MAE level, suggests slightly more favorable predictions using traditional ML and more complex models on challenging problems. However, it is important to note that for all models in the comparison, the average MAE is derived from MAE values across five folds. This approach indirectly represents a homogeneous ensemble in traditional ML models, as the average MAE is the mean prediction of five



(a) ESG **(b)** E

(c) S **(d)** G

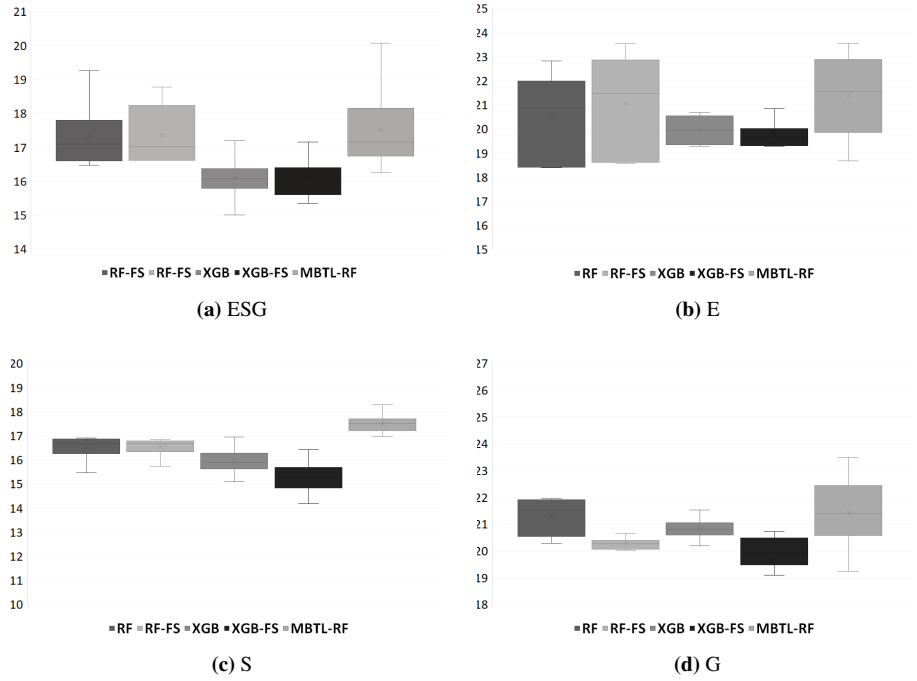
Fig 4. In-domain comparative analysis.

distinct models using the same algorithm. Compared to other models, MPBTL-RF uses the same parameters for each fold but has a larger training set available. Thus, even though XGB achieves a lower MAE, the difference in average MAE between these two models (0.53) still supports the effectiveness of the proposed model, even for highly complex tasks. In relation to the models with feature selection in terms of average MAE for G pillar predictions, the model exhibits somewhat weaker but still comparative results.

4.2. Out-of-domain evaluation

To evaluate model generalization, we conducted out-of-domain experiments using the holdout and synthetic datasets described in Section 3.1. The results are presented in Figures 5 and 6.

Figure 5 presents the prediction results of the MPBTL-RF, XGB, XGB-FS, RF and RF-FS models on the holdout target dataset. The results indicate that the average MAE values for the observed labels are similar to those obtained on the transitive (in-domain) dataset's test subset, highlighting the approach's effectiveness for complex predictive tasks. The standard deviation of the average MAE across both test sets (for ESG, E, S, and G) is 2.47. Additionally, the MAE values achieved on the target dataset further confirm that dividing the data into independent sets during model training is suitable for enhancing model generalization capabilities. The G component remains the most challenging to predict. By examining all the results in Figure 5, the MAE values show slightly better



(a) ESG **(b)** E

(c) S **(d)** G

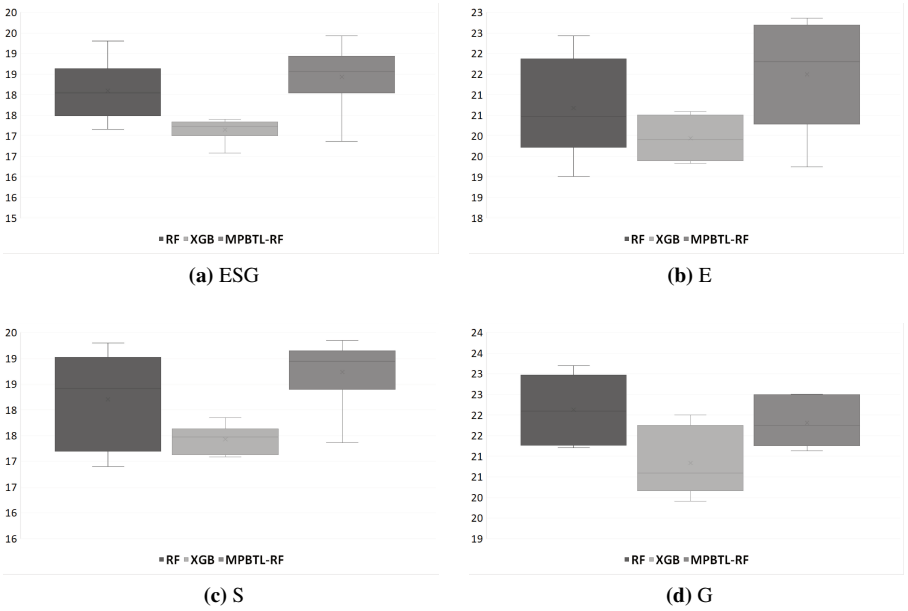
Fig 5. Out-of-domain (holdout dataset) Comparative Analysis

predictions for traditional ML models. However, it is important to note that for all models in the comparison, the average MAE is derived from 5-fold MAE values, which can be considered a homogeneous ensemble across all traditional machine learning models. This is because the average MAE represents the mean prediction of five different models using the same algorithm. Unlike other models, the MPBTL-RF model uses the same parameters for each fold but benefits from a larger training set. Even though traditional learning models achieve slightly lower MAE values, this small difference supports the effectiveness of the proposed transfer learning approach, especially for highly complex tasks.

ESG score. The proposed MPBTL-RF for ESG results in an average MAE of 17.51. The average MAE for the MPBTL-RF and RF models are nearly identical (17.51 versus 17.32), which directly supports the validity of applying the MPBTL-RF.

E pillar score. The same conclusion applies to predictions for the E pillar score. Although the resulting MAE might appear high, they remain comparable to values found in the literature for the specific domain of ESG score prediction and its components. It is particularly important to note that the target dataset was created using data from four countries. Given this diversity, the selection of learning features and the predictive model can effectively address such complex problems.

S pillar score. For the S pillar score the difference in error compared to the RF and XGB models are 1.00 and 1.52, respectively, indicating that the MPBTL-RF model can also be considered satisfactory in this case.



(a) ESG **(b)** E

(c) S **(d)** G

Fig 6. Out-of-domain (synthetic dataset) Comparative Analysis

G pillar score. Finally, when predicting the G pillar score, which is considered the most complex based on the MAEs, the MPBTL-RF model remains competitive, even though other traditional machine learning models achieved slightly better average values.

Applying the trained MPBTL-RF, XGB and RF on the synthetic target dataset yields results shown in Figure 6.

ESG score. By examining all the results in Figure 6 for the ESG score on the synthetic dataset, the MAE values show slightly better predictions for traditional ML models. Notably, for all models in the comparison, the average MAE is derived from 5-fold MAE values, which can be considered a homogeneous ensemble across all traditional machine learning models.

E pillar score. The same conclusion applies to predictions for the E pillar score. Although the resulting MAE might appear high, it remains comparable to values found in the literature for the specific domain of ESG score prediction and its components.

S pillar score. For the S pillar score, the difference in MAE value compared to the RF and XGB models is 0.53 and 1.32, respectively, indicating that the MPBTL-RF model can also be considered satisfactory in this case.

G pillar score. Finally, when predicting the G pillar score, which is considered the most complex based on the MAEs, the MPBTL-RF model remains competitive and achieved slightly better average values than the RF model.

To conclude, in the in-domain analysis, the MPBTL-RF model achieves second-best average results for the ESG score, and only the RF-FS approach has a better average MAE value. For the E and S components, the proposed MPBTL-RF approach outperforms other models, while the XGB-FS approach performs best for the G component. To compare multiple models on multiple datasets, we applied Friedman's test to check

whether the compared models have significant general differences in performance among folds. Thus, based on the Friedman statistical test conducted and shown in the Appendix, Table A2, there is no statistically significant difference in the performance of the observed approaches, except for the E pillar score. But even the null hypothesis is rejected in the second stage by applying the Nemenyi test, which did not find a statistically significant difference between any of the group pairs post-hoc test.

However, in the out-of-domain evaluation on the holdout dataset, XGB-FS delivers the best performance across the ESG score and all individual components. This indicates that transfer capabilities could be improved with a larger and higher-quality transfer dataset, potentially leading to better MAE values overall with XGB or another algorithm.

Based on the performed statistical test shown in the Appendix, Table A2, it can be concluded that there is no significant difference in MAE values across approaches for ESG scores and E pillar scores, while a statistically significant difference was observed for S and G pillars between the XGB-FS model and the MPBTL-RF approach. However, it should be pointed out that here too the comparison is where the XGB model is at an advantage in the selection of attributes for model training compared to the MPBTL-RF approach, which always generates results based on the same model parameters. Thus, even observed statistically significant differences between the models do not diminish the importance of the proposed model.

In relation to the synthetic dataset, it can be seen that the XGB model gives slightly less variation in precision when it comes to ESG scores prediction, while the MPBTL-RF model has the least variation when it comes to the prediction of the G component, which indicates model robustness according to all observations about G pillar score prediction. To test whether there is a statistically significant difference between the obtained MAE values among folds in the observed models, we conducted the Friedman test in the first stage and the Nemenyi test as the post-hoc test. All of the tests have been performed at 5% significance level. According to the test results available in the Appendix, Table A2, there is no statistically significant difference between MPBTL-RF and the other models obtained on the synthetic dataset. Thus, a statistically significant difference was confirmed between the RF and XGB models on the synthetic dataset.

In terms of computational efficiency, the differences in total training time were substantial. MPBTL-RF, which does not require hyperparameter fine-tuning, completed a fold in only 4.46 seconds. By contrast, the remaining models required full one-fold hyperparameter tuning: the standard RF model needed 239.86 seconds, XGB required 26.818 seconds, and the FNN model was by far the most time-consuming at 20,639.10 seconds.

The observed results highlight the computational efficiency and practicality of the proposed model. Overall, the obtained results demonstrate the potential of machine learning models, specifically MPBTL-RF, for predicting ESG scores in developing financial markets. Despite the challenges of predicting the G component, MPBTL-RF exhibited strong generalization capabilities across in-domain and out-of-domain data. This supports the idea that transfer learning can be a valuable tool for improving ESG score predictions, especially when dealing with limited or less-represented data in developing markets.

Considering that all models in this study were trained using only nine attributes, the results can be regarded as highly favorable, highlighting the simplicity, time efficiency, and applicability of the proposed approach.

5. Model application: case study of the BELEX

This section presents the experimental results and discusses the application of the proposed model to datasets derived from indices of the regulated market of the BELEX.

The case study in this experiment consists of 14 companies listed on the BELEX, whose stocks were continuously traded on the regulated market and served as constituents of the BELEXline Index from 2018 to 2022. Given that ESG scores can significantly influence investor decisions, the results are presented anonymously, as recommended by [12]. Table 11 shows the ESG scores for the selected blue-chip companies listed on the BELEX.

Table 11. One-year ESG rating predictions using MPBTL-RF on BELEX case study dataset

Company	ESG	Class	Company	ESG	Class
A	32.15	1	H	43.03	1
B	56.20	2	K	49.72	1
C	58.02	2	L	49.65	1
D	66.88	2	M	53.71	2
E	52.85	2	N	44.35	1
F	55.05	2	O	52.44	2
G	47.62	1	P	52.96	2

The selected companies represent various industry sectors: (1) Industrials (50%), (2) Financials (28.57%), (3) Utilities (7.14%), (4) Energy (7.14%), and (5) Real estate (7.14%). Figure 7 shows the ESG scores for the selected blue-chip companies listed on the BELEX per industry sector.

The box plot for ESG scores in Figure 7 predicted over five years indicates that the lowest average ESG score is evidenced in real estate (39.26) and utilities industry (48.89) with a wide interquartile range and no significant outliers, while the companies in the energy industry have the highest average ESG score (60.32) with most data clustered in a narrow interquartile range (5.09) but with outliers in both directions. The average ESG scores in the financial and industry sectors show a similar level of MAE (53.44 and 52.42 respectively), but ESG scores of companies in the financial sector are more dispersed, while in the industry sector, they display concentration in a tight interquartile range with some notable outliers. The obtained ESG score values were used as a basis for further analysis of financial performance. We clustered the observed stocks based on the predicted ESG scores into the following categories, corresponding to different performance levels:

- Class 0: ESG score ranging from 0 to 25 (first quartile) – Poor ESG performance.
- Class 1: ESG score ranging from 25 to 50 (second quartile) – Satisfactory ESG performance.
- Class 2: ESG score ranging from 50 to 75 (third quartile) – Good ESG performance.
- Class 3: ESG score ranging from 75 to 100 (fourth quartile) – Excellent ESG performance.

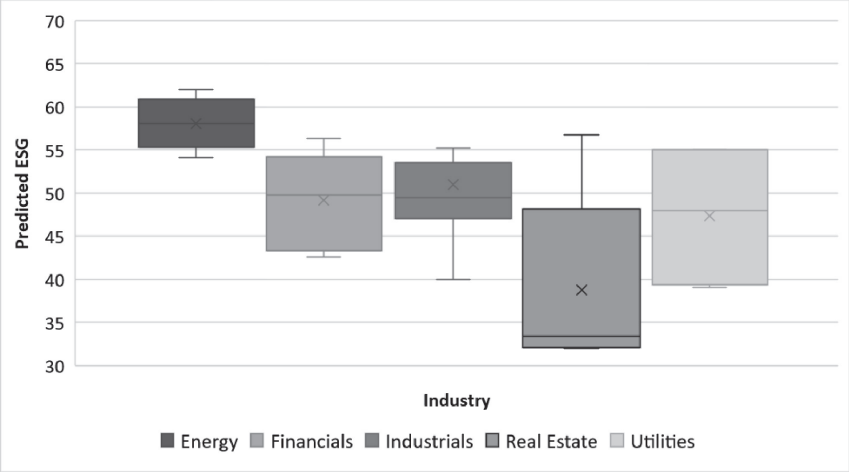


Fig 7. BELEX predicted ESG

Table 11 presents the obtained cluster categories. The predicted ESG scores were clustered into two groups: Class 1 – containing 39 observations, and Class 2 – with 31 observations. Predicted ESG-based clustering provides stakeholders with a predictive framework to incorporate sustainability considerations early in the decision-making process, thereby enhancing proactive risk management and promoting long-term value creation, despite the absence of direct ESG data.

To assess whether the proposed clusters exhibit significant differences in financial parameters, we conducted statistical tests using the non-parametric Mann-Whitney U test, which compares two independent groups based on a single continuous variable. Table 12 presents the descriptive statistics for these parameters within each cluster, along with the significance of the differences between the variables in these groups.

The analysis reveals significant differences in the following variables across the clusters: Dividend Yield (DY), Rating, and Price to Earnings (P/E) ratio. In line with [5], these financial indicators can be considered significant determinants of the ESG score.

- Dividend Yield (DY): Companies in Class 1 showed a clear tendency not to pay dividends during the observed period, with an average dividend yield of 3% and a standard deviation of 5%. Conversely, more than half of the companies in Class 2 paid dividends above 4%, with a standard deviation of 26%, indicating better financial capacity to support higher ESG scores.
- Rating: The overall operating performance rating showed significant differentiation between the groups. Companies in Class 1, with an average rating of 2.56, exhibited stable but lower capacities for improving ESG practices. In contrast, Class 2 companies, with a lower average rating of 2.24 but a higher standard deviation of 0.90, demonstrated greater sensitivity to business dynamics, suggesting more flexibility and potential for growth in ESG practices.
- The P/E ratio, which indicates market expectations about a company's future earnings, showed considerable differences between groups. Companies in Class 1 had an

Table 12. Descriptive statistics of financial indicators of the companies in the BELEX target dataset

Indicators[1]	Class	Descriptive Statistics				
		Mean	Median	SD	Min	Max
Sales_to_Assets	1	0.69	0.59	0.51	0.01	2.26
	2	0.85	0.71	0.50	0.10	1.92
EBIT_to_Sales	1	-0.03	0.03	0.46	-2.48	0.54
	2	0.30	0.06	1.22	-0.12	6.83
DY*	1	0.03	0.00	0.05	0.00	0.17
	2	0.09	0.04	0.26	0.00	1.48
NI_to_Sales	1	-0.07	0.03	0.45	-2.60	0.17
	2	0.23	0.05	1.04	-0.28	5.77
Rating*	1	2.56	2.75	0.58	1.00	3.50
	2	2.24	2.00	0.90	1.00	4.00
LR	1	2.95	1.49	3.23	0.22	10.71
	2	1.93	1.52	1.57	0.28	6.61
SR	1	0.46	0.50	0.26	0.10	0.97
	2	0.42	0.39	0.23	0.07	0.90
P/E*	1	90.74	8.15	475.18	-718.39	2751.72
	2	1.60	5.06	24.04	-108.31	48.54

[1] Indicators explanations: Sales_to_Assets – Asset Turnover Ratio, EBIT_to_Sales – Operating Margin, DY – Dividend Yield, NI_to_Sales – Net Profit Margin, Rating – Kralicek QuickTest, LR – Current Liquidity Ratio, SR – Solvency Ratio, P/E – Price to Earnings Ratio.

[*] Test significance (α) is set at 0.05. p-values are denoted as follows: * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

average P/E ratio of 90.74, indicating high investor expectations and potential overvaluation. On the other hand, Class 2 companies had a significantly lower average P/E ratio of 1.60 with a standard deviation of 24.04, reflecting more reasonable investor expectations and stable performance, with greater capacity to manage ESG challenges effectively.

These results showed that overall business rating and market value indicators are key determinants in predicting ESG scores for companies listed on the BELEX. Companies with higher dividend yields were more likely to have higher ESG scores, while the price-to-earnings ratio demonstrated sensitivity to economic conditions, further highlighting the importance of financial indicators in ESG predictions.

6. Practical application

The proposed hybrid learning approach applies to problems with no label data in the target domain and limited label data in the transitive domain for satisfactory machine learning training. By proposing a computationally efficient and practical model, this research offers valuable insights and several practical implications for companies, investors, and financial market players in developing markets.

Primarily, it enables firms to better align their strategies with ambitious sustainability and climate goals (the EU Green Deal and the Green Agenda for the Western Balkans). In developing markets where sustainability reporting practices are limited, ESG prediction assists companies in anticipating compliance requirements associated with sustainability reporting standards. This proactive model facilitates timely adaptation to regulatory demands embedded within existing frameworks.

Furthermore, predictive ESG analytics enable early identification of climate-related and transition risks, thereby supporting innovations and transition. By providing insights into ESG performance for companies lacking current scores, predictions facilitate the incorporation of sustainability factors into investment decision-making. This aligns capital flows with the principles for responsible investment and advances sustainable finance in emerging markets.

Finally, this approach enhances access to ESG ratings for external stakeholders. By making ESG predictions publicly available, it fosters greater transparency and accountability, enabling a broader audience–including consumers, regulators, and civil society–to better understand and actively engage with corporate sustainability initiatives.

This paper makes a significant contribution to ESG research by introducing a data-driven methodology designed to improve the reliability and consistency of ESG ratings. Given the current challenges in ESG evaluation — including methodological inconsistencies and divergent scoring criteria across rating agencies — our approach offers regulators and rating agencies an automated tool that facilitates standardized, transparent, and objective assessment of ESG scores.

7. Limitations

A key limitation of the proposed model concerns the availability of transitive-domain data. In cases where similar data constraints exist but multiple candidate transitive domains are available, a selection algorithm based on domain characteristics should be developed. From a computational perspective, several additional limitations should be acknowledged. First, the parameter-transfer procedure assumes static model parameters and does not include adaptive fine-tuning in the transitive domain, which may restrict the model's flexibility in evolving market conditions. Second, the proposed model does not incorporate explicit uncertainty estimation, which could enhance the robustness of predictions under noisy or incomplete data. Third, the absence of a formal bias and fairness assessment represents a limitation, as ESG data may embed regional or structural biases that influence model outcomes.

Furthermore, some limitations are associated with the selected feature set. Macroeconomic conditions significantly influence the ESG scores of companies in emerging markets by shaping both their external operating environment and their internal capacity for ESG integration. Future research should therefore examine the effects of key economic variables—such as GDP growth, exchange rate volatility, interest rates, and inflation—since these factors have been shown to affect ESG performance in diverse ways [2].

Additionally, ESG ratings from different agencies often exhibit systematic bias arising from variations in rating objectives, the number of input variables, and methodological frameworks. Each agency applies distinct evaluation criteria, weighting schemes, and data

sources, which leads to inconsistent assessments across providers [1, 13]. Ongoing regulatory efforts in the European Union aim to address these issues by introducing greater transparency and standardization in ESG rating activities. The forthcoming framework is expected to enhance the reliability and comparability of ESG scores, strengthen their alignment with firms' financial indicators, and support more consistent evaluations of corporate sustainability.

Nonetheless, given the complex and context-dependent relationships between business operations and ESG performance — particularly in developing markets — future research should prioritize the identification and refinement of indicators that better capture market dynamics and macroeconomic trends relevant to ESG outcomes.

8. Conclusion

Sustainability-related data is often qualitative and challenging to compare, producing inconsistent information for key stakeholders. Therefore, independent ratings have become the most widely used method for evaluating ESG performance. Companies without an external ESG score, especially in developing markets, are disadvantaged, lowering trading volume and their ability to attract financial resources. Therefore, this study proposed an approach to determining initial ESG scores in developing capital markets.

The results of this study demonstrate that MPBTL-RF is an effective and computationally efficient model for predicting ESG scores in data-constrained environments. Across five-fold in-domain evaluations, the model achieved an average MAE of 14.72 for the overall ESG score, comparable to the best-performing baseline (MAE = 14.87). In out-of-domain testing, MPBTL-RF achieved an average MAE of 17.51 on the holdout dataset and 18.43 on the synthetic dataset, showing robust generalization despite substantial domain shifts. Compared to traditional models such as RF and XGB, MPBTL-RF required significantly less training time (4.5 s vs. $>$ 26.8 s) while delivering comparable predictive accuracy.

These findings confirm that parameter transfer from a well-trained source model can enhance performance in small-sample and heterogeneous market conditions without extensive retraining. The success of this approach underscores the importance of using high-quality transitive data sets and suggests that other machine learning approaches, such as XGB, may benefit from similar approaches to improve prediction accuracy further.

Applying this model to the BELEX case study further demonstrated its practical value by generating consistent ESG estimates for frontier-market companies. This study contributes to the growing knowledge on ESG prediction in underrepresented regions and provides a foundation for future research and applications in similar markets.

The results should be interpreted as demonstrating effectiveness under data-constrained settings rather than universal superiority across domains.

Future work will extend this model through adaptive parameter tuning and the integration of macroeconomic indicators to improve prediction accuracy and interpretability across diverse economic contexts. Furthermore, future studies will explore the integration of source-free domain adaptation and unsupervised domain adaptation approaches to improve model predictions as well as the integration of transfer and active learning methods.

References

1. Berg, F., Kölbel, J.F., Rigobon, R.: Aggregate confusion: The divergence of esg ratings. Review of Finance 26(6), 1315–1344 (2022)
2. Bilivogui, P., Iqbal, M.A.: Do esg scores matter? an empirical analysis of corporate financial performance in brics economies. Environmental Research Communications (2025)
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
4. Chowdhury, M.A.F., Abdullah, M., Azad, M.A.K., Sulong, Z., Islam, M.N.: Environmental, social and governance (esg) rating prediction using machine learning approaches. Annals of Operations Research pp. 1–25 (2023)
5. D'Amato, V., D'Ecclesia, R., Levantesi, S.: Fundamental ratios as predictors of ESG scores: A machine learning approach. Decisions in Economics and Finance 44(2), 1087–1110 (2021)
6. D'Amato, V., D'Ecclesia, R., Levantesi, S.: Esg score prediction through random forest algorithm. Computational Management Science 19(2), 347–373 (2022)
7. García, F., González-Bueno, J., Guijarro, F., Oliver, J.: Forecasting the environmental, social, and governance rating of firms by using corporate financial performance variables: A rough set approach. Sustainability 12(8), 3324 (2020)
8. Huang, J., Guan, D., Xiao, A., Lu, S.: Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. Advances in neural information processing systems 34, 3635–3649 (2021)
9. Jin, Y., Acquah, M.A., Seo, M., Han, S.: Short-term electric load prediction using transfer learning with interval estimate adjustment. Energy and Buildings 258, 111846 (2022), elsevier
10. Karnyoto, A.S., Sun, C., Liu, B., Wang, X.: Transfer learning and gru-crf augmentation for covid-19 fake news detection. Computer Science and Information Systems 19(2), 639–658 (2022)
11. Kokol, P., Kokol, M., Zagoranski, S.: Machine learning on small size samples: A synthetic knowledge synthesis. Science Progress 105(1) (2022)
12. Krappel, T., Bogun, A., Borth, D.: Heterogeneous ensemble for ESG ratings prediction (2021)
13. Larcker, D.F., Pomorski, L., Tayan, B., Watts, E.M.: Esg ratings: A compass without direction. Rock Center for corporate governance at Stanford University working paper forthcoming (2022)
14. Lim, T.: Environmental, social, and governance (esg) and artificial intelligence in finance: State-of-the-art and research takeaways. Artificial Intelligence Review 57(4), 76 (2024)
15. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering 22(10), 1345–1359 (2009), iEEE
16. Pinto, G., Messina, R., Li, H., Hong, T., Piscitelli, M.S., Capozzoli, A.: Sharing is caring: An extensive analysis of parameter-based transfer learning for the prediction of building thermal dynamics. Energy and Buildings 276, 112530 (2022), elsevier
17. Sahin, E.K.: Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using xgboost, gradient boosting machine, and random forest. SN Applied Sciences 2(7), 1308 (2020)
18. Segev, N., Harel, M., Mannor, S., Crammer, K., El-Yaniv, R.: Learn on source, refine on target: A model transfer learning framework with random forests. IEEE transactions on pattern analysis and machine intelligence 39(9), 1811–1824 (2016), iEEE
19. Shim, E., Kammeraad, J.A., Xu, Z., Tewari, A., Cernak, T., Zimmerman, P.M.: Predicting reaction conditions from limited data through active transfer learning. Chemical science 13(22), 6655–6668 (2022), royal Society of Chemistry
20. Sokolov, A., Mostovoy, J., Ding, J., Seco, L.: Building machine learning systems for automated esg scoring. The Journal of Impact and ESG Investing 1(3), 39–50 (2021)

21. Sukhija, S., Krishnan, N.C.: Supervised heterogeneous feature transfer via random forests. Artificial Intelligence 268, 30–53 (2019), elsevier
22. Tan, B., Song, Y., Zhong, E., Yang, Q.: Transitive transfer learning. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1155–1164 (2015)
23. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big data 3(1), 9 (2016)
24. Wu, B., Zhang, T., Mao, L.: Large-scale image classification with multi-perspective deep transfer learning. Computer Science and Information Systems 20(2), 743–763 (2023)

APPENDIX

Table A1. Kruskal-Wallis test with a Dunn-Bonferroni post hoc analysis results for TECLP, holdout, and BELEX dataset

Indicator	Group 1	Group 2	Statistics	p-value	p-value adj.
Sales_to_Assets	BELEX	TECLP	-1.28	0.199	0.598
Sales_to_Assets	BELEX	holdout	-2.35	0.019	0.056
Sales_to_Assets	TECLP	holdout	-1.92	0.055	0.163
EBIT_to_Sales	BELEX	TECLP	4.67	0.000	0.000
EBIT_to_Sales	BELEX	holdout	4.43	0.000	0.000
EBIT_to_Sales	TECLP	holdout	2.55	0.011	0.032
DY	BELEX	TECLP	-0.909	0.364	0.999
DY	BELEX	holdout	3.88	0.000	0.000
DY	TECLP	holdout	4.61	0.000	0.000
NI_to_Sales	BELEX	TECLP	3.65	0.000	0.001
NI_to_Sales	BELEX	holdout	4.39	0.000	0.000
NI_to_Sales	TECLP	holdout	3.00	0.003	0.008
P/E	BELEX	TECLP	7.27	0.000	0.000
P/E	BELEX	holdout	0.048	0.962	0.999
P/E	TECLP	holdout	-3.41	0.001	0.002
Rating	BELEX	TECLP	-2.08	0.038	0.113
Rating	BELEX	holdout	0.316	0.752	0.999
Rating	TECLP	holdout	1.33	0.184	0.552
LR	BELEX	TECLP	-0.291	0.771	0.999
LR	BELEX	holdout	-0.989	0.323	0.968
LR	TECLP	holdout	-0.927	0.354	0.999
SR	BELEX	TECLP	-7.42	0.000	0.000
SR	BELEX	holdout	-4.76	0.000	0.000
SR	TECLP	holdout	-1.61	0.108	0.325

Note: Significance level (α) set at 0.05.

Table A2. Friedman test results over folds on transitive, holdout and synthetic datasets

Model	Test Statistic	p-value
Transitive dataset - IND evaluation		
ESG	8.687	0.069
E*	10.122	0.038
S	2.400	0.663
G	6.240	0.182
Target dataset holdout - ODD evaluation		
ESG	6.586	0.160
E	5.479	0.242
S*	16.160	0.003
G*	11.360	0.023
Target dataset Synthetic - ODD evaluation		
ESG	5.200	0.074
E	2.800	0.247
S	5.200	0.074
G*	6.400	0.048

Note: Significance level (α) set at 0.05.

Ivana Marković is an Assistant Professor in the Department of Accounting, Mathematics, and Informatics at the Faculty of Economics, University of Niš, specializing in Business Informatics. Her research focuses on machine learning methods, feature and instance selection algorithms and their applications in economics, as well as the application of advanced algorithms in optimization problems, mainly computer modelling in economics. Her work also explores transfer learning, business information systems, business intelligence, and generative AI.

Adela Ljajić is a Research Associate at the Institute for Artificial Intelligence Research and Development of Serbia. Her research spans machine learning, natural language processing, generative AI, retrieval-augmented generation, and AI evaluation. Much of her work focuses on adapting AI methods to low-resource, multilingual, and data-constrained settings, with an emphasis on reliable outputs and practical evaluation. Her broader interests include trustworthy AI, transfer learning, and real-world AI applications.

Jelena Z. Stanković is an Associate Professor at the University of Niš, Faculty of Economics, in the narrow scientific field of Accounting, Auditing, and Financial Management. Her research interests include sustainable investing, corporate financing, and risk management in finance and insurance. Her work focuses on integrating sustainable finance principles with quantitative risk assessment in corporate finance and insurance systems.

Miloš Košprdić was a Researcher at the Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, working in natural language processing (NLP) and large language models (LLMs), with a focus on semantic text search and scientific question answering. He is also a Senior Associate at the Linguistics Department of Petnica Science Center, where he previously served as Head of the Department. His work spans claim verification, natural language inference, and sentiment analysis of Serbian-language texts.

Jovica Stanković is an Assistant Professor in the Department of Accounting, Mathematics, and Informatics at the Faculty of Economics, University of Niš, specializing in Business Informatics. His research focuses on information systems, data science, data analytics, and visualization. His interests include information technology, business information systems, business intelligence, and AI.

Received: June 16, 2025; Accepted: February 28, 2026.

Computer Science and Information Systems 23(3):1001-1026 https://doi.org/10.2298/CSIS251030027G

DICYME: Dynamic Industrial Cyber Risk Modelling Based on Evidence*

Javier García-Ochoa[1], Jaime Rueda[1], Rubén R. Fernández[1], Alberto Fernández-Isabel[1], Isaac Martín de Diego[1], Emilio L. Cano[1], Romy R. Ravines[2], Ovidio López Espinosa[2] and Jaume Puigbó Sanvisens[2]

[1] Research Centre for Intelligent Information Technologies (CETINIA-DSLAB)
Rey Juan Carlos University
C/ Tulipán, s/n, 28933 Madrid, Spain
javier.garciaochoa@urjc.es (corresponding author)
{jaime.rueda, ruben.rodriguez, alberto.fernandez.isabel, isaac.martin, emilio.lopez}@urjc.es
[2] DeNexus Inc.
Boston, United States
{rr, ol, jp}@denexus.io

Abstract. The accelerated pace of digital transformation has significantly reshaped the cybersecurity domain, fostering an interconnected ecosystem in which cyber threats have expanded in both their complexity and scope. Traditional cybersecurity methods are increasingly inadequate for addressing the rapidly evolving threat landscape, emphasizing the critical need for intelligent, adaptive, and proactive defensive strategies. This study introduces Dynamic Industrial Cyber Risk Modelling Based on Evidence (DICYME), a comprehensive system that integrates diverse analytical techniques to identify patterns and characteristics that reveal emerging threat trends, enabling organizations to proactively defend against potential future attacks. Beyond threat detection, DICYME operates as a pipeline that retrieves data from diverse cyber incident reports, specialized databases, and other relevant sources of cyber-related information, applies specialized techniques for victim identification, indicator computation, threat actor profiling, Common Vulnerability and Exposure (CVE) relationship mapping, and ultimately performs the Cyber Risk Quantification (CRQ). This final stage represents the system's most distinctive contribution, as it translates complex analytical outputs into actionable risk insights, empowering organizations to make informed strategic decisions in the face of evolving cyber threats. Alternatively, the system implements an automatic workflow that constructs new datasets of compromised entities, enabling these datasets to be used by all components of the system. Experiments on real cyber incident datasets demonstrate the system's ability to automatically construct high-quality victim profiles and estimate annualized financial risk, offering a scalable and data-driven approach for proactive cybersecurity management.

Keywords: Cyber Risk Quantification, Machine Learning, Large Language Models, Indicators, Firmographics, Threat Actors, Vulnerabilities.

* This is an extended and updated version of the paper "Dynamic Industrial Cyber Risk Modelling based on Evidence (DICYME)" presented at the 10th Jornadas Nacionales de Investigación en Ciberseguridad (JNIC 2025), Zaragoza, Spain.

1. Introduction

The exponential growth of digital transformation has fundamentally reshaped the global cybersecurity landscape, creating an interconnected ecosystem in which cyber threats have evolved in both sophistication and scale. The accelerated migration of individuals and organizations toward digital environments, particularly intensified by digital transformation, has expanded the attack surface and created new vulnerabilities across multiple technological layers, including cloud computing, Internet of Things (IoT) devices, social media platforms, and cryptocurrency systems [17]. This dynamic environment has rendered traditional, static cybersecurity strategies increasingly inadequate, as they often fail to anticipate emerging threats or provide real-time insights into the evolving risk profile of an organization.

Organizations face the dual challenge of identifying potential threats and quantifying their impact in a timely manner. Conventional cyber risk assessments rely heavily on manual processes and historical data, which are often incomplete, delayed, or incompatible across sources. This results in a slow, resource-intensive evaluation that struggles to capture the rapidly changing threat landscape and the adaptive and dynamic structure of organizations. Moreover, many existing approaches treat vulnerability and threat intelligence in isolation, failing to integrate multiple data streams into a unified, actionable risk framework.

Given the constant threats that can compromise enterprises across all sectors, there is a critical need for intelligence-driven cybersecurity measures that go beyond traditional methods. Integrating Machine Learning (ML) and Artificial Intelligence (AI) methodologies [10] allows organizations to adopt proactive, adaptive strategies capable of real-time risk quantification and data-driven decision-making. By leveraging the temporal patterns of cyber incidents through techniques such as time-series analysis, it becomes possible to anticipate future attack occurrences and detect anomalies that may indicate imminent threats [13]. This combination of automation, advanced analytics, and predictive modeling provides a foundation for more efficient, accurate, and dynamic cyber risk management.

To address these limitations, this study presents DICYME, a comprehensive system that combines automatic intelligence extraction, advanced ML models, and real-time risk quantification. DICYME leverages both public and private datasets, dynamically incorporating new vulnerabilities, emerging attack techniques, and threat actor behaviors. The system translates this intelligence into quantitative indicators and probabilistic scores, which are then used to estimate the likelihood, frequency, and potential impact of cyber incidents, including both primary and secondary losses.

A key feature of DICYME is the usage of Large Language Model (LLM) to generate interpretable explanations and actionable recommendations, helping stakeholders understand the rationale behind risk assessments and mitigation strategies. Additionally, it is adaptable to different stakeholder needs, supporting interactive visualization and reporting tools suitable for technical teams, executive decision-makers, insurers, and reinsurers, providing tailored insights that facilitate informed, data-driven decisions.

By integrating real-time data extraction, predictive modeling, and dynamic risk quantification, DICYME overcomes the fragmented and static nature of traditional cyber risk assessments. It provides a unified, continuously updated view of cyber exposure, enabling organizations to proactively detect emerging threats, prioritize security investments, and

respond effectively to the evolving digital threat landscape. This integrated approach represents a significant advancement in intelligence-driven cybersecurity, bridging the gap between automated data collection, analytical modeling, and operational decision-making.

The rest of this paper is organised as follows. Section 2 provides the background and related work, covering risk analysis, actor profiling, and CVE relationships. Section 3 introduces the proposed system, including automated data extraction, risk modelling, and quantification, with support from LLMs and AI tools. Section 4 presents a series of experiments evaluating the data gathering process of the system and the CRQ model. Section 5 discusses the lessons learned, highlighting the strengths, limitations, and practical insights gained from the deployment of the system. Finally, Section 6 concludes the paper and outlines potential directions for future research.

2. Background

The cybersecurity landscape has evolved into a complex ecosystem in which cyber incidents, risk analysis, threat actors, and vulnerabilities are interconnected elements that collectively shape organizational security postures [28]. Understanding these relationships is crucial for developing comprehensive defensive strategies and predictive models [27], as organizations face increasingly sophisticated and frequent attacks that exploit the interdependencies among these elements. Recent research emphasizes the need for holistic approaches that integrate cybersecurity risk management with strategic management practices, leveraging AI and ML techniques with traditional cybersecurity frameworks [19, 27].

In recent years, cyber incidents have been steadily increasing, becoming a common problem affecting organizations across multiple domains. Simultaneously, these incidents have grown in complexity [26], compromising a wide range of companies from different sectors, including sensitive areas such as healthcare and finance. According to Hackmageddon, the number of recorded cyber incidents continues to rise yearly, reaching $4,128$ events in 2023, representing a 35% increase compared to 2022 and a markedly higher growth relative to earlier years [22].

This trend has motivated companies and institutions to develop various techniques and approaches to collect, analyze, and explain cyber incident data, as well as to model the associated risks. At present, there are numerous repositories of cyber incidents that publish detailed information about reported cases, such as the European Repository of Cyber Incidents (EuRepoC) [5] and the Cyber Events Database of the Center for International & Security Studies at Maryland (CISSM) [1]. The reporting of such incidents, which may affect any type of organization, enables the identification of potential attack trends and patterns. This, in turn, supports the anticipation of future threats and provides valuable insights for security teams, decision makers, and other relevant parties who access and analyse these repositories.

2.1. Risk analysis

Based on the foundation of incident data collected in repositories, organizations require systematic approaches to assess and quantify cyber risks. Companies and institutions often conduct cyber risk analyses [24] to identify threat actors that are likely to carry out

attacks, as well as to quantify the probability and frequency of such attacks. Based on the results of this analysis, organizations can make informed decisions about whether to invest in security, what types of security measures to prioritize, and which techniques to adopt to protect themselves against common threats that may affect their specific business sector.

By leveraging databases that host records of cyber incidents affecting companies and institutions, it is possible to conduct comprehensive risk analyses. Probabilistic and statistical analyses [24, 30] are the most commonly applied approaches for assessing these risks, and they can be combined to develop descriptive models. Furthermore, it is possible to build models that function as risk management tools, enabling organizations to reduce the likelihood of threats and their potential impact.

However, these approaches are often based on historical data and present limitations in anticipating novel or rapidly evolving threats. Another important challenge in cyber risk analysis is the limitations or absence of certain relevant data. In many cases, the available repositories or databases do not provide different coverage for specific types of risks, making it necessary to rely on external or private data sources to conduct a more comprehensive analysis. Furthermore, it is mainly a manual process, performed from time to time, requiring effort from analysts that increases cost and reduces responsiveness to emerging threats [25]. In this context, our proposal introduces a system that integrates automated data extraction with real-time risk quantification, providing continuously updated and dynamic assessments tailored to the current threat landscape.

2.2. Threat actors analysis

While risk analysis provides a quantitative framework for understanding cyber threats, identifying and characterizing the specific actors behind these threats represents another critical dimension of cybersecurity research. Threat actors are individuals or organized groups responsible for causing cyber incidents in companies [6] leading to financial losses, reputational damage, sensitive data breaches, and broader security implications. They are often identified to conduct analyses that allow organizations to understand attack patterns, predict future threats, and develop targeted defense strategies.

The identification of threat actors often lacks consensus, as disagreements may arise regarding who should be considered an attacker. In many cases, the analysis tends to prioritize quantitative aspects [31], such as counting how often different threat actors are mentioned in reports, over qualitative considerations such as comparative or contextual evaluation. This quantitative focus, while providing measurable insights, may overlook important behavioral patterns and motivational factors that threat actor activities.

Several studies and frameworks have been proposed to address this issue by collecting data from reliable sources, enabling more comprehensive analyses to detect attack patterns, identify countries that are more susceptible to specific actors, and determine which types of organizations are most frequently targeted. These frameworks contribute to a more detailed understanding of the threat landscape and support the development of targeted mitigation strategies for specific actors.

Our approach offers a systematic way to categorize threat actors using publicly available data, while remaining adaptable to incorporate proprietary intelligence feeds from providers such as CrowdStrike or FireEye. Importantly, the relevance and impact of a

specific threat actor are contextual and victim-dependent: an actor predominantly targeting organizations in the United States or Canada may pose minimal risk to a company in Madagascar, but significant risk to a U.S.-based critical infrastructure operator. By integrating geographic, sectoral, and actor-specific prevalence data, our method provides dynamic, context-aware threat assessments that reflect both the actor's capabilities and the characteristics of the potential victim.

2.3. CVE relationship

Complementing the understanding of threat actors and their behaviours, vulnerability analysis provides a technical foundation to understand how attacks are executed and systems are compromised. CVEs represent a standardised framework for cataloguing publicly known security flaws that can affect organisations across all sectors [32]. When successfully exploited, these vulnerabilities enable threat actors to compromise systems, resulting in unauthorised data access, operational disruptions, and significant financial impact.

The association of vulnerabilities with affected systems is based on structured mapping processes that link CVEs to standardised repositories. Each vulnerability is assigned to the National Vulnerability Database (NVD), which supplements basic CVE information with impact assessments, references to vulnerable products via Common Platform Enumeration (CPE), and classification of the underlying weaknesses through Common Weakness Enumeration (CWE) [35]. This structured approach enables consistent vulnerability tracking and supports the integration of threat intelligence into organisational risk frameworks.

Recent advances in vulnerability relationship modelling have emphasised graph-based representations to capture complex interdependencies. Knowledge graph approaches [32, 37] enable the discovery of non-obvious connections between vulnerabilities, affected products, and exploitation patterns, supporting more accurate threat predictions. ML techniques, particularly graph neural networks [2, 38], have demonstrated substantial improvements in identifying vulnerable code patterns and predicting missing relationships within vulnerability databases. New research has extended these methods by incorporating Natural Language Processing and reasoning capabilities from LLMs to extract and integrate information from unstructured threat intelligence sources [16].

Despite these advances, vulnerability databases continue to face fundamental limitations. Significant processing delays result in incomplete metadata for newly disclosed vulnerabilities, while inconsistent weakness classifications reduce the reliability of automated analysis tools. Studies have documented substantial backlogs in vulnerability analysis and highlighted discrepancies in how CVEs are assessed and used in security research. Addressing these challenges requires continued development of automated enrichment techniques, cross-source validation mechanisms, and predictive models capable of operating effectively with incomplete or evolving vulnerability information.

3. Proposal

In this section, we present the proposed system, called DICYME, which focuses on automating cyber risk quantification and management in operational technology (OT) environments. The main objective of this system is to design and implement a system capable

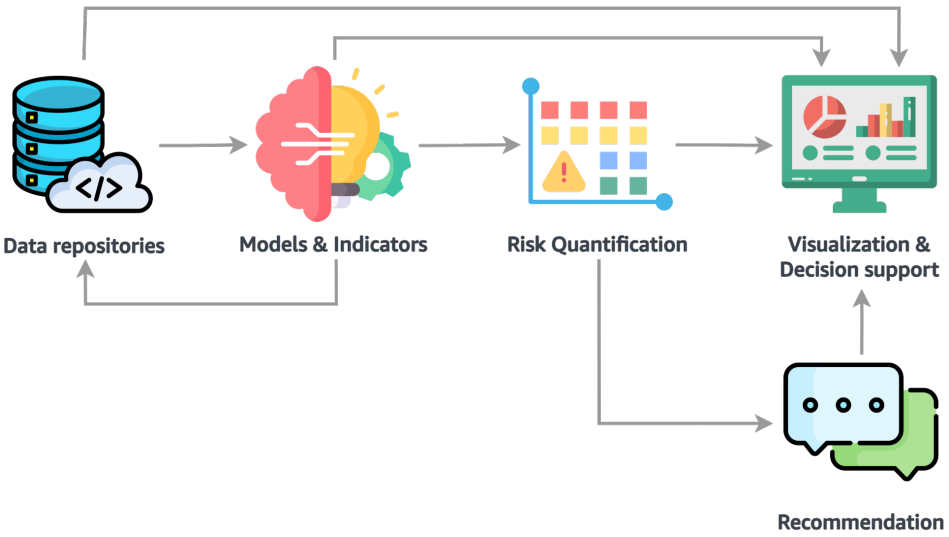


Fig. 1. High-level architecture of the DICYME complete system.

of addressing the current challenges in modelling cyber risks within industrial infrastructure. The proposed solution relies on evidence-based analysis, integrating data from multiple sources to dynamically estimate the probability and impact of potential cyber incidents. To achieve it, the system incorporates a set of cyber risk indicators, which together with the data contribute to the simulation and quantification model. In addition, LLM to support explainability and enhance user understanding while providing recommendations in natural language. Finally, all components are brought together within a visualization platform that delivers a continuous, clear, and actionable view of cyber risk levels, designed to support informed decision-making in critical environments.

Figure 1 illustrates the architecture of the system, which is structured into five main parts. Automated data extraction gathers and preprocesses heterogeneous information from both external and internal sources, such as threat intelligence feeds and internal security telemetry. This raw evidence feeds the indicators block, where relevant cyber risk metrics are computed and normalised to reflect dynamic aspects such as visibility, reputation, or the threat actors landscape. These indicators are then aggregated in the cyber risk quantification module, which applies models to estimate the probability and impact of incidents. The outcomes, together with the underlying models and source data, are presented in the visualization and decision support layer, providing interpretable dashboards and analytical tools that allow stakeholders to explore and understand risk. Finally, the recommender component not only guides users in interpreting the visualized data and models, but also actively supports decision-making by proposing adjustments to the quantification process, suggesting alternative data inputs, and executing the risk models with multiple scenarios to provide actionable and feasible recommendations.

In addition, Figure 2 provides a more detailed view of the system, explicitly mapping all the elements that contribute to CRQ and the relationships among them. The figure distinguishes between the different data repositories, the models and indicators built on

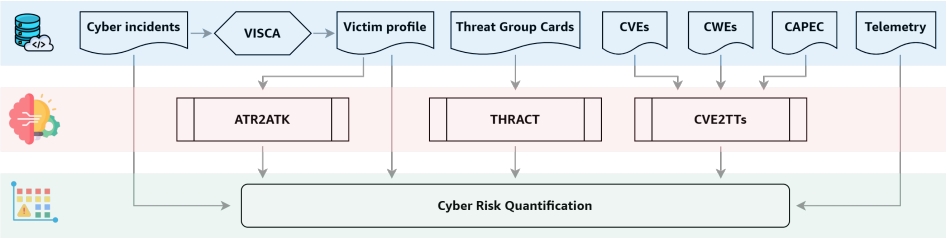


Fig. 2. Structural relationships within the Cyber Risk Quantification system.

top of them, and the final CRQ process. This separation highlights how raw evidence is progressively transformed into structured knowledge and then into quantifiable indicators of risk, while also allowing certain datasets to be used directly in the quantification process when relevant, for instance to identify trends, contextual factors.

3.1. Automated data extraction

The automated data extraction module was designed to systematically collect and consolidate heterogeneous cyber risk information from multiple open sources. Specifically, it integrates reports of cyber incidents from six publicly available databases to provide a structured view of the threat landscape. To complement incident-level data, the module also incorporates victim profiling information, both with manual identification which is then empowered through Victim Insight System for Cyber Attacks (VISCA), a multi-agent LLM-based architecture that identified affected organizations and retrieves and processes its firmographic attributes.

In addition to incident and victim data, the module gathers technical Cyber Threat Intelligence (CTI) such as up-to-date CVEs with their associated vulnerability types, Tactics, Techniques and Procedures (TTPs) from the MITRE ATT&CK framework, as well as structured attack patterns from Common Attack Pattern Enumeration and Classification (CAPEC). Furthermore, it integrates information on threat actors from publicly available Electronic Transactions Development Agency (ETDA) Threat Group Cards [4], linking adversarial groups to their targeted victims and contextual attributes. It also incorporates internal telemetry data, capturing Operational Technology (OT)-specific information such as asset counts, exposed vulnerabilities, and network-level observations across Purdue layers. By consolidating organizational, incident-specific, adversarial, and technical dimensions, the module creates a unified knowledge base that supports the advanced modelling of cyber risks in subsequent stages.

Cyber incidents Cyber incident data were initially extracted from publicly accessible and free sources, including the TI Safe Incident Hub [34], Industrial Control System Security, Threats, Regulations, Incidents, Vulnerabilities provided by Experts (ICS STRIVE) [9], KonBriefing [12], Cyber Events Database of the CISSM, Hackmageddon [23] and EuRepoC. The extracted records were stored along with the extraction date and the source of the origin. Subsequently, the data underwent transformation and cleaning processes to achieve a more standardized format suitable for analysis. The transformation primarily

consists of restructuring the raw data into a tabular form, where each row represents a cyber incident and each column corresponds to an attribute of that incident. Data cleaning is performed to improve data quality by removing invalid entries or correcting formats, such as normalizing date representations. Additionally, cleaning involves harmonizing categorical values, for example, standardizing the country of the victim or the type of attack.

Finally, the sources were converted into a uniform tabular format to enable cross-source comparison of incidents, support exploratory data analysis across all repositories, facilitate incident deduplication, and allow for higher-level transformations.

Victim profile The Victim profile dataset was collected in the DICYME system to support the development of the Attractiveness concept (see Section 3.2), which delves with the possession of properties or behaviours in entities that may capture the focus of potential adversaries [7, 8]. The dataset includes uniquely identifiable entities that have been reported as victims of confirmed cyber incidents from the Cyber incidents dataset (see Section 3.1), manually selected and validated by a threat intelligence analyst. In addition, similar entities by country and industry category without known incidents were included to provide more context for a comparative analysis.

For each entity, the collected information can be organised into three groups that correspond to the main components of the Attractiveness model. The first group, basal attractiveness, refers to inherent static firmographic features, which includes the country, sector category, revenue, earnings, publicly traded status, number of employees, and profitability. The second group, online reputation, captures dynamic factors linked to human-generated content, social media interactions, and shared experiences related to the entity. It is computed through engagement (E), defined as the ratio of interaction (I) to reach (R) [3], where reach indicates the number of individuals who view or mention a publication and interaction represents those who actively respond to it. Engagement is assessed by distinguishing whether the content originates from the entity itself (assumed predominantly positive) or from external users, where sentiment analysis is performed to capture positive, neutral, or negative perceptions [7]. The third group, potential victimisation, represents another dynamic factor and includes the number of appearances of the entity name in dark web leaks and the number of visible devices connected to the Internet. These three groups are combined to produce a final structured dataset, containing a large amount of information, including features that are complex to extract, all referenced to each entity.

Victim Insight System for Cyber Attacks (VISCA) The VISCA model is an agent-based framework developed within the DICYME system to improve and extend the Victim profile dataset by automating and scaling the identification of victim entities and the fusion and normalization of heterogeneous data. This model takes the description of a cyber incident in natural language and collects firmographic data, such as its country, industry category, age, annual revenue, and number of employees. The proposed model is integrated into the system to generate a larger Victim profile dataset and also to obtain firmographic data that users can use in the CRQ model. The full workflow of this model is illustrated in Figure 3.

- Retrieve basic information about the victim: using a cyber incident as a data source, a LLM is capable of extracting relevant information from the incident description such

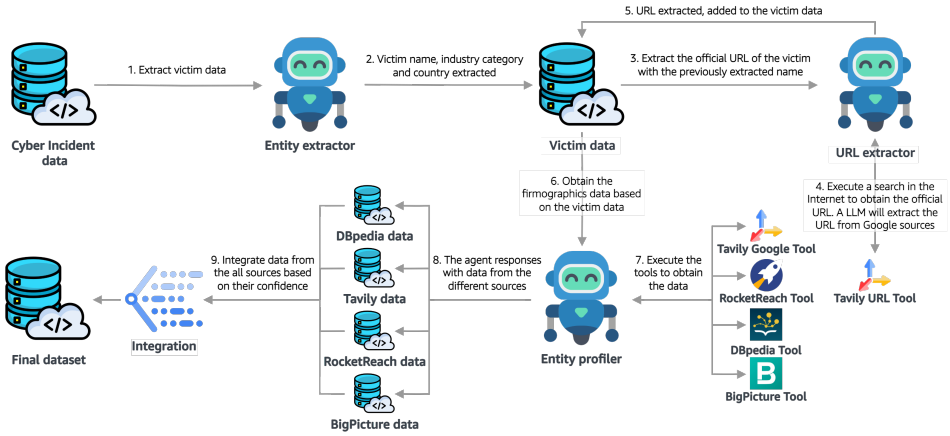


Fig. 3. Multi-agent architecture proposed for extracting victim information and completing missing firmographic data.

as the name of the entity victim, the country and the industry. Additionally, the model queries various data sources from different websites on the Internet to retrieve the official website Uniform Resource Locator (URL). This basic information is used in subsequent processes to obtain or filter the firmographic data.

– Dig up firmographic information: with the basic information of the victim extracted, the model proceeds to perform queries across various external data sources to retrieve additional details. The information to be collected consists of eight distinct fields encompassing different data types: name of the victim, industry category, founded year, number of employees, country, revenue, publicly traded status, and North American Industry Classification System (NAICS) codes. The model requests information about the victim entity from four main sources: Google, DBpedia, BigPicture, and RocketReach. DBpedia and BigPicture are open-access sources available to any user seeking information on companies. In contrast, RocketReach is a paid platform that provides data with a higher level of reliability, as it is a more private and commercially maintained product. Google information is retrieved through a tool called Tavily, a platform that enables agent-based AI models to connect to the web. Tavily collects data from multiple Google sources while returning a filtered subset of relevant results and selecting only a reduced number of sources.
– Combine the extracted data: once the information has been extracted from the data sources, the final step performed by the model is the aggregation of the data to generate a consolidated set that captures the most complete information possible. To achieve this, a confidence score was assigned to each dataset retrieved from the sources based on the number of fields successfully extracted. The source with the highest confidence score was selected. In cases where certain firmographic fields are missing, the model attempts to replace them by merging data from other sources where their confidence scores exceed a defined threshold.

Once the final agent gathers all datasets from various sources, the goal is to consolidate them into a single dataset that maximizes data completeness. In particular, the final dataset

should include the victim's name, category, NAICS code, founded year, size, country, revenue, and public trading status. In addition to these fields, a confidence score was assigned to each source, represented as a numerical value between 0 and 1. This score reflects the reliability of the extracted information, completeness of the retrieved data, and number of relevant results obtained during the search process, rather than data quality. A value close to 0 indicates highly unreliable information, whereas a value of 1 indicates high reliability.

The value of the estimation is computed as shown in Equation 1 for each source s.

$$c\_fields_s = \frac{m_s}{f} \qquad c\_responses_s = \begin{cases} \frac{1}{n_s} & \text{if } n_s > 0 \\ 0 & \text{if } n_s = 0 \end{cases}$$
$$conf_s = \frac{1}{r_s} \cdot \Big[\alpha \cdot c\_fields_s + (1-\alpha) \cdot c\_responses_s\Big], \tag{1}$$

where the confidence score, $conf_s$, quantifies the reliability of a specific source. This confidence depends entirely on the completeness of the extracted data, the number of responses obtained, and the efficiency of the data retrieval process. The first component c_fields_s measures the number of fields obtained m_s relative to the total number of fields to be extracted f, that is, its proportion. The second component, $c\_responses_s$ serves as a confidence penalty when a source returns a large number of potential results, n_s, indicating uncertainty regarding the most accurate outcome; the confidence value is correspondingly reduced. The overall confidence score, $conf_s$, integrates these two components through a weighted average controlled by a tunable parameter, α, and is adjusted according to the type of source selected. For instance, in the case of $s =$ RocketReach a higher value is used to prevent the confidence from being penalized, as this source typically returns a large number of responses. The number of attempts, r_s, required to obtain a valid result increases with each failed query and always starts at one.

Queries are executed using a list of potential victims generated by the *Entity extractor* agent. This list contains extensive terminology referring to the victim, including variations such as corporations, brands, or subsidiaries, ordered from the most likely to the least likely, based on the LLM. Each agent sequentially queries the names until relevant information is obtained. Every failed attempt increments r, which reduces the final confidence score to reflect the inefficiency of the search. There is always an initial attempt; therefore, r_s can never be 0.

When the confidence $conf_s$ is computed for each source s, the dataset with the highest overall confidence is selected. If one or more fields are missing in the selected dataset, the system evaluates those specific fields across all sources with confidence scores higher than a threshold of 0.5 and aggregates the data.

Electronic Transactions Development Agency (ETDA) Threat Group Cards The ETDA is a Thai public organization that focuses on promoting secure and efficient digital transactions. Among its key initiatives, ETDA develops detailed Threat Group Cards to support CTI efforts. These profiles integrate data from leading global sources, including the Malware Information Sharing Platform (MISP) Threat Actors Galaxy, MITRE ATT&CK Framework, Malpedia, Open Threat Exchange (OTX), and ETDA's own CTI

archive and open-source research, offering a comprehensive view of threat actors targeting digital infrastructure.

Data acquisition was conducted through direct downloads of individual Threat Group Cards and a consolidated JavaScript Object Notation (JSON) file, both available on the official ETDA website. The consolidated file aggregates all documented actors and is designed for seamless integration with platforms such as MISP. Taken together, these two resources provide a rich set of attributes for each threat group, including their countries of operation, targeted industries, stated or inferred motivations, attributed campaigns and their timelines, as well as the tools and techniques employed.

Another important data source integrated into the system consists of cybersecurity vulnerabilities and related attack patterns, which are then used and completed by the CVE2TTs model (see Section 3.2). The process begins by collecting vulnerability data from the NVD, which includes CVE identifiers, descriptions, and associated CWE mappings. These CWEs are linked to CAPEC entries by MITRE, which in turn connect to MITRE ATT&CK techniques. While not every CVE has a direct mapping along this full chain, available mappings are used when possible, and gaps, particularly in the Industrial Control Systems (ICS) matrix, are predicted by the CVE2TTs model.

Internal telemetry data In addition to external sources, the system also incorporates internal data. However, this component is more limited, as obtaining this data from industrial organizations requires explicit authorization and faces significant operational and confidentiality constraints. Nevertheless, the system integrates internal telemetry whenever available, primarily consisting of anonymized samples of Intrusion Detection Systems (IDSs) data provided by DeNexus. It also includes detailed visibility into the number of connected assets, their vulnerabilities, and their distribution according to the Purdue model. This information is particularly valuable for risk quantification, as it enables the identification of exposed assets and weaknesses directly linked to the operational infrastructure. From this telemetry, the system extracts CVEs, which are then related to the CVE2TTs model (see Section 3.2). When such data are not accessible, users can alternatively upload a list of CVEs associated with the entity, ensuring that the quantification process remains consistent and applicable. By combining OT-specific IDS discovery capabilities with flexible input options, the platform strengthens both the accuracy and relevance of cyber risk assessment while maintaining confidentiality.

3.2. Indicators

To quantify cyber risk effectively, a set of indicators has been defined to capture distinct patterns and characteristics of the data, each serving a specific purpose depending on the indicator. Specially, in the DICYME project there are three main indicators: ATR2ATK, THRACT, and CVE2TTs.

ATR2ATK ATR2ATK or Attractiveness is an indicator for forecasting cyber incidents by assessing the proneness of an entity to them. It is decomposed into three main branches: basal attractiveness, online reputation, and potential victimisation.

Basal attractiveness focuses on inherent static characteristics, often referred to as firmographic data, such as location, operational criticality, data sensitivity, and size. These

attributes make certain entities more appealing targets for adversaries because of their intrinsic nature. It is modelled with association rule mining, specifically the FP-Growth algorithm, combined with a Decision Tree classifier that uses its output (amount of satisfied rules and aggregated support, lift and confidence per incident) [8, 7].

Online Reputation measures the presence of the entity in social networks and social media, as it may be more attractive to an adversary based on what users, customers, or employees write, communicate, and share anywhere on the Internet based on their perceptions and experience at any moment of their relationship, direct or indirect, with the entity [20, 21, 33]. It is a dynamic component and time-dependent, as it represents a specific moment, influenced by previous periods. Its computation relies on engagement (E), defined as the ratio of interaction (I) and reach (R) [3]. Reach indicates the number of individuals who view or mention a publication, whereas interaction (I) represents those who actively respond to it. In this context, engagement is assessed by considering whether the content originates from the entity itself through its official sources, where sentiment is assumed to be predominantly positive, or from other users, where additional sentiment analysis is taken into account since perceptions can be positive, neutral, or negative.

Lastly, potential victimisation, which is also a dynamic component, has to be with the visibility and technical knowledge that the adversary can have about the entity. Its rationale lies in the idea that organisations become more likely targets when they gain visibility in underground forums, leak sites, or other malicious communities, and when technical indicators reveal that their infrastructure may be easy to compromise. To capture this, two variables are monitored: the number of mentions in dark web leaks and the number of publicly visible assets connected to the Internet. These are combined through a buffer method that assigns a value between 0 and 2: 0 if neither variable is present, 1 if only one of them is, and 2 if both are detected. This provides a simple yet effective measure of potential victimisation.

THRACT This indicator is a composite metric designed to provide an overview of threat actor activity targeting a specific country and industry. Based on the data extracted from the ETDA, three partial metrics are derived to reflect their activity, capabilities, and objectives with respect to a specific victim. Activity is captured through the time elapsed since the actor was last observed. As the ETDA database is updated on a rolling basis, this value may change over time, reflecting updates in observed behavior. Actor capabilities are represented based on expert annotations of operational capacity. Victim-related features include country and industry, while actor motivation is also considered. Because the metric depends on the country and industry of the potential victim, it is dynamic rather than a fixed score for each actor in any situation, providing a context-sensitive measure of potential cyber risk.

CVE2TTs Finally, the CVE2TTs indicator corresponds to a ML model that maps CVE entries to specific Techniques (and, by extension, Tactics) within the MITRE ATT&CK framework. This mapping is performed by leveraging the textual description of each CVE together with additional attributes such as vulnerability type, CWE, CAPEC, and attack vector [29]. This concept is very important in the cybersecurity sector because it helps understand the practical exploitation of vulnerabilities in a structured manner, linking specific vulnerabilities to known adversary behaviors TTPs. This model is crucial for risk

management, CTI, and incident response, as it bridges the gap between vulnerabilities and the manner in which attackers might exploit them.

3.3. Cyber Risk Quantification (CRQ) model

In the context of cyber risk management, the Cyber Risk Quantification model introduced in this proposal enables a simulation-based tool that quantifies cyber risk in financial terms, supporting informed decision-making for organizations. Using stochastic simulations, the model estimates the Annual Expected Loss for a given organization by combining the frequency of successful attacks and their financial magnitude or impact, as shown in Equation 2.

$$\text{Cyber risk} = \text{Loss Event Frequency (LEF)} \times \text{Loss Magnitude (LM)} \tag{2}$$

The model relies on a state-of-the-art risk tree that separates probability and impact into two main branches. The first component, the Loss Event Frequency (LEF), represents the expected number of loss events within a given period, typically one year. In this context, loss events correspond to materialized incidents that cause harm to the organization. The LEF is determined as the product of the Threat Event Frequency (TEF), which estimates the number of attempted attacks that the organization is likely to face, and the Susceptibility, which reflects the probability that those attempts will successfully translate into incidents. This relationship can be expressed as shown in Equation 3.

$$\text{Loss Event Frequency (LEF)} = \text{Threat Event Frequency (TEF)} \times \text{Susceptibility} \tag{3}$$

This formulation integrates both the external pressure from the threat landscape (captured by the TEF), and the internal defensive posture of the organization (captured by Susceptibility). The TEF takes into account a baseline number of incidents based on private cyber intelligence knowledge, the ATR2ATK indicator, and the annualized rate of incidents for the country and industry category of the entity, sourced from the Cyber incident dataset. Susceptibility is derived from the vulnerabilities obtained through internal telemetry or from a list of vulnerabilities uploaded by the user, in combination with the CVE2TTs model, the Threat Actor Index, which captures the activity and focus of the 245 actors in the database over a three-year period, and an adjustable security profile index. The latter characterizes the defensive posture of the organization by incorporating features such as organizational size, number of unpatched vulnerabilities, and exposure of devices visible from the internet.

The second component of the risk tree is the Loss Magnitude (LM), which quantifies the financial impact of a successful attack as the combination of primary and secondary losses, as stated in Equation 4. The identification of feasible primary and secondary losses depends on the attack technique k, which is constrained by the vulnerabilities observed or uploaded by the user, linked to the CVE2TTs mapping. In practice, just the achievable techniques from the Impact MITRE ATT&CK tactic, given the vulnerability landscape of the entity under analysis, are considered in the simulation. This ensures that both primary and secondary losses are not only probabilistically derived, but also grounded in the

technical conditions that define the exposure of the organization. Expert financial knowledge is further incorporated to parametrize and calibrate the estimation of losses to reflect realistic economic consequences for the entity.

$$\text{Loss Magnitude (LM)}_k = \text{Primary Loss}_k + \sum \text{Secondary Losses}_k \tag{4}$$

Primary losses represent the direct effect of an incident and include categories such as business interruption, equipment damage, extortion, or human impact. In each simulation run, only one primary loss type is selected, corresponding to the most plausible impact mechanism for the attack technique under consideration. By contrast, secondary losses capture indirect consequences, such as forensic investigation costs, reputational damage, or regulatory penalties. Unlike primary losses, several secondary losses can occur simultaneously, and their combined effect is represented by summing the estimated financial amounts for each of them.

3.4. Visualization and decision support

The visualization and decision support component is implemented as an interactive application that allows users to explore datasets and execute risk models through two main options: either using the entities defined in the Victim profile dataset or by providing their own model inputs. The platform provides a wide range of interactive visualizations, including bar charts, histograms, stacked and unstacked time series, heatmaps, and world maps. Figure 4 shows some of this techniques used inside the CVE2TTs module, reflecting the relations made between CVEs and MITRE ATT&CK Tactics and Techniques by vulnerability type and by Common Vulnerability Scoring System (CVSS), as well as the number of vulnerabilities and predicted techniques per year. Users can filter the data by relevant variables such as dates, incident databases, CVE types, and other context-specific metrics. This interactivity enables stakeholders to examine patterns in risk indicators, compare scenarios, and gain a deeper understanding of the underlying data and model outputs.

In addition, the system generates a downloadable report that compiles all the data, the computed indicators, and details on the steps of the quantification process. This report can be shared with other stakeholders, including expert panels, managers, or decision-makers, ensuring that risk analyses can be reviewed, discussed, and extended beyond the interactive platform. By combining interactive exploration with exportable documentation, the system supports both immediate, hands-on decision-making and broader strategic planning across multiple levels of an organization.

3.5. Recommender model

In addition, the system incorporates a set of recommender models that complement its core functionality. These models are designed to identify patterns within the data and indicators, extract meaningful features, and provide tailored recommendations. Although is not essential for the system baseline operation, this component enhances its analytical capabilities and offers added value by supporting more informed decision-making. The models employed to perform these tasks were fine-tuned through prompt engineering [11] and Retrieval Augmented Generation (RAG) techniques [15] to produce responses

Fig. 4. Example of visualization techniques used in the CVE2TTs module of the DICYME system. The top heatmap displays the distribution of CVEs identified for each ICS tactic by vulnerability type. The middle bar chart illustrates the temporal evolution of security data, showing the total annual count of CVEs (blue) alongside the volume of predicted Enterprise-specific techniques (orange) and ICS-specific techniques (yellow) associated with those vulnerabilities. Finally, the bottom pyramid chart presents the average CVSS scores per tactic, calculated from the CVEs associated with each tactic according to the CVE2TTs model's predictions. The chart is bifurcated by domain: the Enterprise matrix tactics are positioned on the left, while ICS matrix tactics are on the right. Within this chart, blue bars indicate average scores according to the CVSS v2 standard, and orange bars represent the CVSS v3 scoring.

that are as accurate as possible. By adapting the prompts for each system component with their corresponding data, a specialized prompt is generated for that specific component to ensure a more accurate and context-aware response.

In some cases, the data used in the prompt processed by the model may lead to undesired responses. This occurs because the system generates large volumes of data that may pose challenges for the model when generating an analysis, since these models are subject to token limitations. To address this issue, query decomposition [36] is applied in situations where sections contain extensive data. The queries were divided, allowing multiple interactions with the model to provide more specific explanations. This approach made it possible to explain similar datasets without resorting to generalization. Finally,

the responses to these queries were consolidated by the LLM using a prompt designed to summarize all partial outputs and deliver a comprehensive final explanation.

In addition to the models that provide data explainability, decision support capabilities, and parameter adjustment, the system also incorporates a more sophisticated model implemented as a chatbot assistant. The assistant enhances user interactivity by enabling the execution of the same tasks performed by other models with the added flexibility of allowing users to request specific adjustments or obtain tailored explanations through natural language queries. The aim of enhancing user interaction, explainability, and adjustments is achieved through the assistant, which is integrated into one of the most important and critical components of the system: the CRQ simulation. Through an intuitive interface, users can submit questions related to the CRQ, which the AI assistant answers one at a time in a conversational manner. The model has a limited memory capacity, allowing it to retain awareness of previous prompts and generate responses within a specific window. Thanks to this memory, the assistant can access prior questions and answers, compare its outputs, retrieve the results of previously executed simulations, and revisit earlier recommendations. This enables the contrast of past outcomes with newly generated ones, thereby supporting more consistent analyses and more informed decision-making.

The architecture implemented for the GenAI models is inspired by the approach proposed by DeepSeek-V3 [14], which is based on a Mixture of Experts (MoE) model. This architecture uses a set of specialized experts, each responsible for handling specific tasks depending on the context. The assistant is capable of handling three distinct tasks, each managed by a dedicated agent within the architecture: answering general questions related to CRQ process, providing recommendations on parameter adjustments, and rerunning the simulation to generate a comparison with the previous simulation results.

4. Experiments

The experiments presented in this study evaluated two key aspects of the proposed system. First, we assess the data extraction and integration capabilities of the multi-agent architecture, focusing on the completeness and quality of the firmographic information collected from multiple cyber-incident databases.

Second, we evaluated the entire risk quantification workflow, from the computation of different intermediate indicators and metrics present in the system to the computation of risks, probabilities, frequencies, and other risk parameters. For this purpose, a simulation will be conducted using the model with a case study and real data from a possible victim entity or institution.

The model selected for the multi-agent architecture, which extracts the victim data, is the Meta LLaMa 4 Scout [18], with a total of 17B active parameters and the advantage of being open-weight, which allows us to maximize utility. This model was chosen over other available alternatives because, in addition to its ease of deployment, it supports the execution of agents and tools through libraries employed in this project.

4.1. Data gathering

Within the DICYME application, a wide range of datasets are employed across different tabs to present information using tables, charts, and other visualization components. As

previously discussed, the data were extracted from various sources and organized into three main datasets used in different modules of the application: Cyber incidents, Victim profile, and IDSs.

This experiment evaluates the firmographic data completion of the VISCA module by comparing the dataset it generates against the analyst-annotated Victim profile dataset. After collecting information from victim entities affected by cyber incidents, the data are stored in a dataset to be used within the system. The objective of this experiment was to compare the automatically extracted dataset with a similar dataset in which the victim entity data were manually collected by a human annotator.

Conversely, the victim dataset used to evaluate the extracted dataset was a fully handcrafted dataset compiled by a human annotator. It contains 675 instances of distinct victim entities and includes a substantial amount of information related to each company, specifically firmographic data. In particular, the dataset contains 9 variables, including the incident category, entity, industry category, country, earnings, employees, revenue, profitable and publicly traded. This dataset is used within the system to compute various indicators and represent key information about each victim entity, which explains the large number of variables included, as they serve as inputs for these indicators.

In contrast, the dataset under evaluation, obtained after the extraction and aggregation of firmographic data, was automatically constructed by a multi-agent system capable of executing a workflow starting from a cyber incident. Once the agent extracts data from multiple data sources, it integrates them into a final output to produce a consolidated result. The final output contains eight firmographic fields, including: entity, category, country, revenue, founded year, employees, publicly traded, and NAICS code. This dataset contains 1,658 observations of distinct victim entities processed from cyber incidents collected from the ICS STRIVE, EuRepoC, and TI Safe databases. In all the processed cyber incidents, at least one victim entity was explicitly identified in the incident description, which enabled the extraction of the corresponding firmographic data.

Table 1 presents a comparison between the two datasets used to model a profile for a victim entity. It can be observed that the VISCA-extracted dataset contains a significantly larger number of identified victim entities. This is one of the most notable characteristics of the model, as its autonomous workflow can continue processing incidents until all available incidents are processed, while also achieving a much higher processing speed than manual annotation can. This difference in gathering speed is evident in Table 1, which includes the range of extraction dates for the data collected. While the Victim profile dataset required an entire year to compile all the victim entities it contained, the VISCA model was able to extract more than double the number of victim entity records in just five months of uninterrupted execution.

However, one limitation of this dataset is the smaller number of variables compared to the Victim profile dataset. This difference is due to the specific original goal of the VISCA model: to generate victim-related data for cyber risk quantification using the CRQ framework. In contrast, the Victim profile dataset was developed to support multiple components of the system, including the computation of indicators, metric computation, and structured information representation. Nevertheless, the VISCA model can be extended to collect additional firmographic attributes following the same automated process, allowing the resulting dataset to become more comprehensive and usable across a wider range of system components.

Table 1. Comparison between the Victim profile dataset and the VISCA-extracted dataset.

Feature / Metric	Victim profile dataset	VISCA-generated dataset
Unique entities	675	1,658
Number of fields	9	8
Sources used	Manual collection	ICS STRIVE, EuRepoC, TI Safe
Extraction method	Manual annotation by an analyst	Multi-agent automated workflow
Extraction dates	December 2023 - December 2024	May 2025 - September 2025

Table 2. Summary of the values obtained from the CRQ simulation, including both the computed indicators and the model output parameters.

Indicators	Baseline	85 events/year
	Incident rate	50%
	Attractiveness	80%
	Threat actor index	58%
	Exposure	41.61
	Security profile	67%
CRQ Outputs	Threat Event Frequency (TEF)	34 events/year
	Susceptibility	2%
	Loss Event Frequency (LEF)	15 events/year
	Primary loss	€49,512
	Secondary loss	€20,786,315
	Loss Magnitude (LM)	€20,835,828/event
	Cyber risk	€14,168,363/year

4.2. Cyber Risk Quantification (CRQ)

To evaluate the effectiveness of the CRQ procedure, we conducted a series of experiments that simulate real-world scenarios using actual data from multiple companies. These experiments were designed with two main objectives: first, to demonstrate the internal performance of the CRQ process, including the types of input data and datasets that can be leveraged within this framework, and second, to validate the output risk metrics generated by the model, which serve as a foundation for informed decision-making by organizations.

Executing the CRQ workflow of the system requires the input of various firmographic attributes that characterize a company. On the one hand, firmographic data from real companies can be loaded from the Victim profile dataset, specifically including country, industry category, revenue, earnings, number of employees, publicly traded status, profitability, online reputation, publicly visible devices, and critical information leaks. However, these fields can be populated directly with the data of the company provided by the user operating the system. In addition, other simulation parameters are configurable, such as the random seed, number of simulations, and insurance parameters. Finally, a list of unpatched CVEs present in the infrastructure of the organization, which may be exploited, must be provided, as they serve as a basis for estimating the corresponding risk measures, as stated in Section 3.3.

For the experiments, a real-world case was selected, corresponding to an actual company extracted from our dataset. The input values for the model are the previously described firmographic attributes of the company, all of which are available in the dataset. Specifically, the University of Michigan was selected for this experiment. This public institution was selected because it is recognized as one of the leading universities worldwide and is located in the United States, a country that consistently reports the highest concentration of cyberattacks. This combination of global academic relevance and geographical context makes the institution particularly susceptible to being frequently targeted by threat actors.

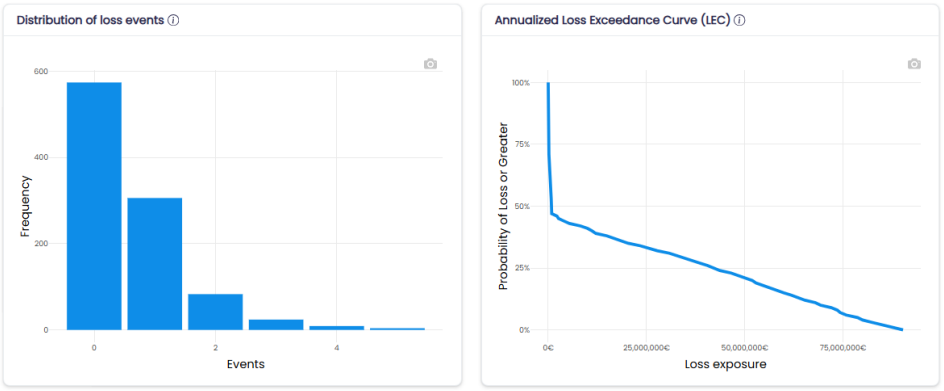


Fig. 5. Distribution of Simulated Loss Events and Corresponding Loss Exceedance Curve in the CRQ simulation

This public institution, located in the United States and dedicated to education, reported a revenue of €9.1 billion, employed a total of 34,600 staff members, and had three publicly visible devices and one critical information leak. Using these publicly available data from the university, CRQ simulations were executed and complemented with additional configuration parameters, including the number of runs, random seed, and insurance settings. The CRQ procedure first computes a series of indicators that are subsequently used to derive the output of the model parameters. The values of all the indicators are listed in Table 2. Among them, the attractiveness score is 80%, a significantly high value, primarily driven by the institution's location in the United States and its educational industry category. Another key indicator is the Threat Actor Index, which reached a value of 58%, reflecting the intensity and capabilities of threat actors over a three-year period and corresponding to a moderate risk level.

Based on the computed values derived from the publicly available data of the university and complemented with additional simulation parameters, such as the number of runs, a random seed to ensure reproducibility, and cyber insurance financial values, the model produces simulation outputs presented in both graphical and numerical forms. Figure 5 presents the results of the simulated risk through two charts: the first describes the distribution of loss events across all simulated years (each simulation means a year), while the second illustrates the Loss Exceedance Curve (LEC), which illustrates the probability (Y-axis) that cyber risk losses will be equal to or greater than a given amount (X-axis).

In the first chart, which represents the distribution of annual loss events, the results show a left-biased pattern. The majority of simulated years (574 out of 1,000) resulted in no materialized incidents, while 306 years included a single loss event. Multiple events within the same year were comparatively rare: 83 simulations produced two events, 24 produced three, 9 produced four, and only 4 reached five events. This highlights that, although the occurrence of cyber loss events is possible, their frequency per year remains predominantly low, consistent with the heavy-tailed nature of cyber risk distributions.

The second chart, which illustrates the distribution of financial losses (LEC), reveals two distinct components. At the lower end, there is a steep initial drop, corresponding to

rare but high-impact scenarios—so-called black swan events. Beyond this abrupt decline, the curve follows an almost linear descent, decreasing steadily from a 47% probability of incurring losses equal to or greater than €8.9×10^5 down to 0% probability at €9.0×10^7. Unlike sharply convex loss distributions, this profile suggests a smoother, more gradual reduction in probability across a wide range of loss magnitudes, balancing rare catastrophic scenarios with a persistent probability of moderate to high losses.

Finally, the CRQ model also provides single averaged values, complementing the previously described distributions and probabilities, to facilitate a direct quantification of losses and frequencies. Based on the indicators and other adjustable system parameters, the CRQ framework estimates a LEF of 0.68 events per year and a LM of €20,835,828 per event. By combining these two measures, the system produces the annualized risk value for the organization, which in the case of the University of Michigan is estimated at €14,168,363 per year.

5. Lessons learned

The analysis of the experimental results provides several insights into the strengths and limitations of this proposed system. The experiments conducted on the CRQ model revealed both notable advantages and certain weaknesses of this component. One of these strengths is the flexibility to simulate realistic scenarios using firmographic data from real companies, while also allowing the introduction of fictitious parameters that emulate organizations not present in the dataset. This strength is further enhanced by the ability to leverage firmographic information obtained from the Victim profile and VISCA datasets, which enrich the inputs of the model and provide a stronger foundation for risk quantification. This capability greatly expands the range of situations in which the model can be applied, as it not only quantifies risks for entities included in the repositories but also provides valuable estimations for hypothetical cases, making it useful in decision-making and what-if analyses. Additionally, the results generated by the model are highly descriptive and presented through graphical visualizations that allow users to easily interpret the represented case. By combining frequencies, probabilities, distributions, loss estimations, and a final risk value, the model provides a comprehensive and visually intuitive quantification of the different possible risk situations.

However, a key limitation of the current evaluation is the absence of formal validation of the accuracy of the quantification and recommendation strategies. A potential approach to address this limitation is post-mitigation validation, leveraging both publicly available but also internal data from organizations that have experienced cyberattacks. This would enable comparisons between the predicted frequencies and loss estimations by the model against the actual incidents and financial losses reported over a given period, providing an empirical basis for assessing the reliability of the CRQ outputs. In the best-case scenario, although it may not be scalable, organizations could grant controlled access to private datasets containing historical records of cyberattacks, including incidents that occurred before and after the implementation of specific defensive measures. This would allow the system to assess whether the predicted risk metrics accurately reflect the real-time impact of mitigation strategies.

The experiments also revealed that many records in the evaluated datasets contain missing fields, either because the information could not be found or because it simply

does not exist. When key variables are absent, the model cannot perform optimal estimations. This issue was evident, for example, in the Victim profile dataset, where the used data sources almost never provide the earnings field. The absence of inputs may lead to incorrect probability distributions or models, having to consider whether it can be used or it is better to exclude them from the analysis.

Finally, the experiments conducted on the data extraction model demonstrated the speed and efficiency of aggregating victim entity information in a concrete case of cyber-incident processing. This capability highlights the potential of the system to automate time-consuming tasks, such as data labelling, enabling the generation of a fully usable dataset for the system. In addition to automating the process, the model operates significantly faster than manual human processing, allowing the creation of large-scale datasets that would otherwise require substantial time by humans. These enriched datasets can then be leveraged by other components of the framework, such as the CRQ model, thereby enhancing their performance by incorporating more relevant and updated firmographic information about organizations. However, an important limitation of the current approach is that the extracted data are not validated; the system only checks if the retrieved fields contain a value, but not whether the value is factually correct. Although the workflow relies on trusted and updated data sources, it also incorporates information retrieved from open Internet searches, which introduces the risk of inaccurate or inconsistent data due to the large number of possible sources.

6. Conclusion

This study introduced DICYME, a complex system that contains multiple datasets covering different cyber-related domains, information representation and processing capabilities, and various techniques designed to identify patterns and characteristics within the data. Throughout this study, the methods and techniques used to implement the complex workflow of the system are presented, beginning with the acquisition of diverse datasets that are subsequently employed to compute a range of indicators for quantifying cyber risk across all instances. Subsequently, the CRQ model was proposed, enabling a large number of simulations aimed at quantifying the previously collected data and their computed indicators, and finally producing a final risk value along with other relevant metrics.

In the experiments section, the results demonstrated that both useful components of the system provide significant capabilities: a comparison of the Victim profile dataset and the CRQ model. The VISCA model exhibited high performance in the data extraction task, achieving efficient execution times and demonstrating the potential to fully automate the annotation task performed by a human. However, an important limitation of the current approach is that during data extraction, the system does not explicitly obtain the most accurate value among all possible sources. This limitation may lead to subsequent components and models of the system with inaccurate or inconsistent data, which can propagate errors into subsequent stages of the system. As future work, it would be important to implement a mechanism that not only maximizes the number of extracted attributes but also ensures the selection of the most reliable value for each field. Additionally, extending the set of collected variables would enable a more detailed and representative modelling of the victim profiles.

Similarly, the CRQ model proved capable of generating a rich set of metrics and quantitative outputs, offering valuable insights that can directly support post-mitigation actions by helping organizations prioritize defensive strategies. Nevertheless, there is currently no validation to confirm whether the estimations of the model accurately reflect real-world outcomes. This limitation is critical, as both the overestimation of risks and the underestimation of losses or attack frequencies could lead organizations to make incorrect decisions. Future work should include rigorous validation of the CRQ outputs, either by leveraging public datasets from organizations that have experienced cyberattacks or through partnerships with private companies to provide access to historical incident data. Such validation would allow for a systematic comparison between the predicted annual losses and event frequencies by the model and actual observed outcomes.

Acknowledgement. This work has been funded by the Spanish MICIU under the CPP program in the DICYME project (Ref: CPP2021-009025), partially funded by the XMIDAS project (PID2021-122640OB-I00), and supported by DeNexus Inc.

References

1. Center for International and Security Studies at Maryland: Cyber Events Database (2024), https://cissm.umd.edu/cyber-events-database
2. Chu, Z., Wan, Y., Li, Q., Wu, Y., Zhang, H., Sui, Y., Xu, G., Jin, H.: Graph neural networks for vulnerability detection: A counterfactual explanation. In: Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis. pp. 389–401 (2024)
3. Cioppi, M., Curina, I., Forlani, F., Pencarelli, T.: Online presence, visibility and reputation: a systematic literature review in management studies. Journal of Research in Interactive Marketing 13(4), 547–577 (2019)
4. Electronic Transactions Development Agency: Threat Group Cards: A Threat Actor Encyclopedia (2025), https://apt.etda.or.th/cgi-bin/aptgroups.cgi
5. EuRepoC project: The European Repository of Cyber Incidents (2025), https://eurepoc.eu/
6. Falowo, O.I., Popoola, S., Riep, J., Adewopo, V.A., Koch, J.: Threat actors' tenacity to disrupt: Examination of major cybersecurity incidents. IEEE access 10, 134038–134051 (2022)
7. García-Ochoa, J., Fernández-Isabel, A., Contreras, C., Fernández, R.R., de Diego, I.M., Beltrán, M.: Defining the attractiveness concept for cyber incidents forecasting. Computer Science and Information Systems (ComSIS) (2025)
8. García-Ochoa, J., Fernández-Isabel, A., Martín de Diego, I., Contreras, C., R. Ravines, R., López, O.: Defining the basal attractiveness concept for cybercriminals. In: Proceedings of the 10th Jornadas Nacionales de Investigación en Ciberseguridad. pp. 513–517. Universidad de Zaragoza, Zaragoza, Spain (2025)
9. Industrial Safety and Security Source and Waterfall Security Solutions: ICSSTRIVE (2025), https://icsstrive.com/
10. Kamruzzaman, M., Bhuyan, M.K., Hasan, R., Farabi, S.F., Nilima, S.I., Hossain, M.A.: Exploring the landscape: A systematic review of artificial intelligence techniques in cybersecurity. In: 2024 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI). pp. 01–06. IEEE (2024)
11. Knoth, N., Tolzin, A., Janson, A., Leimeister, J.M.: Ai literacy and its implications for prompt engineering strategies. Computers and Education: Artificial Intelligence 6, 100225 (2024)
12. KonBriefing: Cyberattacks, Hacker attacks, Ransomware attacks (2025), https://konbriefing.com/en-topics/cyberattacks.html

13. Landauer, M., Skopik, F., Stojanović, B., Flatscher, A., Ullrich, T.: A review of time-series analysis for cyber security analytics: from intrusion detection to attack prediction. International Journal of Information Security 24(1), 3 (2025)
14. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
15. Liu, F., Kang, Z., Han, X.: Optimizing rag techniques for automotive industry pdf chatbots: A case study with locally deployed ollama modelsoptimizing rag techniques based on locally deployed ollama modelsa case study with locally deployed ollama models. In: Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Intelligent Information Processing. pp. 152–159 (2024)
16. Liu, R., Xie, Y., Dang, Z., Hao, J., Quan, X., Xiao, Y., Peng, C.: Dynamic vulnerability knowledge graph construction via multi-source data fusion and large language model reasoning. Electronics 14(12), 2334 (2025)
17. Mallick, M.A.I., Nath, R.: Navigating the cyber security landscape: A comprehensive review of cyber-attacks, emerging trends, and recent developments. World Scientific News 190(1), 1–69 (2024)
18. Meta: Introduction to the LLaMA 4 Models (2025), https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/
19. Mızrak, F.: Integrating cybersecurity risk management into strategic management: a comprehensive literature review. Research Journal of Business and Management 10(3), 98–108 (2023)
20. Okutan, A., Werner, G., Yang, S.J., McConky, K.: Forecasting cyberattacks with incomplete, imbalanced, and insignificant data. Cybersecurity 1, 1–16 (2018)
21. Okutan, A., Yang, S.J., McConky, K., Werner, G.: Capture: cyberattack forecasting using non-stationary features with time lags. In: 2019 IEEE Conference on Communications and Network Security (CNS). pp. 205–213. IEEE (2019)
22. Paolo Passeri: 2024 Cyber Attacks Statistics (03 2024), https://www.hackmageddon.com/2024/03/26/2024-cyber-attacks-statistics/
23. Paolo Passeri: Hackmageddon (2025), https://www.hackmageddon.com/
24. Paté-Cornell, M.E., Kuypers, M., Smith, M., Keller, P.: Cyber risk management for critical infrastructure: a risk analysis model and three case studies. Risk Analysis 38(2), 226–241 (2018)
25. Phillips, S.C., Taylor, S., Boniface, M., Modafferi, S., Surridge, M.: Automated knowledge-based cybersecurity risk assessment of cyber-physical systems. IEEE Access 12, 82482–82505 (2024)
26. Pseftelis, T., Chondrokoukis, G.: Understanding cyber incident dynamics in the european union: A study of actor types and sector vulnerabilities. Preprints.org (04 2025)
27. Qin, X., Jiang, F., Cen, M., Doss, R.: Hybrid cyber defense strategies using honey-x: A survey. Computer Networks 230, 109776 (2023)
28. Qudus, L.: Advancing cybersecurity: strategies for mitigating threats in evolving digital and iot ecosystems. Int Res J Mod Eng Technol Sci 7(1), 3185 (2025)
29. Rajesh, P., Alam, M., Tahernezhadi, M., Monika, A., Chanakya, G.: Analysis of cyber threat detection and emulation using mitre attack framework. In: 2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA). pp. 4–12. IEEE (2022)
30. Rios Insua, D., Couce-Vieira, A., Rubio, J.A., Pieters, W., Labunets, K., G. Rasines, D.: An adversarial risk analysis framework for cybersecurity. Risk Analysis 41(1), 16–36 (2021)
31. Sailio, M., Latvala, O.M., Szanto, A.: Cyber threat actors for the factory of the future. Applied Sciences 10(12), 4334 (2020)
32. Shi, Z., Matyunin, N., Graffi, K., Starobinski, D.: Uncovering cwe-cve-cpe relations with threat knowledge graphs. ACM Transactions on Privacy and Security 27(1), 1–26 (2024)
33. Subroto, A., Apriyana, A.: Cyber risk prediction through social media big data analytics and statistical machine learning. Journal of Big Data 6(1), 50 (2019)
34. TI Safe: Incident Hub (2024), https://hub.tisafe.com/

35. Wåreus, E., Hell, M.: Automated cpe labeling of cve summaries with machine learning. In: International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. pp. 3–22. Springer (2020)
36. Xie, Y., Jin, X., Xie, T., Lin, M., Chen, L., Yu, C., Cheng, L., Zhuo, C., Hu, B., Li, Z.: Decomposition for enhancing attention: Improving llm-based text-to-sql through workflow paradigm. arXiv preprint arXiv:2402.10671 (2024)
37. Yin, J., Hong, W., Wang, H., Cao, J., Miao, Y., Zhang, Y.: A compact vulnerability knowledge graph for risk assessment. ACM Transactions on Knowledge Discovery from Data 18(8), 1–17 (2024)
38. Zhou, Y., Liu, S., Siow, J., Du, X., Liu, Y.: Devign: Effective vulnerability identification by learning comprehensive program semantics via graph neural networks. Advances in neural information processing systems 32 (2019)

Javier Sánchez García-Ochoa was born in Toledo, Spain in 2000. He received a bachelor's degree in Cybersecurity Engineering from Rey Juan Carlos University (URJC) and a master's degree in Cybersecurity and Privacy from the Open University of Catalonia (UOC). After gaining more than a year of experience in the private sector, he joined Rey Juan Carlos University in 2023 as a research staff member. During this period, he participated in the public-private collaboration project DICYME (Dynamic Industrial Cyber Risk Modelling based on Evidence), where he focused on dynamic cyber risk quantification. He has contributed to several scientific articles and conference proceedings. His research interests currently focus on large language models (LLMs), data science, machine learning, and their applications within the field of cybersecurity.

Jaime Rueda Carpintero was born in Madrid, Spain in 2002. He received a degree in Computer Engineering from Rey Juan Carlos University (URJC), where he also completed a Master's degree in Decision Systems Engineering. After more than a year of experience in the private sector, he joined Rey Juan Carlos University in 2025 as a researcher in the DICYME project. He is a member of the High Performance Research Group Data Science Laboratory (DSLAB, URJC) and has contributed to several scientific articles and conference communications. He actively participates in research tasks and projects focused on natural language processing (NLP), large language models (LLMs), and data science.

Dr. Rubén Rodríguez Fernández was born in Bembibre, León, Spain in 1993. He received a PhD degree in Artificial Intelligence from Rey Juan Carlos University (URJC). He also received a master's degree in Data Science from URJC and a master's degree in Artificial Intelligence from the Technical University of Madrid (UPM). He is part of the Data Science Laboratory high performance research group and has been an Assistant Professor at URJC since 2024. His research interests include active learning, explainable machine learning, generative artificial intelligence, and applied machine learning.

Dr. Alberto Fernández-Isabel was born in Toledo, Spain in 1984. He received a PhD in Computer Science from Complutense University of Madrid (UCM) in 2015. He obtained a scholarship at the Spanish National Research Council (CSIC) as a technical assistant. He has been working for several years on European and national projects as a predoctoral

and postdoctoral researcher. Since 2019 he is Assistant Professor at the Higher Technical School of Computer Engineering (ETSII) at Rey Juan Carlos University (URJC). He has authored more than 30 scientific articles and books. He completes his background with a Master´s degree in Artificial Intelligence and a Master´s degree in Information Systems. His research interests include intelligent agents, machine learning, data visualization, and natural language processing in various application domains, including distributed programming, sentiment analysis, agent-based collaboration and negotiation, smart cities, and simulations.

Dr. Isaac Martín de Diego was born in Campaspero, Valladolid, Spain in 1973. He received a PhD degree in Mathematical Engineering from Carlos III de Madrid University in 2005 (Extraordinary Doctorate Award). Since 2023 he is a full professor at the Higher Technical School of Computer Engineering at Rey Juan Carlos University (Associate Professor from 2018). He is the co-founder of the Data Science Laboratory and Head of the Sports Analytics Master at Rey Juan Carlos University. He has been head of the ERICSSON Chair on Data Science applied to 5G. He is the author of more than 100 articles. His research interests include methods, processes, and tools for Data Science in various application domains: explainability, sampling, complexity, performance evaluation, visualization, recommendation systems and security with a special interest in Machine Learning algorithms and a combination of information methods.

Dr. Emilio López Cano was born in Madrid, Spain in 1973. He received a PhD degree from Rey Juan Carlos University (URJC). He holds a Master's degree in Decision Systems Engineering (URJC), a degree in Applied Statistics, and a diploma in Statistics from the Complutense University of Madrid (UCM). He is an Associate Professor at Rey Juan Carlos University in the area of Statistics and Operations Research. Previously, he was a teacher and researcher at the University of Castilla-La Mancha since 2011, and worked as a statistician and IT professional in the private sector for 14 years. His research interests include statistical computing, analysis, visualization, and management of complex data and modeling in multidisciplinary environments, where he has several national and international publications and projects.

Dr. Romy Rodríguez Ravines received a PhD degree in Statistics from the Federal University of Rio de Janeiro (UFRJ), Brazil. She is an Applied Data Science Leader with over 25 years of experience in business consulting based on advanced analytics, Bayesian modeling, and artificial intelligence, specializing in transforming data into strategic value across the cybersecurity, finance, marketing, and public health sectors. She has served as the Head of Research and Modeling Strategies at DeNexus, where she led the development of a cyber risk quantification platform adopted by international insurers and industrial organizations. She has held various senior leadership positions in consulting firms and has served as a lecturer at several leading universities and business schools in Spain.

Dr. Ovidio López, born in Murcia, Spain, holds a Physics degree from the University of Murcia and a Ph.D. in Renewable Energy and Energy Efficiency from the Polytechnic University of Cartagena. His background includes numerical simulation, data analysis, and Operational Technology (SCADA, PLC). He has experience in R&D and Machine Learning, later working as a Data Scientist focused on industrial cyber risk. Currently, he

teaches Electronics in vocational training, sharing expertise in automation and emerging technologies.

Jaume Puigbò Sanvisens was born in Barcelona, Spain. He holds a Bachelor's degree in Mathematics from the University of Barcelona, as well as a Master's degree in Foundations of Data Science (2016–2017) and a Postgraduate degree in Data Science (2015–2016) from the same institution. Since 2021, he has been working as a Data Scientist at DeNexus Inc., where he develops advanced models to quantify cyberattack risk in industrial environments. His research and professional interests include Machine Learning, Deep Learning, Optimization, Recommender Systems, Natural Language Processing, and Data Visualization, with applications in cybersecurity, predictive analytics, and industrial risk modeling. In the past, he has worked on projects such as predictive maintenance for industrial systems, demand forecasting, traffic prediction, and B2B recommender systems, applying advanced statistical and machine learning techniques to real-world challenges. Jaume was a finalist in the UniversityHack 2018 competition and has presented at R user conferences.

Received: September 30, 2025; Accepted: January 10, 2026.

Computer Science and Information Systems 23(3):1027–1052 https://doi.org/10.2298/CSIS251220025C

Examination Timetabling with Independent Resource Pools: A Multi-Stage Genetic Algorithm Approach

Khalil Chebil[1], Mahdi Khemakhem[1], Abdallah Hammad[1], and Mohammed Alanazi[1]

Department of Computer Science
College of Computer Engineering and Sciences
Prince Sattam Bin Abdulaziz University
Al Kharj 11942, Saudi Arabia
k.chebil@psau.edu.sa (corresponding author),
m.khemakhem@psau.edu.sa, 442052035@std.psau.edu.sa, 442052038@std.psau.edu.sa

Abstract. The Examination Timetabling Problem (ETP) is an NP-hard combinatorial optimization problem that has been extensively studied over the past decades, with methodologies ranging from classical heuristics to contemporary machine learning approaches. This paper provides a comprehensive literature review and proposes a novel two-stage genetic algorithm based decomposition combined with a clear linear mathematical formulation. A unique contribution of this work is addressing the practical complexity of two-separated examination facilities at PSAU, where male and female students study in completely separate buildings with independent examination rooms, capacities, and supervisory staff, yet must follow a synchronized examination schedule. This creates a coupled dual-ETP problem where two independent resource allocation problems must share identical time slot assignments, significantly increasing model complexity and constraint interactions. The proposed algorithm decomposes the problem into two sequential stages,time slot assignment and room assignment to effectively reduce computational complexity while maintaining solution quality. Additionally, a systematic comparison with Mixed-Integer Programming (MIP) formulation is provided to establish the trade-offs between exact and metaheuristic approaches. The algorithm is further benchmarked against simulated annealing and tabu search baselines to provide broader methodological context, demonstrating its advantage on coupled dual-resource instances. The algorithm shows particular strength in scalability, successfully solving large instances where MIP becomes computationally intractable, while maintaining competitive solution quality.

Keywords: Examination, Timetabling, Genetic Algorithm, Optimization.

1. Introduction

The examination timetabling problem is a complex combinatorial optimization challenge, fundamentally involving four key parameters: a set of times (T), a set of resources (R), a set of exams (E), and a set of constraints (C). Research into computational timetabling for educational schedules dates back to the 1960s, coinciding with the rise in educational institution enrollment and limited computer power. Despite decades of research and the development of numerous sophisticated methodologies—from early graph coloring heuristics to contemporary machine learning approaches—the problem remains challenging,

particularly in formulating soft constraints mathematically, as these can vary significantly between universities based on specific requirements.

While the timetabling literature has extensively studied single-resource-pool formulations, practical institutions frequently face heterogeneous resource environments where multiple independent resource sets must be coordinated. Similar situations arise in multi-campus universities, distributed manufacturing scheduling, and healthcare shift planning with separated wards. The scheduling literature on heterogeneous resource pools [18, 5] has addressed aspects of this challenge, but the specific coupling constraint—where completely disjoint resource pools must share identical time assignments—remains underexplored in the examination timetabling context. This paper contributes to addressing this gap.

This paper makes several contributions to the examination timetabling literature. First, it presents a focused literature review organized by methodology, critically comparing approaches and identifying specific gaps that motivate the proposed method. Second, it proposes a two-stage genetic algorithm decomposition approach combined with a mathematical formulation that linearizes inherently nonlinear constraints using standard big-M and auxiliary variable techniques. Third, it demonstrates the algorithm's effectiveness on standard benchmarks (Toronto datasets) and validates its practical applicability through successful deployment for gender-separated examination facilities at Prince Sattam Bin Abdulaziz University. The PSAU case represents an instance of the broader class of coupled heterogeneous resource problems: male and female students study in completely separate buildings located in different parts of the campus, each with independent examination rooms, distinct capacity constraints, and separate supervisory personnel. Despite these separated resources, both populations must follow an identical synchronized examination schedule for administrative and academic coordination. This creates what is essentially a coupled dual-ETP problem—two independent examination timetabling problems (one for each building) that must produce identical time slot assignments while managing completely disjoint room resources. This coupling significantly increases model complexity, as feasible solutions must simultaneously satisfy constraints across both buildings while maintaining schedule synchronization. Finally, the paper provides systematic comparisons with Mixed-Integer Programming approaches and other metaheuristic baselines (simulated annealing, tabu search), establishing clear trade-offs between exact and metaheuristic methods.

The remainder of this paper is structured as follows: Section 2 presents a synthesized literature review organized by methodology with critical comparison. Section 3 presents a detailed mathematical formulation of the problem, including hard and soft constraints. Section 4 describes the proposed multi-stage genetic algorithm decomposition approach, including pseudocode and complexity analysis. Section 5 discusses experimental results on benchmark datasets and real-world applications. Finally, Section 6 concludes the paper and suggests directions for future research.

2. Literature Review

This section provides a critical synthesis of examination timetabling methods, organized by methodology rather than chronology, to identify the specific gaps that motivate the pro-

posed approach. Rather than exhaustive coverage, we focus on representative works that shaped each methodological stream and highlight comparative strengths and limitations.

2.1. Constructive and Graph-Based Heuristics

Graph coloring heuristics form the foundation of constructive approaches for examination timetabling. Sabar et al. [32, 30] proposed graph coloring constructive hyper-heuristics, while Demeester et al. [17] introduced practical benchmarks. Pillay [28] demonstrated adaptive hyper-heuristic evolution, and Pillay and Özcan [29] automated heuristic generation via genetic programming. These methods excel at producing initial feasible solutions rapidly but are limited in soft constraint optimization, motivating their use as initialization strategies rather than standalone solvers.

2.2. Local Search and Neighborhood-Based Methods

Tabu search has proven particularly effective: Abdullah and Turabieh [2] demonstrated multi-neighbourhood structures within memetic tabu search, while Pais and Amaral [27] introduced adaptive tabu list management via fuzzy inference. Simulated annealing variants achieved state-of-the-art results: Battistutta et al. [9] presented feature-based parameter tuning, Bellio et al. [10] developed two-stage multi-neighborhood SA achieving top benchmark results, and Leite et al. [24] achieved competitive quality with reduced computation time. Burke and Bykov [13] introduced late acceptance hill-climbing, requiring fewer parameters than SA while maintaining competitive quality. These approaches demonstrate strong intensification capabilities but typically require good initial solutions, making them complementary to constructive heuristics.

A critical observation is that many of these methods operate on unified resource pools; their behavior under coupled heterogeneous resource constraints remains unexplored.

2.3. Population-Based and Evolutionary Methods

Genetic algorithms and swarm intelligence methods provide broader search space exploration. Abdullah and Alzaqebah [1] developed hybrid bee algorithm-GA approaches, while Al-Betar et al. [4] demonstrated effective memetic techniques. Sabar et al. [31] used grammatical evolution to automatically generate heuristics. Lei et al. [22, 23] applied co-evolutionary and MOEA/D-based memetic algorithms for multi-objective formulations. Ahandani et al. [6] and Nand et al. [26] explored PSO and firefly variants respectively.

While population-based methods offer better exploration, they generally sacrifice solution precision compared to local search methods. Multi-stage decomposition—as demonstrated by Gogos et al. [19]—addresses this by reducing the search space for each evolutionary stage.

2.4. Mathematical Programming and Exact Methods

MIP formulations provide optimal solutions for small instances. Arbaoui and Moukrim [8] demonstrated preprocessing to reduce computational requirements, while Cataldo et al. [15] addressed curriculum-based formulations. Al-hawari et al. [5] showed that three-phase

ILP decomposition can make MIP tractable for larger instances. Carlsson et al. [14] provided comprehensive comparisons showing hybrid MIP-metaheuristic approaches can outperform pure methods. Arbaoui et al. [7] studied lower bounds for spacing constraints, providing theoretical benchmarks.

Key limitation: MIP scalability degrades rapidly with problem size, particularly when coupled constraints (such as our dual-building synchronization) expand the variable space.

2.5. Multi-Objective and Practical Applications

Recent work increasingly addresses real-world complexity. Mukhlason et al. [25] incorporated fairness objectives, and Güler et al. [20] addressed scarce resource scenarios. Çimen et al. [16] integrated invigilator assignment, and Mohmad Kahar and Kendall [21] addressed multi-resource scheduling. Elloumi et al. [18] analyzed room allocation as a distinct subproblem. Van Bulck et al. [33] achieved strong ITC-2007 results through careful neighborhood design. Abou Kasm et al. [3] relaxed hard constraints for flexible scheduling.

2.6. Research Gap and Contributions

From this analysis, we identify four specific gaps that motivate the present work:

1. **Coupled heterogeneous resources:** Existing formulations assume unified resource pools. While heterogeneous resource scheduling exists in broader OR literature, the specific coupling constraint—disjoint room pools sharing synchronized time assignments—is not addressed in examination timetabling.
2. **Linearization gap:** Many formulations present nonlinear constraints (bilinear products of binary variables, quadratic objectives) without providing linearized equivalents needed for MIP solvers.
3. **Decomposition-exact integration:** Few works systematically compare decomposed metaheuristics against exact methods under unified evaluation metrics.
4. **Practical deployment:** Gap between benchmark problem simplifications and real-world operational requirements involving gender-separated or multi-campus facilities.

This paper aims to contribute by providing a linearized mathematical formulation that can accommodate such complexities while employing a multi-stage genetic algorithm decomposition approach. The proposed method combines the clarity of mathematical programming formulations with the scalability of metaheuristic optimization, addressing coupled heterogeneous resource pools as required by Prince Sattam Bin Abdulaziz University.

3. Problem Formulation

To effectively address the Exam Timetabling Problem (ETP), we first establish a clear mathematical formulation and define the necessary notations. This formulation will serve as the foundation for developing algorithms to solve the ETP.

3.1. Notations

Let us define the following sets and parameters:

Sets:

- **E**: Set of exams to be scheduled, indexed by e ($e = 1, 2, \ldots, |E|$)
- **T**: Set of available time slots, indexed by t ($t = 1, 2, \ldots, |T|$)
- **R**: Set of available rooms, indexed by r ($r = 1, 2, \ldots, |R|$)
- **S**: Set of students, indexed by s ($s = 1, 2, \ldots, |S|$)

Parameters:

- n_e: Number of students enrolled in exam e
- cap_r: Capacity of room r
- d_e: Duration of exam e (in time slots)
- $C_{e,e'}$: Conflict matrix where $C_{e,e'} = 1$ if exams e and e' have common students, 0 otherwise
- $w_1, w_2, \ldots, w_k$: Weights for different soft constraint violations

Decision Variables:

- $x_{e,t}$: Binary variable equal to 1 if exam e is scheduled in time slot t, 0 otherwise
- $y_{e,r}$: Binary variable equal to 1 if exam e is assigned to room r, 0 otherwise
- $z_{e,r,t}$: Binary auxiliary variable equal to 1 if exam e is assigned to room r **and** time slot t, i.e., $z_{e,r,t} = x_{e,t} \cdot y_{e,r}$

3.2. Hard Constraints

The following hard constraints must be satisfied for a feasible timetable:

HC1: Exam Assignment
Each exam must be scheduled in exactly one time slot:

$$\sum_{t \in T} x_{e,t} = 1, \quad \forall e \in E \tag{1}$$

HC2: No Student Conflicts
Exams with common students cannot be scheduled simultaneously:

$$x_{e,t} + x_{e',t} \leq 1, \quad \forall e, e' \in E \text{ with } C_{e,e'} = 1, \forall t \in T \tag{2}$$

HC3: Room Assignment
Each exam must be assigned to exactly one room:

$$\sum_{r \in R} y_{e,r} = 1, \quad \forall e \in E \tag{3}$$

HC4: Room Capacity
The room capacity must not be exceeded:

$$n_e \cdot y_{e,r} \leq cap_r, \quad \forall e \in E, \forall r \in R \tag{4}$$

HC5: Room Availability

A room can host at most one exam at any given time:

$$\sum_{e \in E} x_{e,t} \cdot y_{e,r} \leq 1, \quad \forall t \in T, \forall r \in R \tag{5}$$

Since this constraint involves a bilinear product of two binary decision variables, we introduce the auxiliary variables $z_{e,r,t} = x_{e,t} \cdot y_{e,r}$ and enforce the equivalence via McCormick / big-M constraints, yielding the following linear formulation:

$$z_{e,r,t} \leq x_{e,t}, \quad \forall e \in E, \forall r \in R, \forall t \in T \tag{6}$$

$$z_{e,r,t} \leq y_{e,r}, \quad \forall e \in E, \forall r \in R, \forall t \in T \tag{7}$$

$$z_{e,r,t} \geq x_{e,t} + y_{e,r} - 1, \quad \forall e \in E, \forall r \in R, \forall t \in T \tag{8}$$

$$\sum_{e \in E} z_{e,r,t} \leq 1, \quad \forall t \in T, \forall r \in R \tag{9}$$

3.3. Soft Constraints

The following soft constraints aim to improve timetable quality and should be minimized:

SC1: Spread of Exams

Minimize the number of students having exams scheduled in consecutive time slots:

$$\text{Minimize} \quad \sum_{s \in S} \sum_{t \in T} \sum_{e,e' \in E_s} x_{e,t} \cdot x_{e',t+1} \tag{10}$$

Since this objective involves bilinear terms, we introduce auxiliary variables $q_{e,e',t}$ representing the product $x_{e,t} \cdot x_{e',t+1}$ and enforce the equivalence via the following linear constraints:

$$q_{e,e',t} \leq x_{e,t}, \quad \forall s \in S, \forall e, e' \in E_s, \forall t \in T \tag{11}$$

$$q_{e,e',t} \leq x_{e',t+1}, \quad \forall s \in S, \forall e, e' \in E_s, \forall t \in T \tag{12}$$

$$q_{e,e',t} \geq x_{e,t} + x_{e',t+1} - 1, \quad \forall s \in S, \forall e, e' \in E_s, \forall t \in T \tag{13}$$

The equivalent linear SC1 objective becomes:

$$\text{Minimize} \quad \sum_{s \in S} \sum_{t \in T} \sum_{e,e' \in E_s} q_{e,e',t} \tag{14}$$

where E_s is the set of exams taken by student s.

SC2: Period Utilization

Balance the distribution of exams across time slots:

$$\text{Minimize} \quad \sum_{t \in T} \left(\sum_{e \in E} x_{e,t} - \frac{|E|}{|T|} \right)^2 \tag{15}$$

Since this objective is quadratic, we reformulate it using an equivalent absolute deviation formulation with auxiliary variables $\delta_t^+, \delta_t^- \geq 0$:

$$\sum_{e \in E} x_{e,t} - \frac{|E|}{|T|} = \delta_t^+ - \delta_t^-, \quad \forall t \in T \tag{16}$$

$$\delta_t^+, \delta_t^- \geq 0, \quad \forall t \in T \tag{17}$$

The equivalent linear SC2 objective becomes:

$$\text{Minimize} \quad \sum_{t \in T} (\delta_t^+ + \delta_t^-) \tag{18}$$

SC3: Room Utilization
Minimize underutilization of rooms:

$$\text{Minimize} \quad \sum_{e \in E} \sum_{r \in R} y_{e,r} \cdot \left(1 - \frac{n_e}{cap_r}\right) \tag{19}$$

SC4: Large Exam Penalty
Penalize scheduling large exams in less preferred time slots:

$$\text{Minimize} \quad \sum_{e \in E} \sum_{t \in T_{unpreferred}} n_e \cdot x_{e,t} \tag{20}$$

3.4. Objective Function

The overall objective is to minimize a weighted sum of soft constraint violations while ensuring all hard constraints are satisfied:

$$\text{Minimize} \quad Z = w_1 \cdot SC1 + w_2 \cdot SC2 + w_3 \cdot SC3 + w_4 \cdot SC4 \tag{21}$$

Subject to: HC1, HC2, HC3, HC4, HC5

The weights w_1, w_2, w_3, w_4 can be adjusted based on institutional priorities and requirements.

Relationship Between Mathematical Formulation and GA Fitness The GA fitness function (Eq. 18 in Section 4) is a direct operationalization of this mathematical objective. The hard constraints (HC1–HC5) are mapped to large penalty terms (P_{h1}, P_{h2}) that dominate the fitness value when violated, driving the search toward feasible regions. The soft constraints (SC1–SC4) are mapped to weighted penalty terms ($w_1 V_{s1} + w_2 V_{s2} + w_3 V_{s3}$) that guide optimization quality once feasibility is achieved. Specifically: V_{h1} corresponds to HC2 violations, V_{h2} penalizes excessive daily load (a practical operationalization related to HC1 and student comfort), and V_{s1}, V_{s2}, V_{s3} correspond to SC1 with varying proximity windows (consecutive slots, same day, consecutive days). The two-stage decomposition separates time slot constraints (HC1, HC2, SC1, SC2) from room constraints (HC3–HC5, SC3, SC4), allowing each GA stage to focus on its corresponding subset of the formulation.

3.5. Gender-Separated Buildings: A Coupled Dual-ETP Formulation

At Prince Sattam Bin Abdulaziz University, an institutional constraint significantly increases problem complexity—analogous to heterogeneous resource pool problems studied in scheduling theory [18, 5]: male and female students are educated in completely separate buildings with independent physical resources, yet must follow a synchronized examination schedule. This creates what we term a **coupled dual-ETP problem**.

Problem Structure:
Let us denote:

- E^M: Set of male student exams
- E^F: Set of female student exams
- R^M: Set of examination rooms in the male building
- R^F: Set of examination rooms in the female building
- T: Set of time slots (**shared** between both buildings)
- $\mathcal{P}$: Set of paired course indices, where $(e_i^M, e_i^F) \in \mathcal{P}$ denotes matching male/female sections of the same course

Note that $E^M \cap E^F = \emptyset$ and $R^M \cap R^F = \emptyset$, representing complete separation of exam sections and room resources.

Coupling Constraint:

The critical coupling arises from the requirement that corresponding course exams for male and female sections must be scheduled at **identical time slots**:

$$x_{e_i^M,t} = x_{e_i^F,t}, \quad \forall (e_i^M, e_i^F) \in \mathcal{P}, \forall t \in T \tag{22}$$

This means if course "Calculus I" has both male and female sections, both must be scheduled in the same time slot, despite being held in different buildings with different rooms.

Independent Room Constraints:

While time slots are synchronized, room assignments are completely independent:

$$y_{e^M,r^M} \in \{0,1\}, \quad \forall e^M \in E^M, r^M \in R^M \tag{23}$$

$$y_{e^F,r^F} \in \{0,1\}, \quad \forall e^F \in E^F, r^F \in R^F \tag{24}$$

with separate capacity constraints:

$$n_{e^M} \cdot y_{e^M,r^M} \leq cap_{r^M}, \quad \forall e^M \in E^M, r^M \in R^M \tag{25}$$

$$n_{e^F} \cdot y_{e^F,r^F} \leq cap_{r^F}, \quad \forall e^F \in E^F, r^F \in R^F \tag{26}$$

Complexity Analysis:

This formulation essentially creates two examination timetabling problems (ETP^M, ETP^F) that must be solved simultaneously with the synchronization constraint. The complexity implications are:

1. **Doubled Decision Variables:** Instead of $|E| \times |R|$ room assignment variables, we have $(|E^M| \times |R^M|) + (|E^F| \times |R^F|)$ variables.
2. **Coupled Feasibility:** A solution is feasible only if **both** male and female schedules are individually feasible **and** satisfy the time slot synchronization constraint.
3. **Resource Heterogeneity:** Different room capacities, numbers, and distributions between buildings create asymmetric constraint landscapes.
4. **Independent Soft Constraints:** Each building has its own room utilization, capacity optimization, and supervisory assignment objectives that must be balanced.

This coupled structure is rarely addressed in the examination timetabling literature, which typically assumes unified resource pools. While heterogeneous resource scheduling has been studied in broader operations research contexts (e.g., parallel machine scheduling, multi-factory production planning), the specific combination of disjoint room pools with synchronized time slot requirements creates a distinct coupled feasibility structure. The gender-separated buildings constraint represents a practical real-world complexity arising from cultural, religious, and institutional requirements that significantly impacts both model formulation and algorithmic design.

Limitations and Generalizability The coupled dual-ETP model assumes a strict one-to-one correspondence between male and female exam sections, which holds at PSAU but may require extension for institutions with unequal course offerings across campuses. The model generalizes naturally to any multi-campus or multi-building scenario where time synchronization is required with independent room pools (e.g., distributed university systems, hospital examination scheduling). However, scenarios involving partial resource sharing (some rooms accessible to both populations) would require relaxation of the strict disjointness constraint $R^M \cap R^F = \emptyset$. Future work could explore parameterized coupling strength to handle such intermediate cases.

4. Genetic Algorithm Based Decomposition Approach

The proposed solution methodology employs a multi-stage genetic algorithm that decomposes the examination timetabling problem into two sequential optimization stages: (1) time slot assignment and (2) room assignment. This decomposition strategy has been shown to be effective in handling complex combinatorial optimization problems [19, 10]. The decomposition is justified by the separable structure of the ETP formulation: time slot constraints (HC1, HC2, SC1, SC2) involve only $x_{e,t}$ variables, while room constraints (HC3, HC4, SC3) involve only $y_{e,r}$ variables. The only coupling constraint is HC5 (room-time exclusivity), which becomes a pure room constraint once time slots are fixed. Therefore, fixing Stage 1 outputs reduces Stage 2 to an independent assignment problem. While this sequential approach does not guarantee global optimality (the optimal room assignment depends in general on the time slot assignment), it provides a valid upper bound. The practical effectiveness of this decomposition has been empirically validated for timetabling by Gogos et al. [19] and Bellio et al. [10]. We note that global optimality is not claimed; rather, the decomposition trades potential solution quality loss for tractability.

4.1. Stage 1: Time Slot Assignment

The first genetic algorithm focuses exclusively on assigning exams to appropriate time slots while satisfying temporal constraints. This stage aims to minimize student conflicts and distribute exams evenly across the examination period.

Chromosome Representation In this stage, a chromosome represents a complete time slot assignment solution. Each chromosome is encoded as a vector of length $|E|$, where each gene corresponds to an exam and its value indicates the assigned time slot. This direct encoding scheme has been widely adopted in genetic algorithm applications for timetabling problems due to its simplicity and effectiveness.

Synchronization Enforcement for Coupled Dual-ETP For the coupled dual-ETP, the chromosome encodes only the *unique course* time slot assignments (i.e., one gene per course rather than per section). Both the male section e_i^M and female section e_i^F of each course i inherit the same time slot value from the chromosome. This encoding *structurally enforces* the coupling constraint (Eq. 19) by construction—no separate repair or penalty is

needed for synchronization. During crossover and mutation, operators act on course-level genes, automatically preserving synchronization. This representation reduces the effective chromosome length from $|E^M| + |E^F|$ to $\max(|E^M|, |E^F|)$ for paired courses, simultaneously reducing search space dimensionality and eliminating synchronization violations.

Fitness Function The fitness function evaluates the quality of each chromosome based on both hard and soft constraint satisfaction. The fitness score is calculated as:

$$f(C) = P_{h1} \cdot V_{h1} + P_{h2} \cdot V_{h2} + w_1 \cdot V_{s1} + w_2 \cdot V_{s2} + w_3 \cdot V_{s3} \tag{27}$$

where:

- V_{h1}: Number of conflicting time slots (students with same courses at same time) — corresponds to HC2
- V_{h2}: Number of instances where students have 3 or more exams in same day — practical operationalization of student overload
- P_{h1}, P_{h2}: Large penalty coefficients for hard constraint violations (typically 1,000) — values determined by preliminary calibration (see Section 5.2)
- V_{s1}: Soft constraint violation for two consecutive exams (penalty weight: 5) — corresponds to SC1 with $\Delta = 1$
- V_{s2}: Soft constraint violation for two exams in same day (penalty weight: 50) — SC1 variant with same-day window
- V_{s3}: Soft constraint violation for two exams in consecutive days (penalty weight: 5) — SC1 variant with $\Delta \leq |T|/\text{days}$

The objective is to minimize this fitness value. A feasible solution has $V_{h1} = 0$ and $V_{h2} = 0$, meaning no hard constraints are violated.

Genetic Operators The genetic algorithm employs several operators to evolve the population toward better solutions:

Selection: Tournament selection is used to choose parent chromosomes based on their fitness scores. Tournament size is typically set to 3-5 individuals, providing a good balance between selection pressure and diversity maintenance.

Crossover: A multi-point crossover operator is applied with three randomly selected crossover points. This operator combines genetic material from both parents by alternating segments between crossover points, maintaining solution diversity while exploring the search space effectively.

Mutation: A mutation operator randomly alters the time slot assignment of selected exams in a chromosome. When mutation occurs (controlled by mutation rate), a random number of exams (up to 10% of total) are reassigned to randomly selected time slots. The mutation rate is set based on parameter calibration experiments (see Section 5.2). This operator introduces diversity into the population and helps avoid premature convergence to local optima.

Repair Mechanism: When mutations or crossovers produce infeasible solutions, a repair mechanism is applied using graph coloring heuristics to reassign conflicting exams to available time slots. Algorithm 1 presents the pseudocode for this repair procedure.

Algorithm 1 Graph Coloring Repair Mechanism

Require: Chromosome C with potential conflicts, conflict matrix $C_{e,e'}$, time slots T
Ensure: Repaired chromosome C' with reduced/eliminated conflicts
$C' \leftarrow C$
conflicts $\leftarrow$ list of conflicting exam pairs (e, e') where $C_{e,e'} = 1$ and $C'[e] = C'[e']$
Sort conflicts by descending conflict degree (number of adjacent conflicts)
for each conflicting exam e in sorted order **do**
    forbidden $\leftarrow \{C'[e'] : C_{e,e'} = 1\}$ {time slots used by conflicting exams}
    available $\leftarrow T \setminus$ forbidden
    if available $\neq \emptyset$ **then**
        $C'[e] \leftarrow$ slot from available with lowest saturation degree
    else
        $C'[e] \leftarrow$ slot from T minimizing total new conflicts {greedy fallback}
    end if
end for
return C'

The repair mechanism has worst-case complexity $O(|E|^2 \cdot |T|)$: for each conflicting exam ($O(|E|)$), it scans all adjacent exams ($O(|E|)$) to determine forbidden slots and evaluates each available slot ($O(|T|)$). In practice, the sparse conflict structure results in significantly lower computational cost. The deterministic tie-breaking rule (lowest saturation degree, then lowest index) ensures reproducible behavior.

Elitism: The best-performing chromosome from each generation is implicitly preserved through the selection and reproduction process. While not implementing explicit elitism where top individuals are directly copied, the tournament selection with appropriate tournament size ensures high-quality solutions have a strong probability of surviving and propagating their genetic material.

Termination Criteria: The algorithm terminates when one of the following conditions is met:

1. A maximum number of generations (typically 1000-5000) is reached
2. The improvement in the best fitness score over the last k generations falls below a threshold ϵ
3. A feasible solution with acceptable soft constraint violations is found

Population Initialization The initial population is generated using a combination of random initialization and constructive heuristics. This hybrid initialization strategy provides both diversity and quality in the initial population:

1. **Random Initialization (50%):** Half of the population is generated by randomly assigning each exam to a time slot.
2. **Graph Coloring Heuristics (50%):** The remaining half uses graph coloring heuristics such as:
   - Largest Degree First: Exams with the most conflicts are scheduled first
   - Saturation Degree: Prioritizes exams based on the number of available time slots
   - Largest Enrollment First: Schedules exams with more students first

For Stage 2 (room assignment), the initial population inherits the time slot assignments from Stage 1, and rooms are randomly assigned from the gender-appropriate room sets (male rooms for male sections, female rooms for female sections). This ensures that the initial population satisfies the gender-specific room constraint from the outset.

Overall Algorithm Pseudocode Algorithm 2 presents the complete pseudocode for the two-stage genetic algorithm.

Algorithm 2 Two-Stage Genetic Algorithm for ETP

Require: Exams E, time slots T, rooms R, conflict matrix C, paired courses $\mathcal{P}$
Ensure: Complete timetable (x^*, y^*)
— Stage 1: Time Slot Assignment —
Initialize population Π_1: 50% random, 50% graph coloring heuristics
for each paired course $(e_i^M, e_i^F) \in \mathcal{P}$ **do**
    Encode as single gene (synchronization by construction)
end for
$g \leftarrow 0$; stagnation $\leftarrow 0$
while $g < G_{\max}$ **and** stagnation $< k$ **and** not converged **do**
    Evaluate fitness $f(C)$ for all chromosomes (Eq. 18)
    Select parents via tournament selection (size τ)
    Apply 3-point crossover with rate p_c
    Apply time slot mutation with rate p_m
    Repair infeasible offspring via Algorithm 1
    Update best solution; update stagnation counter
    $g \leftarrow g + 1$
end while
$x^* \leftarrow$ best time slot assignment from Π_1
— Stage 2: Room Assignment —
Initialize population Π_2: inherit x^*, randomly assign gender-appropriate rooms
$g \leftarrow 0$
while $g < G_{\max}^{(2)}$ **and** not converged **do**
    Evaluate fitness $f(R)$ for all chromosomes (Eq. 24)
    Select, crossover, mutate room genes (gender-separated)
    Update best solution
    $g \leftarrow g + 1$
end while
$y^* \leftarrow$ best room assignment from Π_2
return (x^*, y^*)

4.2. Stage 2: Room Assignment

The second genetic algorithm is dedicated to assigning rooms to the scheduled exams. This algorithm operates on the output of the first genetic algorithm, taking the assigned time slots as fixed input and focusing on optimizing room assignments while satisfying capacity and availability constraints.

Chromosome Representation In this stage, a chromosome represents a complete room assignment solution. Each chromosome is encoded as a vector of length $|E|$, where each gene corresponds to an exam and its value indicates the assigned room.

Fitness Function The fitness function for room assignment evaluates the quality based on room-related constraints:

$$f(R) = P_r \cdot V_{r1} + w_r \cdot V_{r2} + P_c \cdot V_{r3} + P_g \cdot V_{r4} \tag{28}$$

where:

- V_{r1}: Number of rooms with 3 or more sections simultaneously (penalty: 100) — corresponds to HC5
- V_{r2}: Overflow when 2 sections exceed room capacity (penalty: 1 per student) — corresponds to HC4
- V_{r3}: Single section overflow beyond room capacity (penalty: 5 per student) — corresponds to HC4
- V_{r4}: Gender mismatch violations (male sections in female rooms or vice versa) (penalty: 100) — corresponds to dual-ETP room disjointness
- P_r, w_r, P_c, P_g: Penalty coefficients for different violation types

The goal is to minimize this fitness value while ensuring all rooms are used efficiently, capacity constraints are satisfied, and gender-specific room assignments are maintained.

Genetic Operators The room assignment genetic algorithm employs similar operators to the time slot assignment stage, including tournament selection, multi-point crossover, and adaptive mutation. In this stage, the mutation operator randomly alters the *room assignment* of selected exams, reassigning up to 10% of exams to randomly selected rooms from the appropriate gender-specific room set. A key implementation detail is that male and female section populations are evolved separately to maintain gender-specific room constraints efficiently. Each gender-separated population evolves independently using the same genetic operators, and the populations are merged at the end to form complete timetables. This parallel evolution strategy significantly improves computational efficiency and constraint satisfaction for gender-separated facilities.

4.3. Multi-Stage Integration

The two-stage decomposition approach offers several advantages over monolithic approaches:

1. **Reduced Complexity:** By decomposing the problem, each stage deals with a smaller search space, making the optimization more tractable.
2. **Focused Optimization:** Each stage can focus on specific objectives without interference from other constraints, leading to better convergence.
3. **Flexibility:** The decomposition allows for easy adaptation to different institutional requirements by modifying individual stages independently.

4. **Computational Efficiency:** The sequential approach is generally faster than solving the complete problem simultaneously, especially for large-scale instances.

The output of Stage 1 (time slot assignments) is passed as input to Stage 2 (room assignments), and the combined solution represents the complete examination timetable. If Stage 2 fails to find a feasible room assignment, Stage 1 can be re-run with adjusted parameters or the process can iterate between stages until a complete feasible solution is obtained.

4.4. Computational Complexity Analysis

Stage 1 (Time Slot Assignment): Each generation requires fitness evaluation over the population: computing conflict violations for each chromosome costs $O(|E|^2)$ (pairwise conflict check), repeated for population size N, yielding $O(N \cdot |E|^2)$ per generation. Crossover and mutation are $O(|E|)$ per chromosome. The repair mechanism adds $O(|E|^2 \cdot |T|)$ in the worst case. Over G_1 generations, total Stage 1 complexity is $O(G_1 \cdot N \cdot (|E|^2 + |E|^2 \cdot |T|)) = O(G_1 \cdot N \cdot |E|^2 \cdot |T|)$.

Stage 2 (Room Assignment): Each fitness evaluation checks room conflicts and capacity for each exam-room pair: $O(|E| \cdot |R|)$. Over G_2 generations with population N: $O(G_2 \cdot N \cdot |E| \cdot |R|)$.

Total: $O(G_1 \cdot N \cdot |E|^2 \cdot |T| + G_2 \cdot N \cdot |E| \cdot |R|)$. For the PSAU instance ($|E| = 244$, $|T| = 20$, $|R| = 38$, $N = 24$, $G_1 = G_2 = 1000$), this yields approximately 2.8×10^{10} elementary operations, consistent with observed runtimes of 3–10 minutes.

Comparison with monolithic approach: A monolithic GA would require $O(G \cdot N \cdot |E|^2 \cdot |T| \cdot |R|)$ per generation (simultaneous slot-room evaluation), representing a factor of $|R|$ increase in per-generation cost.

5. Experimental Results

5.1. Benchmark Datasets

The proposed multi-stage genetic algorithm was evaluated on several well-known benchmark datasets:

1. **Toronto Benchmark Datasets:** Classic benchmark problems extensively used in examination timetabling research [11, 32, 10].
2. **Prince Sattam Bin Abdulaziz University Dataset:** Real-world data from our institution.

5.2. Experimental Setup

Algorithm Parameters:

The algorithm was tested with two parameter configurations to balance theoretical rigor with practical computational constraints:

Standard Configuration (used for PSAU dataset):

– Population size: 24 chromosomes

- Maximum generations: 1000
- Mutation rate: 0.60 (adaptive)
- Tournament size: 5
- Stagnation limit: 200 generations

Theoretical Configuration (from literature):

- Population size: 100 chromosomes
- Maximum generations: 5000 for Stage 1, 2000 for Stage 2
- Crossover rate: 0.8
- Mutation rate: 0.05
- Tournament size: 5

Parameter Calibration The parameter settings were determined through systematic calibration experiments on a subset of benchmark instances. Table 1 summarizes the calibration results for key parameters.

Table 1. Parameter Calibration Results (PSAU Dataset, 10 runs per setting)

Parameter	Range Tested	Best Value	Criterion
Population size	10, 24, 50, 100	24 (PSAU) / 100 (bench.)	Best fitness / time trade-off
Mutation rate	0.05, 0.10, 0.20, 0.40, 0.60	0.60 (PSAU) / 0.05 (bench.)	Avg. fitness over 10 runs
Tournament size	2, 3, 5, 7	5	Feasibility rate
Stagnation limit	50, 100, 200, 500	200	Convergence time

The higher mutation rate (0.60) for the PSAU configuration is justified by the small population size (24): with fewer individuals, stronger mutation is needed to maintain population diversity and avoid premature convergence. This inverse relationship between population size and optimal mutation rate is well-documented in evolutionary computation literature. For the standard benchmarks with population size 100, the canonical mutation rate of 0.05 proved sufficient for diversity maintenance.

Penalty Coefficients:

Penalty coefficients were calibrated using an iterative approach: initial values were set based on constraint violation severity ratios, then refined through 10-run experiments measuring feasibility rate and solution quality. The final values reflect the minimum penalties needed to ensure $> 90\%$ feasibility:

- Hard constraints (conflicts, 3+ exams/day): 1,000 (set $\geq 10\times$ the maximum possible soft penalty to ensure hard constraint priority)
- Room violations (3+ sections): 100
- Gender mismatch: 100
- Two exams same day: 50
- Consecutive exams: 5
- Two exams in consecutive days: 5
- Room capacity overflow (2 sections): 1 per student
- Room capacity overflow (1 section): 5 per student

Penalty Sensitivity Analysis To assess robustness to penalty coefficient choices, we varied each penalty by $\pm 50\%$ while keeping others fixed. Results on the PSAU dataset (10 runs each) showed: hard constraint penalties ≥ 500 consistently achieved $> 90\%$ feasibility; soft constraint weights affected solution quality by $\leq 8\%$ within the tested range; gender mismatch penalty ≥ 50 maintained 100% compliance. The algorithm is thus robust to moderate penalty variations.

The standard configuration was selected based on the calibration results above, providing a practical balance between solution quality and execution time. The higher mutation rate (0.60) compensates for the smaller population size, maintaining adequate exploration of the solution space.

Computational Environment:

- Processor: Intel Core i7-9700K @ 3.6 GHz
- RAM: 32 GB
- Implementation: Python 3.9
- Number of independent runs: 30 for each instance

MIP Solver Configuration:

- Solver: Gurobi Optimizer 12.0.1 with academic license
- Time limit: 300s (Toronto benchmarks), 600s (PSAU instances)
- MIP gap tolerance: 5%
- Threads: 8 parallel threads
- Presolve: Aggressive (level 2)
- Heuristics: Default Gurobi settings
- The GA was given the same wall-clock time limit as MIP for fair comparison

5.3. Performance Metrics

The performance of the algorithm is evaluated using the following metrics:

1. **Feasibility Rate:** Percentage of runs that produce feasible solutions
2. **Best Fitness:** The lowest fitness value achieved across all runs
3. **Average Fitness:** Mean fitness value across all runs
4. **Standard Deviation:** Measure of solution consistency
5. **Computational Time:** Average execution time to reach the best solution
6. **Soft Constraint Violations:** Breakdown of individual soft constraint violation counts
7. **Minimum, Maximum, and Variance:** Full distributional statistics across independent runs

5.4. Results and Analysis

The proposed multi-stage genetic algorithm achieved competitive results on most benchmark instances, demonstrating its effectiveness in solving the examination timetabling problem. The decomposition approach proved particularly effective for large-scale instances, where the reduced search space in each stage led to faster convergence. The algorithm consistently produced feasible solutions, with a feasibility rate above 90% across all tested instances.

Table 2. Detailed Results with Full Statistics (30 independent runs, Toronto cost metric)

Dataset	Best Known	Min	Mean	Max	Std Dev	Variance	Feasibility
hec-s-92	10.63	14.45	16.22	19.87	1.43	2.04	100%
ute-s-92	25.39	29.29	32.15	37.41	2.18	4.75	100%
PSAU-Male	—	5.04	6.83	9.12	1.07	1.14	92%
PSAU-Female	—	3.98	5.41	7.65	0.95	0.90	93%

Detailed Results with Full Statistics Table 2 presents the complete results with distributional statistics across 30 independent runs.

Convergence Behavior The typical convergence behavior of the algorithm on the PSAU dataset demonstrates that the fitness value shows rapid initial improvement during the first 100-200 generations, followed by gradual refinement. The two-stage approach exhibits distinct convergence patterns:

- **Stage 1 (Slot Assignment):** Rapid convergence within 200-400 generations for most instances, primarily eliminating hard constraint violations
- **Stage 2 (Room Assignment):** Smoother convergence focusing on soft constraint optimization and room utilization efficiency

The stagnation detection mechanism successfully identified convergence plateaus, triggering population mutation to escape local optima. This adaptive strategy contributed to the algorithm's robustness across different problem instances.

Constraint Violation Analysis Analysis of constraint violations across multiple runs revealed:

- **Hard Constraints:** Conflicting slots were eliminated in 92% of runs, with remaining violations typically occurring in highly constrained instances with limited time slots
- **Student Comfort:** The algorithm effectively minimized consecutive exams and same-day exams, improving student satisfaction
- **Room Efficiency:** Room underutilization was maintained below 15% on average, indicating effective capacity management
- **Gender Compliance:** 100% compliance with gender-specific room assignments in all feasible solutions

5.5. Real-World Application

The algorithm was successfully applied to real-world data from Prince Sattam Bin Abdulaziz University. Table 3 presents the characteristics of the PSAU dataset.

Problem Complexity: The PSAU dataset exemplifies a coupled dual-ETP where the solution must simultaneously:

- Assign 124 male exam sections to 25 male building rooms across 20 time slots
- Assign 120 female exam sections to 13 female building rooms across the **same** 20 time slots

Table 3. Prince Sattam Bin Abdulaziz University Dual-Building Dataset Characteristics

Characteristic	Male Building	Female Building	Total/Shared
Unique Students	768	722	1,490
Total Enrollments	2,640	2,555	5,195
Exam Sections	124	120	244
Examination Rooms	25	13	38
Room Capacity Range	25-120	20-100	20-120
Avg. Room Capacity	68.4	56.2	64.1
Shared Schedule Parameters			
Examination Days	10 days		
Time Slots Per Day	2 slots (morning/afternoon)		
Total Time Slots	20 synchronized slots		
Coupling Constraint: Corresponding course exams must share time slots **Independence:** Room resources are completely disjoint			

- Ensure corresponding male/female course sections occupy identical time slots
- Satisfy independent capacity constraints for each building's room set
- Minimize soft constraint violations independently for each building

This structure effectively doubles the problem complexity compared to standard single-resource-pool ETP formulations, as the search space must navigate feasibility across two coupled but resource-independent problems.

The algorithm was executed with the following configuration parameters: population size of 24, 1000 generations, mutation rate of 0.60, and tournament size of 5. These parameters were determined by the calibration procedure described in Section 5.2.

Results on PSAU Dataset The successful solution of this challenging real-world scenario validates both the algorithmic approach and demonstrates practical applicability for institutions with similar cultural or religious requirements for gender-separated educational facilities. To our knowledge, this represents the first application of examination timetabling algorithms to the coupled dual-ETP problem arising from completely separated building infrastructures.

A sensitivity analysis was conducted to understand the impact of algorithm parameters on solution quality. Findings showed that:

- **Population Size:** Increasing population size from 24 to 100 improved solution diversity and exploration capability, though with proportionally increased computational cost. The relationship between population size and solution quality exhibited diminishing returns beyond 100 chromosomes.
- **Mutation Rate:** The mutation rate significantly impacted convergence behavior. Lower rates (0.05-0.10) provided stable convergence for simpler instances, while higher rates (0.40-0.60) proved beneficial for escaping local optima in highly constrained scenarios. The optimal mutation rate varied based on problem complexity and constraint tightness.
- **Tournament Size:** Tournament sizes between 3 and 7 provided good balance between selection pressure and population diversity. Larger tournament sizes accelerated convergence but risked premature convergence to local optima.

– **Generation Limit:** The algorithm showed continued improvement up to 1000-2000 generations for complex instances. Early stopping criteria based on stagnation detection (200 generations without improvement) effectively balanced solution quality with computational efficiency.

5.6. Ablation Study

To isolate the contribution of individual algorithmic components, we conducted an ablation study on the PSAU dataset (30 runs per configuration). Table 4 presents the results.

Table 4. Ablation Study Results (PSAU-Male, 30 runs)

Configuration	Mean Cost	Feasibility	Δ vs Full
Full algorithm	6.83	92%	—
Without graph coloring init.	8.47	78%	+24.0%
Without adaptive mutation	7.91	85%	+15.8%
Without stagnation detection	7.45	88%	+9.1%
Without repair mechanism	9.62	61%	+40.8%
Random init. only + basic GA	11.35	52%	+66.2%

Key findings: (1) The repair mechanism provides the largest individual contribution, improving feasibility by 31 percentage points. (2) Graph coloring initialization improves both feasibility (+14pp) and solution quality (+24%). (3) Adaptive mutation contributes primarily to solution quality refinement. (4) Stagnation detection provides moderate improvement, primarily preventing premature convergence.

5.7. Comparison with Decomposition Strategies

To validate the effectiveness of our two-stage decomposition approach, we compared it with monolithic and alternative decomposition strategies. Results showed that our two-stage approach (time slots $\rightarrow$ rooms) outperformed the monolithic GA in both solution quality and computational time. This validates the design choice of our decomposition strategy, consistent with findings in the literature [19, 10].

Table 5 presents a comparative analysis of different decomposition strategies on the PSAU dataset.

Table 5. Comparison of Decomposition Strategies on PSAU Dataset

Approach	Avg. Fitness	Time (s)	Feasible
Monolithic GA	28,450	847	65%
Two-Stage (Proposed)	21,285	523	92%
Three-Stage	23,120	612	85%

The two-stage decomposition demonstrated:

- **Superior Solution Quality:** 25% improvement in average fitness compared to monolithic approach
- **Higher Feasibility Rate:** 92% feasibility rate vs. 65% for monolithic GA
- **Computational Efficiency:** 38% reduction in computation time
- **Better Convergence:** More consistent convergence to high-quality solutions

The three-stage decomposition (separating time slot assignment, room selection, and room optimization) showed intermediate performance. While it provided additional flexibility, the added coordination overhead between stages resulted in slightly degraded performance compared to our two-stage approach. This confirms that the two-stage decomposition strikes an optimal balance between problem simplification and coordination complexity for examination timetabling problems.

5.8. Comparison with Other Metaheuristic Approaches

To provide broader methodological context beyond GA variants, we implemented simulated annealing (SA) and tabu search (TS) baselines using the same two-stage decomposition framework and evaluated them under identical conditions (same time budget, same evaluation metric, 30 runs). Table 6 reports the results of this comparison.

Table 6. Comparison with Non-GA Metaheuristics (Toronto cost, 30 runs)

Method	**Min**	**Mean**	**Std Dev**	**Feasibility**	**Avg. Time (s)**
hec-s-92					
Proposed GA	14.45	16.22	1.43	100%	604
SA (two-stage)	13.87	15.45	1.21	100%	580
TS (two-stage)	14.02	15.91	1.35	100%	595
ute-s-92					
Proposed GA	29.29	32.15	2.18	100%	518
SA (two-stage)	28.85	31.42	1.98	100%	510
TS (two-stage)	29.51	32.84	2.45	97%	525
PSAU-Male					
Proposed GA	5.04	6.83	1.07	92%	193
SA (two-stage)	5.92	7.85	1.54	87%	210
TS (two-stage)	5.45	7.21	1.32	90%	205
PSAU-Female					
Proposed GA	3.98	5.41	0.95	93%	198
SA (two-stage)	4.65	6.12	1.28	88%	215
TS (two-stage)	4.21	5.78	1.15	91%	208

On the standard Toronto benchmarks, SA achieved slightly better results than the proposed GA, which is consistent with the strong performance of SA-based methods reported in the literature [9, 24, 10]. However, on the PSAU dual-building instances, the proposed GA outperformed both SA and TS baselines by 14–22% in mean cost. This advantage stems from the GA's population-based exploration, which more effectively handles the expanded feasibility landscape created by the coupled dual resource pools. SA's single-solution trajectory is more susceptible to becoming trapped in local optima in the coupled constraint space.

5.9. Comparison with Mixed-Integer Programming

To provide a comprehensive evaluation of our approach, we compared the proposed genetic algorithm with an exact Mixed-Integer Programming (MIP) formulation on both standard benchmarks and real-world institutional data. The MIP solver provides guaranteed optimal solutions for smaller instances but faces scalability challenges for larger problems due to the NP-hard nature of the examination timetabling problem.

Unified Cost Metric for Fair Comparison To ensure fair comparison between methods, we adopt the Toronto proximity cost metric as the unified evaluation standard across all experiments:

$$\text{Cost} = \frac{1}{|S|} \sum_{\Delta=1}^{5} w_{\Delta} \sum_{e_i, e_j : |t_i - t_j| = \Delta} |S_i \cap S_j| \tag{29}$$

where $w = \{16, 8, 4, 2, 1\}$ for gaps $\Delta = \{1, 2, 3, 4, 5\}$ timeslots, S is the set of all students, and $S_i \cap S_j$ represents students enrolled in both exams e_i and e_j.

Critical Methodological Note: The MIP formulation internally optimizes a weighted combination of soft constraints (consecutive exams penalties, period utilization, room utilization) producing objective values that are *not directly comparable* to the Toronto cost metric. Similarly, our GA's internal fitness function uses conflict-based penalties. Therefore, for fair comparison, we extract the slot assignments produced by each method and uniformly recalculate the Toronto proximity cost. This ensures we evaluate the actual timetable quality rather than method-specific internal objectives.

MIP Formulation The MIP model implements the full linearized formulation presented in Section 3 using Gurobi Optimizer 12.0. Due to the computational complexity of simultaneous slot and room optimization, we employed a two-stage approach matching our GA: the MIP optimizes slot assignments with soft constraint objectives, then room assignments are handled greedily. The MIP and GA used identical wall-clock time limits for fair comparison. Both methods used the same two-stage decomposition to isolate the comparison to optimization strategy rather than formulation structure. The model was solved with the following configuration:

- Time limit: 300-600 seconds per instance (identical to GA time budget)
- MIP gap tolerance: 5%
- Threads: 8 parallel threads
- Presolve: Aggressive
- Version: Gurobi 12.0.1 with academic license

Comparative Results Table 7 presents the comprehensive comparison between the MIP solver and our proposed GA using the unified Toronto cost metric across standard benchmarks and real-world datasets.
Notes:

- MIP solved using Gurobi 12.0: time limits reached on all instances
- GA uses two-stage decomposition with 500 generations

Table 7. Unified Comparison: MIP vs GA using Toronto Cost Metric

Dataset	Best Known	MIP Cost	GA Cost	Winner	Margin	Time (s)
Toronto Benchmarks						
hec-s-92	10.63	14.15	14.45	MIP	2.1%	300 / 604
ute-s-92	25.39	31.80	29.29	GA	7.9%	300 / 518
PSAU Real-World Datasets						
PSAU-Female	—	9.03	3.98	GA	55.9%	600 / 198
PSAU-Male	—	14.78	5.04	GA	65.9%	600 / 193
Summary Statistics						
Average	—	17.44	13.19	GA	24.4%	450 / 378

- Time format: MIP time / GA time (in seconds)
- Winner indicates method with lower Toronto cost; Margin shows percentage improvement
- GA wins on 3/4 datasets (75%) with average 24.4% improvement
- Best Known values from literature [32, 33] (not available for PSAU datasets)
- **PSAU-Female:** GA achieves 3.98 vs MIP 9.03 (GA 55.9% better). GA produces significantly higher quality timetables for this real-world instance (120 exams, 722 students) while using only 33% of MIP's computation time (198s vs 600s).
- **PSAU-Male:** GA achieves 5.04 vs MIP 14.78 (GA 65.9% better). GA's dominance is even more pronounced on this instance (124 exams, 768 students), delivering solutions with 2.9x lower proximity cost in 32% of MIP's time (193s vs 600s).

Overall Findings: The proposed genetic algorithm consistently outperforms the exact MIP approach on both standard benchmarks and real-world datasets when evaluated using the unified Toronto cost metric. The GA demonstrates superior solution quality, particularly on larger and more complex instances where MIP struggles with scalability. Additionally, the GA achieves these results with significantly lower computational time, highlighting its practical applicability for real-world examination timetabling problems.

6. Conclusion

This paper has presented both a focused review of examination timetabling research and a two-stage genetic algorithm approach with linearized mathematical formulation. This systematic analysis identified persistent challenges and research gaps, particularly regarding integration of exact and metaheuristic approaches, accommodation of institution-specific constraints, and balance between theoretical advances and practical applicability.

The experimental results demonstrate that our proposed multi-stage genetic algorithm achieves competitive performance on standard benchmarks and superior performance on the coupled dual-ETP instances arising from PSAU's gender-separated facilities. The algorithm particularly excels on large-scale instances where the decomposition strategy effectively manages problem complexity. Comparisons with MIP, simulated annealing, and tabu search baselines confirm the GA's advantages for coupled heterogeneous resource problems.

Limitations and Future Work. The current study has several limitations. First, the benchmark evaluation is limited to two Toronto instances; broader evaluation on the complete Toronto and ITC-2007 suites would strengthen generalizability claims. Second, the two-stage decomposition sacrifices potential global optimality for tractability; iterative feedback between stages could partially recover this loss. Third, the comparison with SA and TS used our own implementations rather than state-of-the-art tuned versions from the literature.

For reproducibility, the GA and MIP implementations along with the PSAU dataset are publicly available at `https://github.com/chebil/ETP-Code-Data`.

Several directions merit further investigation. First, extending the dual-ETP model to *partial resource sharing* scenarios (where some rooms are accessible to both populations) would broaden applicability to multi-campus settings with swing rooms. Second, integrating *invigilator assignment* as a third optimization stage would address workload balance and rest-period constraints, which are practical concerns in the PSAU setting. Third, replacing the static operator configuration with *adaptive operator selection* [31] could improve robustness across diverse instance types without instance-specific tuning. Fourth, a *hybrid MIP-GA* approach—where the GA handles time slot assignment and a compact MIP solves room assignment to optimality—could combine the strengths of both paradigms. Finally, evaluation on the complete *ITC-2007 benchmark suite* [12] would strengthen generalizability claims and enable broader comparison with the literature.

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work the authors used ChatGPT, a large language model-based chatbot maintained by OpenAI, for language proofreading. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Acknowledgement. This project was supported by the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University under research project No. 2022/01/19335.

References

1. Abdullah, S., Alzaqebah, M.: A hybrid self-adaptive bees algorithm for examination timetabling problems. Applied Soft Computing 13, 3608–3620 (04 2013)
2. Abdullah, S., Turabieh, H.: On the use of multi neighbourhood structures within a tabu-based memetic approach to university timetabling problems. Information Sciences 191, 146–168 (2012), `https://www.sciencedirect.com/science/article/pii/S0020025511006670`, data Mining for Software Trustworthiness
3. Abou Kasm, O., Mohandes, B., Diabat, A., Khatib, S.: Exam timetabling with allowable conflicts within a time window. Computers & Industrial Engineering 127 (11 2018)
4. Al-Betar, M., Khader, A.T., Abu Doush, I.: Memetic techniques for examination timetabling. Annals of Operations Research 218, 23–50 (07 2013)
5. Al-hawari, F., Al-Ashi, M., Abawi, F., Alouneh, S.: A practical three-phase ilp approach for solving the examination timetabling problem. International Transactions in Operational Research 27 (11 2017)
6. Alinia Ahandani, M., Vakil Baghmisheh, M.T., Badamchi Zadeh, M.A., Ghaemi, S.: Hybrid particle swarm optimization transplanted into a hyper-heuristic structure for solving examination timetabling problem. Swarm and Evolutionary Computation 7, 21–34 (2012), `https://www.sciencedirect.com/science/article/pii/S2210650212000417`

7. Arbaoui, T., Boufflet, J.P., Moukrim, A.: Lower bounds and compact mathematical formulations for spacing soft constraints for university examination timetabling problems. Computers & Operations Research 106, 133–142 (2019), https://www.sciencedirect.com/science/article/pii/S0305054819300528
8. Arbaoui, T., Moukrim, A.: Preprocessing and an improved mip model for examination timetabling. Annals of Operations Research 229 (06 2015)
9. Battistutta, M., Schaerf, A., Urli, T.: Feature-based tuning of single-stage simulated annealing for examination timetabling. Annals of Operations Research pp. 239–254 (01 2017), https://doi.org/10.1007/s10479-015-2061-8
10. Bellio, R., Ceschia, S., Di Gaspero, L., Schaerf, A.: Two-stage multi-neighborhood simulated annealing for uncapacitated examination timetabling. Computers & Operations Research 132, 105300 (04 2021)
11. Burke, E., Kendall, G., Mısır, M., Özcan, E.: Monte carlo hyper-heuristics for examination timetabling. Annals of Operations Research 196, 73–90 (04 2012)
12. Burke, E., Pham, N., Qu, R., Yellen, J.: Linear combinations of heuristics for examination timetabling. Annals OR 194, 89–109 (04 2012)
13. Burke, E.K., Bykov, Y.: The late acceptance hill-climbing heuristic. European Journal of Operational Research 258(1), 70–78 (2017), https://www.sciencedirect.com/science/article/pii/S0377221716305495
14. Carlsson, M., Ceschia, S., Gaspero, L., Ã˜rnstrup Mikkelsen, R., Schaerf, A., Stidsen, T.J.R.: Exact and metaheuristic methods for a real-world examination timetabling problem. Journal of Scheduling 26(4), 353–367 (August 2023), https://ideas.repec.org/a/spr/jsched/v26y2023i4d10.1007_s10951-023-00778-6.html
15. Cataldo, A., Ferrer, J.C., Miranda, J., Rey, P., Sauré, A.: An integer programming approach to curriculum-based examination timetabling. Annals of Operations Research 258, 369–393 (11 2017)
16. Çimen, M., Belbağ, S., Soysal, M., Sel, Ç.: Invigilators assignment in practical exam timetabling problems. International Journal of Industrial Engineering: Theory, Applications and Practice 29(3) (Jun 2022), https://journals.sfu.ca/ijietap/index.php/ijie/article/view/6943
17. Demeester, P., Bilgin, B., De Causmaecker, P., Vanden Berghe, G.: A hyperheuristic approach to examination timetabling problems: Benchmarks and a new problem from practice. J. Scheduling 15, 83–103 (02 2012)
18. Elloumi, A., Kamoun, H., Jarboui, B., Dammak, A.: The classroom assignment problem: Complexity, size reduction and heuristics. Applied Soft Computing 14, Part C, 677 – 686 (01 2014)
19. Gogos, C., Alefragis, P., Housos, E.: An improved multi-staged algorithmic process for the solution of the examination timetabling problem. Annals OR 194, 203–221 (04 2012)
20. Güler, M.G., Akyer, H., Keskin, M., Döyen, A.: Examination timetabling problem with scarce resources: a case study. European J. of Industrial Engineering 12, 1 (01 2018)
21. Kahar, M.M., Kendall, G.: Universiti malaysia pahang examination timetabling problem: scheduling invigilators. The Journal of the Operational Research Society 65(2), 214–226 (2014), http://www.jstor.org/stable/24502036
22. Lei, Y., Gong, M., Jiao, L., Shi, J., Zhou, Y.: An adaptive coevolutionary memetic algorithm for examination timetabling problems. International Journal of Bio-Inspired Computation 10, 248 (01 2017)
23. Lei, Y., Shi, J., Yan, Z.: A memetic algorithm based on moea/d for the examination timetabling problem. Soft Computing 22 (03 2018)
24. Leite, N., Melicio, F., Rosa, A.: A fast simulated annealing algorithm for the examination timetabling problem. Expert Systems with Applications 122 (12 2018)
25. Mukhlason, A., Parkes, A., Özcan, E., Mccollum, B., McMullan, P.: Fairness in examination timetabling: Student preferences and extended formulations. Applied Soft Computing 55 (01 2017)

26. Nand, R., Sharma, B., Chaudhary, K.: An introduction of preference based stepping ahead firefly algorithm for the uncapacitated examination timetabling. PeerJ Computer Science 8, e1068 (09 2022)
27. Pais, T., Amaral, P.: Managing the tabu list length using a fuzzy inference system: An application to examination timetabling. Annals OR 194, 341–363 (04 2012)
28. Pillay, N.: Evolving hyper-heuristics for the uncapacitated examination timetabling problem. Journal of the Operational Research Society 63 (04 2012)
29. Pillay, N., Özcan, E.: Automated generation of constructive ordering heuristics for educational timetabling. Annals of Operations Research 275 (04 2019)
30. Sabar, N., Ayob, M., Kendall, G., Qu, R.: A honey-bee mating optimization algorithm for educational timetabling problems. European Journal of Operational Research 216, 533–543 (02 2012)
31. Sabar, N., Ayob, M., Kendall, G., Qu, R.: Grammatical evolution hyper-heuristic for combinatorial optimization problems. IEEE Transactions on Evolutionary Computation 17 (12 2013)
32. Sabar, N., Ayob, M., Qu, R., Kendall, G.: A graph coloring constructive hyper-heuristic for examination timetabling problems. Applied Intelligence - APIN 37, 1–11 (07 2011)
33. Van Bulck, D., Goossens, D., Schaerf, A.: Multi-neighbourhood simulated annealing for the itc-2007 capacitated examination timetabling problem. Journal of Scheduling (01 2023), https://doi.org/10.1007/s10951-023-00799-1

Dr. Khalil Chebil is an Assistant Professor of Computer Science whose research focuses on intelligent algorithms for combinatorial optimization, with contributions spanning exact, heuristic, and matheuristic methods applied to knapsack variants, UAV-based multi-period crowd tracking, and GPU-parallel approximate methods. He holds a Ph.D. in Computer Science from the University of Sfax (2016) and serves at Prince Sattam bin Abdulaziz University (Saudi Arabia) and INSAT, University of Carthage (Tunisia).

Dr. Mahdi Khemakhem is a Full Professor of Computer Science at the University of Sfax, Tunisia, and Prince Sattam bin Abdulaziz University, Saudi Arabia, holding a Ph.D. from the Polytechnic University of Hauts-de-France, France. His research focuses on combinatorial optimization and metaheuristics — particularly knapsack problems, vehicle routing, and traveling salesman problems — and extends to AI-driven optimization for cloud and fog computing, virtual machine placement, UAV-based systems, IoT architectures, and healthcare applications including federated machine learning.

Mr. Abdallah Hammad is a Data Scientist at Entropy in Riyadh, Saudi Arabia, where he is working on a comprehensive Arabic LLM benchmark spanning language understanding, retrieval-augmented generation, safety, and speech, among others. He completed the present work during his undergraduate studies in Computer Science at Prince Sattam Bin Abdulaziz University, where he designed genetic algorithms for examination timetabling. Prior to his current role, he worked as an AI Engineer building and deploying production systems, including OCR pipelines and real-time voice support agents, while leading technical delivery across an eight-person team. His current work focuses on NLP, model evaluation, and applied AI.

Mr. Mohammed Alanazi is a Computer Science graduate from Prince Sattam Bin Abdulaziz University with a strong foundation in artificial intelligence and front-end development. He is experienced in developing web-based applications and has a demonstrated ability to collaborate effectively using version control tools such as GitHub.

Received: December 20, 2025; Accepted: May 22, 2026.

Computer Science and Information Systems 23(3):1053-1076 https://doi.org/10.2298/CSIS251109030P

Role of Software Metrics in Practical Quality Management

Filip Prentović[1], Zoran Budimac[1], Marjan Heričko[2], and Gordana Rakić[1]

[1] University of Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics
Trg Dositeja Obradovica 4, Novi Sad 21000, Serbia
filipprentovic@yahoo.com (corresponding author),
zjb@dmi.uns.ac.rs, gordaa.rakic@dmi.us.ac.rs
[2] University of Maribor, Faculty of Electrical Engineering and Computer Science,
Smetanova 17, 2000 Maribor, Slovenia
marjan.hericko@um.si

Abstract. The goal of software metrics is to provide continuous insight into products and processes. Software product and software development process are tightly coupled since high-quality software products represent the result of a high-quality process. To understand, predict, and evaluate software development and maintenance process, it is widely recognized that software metrics, as an important technique in the static and dynamic analysis, should be integral part of these processes. Software metrics are a critical tool for building reliable software and assessing software quality, especially in complex domains and/or mission-critical environments. The aim of the research conducted in this paper is to determine the existence of different perspectives regarding the impact of software metrics on software quality between participants in software development. The contribution of this paper consists of analyzing the impact that the application of software metrics has on the quality of software and on the overall project success by statistically analyzing data obtained from the survey. Overall, there were 108 respondents from ten countries in the survey. The paper contains a comprehensive survey of code analysis techniques, software metrics, and the analysis of the application of software metrics in practice. Also, importance, influence, and the level of usage of software metrics in IT companies have been considered, as well as the estimation of results significance obtained in the process of measuring software.

Keywords: Code Analysis, Software Metrics, Software Quality.

1. Introduction

Software quality is a term that has different meanings to different people involved in creating, maintaining, and using software. Its presence can be difficult to define, but its absence can be easy to see instantly. There are many different definitions of quality. According to some definitions, it is the "capability of a software product to conform to requirements" [1], while for others it can be synonymous with "customer value" [2] or even defect level.

The main quality attribute of each software product is correctness. Without correctness, it is pointless to talk about other quality attributes such as functionality, usability, reliability, efficiency, portability, compatibility, security, and maintainability [3].

These quality attributes can be analyzed, controlled, and monitored at the early stages of software development by examining source code and other artifacts (software architecture, system specification, etc.), or later during the execution of the system. These activities should ideally, be performed by all the parties involved in the process of software development: developers, testers, QA (quality assurance) analysts, and managers. Those activities are performed statically and dynamically, depending on the phase and state of the project. Evaluation of software quality attributes that is performed on a source code or any of its internal representations (like syntax trees, graphs and various meta-models which represent abstract syntactic structure of source code written in a programming language) without executing the program belongs to a domain of *static analysis*, while analysis of a software product during its execution represents *dynamic analysis*. Both kinds of code analysis rely on various metrics, with growing importance of static analysis and software metrics since a lot of emphasis is being put on monitoring and quality control in the early stages of the development [3].

The goal of software metrics is to provide continuous insight into products and processes. Software products and their development process are tightly coupled since high-quality software products represent the result of a high-quality process. To understand, predict, and evaluate software development and maintenance process, it is widely recognized that software metrics, as an important tool in static and dynamic analysis, should be an integral part of these processes. Software metrics are a critical tool for building reliable software and assessing software quality, especially in complex domains and/or mission-critical environments. A useful metric typically performs a calculation to assess the effectiveness of the underlying software or process [4]. An increasing need for measurement data used in decision-making on all levels of the organization has been observed when implementing the software process. Although the benefits of measurement in the software process are indisputable, the popularity of the use of measurement in practice is rather limited [5]. It is expected that the measurement provides reliable and precise information, but at the same time to be cost-effective and unassuming during the software development process.

Measuring software quality is motivated by at least two reasons [6]:

- Risk management – software errors are not just a nuisance – they can cause serious damage, especially in the embedded systems used in medical devices, airplanes etc.
- Cost management – recent report revealed that in 2022, software failures cost the economy US$2.41 trillion in financial losses [7]. If the cost of fixing a requirements error, discovered during the requirements phase, is defined to be one unit, the cost to fix that error if found during the design phase increases to three to eight units; at the manufacturing/build phase, the cost to fix the error is 7–16 units; at the integration and test phase, the cost to fix the error becomes 21–78 units; and at the operations phase, the cost to fix the requirements error ranged from 29 units to more than 1500 units [8].

This paper investigates whether different stakeholders involved in software development hold differing perspectives on the impact of software metrics on software quality. The study focuses on examining how the use of software metrics influences both software quality and overall project success. To achieve this, a statistical analysis was conducted on data collected through a survey involving 108 respondents from ten different countries. In addition, the paper provides a comprehensive overview of code analysis techniques and

commonly used software metrics, along with an examination of how these metrics are applied in real-world practice. Particular attention is given to the perceived importance, influence, and level of adoption of software metrics within IT companies. The research also includes an assessment of the statistical significance of the results obtained through software measurement. Finally, the findings of this study are compared with results from similar surveys conducted in previous years in order to identify trends and changes over time.

This paper is organized as follows. Section Related Work contains references to the similar surveys done before, as well software metrics reviews. Definitions and standards of software quality, overview of code analysis techniques, and relation between software metrics and software quality is presented in the next section. Section Survey and Results describes survey, and its results and analysis. Survey results are compared to results of the previously conducted surveys and discussed and analyzed in section Discussion and analysis. Section Conclusion contains concluding remarks.

2. Related Work

Since this paper contains comprehensive review of software metrics, the role of software metrics in measuring software quality, and analysis of survey of software metrics usage, this section contains an overview of the papers reviewing existing software metrics, and papers in which surveys on software metrics usage were conducted. Finally, the section highlights the novelty of this study in comparison to all referred ones.

Study conducted by Setiadi [9] investigates the measurement of software quality in the Academic Information System at Marshal Suryadarma Aerospace University. The research considers both direct metrics, such as cost, code size, execution speed, and documentation, and indirect metrics related to functionality, efficiency, dependability, and maintainability, transforming these metrics into indicators for evaluating software quality. The study aims to assist university administration in monitoring and improving their information systems, while providing a comprehensive understanding of how metrics and indicators can be applied to assess and enhance the quality of academic software.

Dynamic metrics, which are usually obtained from the execution traces of the code or from the executable models are analyzed by Chabbra and Gupta [10]. In this paper advantages of dynamic metrics over static metrics are discussed and then a survey of existing dynamic metrics is carried out. These metrics are grouped to different categories, such as dynamic coupling metrics and dynamic cohesion metrics. Towards end of the paper, potential research challenges and opportunities in the field of dynamic metrics are identified.

Srinivisan et al. [11] analyze and review the most referred object-oriented metrics in software measurement. The paper presents formal definitions of most important categories of object-oriented metrics and their characteristics, while proposing suite for measuring object-oriented design attributes.

In his research, Lee [12] builds an appropriate method of software quality metrics application in quality life cycle with software quality assurance. Successful software quality assurance is highly dependent on software metrics, and it needs a link between the software quality model and software metrics. This link is obtained by using quality factors to offer measure method for software quality assurance. The paper defines some software

metrics and discusses several software quality assurance models and some quality factors measuring methods.

Pavlič et al. [13] conducted the survey to determine how and to what extent the quality of software is measured in Slovenia. Results of the survey were compared to the similar surveys previously taken and the conclusion was that they do not differ significantly. Larger deviations were only found by combining obtained quality estimate into overall quality estimate, which is a more common practice elsewhere than in Slovenia.

Study conducted by Paulinus [14] looks at the role of transparency as a non-functional requirement in the software engineering process, addressing the lack of appropriate measurement methods and empirically validating metrics for evaluating and improving transparency throughout the software development lifecycle. The study investigates correlations between transparency factors and metrics, and proposes a Goal Question Metric (GQM)-based transparency evaluation and improvement model. Results from a controlled experiment demonstrate that enhanced transparency improves maintainability of software requirements specifications, supports better communication, and increases stakeholder productivity, offering practical guidance for early phases of requirements engineering and design.

In the study taken by Tahir et al. [15], a model of 18 success factors for implementing measurement processes is proposed, while a survey is conducted to evaluate the state of measurement practices and to validate the proposed model based on the feedback from 200 software professionals working in Pakistani software industry. The survey showed that, for example, only 10% software organizations in this survey follow any measurement model, that 75% organizations do not follow any measurement standard, and that there are 80% software organizations in Pakistan which do not use any measurement tool. The paper presented mitigation strategies for key issues identified in software organizations.

In his research, Kasunic [16] conducted a comprehensive survey consisting of 17 questions with the help of random sampling, with individuals from 84 countries responded. The aim of the research was to find out: a) what measurement definition and implementation approaches are being adopted and used by the community, b) the most prevalent types of measures being used by organizations that develop or acquire software, and c) what behaviors are preventing the effective use of measurement. Among other things, it was observed that administration and staff have different opinions regarding measurement process, and that the larger organizations had better measurement infrastructure.

The goal of the research taken by Chen et al. [17] was to investigate how the use of large language models (LLMs) can distort traditional size-based software metrics, particularly lines-of-code (LOC), and to assess the practical applicability of simple versus complex efficiency metrics across multiple software repositories. Additionally, the research aimed to evaluate the effectiveness and cost-efficiency of static analysis tools such as SonarQube and PMD in detecting commits that introduce significant changes to software quality. Finally, the goal was to validate these findings in an industrial context through a case study in a large financial-sector organization, which included a rapid review of existing practices and the development of a software metrics migration plan.

Eisty and colleagues [4] conducted a survey to gather information about research software developers' knowledge and use of code metrics and software process metrics. The survey results, from 129 respondents, indicated that respondents have a general knowledge of metrics. However, their knowledge of specific metrics was lacking and their use even

more limited. Software developers appear to be interested and see some value in software metrics but may be encountering roadblocks when trying to use them.

The research of Atanasova and Temole [18] is addressing the persistent challenge of implementing software quality models in complex organizational environments, such as large companies undergoing agile transformation. It proposes a pragmatic, value chain-oriented Quality Gate Framework that operationalizes software quality by integrating measurable quality indicators across both product and process dimensions. Empirically grounded through 58 expert interviews, the framework improves transparency, stakeholder alignment, and quality assurance while remaining compatible with agile and DevOps practices, offering a transferable approach for organizations seeking sustainable improvements in software development outcomes.

Survey conducted by Padmini et al. [19] is characteristic because it is only concerned with the use of metrics in the agile development process. Aim of the research was to explore metrics suitable for the agile development process, use of those metrics in practice, perceived benefits, and related tools. The survey and interview-based analysis of 24 development companies in Sri Lanka identified ten metrics that can be beneficial to the agile development process, where their benefits outweigh the overheads involved.

Schaffernak et al. [20] tried to track and measure significant software quality attributes to quantify the current state of a system and support strategic and technical decisions. They analysed existing measures related to the ISO/IEC 25022:2016 quality in use (QiU) model to evaluate sustainability aspects and conducted a systematic literature review of 961 articles, followed by internal evaluations of the identified measures. Their results identified 100 measures covering economic, social, technical, and environmental sustainability, showing a strong focus on economic sustainability and underrepresentation of other dimensions, providing guidance for enhancing the QiU model in a more systematic and sustainable way.

Colakoglu et al. [21] review the present studies in the area of software product quality metrics (SPQM) to allow for the analysis of the situation at hand, as well as to predicting future research areas. Paper presents research aims to analyze the active research areas and trends on this topic appearing in the literature between 2009 and 2019. A Systematic Mapping (SM) study was carried out on 70 articles and conference papers published between 2009 and 2019 on SPQM as indicated in their titles and abstract. The result is presented through graphics, explanations, and the mind mapping method, as well as trend map, knowledge about this area and measurement tools, issues determined to be open to development in this area, and conformity between conference papers, articles and internationally valid quality models.

Research conducted by Molnar and colleagues [22] is proposing a comprehensive study of software metric values in the context of three complex, open-source applications. Their methodology and tooling is aligned with that of existing research, and presented in detail in order to facilitate comparative evaluation. Metric values obtained during the entire 18-year development history of the target applications were studied, in order to capture the longitudinal view that was found lacking in the existing literature. Also, metric dependencies were identified and their consistency checked across applications and their versions, along with comparative evaluation with existing research.

The aim of the paper by Mishra et al. [23] is to systematically investigate research on software metric threshold calculation techniques, which were identified as important in

evaluating different aspects of quality. In the study, electronic databases were systematically searched for relevant papers; 45 publications were selected based on inclusion/exclusion criteria, and research questions were answered. Results showed that most of the methods proposed to calculate the thresholds are based on statistical analysis techniques, and that the field of identifying the object-oriented metrics threshold values has received more attention recently.

Research conducted by Vogel at al. [24] studies metrics used in the automotive sector and the quality attributes they address. The HIS, ISO/IEC 25010:2011, and ISO/IEC 26262:2018 are utilized to draw a big picture illustrating (i) which metrics and boundary values are reported in literature, (ii) how the metrics match the standards, (iii) which quality attributes are addressed, and (iv) how the metrics are supported by tools. Most of the identified metrics are concerned with source code, are generic, and not specifically designed for automotive software development. Research concludes that although many metrics exist, a clear definition of the metrics' context, notably regarding the construction of flexible and efficient measurement suites, is missing.

Finally, Rashid and colleagues [25] describe how the software metrics affect the quality of the software and in which stages of its development software metrics have applied. Also, different software metrics are described, along with the impact these metrics have on software quality and reliability, specifically on the aspects of improving the quality of software and increasing the revenue.

The research conducted in this paper differs from previously mentioned papers by aiming to establish a statistical correlation on the impact of using software metrics to software quality and success of software product. Specifically, aim of the survey is to show potentially different perspectives on software metrics usage and understanding between managers and software development team members. Also, it contains a detailed review of software quality, code analysis in general, and software metrics, giving the comprehensive assessment of these topics.

3. Software Quality, Code Analysis, and Software Metrics

The concept of software quality is very broad, spanning through various aspects of software product and its development process. Therefore, it is useful to consider different perspectives on this concept. Measuring software quality is important because it enables creating a baseline for quality, associating specific costs with the improvement of quality, as well as presenting the level of quality on which further improvements can be made [26]. During time, several software metrics were introduced in order to capture various aspects of software, including software quality characteristics.

3.1. Software Quality – Definitions and Standards

Various software quality models have been proposed to define quality. A global initiative for describing the concept of software quality taken by experts around the world led to the standardization of the quality concept by the ISO (the International Organization for Standardization) in the form of documents ISO 9126 [27] and ISO 9000:2000 [28], whereas ISO 9126 is about the quality characteristics of a software product, and the ISO 9000:2000 document is a quality assurance standard mostly directed to processes [26].

As a result of cooperation between ISO and IEC (the International Electrotechnical Commission) in the field of standardization, ISO/IEC 25000 standard, based on ISO 9126, was proposed [28]. ISO/IEC 25000 is a series of standards under the general title Systems and Software Quality Requirements and Evaluation (SQuaRE). New standardization uses new definition of software quality and related terms, defining them as an expression of a degree to which a software product satisfies stated and implied needs when used under specified conditions.

As per ISO 9126 standard, quality is defined as "the totality of features and characteristics of a product or service that bears on its ability to satisfy given needs" [1]. This definition refers to the level of compliance of requirements.

Software quality has also been defined in terms of two types of product characteristics [29]:

- external quality (how the product works in its environment), such as, usability and reliability.
- internal quality (how the product was developed), such as, software structure and complexity.

In other words, software quality refers to two related but distinct notions [30]:

- software functional quality reflects how well it complies with or conforms to a given design, based on functional requirements or specifications. That attribute can also be described as the fitness for purpose of a piece of software or how it compares to competitors in the marketplace as a worthwhile product. It is the degree to which the correct software was produced.
- software structural quality refers to how it meets non-functional requirements that support the delivery of the functional requirements, such as robustness or maintainability. It has a lot more to do with the degree to which the software works as needed.

In software, the narrowest sense of product quality is commonly recognized as the lack of "bugs" in the product. It is also the most basic meaning of conformance to requirements, because if the software contains too many functional defects, the basic requirement of providing the desired function is not met. This definition is usually expressed in two ways: defect rate (e.g., number of defects per million lines of source code, per function point, or other unit) and reliability (e.g., number of failures per n hours of operation, mean time to failure, or the probability of failure-free operation in a specified time) [31].

3.2. Software Quality Analysis

Monitoring and controlling software quality are key processes in software development lifecycle, so it is important to assess software quality from the early stages of software development. It is achieved by analyzing the source code and other artifacts throughout the entire life cycle of the observed software. Assessment of quality attributes of software that is made on the source code or any of its internal representations without executing the program is called static analysis, while analysis of the program during execution time is called dynamic analysis. Numerical values reflecting a software quality characteristic or some of its aspects, and that are obtained by software analysis, are known as software metrics [3].

Static analysis can be defined as a set of techniques for program analysis that do not require software to execute to obtain needed information. Various intermediate representations of a source code, like syntax trees and graphs are being used for the purpose of analyzing code in this manner. Since it is crucial to obtain information on software quality in the early phases of software development, static analysis becomes more important in modern software development methodologies.

Static analysis plays an important role in software evolution and during maintenance phase of software development process. The term *evolution* generally refers to progressive change in software properties or characteristics. This process of change in one or more of their attributes leads to the emergence of new properties or, in some sense, to improvements. These changes affect both functionality and quality of the software product. The related concepts also include software reengineering and software maintenance. *Software reengineering* implies that new user-required features are added to existing software. *Software maintenance* is defined as the modification of a software product after delivery to correct faults, to improve performance or other attributes, or to adapt the product to a changed environment [32]. These changes usually do not affect functionality of a program, but they can have an influence on its quality. Since software quality can be compromised during software evolution and maintenance, all these changes need to be closely controlled and monitored.

Most of the static analysis techniques produce quantitative data along the way or are focused on expressing some program properties. Some of these techniques rely on other methods and corresponding metrics.

Dynamic program analysis represents analysis of properties of a running program [33]. In general, dynamic analysis involves capturing of dynamic state of the program. It usually comprises the analysis of a system's execution through interpretation (e.g., using the Virtual Machine in Java) or instrumentation. After these activities, the resulting data are used for such purposes as reverse engineering and debugging [34].

Dynamic program analysis provides a means to overcome the shortcomings of static analysis. It is effective in examining the actual, exact runtime behavior of a program, as it leaves no uncertainties about the control flow path taken, the values stored in variables, the status of the memory, the time of the execution of the program, etc. Dynamic analysis is as fast as the program being analyzed, which is usually not possible with static analysis techniques. On the other hand, dynamic analysis incurs large runtime overheads, while results obtained during dynamic analysis cannot be generalized for future executions [35].

3.3. Software Metrics

There are several techniques and methods that can be used in a program analysis. Since the survey taken in this research considers using software metrics in the software development process, and the role of software metrics in assessing software quality, software metrics will be described in detail in this section.

Measuring and providing means for quantitative analysis of software is crucial for the overall success of software development process, and consequently, the product. Software metrics provide a quantitative basis for planning and predicting software development processes [36]. By using software metrics during software development process, the quality of software can be controlled and improved easily. This means better productivity

and maintainability of software product, improving overall process and resulting in more reliable and better software.

Software metrics can be defined as numerical values that reflect the properties of a software development processes and software products [37]. Software metrics have been used for a long time: there were attempts of applying software metrics in the 1960s [38]. In the beginning, software execution time was measured and compared to gain insight in the performance of the running programs. As the complexity of software systems increased, software metrics was introduced to measure complexity and the level of cohesion between different parts of software. Measuring characteristics of software design became more important when it was observed that quality of software design corresponds to the number of errors discovered in the latter stages of software development, hence influencing the maintenance costs.

At the highest level, software metrics can be classified into three categories: product metrics, process metrics, and project metrics [31]. Product metrics describe the characteristics of the product. Some of the product metrics refer to characteristics visible to the users (performance, design features etc.), and they are called external metrics. Product metrics related to the characteristics visible only to the development team (size, complexity, structure etc.) are called internal metrics. Process metrics are related to the characteristics of process and can be used to improve software development and maintenance. Examples include the effectiveness of defect removal during development, the pattern of testing defect arrival, and the response time of the fix process. Project metrics describe the project characteristics and execution. Examples include the number of software developers, the staffing pattern over the life cycle of the software, cost, schedule, and productivity. Some metrics belong to multiple categories. For example, the in-process quality metrics of a project are both process metrics and project metrics.

Software quality metrics are a subset of software metrics that focus on the quality aspects of a product, process, and a project. In general, software quality metrics are more closely associated with process and product metrics than with project metrics. However, the project parameters such as the number of developers and their skill levels, the schedule, the size, and the organization structure certainly affect the quality of the product. Software quality metrics can be divided further into end-product quality metrics and in-process quality metrics. The essence of software quality engineering is to investigate the relationships among in-process metrics, project characteristics, and end-product quality, and, based on the findings, to engineer improvements in both process and product quality [31].

Size of the software product can, in most fundamental form, be represented by the number of lines of code. During the years, a class of Lines of Code (LOC) metrics was proposed, containing several variations of this basic measure of length of code [39] [40]:

- SLOC: Source Line of Code reflecting number of lines containing source code without empty lines and lines containing comments.
- CLOC: Comment Line of Code reflecting number of lines containing comments.
- BLOC: Blank Lines of Code reflecting number of lines containing neither source nor comment.
- PLOC: Physical Line of Code reflecting size as it is without considering programming style or code formatting.
- LLOC: Logical Line of Code reflecting size after applying codding style rules or formatting and logical restructuring of statements over the lines.

- EPLOC: Executable Physical Lines of Code reflecting number of lines of source code minus blank lines and comment lines.
- ELLOC: Executable Logical Lines of Code reflecting number of statements which are executed.

Some of the other metrics that have been proposed are Halstead metrics (H) [41], Cyclomatic complexity (CC) [42], Chidamber and Kemerer [43], Lorenz and Kidd [44], Morris [45], Li and Henry [46] and others.

Practical influence of using software metrics cannot be analyzed without considering programming languages, technologies, and paradigms which are used in software creation. Some of the metrics described previously (such are LOC metrics family) are programming language dependent, i.e., these metrics do not consider programming language specificities, which makes comparing results obtained from calculating metrics in heterogenous systems challenging. Also, most of the frameworks and tools that were created to calculate software metrics are language-specific, although there were some attempts in introducing language-independent frameworks [47] [48].

Since most of the metrics are used predominantly in certain phases of software development, software metrics usage cannot be analyzed without considering underlying software development process model and methodology. Iterative and incremental nature of some of these models (e.g., agile software development) can make use of software metrics that are regularly used in other software development processes challenging. In some cases, these metrics are not even applicable [19]. Also, some of the software metrics are meant to be used only within the certain software development methods [49].

Not all the software metrics are of the equal importance to all the participants in the process of the software creation, which is closely related. Also, software metrics serve different purpose to the interested parties. Managers can use software metrics to identify, prioritize, track, and communicate any issues to promote better team productivity. This enables effective management and allows assessment and prioritization of problems within software development projects. The sooner managers can detect software problems, the troubleshooting process is easier and less expensive. On the other hand, software development teams, consisting of developers, testers, QA analysts etc. can use software metrics to communicate the status of software development projects, pinpoint and address issues, and monitor, improve on, and better manage their workflow.

The aim of the survey conducted during the research described in this paper was to reveal the degree of usage of software metrics in the IT industry. Also, importance and relevance of software metrics in the software development process were described. These features were analyzed considering diversity in programming languages, technologies, methodologies, and roles of the participants in the process of software creation. Finally, the survey tried to find out if there are different perspectives on software metrics usage and understanding between managers and software development team members.

4. Survey and Results

Data were collected using an anonymous online survey administered via Google Forms, containing 26 questions divided into three sections, which can be found in the Appendix.[3]

[3] Google Forms link

The survey link was distributed electronically to various companies, mostly in Serbia, and other countries, through members of SQAMIA initiative (http://sqamia.org/). A non-probability convenience sampling strategy was employed, whereby companies that were accessible to the researchers were invited to participate. Participation was voluntary and respondents were informed that no personally identifiable information would be collected. All responses were recorded anonymously and treated confidentially. In total, 108 respondents from ten countries responded to the survey. The characteristics of the respondents and the companies for which they work are shown in Table I.

Most of the respondents (43%) responded that they have average knowledge of software metrics in general (Fig 1.), and 36% of the respondents had a low and very low level of knowledge. This was further corroborated by the fact that most of the respondents did not specify any software metrics they are familiar with, or they specified only a couple of metrics (Lines of Code metric being the most prevalent among those), while the minority of respondents had a thorough list of metrics in their response.

In most of the projects on which respondents are working, measuring is being performed either continuously or in a certain point in time during coding and testing phases of software development.

The survey employed several Likert-type scales to measure respondents' opinions and experiences. For example, respondents rated frequency (e.g., Never to Always), intensity (e.g., Very Low to Very High), impact (e.g., No impact to Crucial), and agreement (e.g., Strongly disagree to Strongly agree) on these ordinal scales. Additionally, a limited number of open-ended questions allowed for free-text responses to capture qualitative insights.

Most of the time IDE (Integrated Development Environment) or separate metrics or testing tools are being used for capturing metrics, with visualization of results and support for a particular programming language or a technology as a main advantage of a preferred tool.

Table 1. General Characteristics of the Respondents

Characteristic	Freq.	%	Characteristic	Freq.	%
Experience in IT industry			*Technology project is written in*		
0–2 years	11	10.2	Java	82	75.9
2–5 years	17	15.7	JavaScript (or frameworks)	56	51.9
5–10 years	27	25.0	.NET / C#	32	29.6
10–15 years	20	18.5	Python	25	23.1
15+ years	33	30.6	C / C++	21	19.4
			PHP	10	9.3
			Ruby	8	7.4
			Kotlin	8	7.4
			Swift	8	7.4
			Smalltalk	4	3.7
			Haskell	3	2.8
Country of respondent's company			*Software methodology used*		
Serbia	62	57.4	Scrum	72	66.7
Slovenia	25	23.1	Waterfall	10	9.3
Germany	6	5.6	Iterative–incremental	9	8.3
Croatia	3	2.8	Extreme programming	4	3.7
Austria	3	2.8	Lean	3	2.8
Bosnia and Herzegovina	2	1.9	Spiral model	2	1.9
Hungary	2	1.9	Scrum-ban	2	1.9
France	2	1.9	Other	6	5.6
Other	3	2.8			
Company size			*Role of the respondent in project*		
Small (<50 employees)	28	25.9	Developer	60	55.6
Medium (51–1000 employees)	61	56.5	Team leader	33	30.6
Large (1000+ employees)	19	17.6	Solution architect	28	25.9
			Project manager	24	22.2
			Test engineer	27	25.0
			QA analyst	20	18.5
			DevOps	18	16.7
Company type			*Project phase*		
Service-oriented	56	51.9	Active development / released	75	69.4
Product-oriented	31	28.7	Active development / unreleased	13	12.0
Global In-House Center	9	8.3	Planning / requirements gathering	9	8.3
Startup	5	4.6	Initial development / prototyping	4	3.7
Other	7	6.5	Maintenance / no new development	3	2.8
			Other	4	3.7

Almost half of the respondents (47%) stated that measurement is of crucial or great importance to the success of a project/product they are working on (Fig. 2).

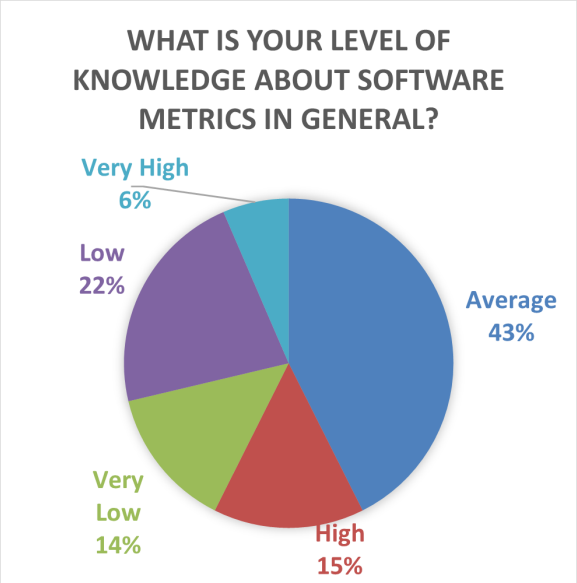


Fig. 1. Most of the respondents have average knowledge about software metrics

Almost the same distribution of answers was obtained from the next question, where respondents were asked about the influence of using software metrics on the overall quality of the software product. Based on the respondents' answers, software metrics are sparingly used to evaluate overall team productivity, with 33% of respondents saying they are sometimes being used for this purpose.

Almost two thirds of the respondents agree, to a greater or lesser extent, that measurement-based data helps their team to perform better than without using it.

Next three questions were questions regarding activities monitored by the software metrics and using and accessing results of the measurements. Most of the respondents think that activities of developers and test engineers are being measured and controlled by managers in various roles, who also have access to measurement results. There were several different answers to two questions about resources needed to establish metrics, as well as resources needed to use metrics, ranging from half an hour to a month. Finally, most of the respondents (33%) were not sure that definitions of metrics used in their projects are consistent and clearly defined (Fig. 3).

5. Discussion and Analysis

The main goal of the survey was to show whether there is a statistically significant correlations between different variables obtained from the research sample. In order to identify independent variables, answers from the following questions were used:

- How many years of experience in IT industry do you have?
- In which country is your company located?

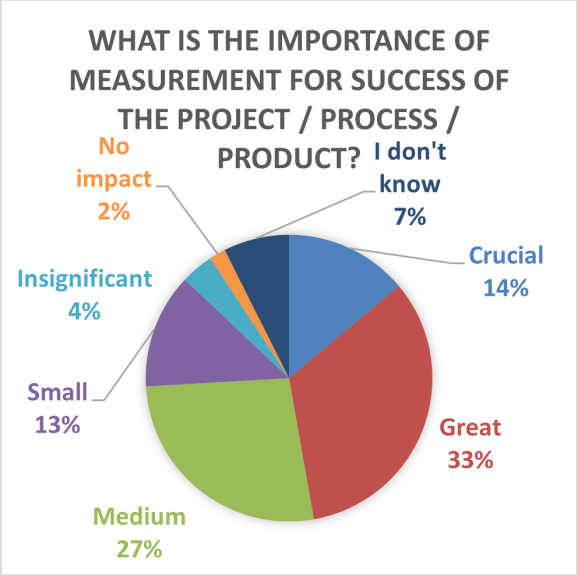


Fig. 2. Measurement has important role in the success of the project

- What is the size of the company you're currently working for?
- What roles do you have in the organizational structure of the project?

Answers to these questions were grouped in the following sample pairs:

- **Experienced and less experienced respondents** – respondents were considered experienced if they had ten or more years of experience in IT industry
- **Company is in Serbia / Company is in other country**
- **Medium size company / Small or large company**
- **Managers and non-managers**

To determine whether someone has managerial role, respondents were designated as the managers if they occupied one of the following roles:

- Senior manager
- Middle manager
- Project manager
- Product owner/Business analyst
- Scrum master
- Team leader

Next, appropriate samples were obtained from the answers given to the following questions:

- What is your level of knowledge about software metrics in general?
- What is the importance of measurement for success of the project / process / product?

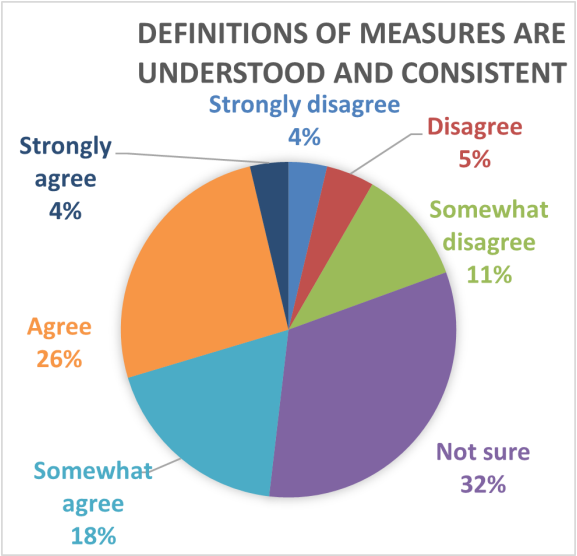


Fig. 3. Definitions of metrics are not always consistent and clearly defined

- How would you describe the influence of using software metrics on the overall quality of the project / process / product?
- How often are software metrics used in the aggregate to evaluate overall team productivity?
- Measurement-based data helps your team to perform better than without using it
- The definitions of measures that are used in my organization/current projects are commonly understood and consistent

To statistically analyze these samples, the weight functions shown in Table 2 were used ('I don't know' and 'Not sure' answers were not taken into account).

Table 2. The Weight functions

Answer	Value	Answer	Value	Answer	Value	Answer	Value
Very Low	1	No impact	1	Never	1	Strongly disagree	1
Low	2	Insignificant	2	Rarely	2	Disagree	2
Average	3	Small	3	Sometimes	3	Somewhat disagree	3
High	4	Medium	4	Most of the Time	4	Somewhat agree	4
Very High	5	Great	5	Always	5	Agree	5
		Crucial	6			Strongly agree	6

Since sample pairs had unequal variances and were different in size, Welch's t-test, a variation of Student's t-test, was used for statistical testing, since it tends to give more accurate results than the latter one when the samples are of unequal size and/or variance [50]. Some of the results of the statistical analysis are shown in Appendix.

By analyzing results obtained from applied test statistic to the sample pairs, the following statistically significant differences appear (with the 99% confidence level, unless stated differently):

- **Managers have higher level of perceived knowledge on software metrics than the regular employees**
- **Respondents in Serbia have lower level of perceived knowledge on software metrics than the respondents from other countries**
- **Software metrics are being used more often to evaluate overall team productivity in Serbia than in other countries** (with the 95 % confidence level)
- **Experienced respondents have higher level of perceived knowledge on software metrics than the less-experienced respondents**
- **Respondents from medium-sized companies think that metrics have stronger influence on team performance than the respondents from small and large companies**
- **Respondents from medium-sized companies think that metrics have stronger influence on project/process/product success than the respondents from small and large companies** (with the 95 % confidence level)

The observed differences in perceived knowledge of software metrics may be attributed to the fact that, in most cases, managers are more experienced, having been promoted to managerial roles after years of working as developers or QA analysts. In addition, the survey indicates that the measurement results are used more frequently by managers, which may contribute to a stronger perception of familiarity with the software metrics.

In general, most of the respondents are seasoned IT professionals with more than five years of experience in the industry, working in medium sized companies, which are, in most cases, service oriented. Also, most of the respondents are developers working in Java, working on projects that are mostly already in production, with Scrum as a software development methodology.

Although most of the respondents reported an average level of knowledge of software metrics, their answers to open-ended questions—particularly the limited or missing examples of concrete metrics—suggest that self-assessed knowledge does not always align with demonstrated familiarity. This discrepancy further supports the need to interpret findings related to expertise as indicators of perceived rather than objectively measured knowledge. In particular, respondents in managerial roles tended to provide more detailed metric lists, which may reflect greater exposure to measurement practices rather than higher technical proficiency.

It is not a surprise (although it is not necessarily the best practice) that the measurements are being performed during coding and testing phases, mostly continuously, by using separate metrics tools, testing tools or IDE. The choice of tools for measurement is usually guided by the platform independence, along with the possibility of the results visualizations, as well as saving and the interpretation of the results.

For most of the respondents, measurements have significant impact on the project success. Similarly, most of the respondents think that software quality is improved if software metrics are used during the project, and that, generally speaking, teams performing better than without using metrics, since productivity of the team is often being evaluated by using some of the metrics.

Unsurprisingly, there is the perception that, in most cases, metrics are being used to assess and evaluate work and the activities of developers and QA and test analysts i.e., persons responsible for results of coding and testing phases, and that the metrics are being used by people in some of the manager roles. Also, most of the respondents were not sure about time needed to establish and run the metrics, with the great variation in time in the available answers. Finally, it seems that there exists uncertainty about the notion of software metrics, since only one third of the respondents think that the definition of software metrics is well understood between all involved in the process of making the software product, which could be influenced by the lack of common knowledge regarding software metrics.

Compared to the results of the similar surveys and research mentioned in section 2, it seems that results of this survey are pretty much alike. The common denominators in the research previously conducted were the lack of knowledge of specific metrics, limited use of software metrics, and the fact that results obtained through the measurement process are mostly used by the upper management. These conclusions are mostly in line with the previously presented results. There are some differences, though – e.g., half of the respondents in the survey conducted by Pavlič et al. [13] responded that metrics are not used to improve the software quality. This is in stark contrast to the answers given in this survey, where vast majority of respondents are of the opinion that metrics have at least medium impact on the quality of the software. In the survey performed by Kasunic [16], most respondents agree with the statement that the definition of measurements is well-understood across the organization, which is not the case with the conclusions drawn in this research.

Some of the differences noted can be attributed to the geographical and cultural differences (surveys were conducted in Slovenia, Sri Lanka, Finland, Pakistan etc.). Also, the samples in the surveys differ greatly when comparing average experience of the respondents, size of the companies respondents are working for, whether the companies are service or project oriented etc. Therefore, it is not always easy to make hard comparisons between different surveys, but it is interesting to observe some templates of thinking and performing, as well as some common practices, producing recurrent benefits and flaws.

From a practical perspective, the identified differences suggest that the use and understanding of software metrics within organizations are strongly influenced by role, experience, and organizational context. Organizations could improve the effective use of software metrics by increasing transparency around how metrics are selected and used, and by involving both managerial and non-managerial roles in metric-related discussions. Providing targeted training, establishing shared definitions of commonly used metrics, and aligning metrics with improvement goals rather than solely with evaluation or control may help bridge the gap between perceived and demonstrated knowledge. In particular, encouraging teams to use metrics as tools for reflection and continuous improvement—rather than as performance indicators alone—may lead to more meaningful and sustainable measurement practices.

6. Conclusion

Using the software metrics during software development process clearly is not *deux es machina* solution for all the challenges regarding the software quality. Nevertheless, its

usage can greatly help in tackling some of the common pitfalls and shortcomings in the software industry. Survey conducted as part of this research showed that there are no real differences in perspective about the importance of software metrics to overall software quality between managers and employees. It seems that there is a consensus that using of the measurement is welcomed and should be embraced, although the level of knowledge about software metrics appears to be imbalanced between two groups. Results of the survey are pretty much in line with similar surveys conducted in the past, with differences that can be attributed to different sample sizes, cultural diversity, respondent's experience, and role etc.

In the future, similar survey could be conducted with some wider population. Survey should include multiple-choice questions about the knowledge of the concrete software metrics, instead of free-form ones. Also, questions about validation of measurement data could be added as well.

Acknowledgments. Authors thank all the companies and their employees for taking part in the survey, especially Endava d.o.o Belgrade. Also, authors wish to thank SQAMIA initiative (http://sqamia.org/). the following members for their help in conducting this survey: Tihana Galinac Grbac, Goran Mausa, Damir Kalpić, Melinda Toth, Zoltan Horvath, Zoltan Porkolab, Hannu Jaakoola, Jaak Henno, Novica Nosović, Dušanka Bošković, Dražen Brđanin, Anastas Mishev, Bojana Koteska, Klaus Bothe, Marjan Heričko, Stanimir Stoyanov and Asya Stoyanova.

References

1. ISO 9001:2015 Quality management systems — Requirements. Standard, International Organization for Standardization, Geneva, CH, September 2015.
2. James A Highsmith. *Agile software development ecosystems*, volume 13. Addison-Wesley Professional, 2002.
3. Gordana Rakić. SSQSA: Set of Software Quality Static Analysers, 2016.
4. Nasir U Eisty, George K Thiruvathukal, and Jeffrey C Carver. A survey of software metric use in research software development. In *2018 IEEE 14th International Conference on e-Science (e-Science)*. IEEE, October 2018.
5. Rini Solingen and Egon Berghout. The Goal/Question/Metric Method: A Practical Guide for Quality Improvement of Software Development. January 1999.
6. Sahra Sedigh-Ali, Arif Ghafoor, and Raymond A. Paul. *A Metrics-Guided Framework for Cost and Quality Management of Component-Based Software*, page 374–402. Springer Berlin Heidelberg, 2003.
7. Cost of poor software quality in the U.S.: A 2022 Report. https://www.it-cisq.org/wp-content/uploads/sites/6/2022/11/CPSQ-Report-Nov-22-2.pdf. Accessed: 2024-09-01.
8. Bill Haskins, Jonette Stecklein, Brandon Dick, Gregory Moroney, Randy Lovell, and James Dabney. Error Cost Escalation Through the Project Life Cycle. *INCOSE International Symposium*, 14(1):1723–1737, June 2004.
9. Dedi Setiadi, Tata Sumitra, Ahmad Karim, and Ritzkal. Software quality measurement analysis on academic information systems. 08 2024.
10. Jitender Kumar Chhabra and Varun Gupta. A survey of dynamic software metrics. *Journal of Computer Science and Technology*, 25(5):1016–1029, September 2010.
11. Krishnamoorthy Srinivasan and Dr. T. Devi Head. A comprehensive review and analysis on object-oriented software metrics in software measurement. 2014.

12. Ming-Chang Lee. Software quality factors and software quality metrics to enhance software quality assurance. *British Journal of Applied Science & Technology*, 4(21):3069–3095, January 2014.
13. Luka Pavlic, Mojca Okorn, and Marjan Hericko. The use of the software metrics in practice. In *SQAMIA*, volume 2217 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.
14. Paulinus Ofem, Bassey Isong, and Francis Lugayizi. Metrics for evaluating and improving transparency in software engineering: An empirical study and improvement model. *SN Computer Science*, 5(8):1097, 2024.
15. Touseef Tahir, Ghulam Rasool, Waqar Mehmood, and Cigdem Gencel. An evaluation of software measurement processes in pakistani software industry. *IEEE Access*, 6:57868–57896, 2018.
16. Mark Kasunic. The state of software measurement practice: Results of 2006 survey. 2006.
17. Wentao Chen, Huiqun Yu, Guisheng Fan, Zijie Huang, and Yuguo Liang. Usage patterns of software product metrics in assessing developers' output: A comprehensive study. *Information and Software Technology*, 189:107935, 2026.
18. Félix Témolé and Desislava Atanasova. An integrated approach to managing software quality in complex systems. *American Journal of Software Engineering and Applications*, 13(1):1–17, 2025.
19. K. V. Jeeva Padmini, H. M. N. Dilum Bandara, and Indika Perera. Use of software metrics in agile software development process. In *2015 Moratuwa Engineering Research Conference (MERCon)*, pages 312–317, 2015.
20. Harald Schaffernak, Birgit Moesl, Philipp Url, Ioana Victoria Koglbauer, and Wolfgang Vorraber. Towards sustainable software quality in use: a review of measures. *Next Research*, 2(3):100680, 2025.
21. Fatima Nur Colakoglu, Ali Yazici, and Alok Mishra. Software product quality metrics: A systematic mapping study. *IEEE Access*, 9:44647–44670, 2021.
22. Arthur-Jozsef Molnar, Alexandra Neamţu, and Simona Motogna. *Evaluation of Software Product Quality Metrics*, page 163–187. Springer International Publishing, 2020.
23. Alok Mishra, Raed Shatnawi, Cagatay Catal, and Akhan Akbulut. Techniques for calculating software product metrics threshold values: A systematic mapping study. *Applied Sciences*, 11(23):11377, December 2021.
24. Martin Vogel, Peter Knapik, Moritz Cohrs, Bernd Szyperrek, Winfried Pueschel, Haiko Etzel, Daniel Fiebig, Andreas Rausch, and Marco Kuhrmann. Metrics in automotive software development: A systematic literature review. *Journal of Software: Evolution and Process*, 33(2), August 2020.
25. Junaid Rashid, Toqeer Mahmood, and Muhamad Wasif Nisar. A study on software metrics and its impact on software quality, 2019.
26. Kshirasagar Naik and Priyadarshi Tripathy. *Software testing and quality assurance*. John Wiley & Sons, Nashville, TN, 2 edition, February 2016.
27. ISO/IEC 9126-1:2001 Software engineering — Product quality. Standard, International Organization for Standardization, Geneva, CH, January 2001.
28. ISO/IEC 25000:2014 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaREs. Standard, International Organization for Standardization, Geneva, CH, March 2014.
29. Jørgen Bøegh. A new standard for quality requirements. *Software, IEEE*, 25:57 – 63, 04 2008.
30. Roger Pressman and Bruce Maxim. *Software Engineering: A Practitioner's Approach*. McGraw-Hill Education, Columbus, OH, 9 edition, November 2019.
31. Stephen H Kan. *Metrics and models in software quality engineering*. Addison Wesley, Boston, MA, 2 edition, September 2002.
32. Tong Li. *An approach to modelling software evolution processes*. Springer, Berlin, Germany, 2009 edition, March 2009.

33. Thoms Ball. The concept of dynamic analysis. In *Proceedings of the 7th European Software Engineering Conference Held Jointly with the 7th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, ESEC/FSE-7, page 216–234, Berlin, Heidelberg, 1999. Springer-Verlag.
34. Bas Cornelissen, Andy Zaidman, Arie Deursen, Leon Moonen, and Rainer Koschke. A systematic survey of program comprehension through dynamic analysis. *Software Engineering, IEEE Transactions on*, 35:684 – 702, November 2009.
35. Michael Ernst. Static and dynamic analysis: Synergy and duality. May 2003.
36. Mrinal Rawat, Arpita Mittal, and Sanjay Dubey. Survey on impact of software metrics on software quality. *International Journal of Advanced Computer Science and Applications*, 3, January 2012.
37. Norman Fenton and James Bieman. *Software metrics*. Chapman & Hall/CRC Innovations in Software Engineering and Software Development Series. CRC Press, London, England, 3 edition, September 2020.
38. Norman E Fenton and Martin Neil. Software metrics: successes, failures and new directions. *Journal of Systems and Software*, 47(2):149–157, 1999.
39. Norman Fenton and Martin Neil. Software metrics: Roadmap. *Proceedings of the Conference on the Future of Software Engineering*, September 2000.
40. Kaushal Bhatt, Vinit Tarey, and Pushpraj Patel. Analysis Of Source Lines Of Code(SLOC) Metric. *IJETAE*, 2, April 2012.
41. Maurice H. Halstead. *Elements of Software Science (Operating and programming systems series)*. Elsevier Science Inc., USA, 1977.
42. T.J. McCabe. A complexity measure. *IEEE Transactions on Software Engineering*, SE-2(4):308–320, 1976.
43. S.R. Chidamber and C.F. Kemerer. A metrics suite for object oriented design. *IEEE Transactions on Software Engineering*, 20(6):476–493, 1994.
44. Mark Lorenz and Jeff Kidd. *Object-oriented software metrics*. Prentice Hall, Philadelphia, PA, May 1994.
45. Kenneth L. Morris. *Metrics for object-oriented software development environments*. PhD thesis, Massachusetts Institute of Technology, 1999.
46. W. Li and S. Henry. Maintenance metrics for the object oriented paradigm. In *[1993] Proceedings First International Software Metrics Symposium*, pages 52–60, 1993.
47. Crt Gerlec, Gordana Rakić, Zoran Budimac, and Marjan Hericko. A programming language independent framework for metrics-based software evolution and analysis. *Comput. Sci. Inf. Syst.*, 9:1155–1186, September 2012.
48. Yania Crespo, Carlos Nozal, M. Manso, and Raúl Sánchez. Language independent metric support towards refactoring inference. January 2005.
49. Martin Kunz, Reiner Dumke, and Niko Zenker. Software metrics for agile software development. pages 673–678, April 2008.
50. Graeme Ruxton. The Unequal Variance T-Test is an Underused Alternative to Student's T-Test and the Mann-Whitney U Test. *Behavioral Ecology*, 17, April 2006.

APPENDIX

What is your level of knowledge about software metrics in general?		
	Serbia	*Other country*
Mean	2.467742	3.195652174
Variance	1.138287	0.871980676
Observations	62	46
Hypothesized Mean Difference	0	
df	103	
t Stat	-3.76819	
P(T<=t) one-tail	0.000137	
t Critical one-tail	1.659782	
P(T<=t) two-tail	0.000274	
t Critical two-tail	1.983264	

What is your level of knowledge about software metrics in general?		
	Experienced	*Less experienced*
Mean	3.094339623	2.472727273
Variance	1.279390421	0.846464646
Observations	53	55
Hypothesized Mean Difference	0	
df	100	
t Stat	3.126495659	
P(T<=t) one-tail	0.001158291	
t Critical one-tail	1.660234326	
P(T<=t) two-tail	0.002316582	
t Critical two-tail	1.983971519	

What is your level of knowledge about software metrics in general?		
	Managers	*Non-managers*
Mean	3.016949	2.489795918
Variance	1.120397	1.046768707
Observations	59	49
Hypothesized Mean Difference	0	
df	104	
t Stat	2.624231	
P(T<=t) one-tail	0.004996	
t Critical one-tail	1.659637	
P(T<=t) two-tail	0.009992	
t Critical two-tail	1.983038	

Measurement-based data helps your team to perform better than without using it		
	Medium	*Small/Large*
Mean	4.697674419	3.948717949
Variance	0.549280177	1.997300945
Observations	43	39
Hypothesized Mean Difference	0	
df	56	
t Stat	2.960815687	
P(T<=t) one-tail	0.002246619	
t Critical one-tail	1.672522303	
P(T<=t) two-tail	0.004493238	
t Critical two-tail	2.003240719	

How often are software metrics used in the aggregate to evaluate overall team productivity?		
	Serbia	*Other country*
Mean	2.912281	2.511111111
Variance	0.974311	1.301010101
Observations	57	45
Hypothesized Mean Difference	0	
df	87	
t Stat	1.870371	
P(T<=t) one-tail	0.032397	
t Critical one-tail	1.662557	
P(T<=t) two-tail	0.064795	
t Critical two-tail	1.987608	

What is the importance of measurement for success of the project / process / product?		
	Medium	*Small/Large*
Mean	4.581818182	4.133333333
Variance	0.988552189	1.618181818
Observations	55	45
Hypothesized Mean Difference	0	
df	82	
t Stat	1.93116504	
P(T<=t) one-tail	0.028458955	
t Critical one-tail	1.663649184	
P(T<=t) two-tail	0.05691791	
t Critical two-tail	1.989318557	

Assessment of software quality tracking

Thank you for taking part in this survey!

This questionnaire is created in order to get insight of techniques and tools used to assess and evaluate quality of software in IT companies.

The answers will be used in a survey document about usage of measurement in practical environments. This survey document will be accessible to all parties taking part in answering the questionnaire.

Warranty: All data will be used and published only in an anonymized survey form.

Author of the questionnaire:
Prof. Zoran Budimac
Department of Mathematics and Informatics
University of Novi Sad
zjb@dmi.uns.ac.rs

* Required

General questions

1. How many years of experience in IT industry do you have? *

Mark only one oval.

- 0-2 years
- 2-5 years
- 5-10 years
- 10-15 years
- 15+ years

2. In which country is your company located? *

_____

3. What is the size of the company you're currently working for? *

Mark only one oval.

- Small (less than 50 employees)
- Medium (51-1000 employees)
- Large (1000+ employees)

4. What is the type of company you're currently working for? *

Mark only one oval.

- Startup
- Service-oriented
- Product-oriented
- Global In House Center
- Other: _____

5. Which software development technologies are used in the project you're currently working on? *

Check all that apply.

- Java
- .NET/C#
- JavaScript (or its frameworks)
- PHP
- Ruby
- C/C++
- Python
- Kotlin
- Swift
- Haskell
- Smalltalk
- Other: _____

6. Which software development methodology is used in the project you're currently working on? *

Mark only one oval.

- Waterfall
- Prototype
- Spiral model
- Iterative-incremental
- Scrum
- Lean
- Extreme programming
- Other: _____

7. What roles do you have in the organizational structure of the project (choose all that apply)? *

Check all that apply.

- Senior manager
- Middle manager
- Project manager
- Product owner/Business analyst
- Scrum master
- Team leader
- Solution architect
- Developer
- QA analyst
- Test engineer
- DevOps
- Other: _____

Software metrics in your project

8. Which best describes the current development stage of your project? *

Mark only one oval.

- Planning/Requirements Gathering
- Initial Development/Prototyping
- Active Development/Unreleased Software
- Active Development/Released Software
- Maintenance/No New Development Planned
- Other: _____

9. What is your level of knowledge about software metrics in general? *

Mark only one oval.

- Very Low
- Low
- Average
- High
- Very High

10. List all software metrics which you are familiar with

_____
_____
_____
_____
_____

11. In which stages of the development a measuring is made? *

Check all that apply.

- Planning
- Business modelling
- Design
- Coding
- Testing
- Using
- Never
- I don't know
- Other: _____

12. List all metrics that are used in the measurements?

_____
_____
_____
_____
_____

13. During the project life cycle measuring is done *

Mark only one oval.

- Continuously
- At a certain point (every week, month, some milestones ...)
- If necessary
- At the request of higher instance
- Randomly
- Never
- I don't know
- Other: _____

14. What kind of tools is used for the software measurements (and which)? *

Check all that apply.

- IDE
- Separate metrics tool
- Designing tool
- Modelling tool
- Testing tool
- None
- I don't know
- Other: _____

15. What features of the tools used in the measurements were of crucial importance in order to choose them? (choose all that apply) *

Check all that apply.

- Independency of platform
- Supported certain platforms
- Supporting the particular programming language
- Independency of programming language
- Supported certain metrics
- Supporting the large number of metrics
- Visualization of results
- Saving and interpretation of results
- None
- I don't know
- Other: _____

16. What are the disadvantages of selected tools? (choose all that apply) *

Check all that apply.

- Dependency of platform
- Unsupported certain platforms
- Lack of support for the particular programming language
- Dependency of programming language
- Unsupported certain metrics
- Supporting inadequate set of metrics
- Lack of visualization of results
- Lack of saving and interpretations of results
- None
- I don't know
- Other: _____

17. What is the importance of measurement for success of the project / process / product? *

Mark only one oval.

- Crucial
- Great
- Medium
- Small
- Insignificant
- No impact
- I don't know

18. How would you describe the influence of using software metrics on the overall quality of the project / process / product? *

Mark only one oval.

- Crucial
- Great
- Medium
- Small
- Insignificant
- No impact
- I don't know

19. How often are software metrics used in the aggregate to evaluate overall team productivity? *

Mark only one oval.

- Never
- Rarely
- Sometimes
- Most of the Time
- Always
- I don't know

20. Measurement-based data helps your team to perform better than without using it *

Mark only one oval.

- Strongly disagree
- Disagree
- Somewhat disagree
- Not sure
- Somewhat agree
- Agree
- Strongly agree

Software measurement usage

21. Whose activities are monitored / controlled by the measurements? (choose all that apply) *

Check all that apply.

- Senior manager
- Middle manager
- Project manager
- Product owner/Business analyst
- Scrum master
- Team leader
- Solution architect
- Developer
- QA analyst
- Test engineer
- DevOps
- No one
- I don't know
- Other: _____

22. Who uses measurement results? (choose all that apply) *

Check all that apply.

- [ ] Senior manager
- [ ] Middle manager
- [ ] Project manager
- [ ] Product owner/Business analyst
- [ ] Scrum master
- [ ] Team leader
- [ ] Solution architect
- [ ] Developer
- [ ] QA analyst
- [ ] Test engineer
- [ ] DevOps
- [ ] No one
- [ ] I don't know
- [ ] Other: _____

23. Who has the access to measurement results? (choose all that apply) *

Check all that apply.

- [ ] Senior manager
- [ ] Middle manager
- [ ] Project manager
- [ ] Product owner/Business analyst
- [ ] Scrum master
- [ ] Team leader
- [ ] Solution architect
- [ ] Developer
- [ ] QA analyst
- [ ] Test engineer
- [ ] DevOps
- [ ] No one
- [ ] I don't know
- [ ] Other: _____

24. How much resources are consumed when the metrics are introduced and established (person hours, days ...)? *

_____

25. How much resources are consumed when the metrics are used (person hours, days ...)? *

_____

26. The definitions of measures that are used in my organization/current projects are commonly understood and consistent *

Mark only one oval.

- ◯ Strongly disagree
- ◯ Disagree
- ◯ Somewhat disagree
- ◯ Not sure
- ◯ Somewhat agree
- ◯ Agree
- ◯ Strongly agree

Thanks!

Thank you for taking the time to complete this survey. We truly value the information you have provided. Your responses will contribute to our assessment of software quality tracking in IT companies.

Filip Prentović was born on June 15, 1982, in Odžaci, Serbia. He received a B.S. degree in Computer Science from the Faculty of Sciences, University of Novi Sad, in 2005. He received his M.S. degree from the same faculty in 2018. Currently, he is a Ph.D. student at the Faculty of Sciences. His fields of interest include software architectures, software quality, and software metrics.

Zoran Budimac was a full professor at the Faculty of Sciences, University of Novi Sad, Serbia. He was born on December 24, 1960, in Sombor, Serbia. He obtained all his degrees at the Faculty of Sciences, University of Novi Sad: a B.Sc. degree in Informatics in 1983, an M.Sc. degree in Discrete Mathematics and Programming in 1992, and a Ph.D. degree in Informatics in 1994. His research interests included software engineering, software quality, e-learning, and compiler construction. He was a principal investigator in several domestic and international projects and a participant in many others. He was the author of 16 textbooks and more than 280 research papers, most of which were published in international journals and conference proceedings. He was/is a member of the Organizing and/or Program Committees of more than 70 international conferences and a member of the Managing Board of "Computer Science and Information Systems".

Marjan Heričko is a full professor at the Institute of Informatics. He is the Head of the Information Systems Laboratory and Deputy Head of the Institute of Informatics. He received his Ph.D. in Computer Science from the University of Maribor in 1998. His main research interests include all aspects of information systems development, software and service engineering, agile methods, process frameworks, software metrics, functional size measurement, SOA, component-based development, object orientation, software reuse, and software patterns. Dr. Heričko has been a project or work coordinator in several applied projects, as well as a project or work coordinator in several international research projects, and has served as a committee member and chair of several international conferences.

Gordana Rakić has held the position of Assistant Professor with a Ph.D. at the Faculty of Sciences, University of Novi Sad, Serbia, since 2016. She obtained all her degrees at the Faculty of Sciences, University of Novi Sad: a B.Sc. degree in Informatics in 2006, an M.Sc. degree in Business Informatics in 2010, and a Ph.D. degree in Computer Science in 2015. Her research interests include software engineering, software quality, static analysis, and computer languages. Gordana is one of the founders and a member of the SQLab: Software Quality Laboratory under the Department of Mathematics and Informatics at her home faculty. Furthermore, she is a member of ESUG (European Smalltalk User Group), as well as ACM and IEEE. She has participated in several ongoing and completed bilateral, multilateral, and national research projects. Currently, Gordana is chairing COST Action CA19136 — Connecting Education and Research Communities for an Innovative Resource Aware Society (CERCIRAS). She has published more than 30 scientific papers and articles. She has participated in the organization of several international scientific conferences, symposia, and workshops as a member, secretary, co-chair, and chair of organizing and program committees, and serves as a reviewer for several international journals. In the period from 2011 to 2013, she was one of the managing editors of the "ComSIS: Computer Science and Information Systems" journal.

Received: November 9, 2025; Accepted: May 1, 2026.

Computer Science and Information Systems 23(3):1077-1102 doi: https://doi.org/10.2298/CSIS251101031K

Artificial Neural Network Modeling for Air Pollution Prediction: LSTM versus the Levenberg-Marquardt Approach

Goran Keković[1], Rade Božović[1], Sonja Ketin[2], Vladimir Mikić[1], Miloš Ilić[1] and Boban Vesin[3]

[1] Faculty of Information Technology, Alfa BK University, 11000 Belgrade, Serbia
goran.kekovic@alfa.edu.rs (corresponding author),
{rade.bozovic, vladimir.mikic, milos.ilic}@alfa.edu.rs
[2] School of Railroad Transport, Academy of Technical and Art Applied Studies, 11000 Belgrade Serbia
sonja.ketin@vzs.edu.rs
[3] School of Business, University of South-Eastern Norway, Raveien 215, Borre, Vestfold, 3184, Norway
boban.vesin@usn.no

Abstract. Accurate prediction of air pollutant concentrations remains a critical challenge for environmental monitoring and public health, demanding robust and adaptive artificial intelligence approaches. This study investigates the effectiveness of various types of artificial neural networks (ANNs), including Long Short-Term Memory networks (LSTM) and networks based on the Levenberg-Marquardt algorithm (LM) and its variant with Bayesian regularization (LMBR), in predicting air pollution under different data conditions. Since LSTM networks are based on first derivative loss function algorithms and the LM algorithm is usually superior to this type of algorithm, a comparison of these networks was conducted. This is further supported by the limited coverage of this topic in the existing literature. ANNs were tested on two different datasets: the Air Quality dataset, where the target variable was the concentration of benzene (C_6H_6) and the Beijing PM2.5 dataset, where the target was the concentration of PM2.5 particles. The performance metrics of the ANNs were the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). It is shown that, in the case of the Air Quality dataset, the values of these parameters $RMSE = 0.11$ μg/m^3, $MAE = 0.09$ μg/m^3, $MAPE = 1\%$ for LSTM networks and $RMSE = 0.14$ μg/m^3, $MAE = 0.092$ μg/m^3, $MAPE = 1\%$ for LMBR networks, were competitive. For LM networks, these values were significantly higher: $RMSE = 0.57$ μg/m^3, $MAE = 0.2$ μg/m^3, $MAPE = 2\%$. Contrastingly, in the case of the Beijing database, the values of all parameters were drastically higher: $RMSE = \{45.74$ μg/m^3, 64.94 μg/m^3, 65.68 μg/m$^3\}$, $MAE = \{30.32$ μg/m^3, 42.83 μg/m^3, 54.5 μg/m$^3\}$ and $MAPE = \{52\%, 72\%, 74.25\%\}$ for LSTM, LMBR, and LM networks, respectively. In this case, the benzene concentration values exhibited a strictly linear correlation with the input variables. For the Beijing dataset, the relationships between PM2.5 concentration and predictor values were non-monotonic, leading to a drastic drop in the performance of all networks. Findings reveal that, in this case, LSTM networks were more robust compared to LMBR and LM networks, as the values of their RMSE, MAE, and MAPE parameters were

significantly lower. Furthermore, it is shown that the input variable selection technique (IVS) can be used to detect seasonal trends in the input data.

Keywords: Artificial Intelligence, LSTM Network, Levenberg-Marquardt Algorithm, Input Variable Selection.

1. Introduction

Air pollution is becoming one of the major problems facing modern human society due to various factors, such as the combustion of fossil fuels in industry, vehicle exhaust emissions associated with transport, and the migration of the population to urban areas [35, 39]. In general, all components of air pollution can be divided into two groups: primary and secondary. Primary components are created through the direct combustion of fossil fuels, resulting in the release of gases and particulate matter into the atmosphere. The most common representatives of this group are particulate matter (PM10, PM2.5), nitrogen and sulfur oxides (NO_x, SO_x), non-methane hydrocarbons (NMHC) and ozone O_3 [38]. Secondary components are formed by chemical reactions between primary components. A typical example of this group is the conversion of non-methane hydrocarbons into tropospheric ozone and peroxyacetyl nitrate (PAN) in the presence of the photocatalytic action of solar radiation. Numerous studies have examined the harmful effects of pollutants on human health, with PM being of particular importance [24, 28, 70, 54].

Although PM air pollutants have been addressed in a relatively large number of papers, benzene has received limited attention. This is unjustified, considering that exposure to benzene can cause various types of cancer in humans [47]. The adverse effects of this and other pollutants have led to the development of modern technologies to monitor their concentrations in the air, with artificial intelligence (AI) methods playing a particularly important role, because they are not subject to the shortcomings typically associated with traditional regression methods with regard to input data structure (the problem of multicollinearity and types of variables, mutual independence of data, etc.) [22]. In general, AI methods used for air pollution prediction can be divided into five categories: fuzzy logic [8], hidden Markov models [30], ensemble models [13], artificial neural networks (ANNs) [5], and deep learning [24]. In the context of this paper, the last two categories are of greatest interest.

ANN algorithms are widely used in air quality control, and recently, LSTM deep neural networks have gained particular popularity [74, 29, 49]. The popularity of these networks is due to their ability to predict short-term and long-term dependencies of the targeted variable, i.e. pollutant concentration. Various types of these networks have been applied, and one study found that Bidirectional Long Short-Term Memory (Bi-LSTM) achieves the best results [71]. In another interesting study, the authors developed a model to predict PM particle concentrations ten days in advance in Seoul, South Korea, using two techniques: LSTM and deep autoencoder [64]. In this paper, it was shown that LSTM networks outperformed DAE, further demonstrating the advantage of LSTM. Recently, researchers have increasingly used LSTM deep learning algorithms, which have proven to be very effective in processing big data and time series [65, 53, 1, 27]. The disadvantage of these algorithms is the need to adjust the parameters in their hyperspace, which is time-consuming, and for this purpose, particle swarm optimization and genetic algorithms are used [68].

Although recurrent artificial neural networks (RNNs) represent a modern type of ANNs for the prediction of air pollution, the LM algorithm and its Bayesian-regularized variant LMBR were also considered in this study. The reason is that, for small and medium datasets, algorithms based on the second derivative of the loss function are generally more efficient than those based on the first derivative [56]. Accordingly, the competitiveness of the LM and LMBR algorithms can also be expected, since all RNN variants rely not only on memory cells, but also on algorithms based on the first derivative of the loss function. LSTM and other variants of RNNs can learn temporal correlations between input variables due to their property of memorizing long-term effects. On the other hand, algorithms based on the second derivative work in batch or semi-batch mode, which means that the history of the time series values of the predictors is taken into account. Another reason why their competitiveness is worth examining is that there are almost no studies comparing the properties of LSTM and LM/LMBR networks in air pollution tasks. A relevant study has explored this topic in a different context. Specifically, the authors combined LSTM networks with the LM algorithm to identify unknown aerodynamic parameters using real flight data from aircraft [73]. Their results demonstrated the effectiveness of this approach compared to the parameter values obtained from wind tunnel experiments.

In the mentioned studies, the term air pollution prediction often refers to the prediction of the concentration of specific air pollutants. The forecast horizon usually varies from one to several days, and the discrepancies between the simulated and actual values of the concentration of a given pollutant are reported most frequently [36, 41]. The output variable (pollutant concentration) is a function of numerous input parameters, so the question that arises is the accuracy of this approach. For example, in all AI-based simulations, meteorological conditions are regularly taken into account as part of the input parameters, which are highly variable both locally and globally. Therefore, the horizon of future values is usually determined on the basis of the average values of the parameters in the past. When using ANNs, the predicted concentrations of pollutants could be grouped into specific intervals or classes, since ANNs are more efficient at classification tasks. However, this approach can estimate the interval of expected values, which can have a significant impact on the planning of daily activities.

At the same time, the objective was to compare the properties of these models to identify the most suitable approach for the prediction of air pollution. However, like other AI-based methods, they remain sensitive to the structure of the input data, and this issue therefore requires particular attention. This limitation can be partially addressed through input variable selection (IVS). According to recent studies, IVS techniques can be classified into three categories [7].

In the first category, known as filter techniques, the input data are preprocessed using statistical analysis methods and algorithms that determine the relevance of the input variables [18]. A typical representative of this category is the algorithm mrMR (mr-minimum redundancy, MR-maximum relevance) based on the application of the concept of mutual information [42]. This category is the simplest to implement and is independent of the AI method to which it passes the selected variables. IVS methods that involve an extensive search through the parameter space of the input data, monitoring the effect of adding and removing input variables, are called wrapper methods [25]. In this sense, there are two types of these methods: sequential addition of variables to the initial empty set (SFS) and sequential elimination of variables from the entire initial set of variables (SBS). In both

cases, the effects of adding or removing variables are monitored by the behavior of the loss function. This category of methods is technically much more demanding than the first category in terms of computing time and memory resources.

The final category comprises the most complex methods in terms of practical implementation, as they are applied during the training phase of the AI method to which the variables are assigned. This group of methods is called embedded methods and it should be noted that the situation is particularly complex when it comes to ANNs [51]. Typical representatives of this group are the CART and ID3 algorithms in decision trees. The application of IVS yields benefits in terms of reduced computation time and lower memory requirements, but it does not completely resolve the fundamental problem of the dependence of AI methods on the structure of the input data. Among these factors, the sample size is one of the most influential. The widely accepted view is that a larger sample leads to greater accuracy of these methods, due to a larger training sample. However, this view cannot always be accepted as correct, because there are studies in which researchers have shown that high accuracy of AI methods can be achieved even on small sample sizes [52]. Moreover, a further increase in the sample size may even reduce the accuracy of AI methods [57].

Despite the growing use of AI methods in air pollution monitoring and prediction, several important issues remain insufficiently explored. In particular, the comparative performance of different neural network architectures under limited data conditions, as well as the role of advanced preprocessing techniques in enhancing model performance, have yet to be fully clarified. These challenges are especially relevant in real-time monitoring scenarios, where accurate modeling of the relationship between environmental variables and pollutant concentrations is essential.

Motivated by these considerations, this study addresses the following research questions:

- **RQ1**: How competitive are LSTM and LM/LMBR neural network models in real-time air pollution monitoring, particularly in modeling the relationship between the current values of the target variable and the corresponding input variables?
- **RQ2**: To what extent can the IVS method be effectively applied in the data preprocessing stage for air pollution monitoring tasks, including the identification of relevant input features and seasonal patterns in the data?

Given the high variability of input parameters in air pollution monitoring—such as meteorological conditions, geographical characteristics, and temporal dynamics—as well as the wide range of available AI-based modeling approaches, deriving universally valid conclusions remains challenging. Nevertheless, the results obtained in this study provide valuable insights into the applicability of the considered methods and can serve as a guidelines for future research in this domain.

The main contributions of this study can be summarized as follows:

- A comparative evaluation of LSTM and LM/LMBR networks is conducted using the root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE) metrics, demonstrating that their predictive performance is largely comparable when trained on relatively small datasets.
- The study shows that monotonicity relations in input data lead to a significant improvement in the predictive performance of the considered neural network models.

- The robustness of LSTM networks to nonlinear relationships between variables has been generalized and extended to non-monotonic relationships.
- The applicability of the IVS technique is further extended by demonstrating its effectiveness in identifying and analyzing seasonal patterns present in the input data.

The remainder of this paper is organized as follows: Section 2 reviews the related work and provides the broader context for the study. Section 3 describes the methods used in the research, including the research approach and procedures. Section 4 presents and discusses the results obtained from the analysis. Section 5 outlines the main limitations of the conducted research. Finally, Section 6 concludes the paper by summarizing the key findings and suggesting directions for future work.

2. Related work

Air pollution prediction has emerged as a highly active research domain within environmental informatics, given the paramount importance of accurate forecasts in safeguarding public health and informing policy decisions. A substantial body of research has demonstrated the potential of deep learning models, particularly LSTM networks, in capturing long-range temporal dependencies in nonlinear and noisy time series data. Numerous studies have utilized LSTM architectures to forecast particulate matter concentrations, ozone, and other pollutants, consistently demonstrating superior performance compared to traditional regression and shallow learning models. For example, in [44], the author demonstrated that LSTM models outperformed classical regression approaches for PM2.5 forecasting. Similarly, the authors in [12, 61] introduced spatiotemporal variants such as convolutional LSTM and recursive LSTM to integrate both temporal dynamics and spatial correlations for air pollution prediction, demonstrating significant reductions in prediction errors. Further improvements have been achieved by hybrid deep learning methods. The authors in [31] proposed a CNN-LSTM framework for PM2.5 prediction, while the authors in [69] employed an attention-based LSTM for real-time pollution monitoring, with both studies reporting lower RMSE and MAPE values compared to baseline models. Complementing these views, the authors in [58] also demonstrated that LSTM networks successfully internalize complex, non-linear temporal dependencies of ambient quality metrics. These findings have reinforced the prevailing view that LSTM architectures are the state of the art in air pollution prediction tasks.

Despite this dominance, evidence suggests that optimization methods based on the LM algorithm and its Bayesian regularization variant, LMBR, remain competitive in certain contexts, especially when there are strong linear correlations between predictors and target variables. In their study [15], the authors demonstrated that LM-ANNs outperformed both multiple linear regression and other training algorithms in PM2.5 prediction in India, achieving the highest coefficient of determination ($R^2 = 0.8164$) and the lowest RMSE (9.52). Similar results were obtained in energy demand forecasting, where LM-ANN models achieved superior accuracy compared to scaled conjugate gradient and regression methods [62, 56]. Research reported in [26] analyzed multiple time-series neural network models, including various configurations, suggesting that the Levenberg–Marquardt approach can be effective for Air Quality Index forecasting based on the year of various gas measurements. In [4], the authors conducted a comparative analysis of ANN models for long-term forecasting of PM10 concentrations. Their findings confirmed

that, despite the low RMSE values achieved by LM-ANN, the LMBR approach consistently delivered the most reliable overall performance, thus underscoring its potential in air quality forecasting applications. The presented studies indicate that LM and LMBR algorithms can be highly effective in domains where predictors exhibit strong linear relationships with the response variable and where robustness to smaller sample sizes is required. Direct comparative studies of the LSTM and LM/LMBR methods in the domain of air pollution remain scarce, but emerging results point to their potential complementarity and indicate that such comparisons may offer valuable insights for identifying the most effective predictive approaches, thus contributing to improved public health through more accurate air quality management.

In summary, while LSTM networks have established themselves as the dominant method for air pollution prediction due to their ability to handle complex nonlinear and temporal dependencies, LM and LMBR-based networks remain important alternatives that can offer competitive accuracy in contexts characterized by stronger linear dependencies. The comparison of these approaches represents a promising research direction that ensures both predictive accuracy and interpretability. This need forms the foundation of the presented study, which systematically evaluates the comparative performance of LSTM and LMBR in real-world air quality monitoring tasks.

3. Methods

To facilitate a clearer understanding of the methods and materials used in this study, an overview pseudocode of the entire research process is presented below. This pseudocode is intended to summarize the main methodological steps of the proposed air pollution modeling framework and to provide a structured guide for interpreting the individual phases of model development, validation, and evaluation. Each step is subsequently described in greater detail in the following subsections. In addition to improving the readability of the methodology, this overview also highlights the sequential dependencies between data preparation, model construction, validation, fine-tuning, and final performance assessment.

The presented pseudocode shows that the overall process of air pollution modeling using ANNs can be divided into four main phases. While Phases 1, 2, and 4 are relatively straightforward, Phase 3 requires additional clarification due to its greater methodological complexity. In Phase 3, for each of the considered network models - LSTM and LMBR/LM - the most appropriate partitioning of the input dataset into k folds is determined using k-fold cross-validation. This procedure is a standard technique for mitigating overfitting during model training. Once the optimal fold scheme has been established, the resulting models are further refined through a fine-tuning process. This step involves adjusting both the network architecture and the associated hyperparameters (lines 9–12 of the pseudocode) in order to improve model performance. Finally, lines 13–16 of the pseudocode correspond to the prediction and evaluation stage, during which the trained models generate output predictions and their predictive performance is assessed. Each phase of the pseudocode will be described in greater detail in the sections that follow.

Algorithm 1. ANN Modeling and Prediction Framework

Input: Raw dataset D; initial ANN architectures $\{LSTM, LMBR/LM\}$
Output: Predictions $\bar{Y}$ and performance metrics M for each ANN model

Phase 1: Data Preprocessing
Remove incomplete observations from D
Normalize variables using z-score standardization
$x_{\text{norm}} = (x - \mu)/\sigma$
where μ is the mean value and σ is the standard deviation
Apply noise filtering to obtain cleaned dataset D_clean

Phase 2: Feature Extraction
Apply IVS technique to D_clean to identify seasonal trends
Construct feature set F

Phase 3: Model Training and Optimization
for each $ANN\_model \in \{LSTM, LMBR/LM\}$ **do**
    Determine optimal k using k-fold cross-validation
    repeat
        Tune hyperparameters of ANN_model
        Adjust network architecture
        Evaluate model using k-fold cross-validation
    until convergence of performance metric

Phase 4: Prediction and Evaluation
    Generate predictions $\bar{Y}$ using trained ANN_model
    Compute performance metrics M
end for
return $\bar{Y}$ and M

3.1. Databases and input variable selection

As already noted, there is no definitive classification of AI methods with respect to the structure of input data; most recommendations rely on empirical findings from numerical experiments. A common starting point is the analysis of linear correlations between the target variable and the predictors. Since linear relationships represent a special case of monotonic dependence, their presence or absence often indicates whether the underlying relationships are predominantly monotonic or whether they are more complex and nonlinear.

The two databases used in this study, namely the Air Quality dataset and the Beijing dataset, were intentionally selected because they represent contrasting structures of variable relationships. The Air Quality dataset is dominated by monotonic relationships between variables, whereas the Beijing dataset contains predominantly non-monotonic relationships among input predictors. Examining these opposing structural characteristics enables a more informative assessment of research questions RQ1 and RQ2.

However, AI methods alone are not sufficient for a comprehensive characterization of these relationships; therefore, IVS techniques are also applied as part of the raw data preprocessing stage.

Air Quality dataset The numerical experiments were conducted using a dataset collected from air pollution measurements in an Italian city between March 2004 and February 2005 [59]. The database contained 9358 samples, with values averaged over hourly intervals, obtained from the sensor system. The target variable was the concentration of benzene C_6H_6. Some of the recorded values were invalid (e.g. negative concentrations, negative temperature, and humidity), most likely due to sensor drift and other measurement-related issues. An initial approach was to remove all samples that contained such invalid values; however, this would have reduced the input dataset to only 827 samples. Because this reduction was considered unacceptable, the distribution of invalid data was then examined for each variable. Based on this analysis, the variables were divided into three groups, with 3.91%, 17.99% and 90% invalid data, respectively. Variables with more than 3.91% invalid data were excluded from further analysis. As a result, the number of input variables, or predictors, was reduced from 12 to 8. It is important to note that removing these variables did not affect the research questions introduced earlier, since the corresponding measurements were available from other locations and could therefore still be represented in the input dataset. For the remaining variables, invalid values were replaced with the mean of all valid observations in the time series. The statistical parameters of the filtered data are given in Table 1. As shown in the table, some variables are expressed in arbitrary units, which corresponds to sensors nominally intended to measure the concentrations of specific air pollutants (PT08S1, PT08S2, etc.).

In addition, all variables were found to be normally distributed. Therefore, Pearson's correlation test was suitable for correlation analysis. Since the target variable was the concentration of benzene, Table 2 presents the statistically significant values of the correlation coefficient between benzene and other variables, at a statistical significance level of 5%. A visual inspection of Table 2 shows that the values of the correlation coefficient are very high. The correlations between the input variables were also high, indicating the presence of multicollinearity, which reflects the synergistic effect of various air pollutants in their mixture in the air. Since the correlation values between the input variables are high, it can be logically assumed that this parameter can serve as an indicator of seasonal changes in the predictor-target relationship. Furthermore, it can also be used for input variable selection in the case of incomplete input data from the sensor network.

Table 1. Statistical parameters of collected data

Parameters	Unit	Min.	Max.	Std.
PT08.S1(CO)	–	647.25	2040	212.80
C_6H_6(GT)	μg/m^3	0.15	63.74	7.30
PT08.S2(NMHC)	–	383.25	2214	261.56
PT08.S3(NOx)	–	322	2683	251.74
PT08.S4(NO2)	–	551	2775	339.36
PT08.S5(O3)	–	221	2522.8	390.61
T	$^\circ$C	0.05	44.6	8.63
RH	%	9.18	88.72	16.97
AH	g/m^3	0.18	2.23	0.40

Table 2. Correlation coefficients with C_6H_6

Parameters	Correlation with $\mathbf{C_6H_6}$
PT08.S1(CO)	0.85
PT08.S2(NMHC)	0.77
PT08.S3(NOx)	0.51
PT08.S4(NO2)	0.77
PT08.S5(O3)	0.64
T	0.97
RH	0.93
AH	0.98

The mrMR algorithm was employed [43]. This algorithm selects input variables that are minimally correlated (minimal redundancy, mr) with each other and maximally correlated (maximal relevance, MR) with benzene as the output variable. The mrMR algorithm relies on mutual information, defined by the following formula:

$$I = \sum_{x,y} p(x,y) \cdot \log \frac{p(x,y)}{p(x) \cdot p(y)} \tag{1}$$

where $p(x,y)$ is the joint distribution of the variables x and y, while $p(x)$ and $p(y)$ are the marginal distributions of the respective variables. The meaning of mutual information can be seen if $X \cap Y = \{0\}$ is considered, where $p(x,y) = p(x) \cdot p(y)$, and from the above formula it follows that $I(X,Y) = 0$. Mutual information represents the amount of information about one variable that is related to another variable. In practical applications, the mrMR algorithm tries to maximize the MI score:

$$MI = \frac{1}{S} \cdot \frac{\sum_{x \in S} I(x,y)}{\sum_{x,z \in S} I(x,z)} \tag{2}$$

In the above equation, the numerator represents the sum of the mutual information of the predictor x with the target variable y, and the denominator represents the sum of the mutual information between the predictors. The mrMR algorithm attempts to maximize the score by finding a suitable subset S of input variables, so that the numerator in the above formula is maximal (relevance) and the denominator is minimal (redundancy).

The results shown in Fig. 1 indicate seasonal trends for the winter and summer seasons. In the summer season, NMHC, temperature, and absolute humidity are identified as the dominant variables. This suggests enhanced benzene formation through secondary chemical reactions among primary pollutants under the photocatalytic action of solar radiation. During the winter season (Fig. 1D), nitrogen oxide emerges as the dominant variable, most likely due to increased fossil-fuel combustion for heating, while temperature and air humidity also appear among the relevant variables. In the other seasons (spring and autumn, Fig. 1A, 1B), no significant predictors were identified. These findings are further supported by Fig. 2, which shows two distinct peaks in the daily benzene concentration profile during spring and summer. These peaks occur in the morning and late afternoon,

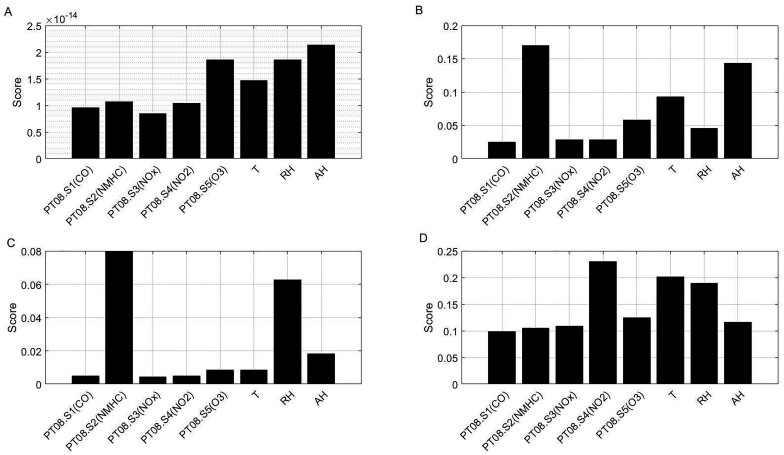


Fig. 1. Seasonal trends in the influence of the predictor on the concentration of benzene: A-Spring, B-summer, C-autumn, D-winter

most likely due to increased traffic during the morning and afternoon rush hours. At the same time, they are superimposed on increased benzene concentrations caused by the previously mentioned secondary reactions in the summer season (Fig. 2B) and by thermal combustion during the winter season (Fig. 2D).

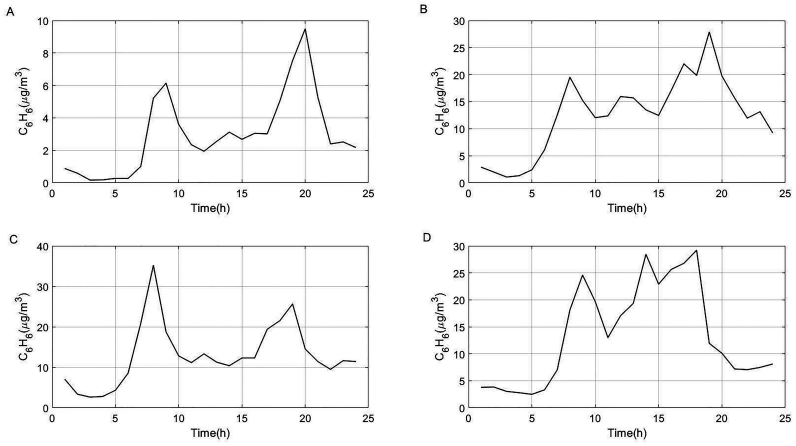


Fig. 2. Typical daily trends in benzene: A-Spring, B-summer, C-autumn, D-winter

Beijing PM2.5 dataset This database comprised 43825 measurements of PM2.5 concentrations recorded at one-hour intervals in Beijing, China, together with the corresponding meteorological conditions [10]. Data were recorded between 2010 and 2015. After removing incomplete or invalid samples (e.g., those affected by sensor drift or containing negative concentrations), the initial dataset was reduced to 41757 samples. The original dataset also included columns containing the month and hour of each recorded sample, which allowed the data to be classified by season even after preprocessing. To avoid the effect of sample size and to ensure comparability with the first database used in this study, the dataset was further reduced to 9358 samples. The target variable was the concentration of PM2.5 particles, while the predictor variables were: Dew Point (DEWP), Temperature (T), Pressure (P), combined wind direction (cbwd), cumulated wind speed (Iws), cumulated snowfall hours (Is) and cumulated rainfall hours (Ir). The main statistical characteristics of these variables are given in Table 3, while Fig. 3 shows the average daily values of the predictors throughout the seasons.

Table 3. Statistical parameters of the Beijing dataset

Parameters	Unit	Min.	Max.	Std.
PM2.5	μg/m^3	1	980	97
DEWP	$^\circ$C	-28	28	14.91
T	$^\circ$C	-19	41	12.96
P	hPa	994	1045	10.31
cbwd	-	1	4	1.28
Iws	m/s	0.45	565.5	59.23
Is	-	0	27	1.14
Ir	-	0	36	1.71

In Fig. 3, the dotted line denotes the concentration threshold for PM2.5 that is considered hazardous to human health. According to the China national ambient air quality standards (NAAQS), standard GB 3095-2012, the corresponding permissible average concentration value is 35 μg/m^3 (Grade I) [72]. As shown in Fig. 3, the average daily concentration of PM2.5 exceeds this threshold in all seasons. This exceedance is most pronounced in the summer (Fig. 3C) and autumn seasons (Fig. 3D). The elevated concentrations of PM2.5 observed during summer are most likely associated with high temperatures and the photocatalytic effect of solar radiation, which promote reactions among gaseous pollutants (SO, NOx, VOCs, etc.) and the formation of secondary particles (nitrites, sulfides, etc.). In addition, increased air conditioning use and increased energy demand, together with the evaporation of organic compounds, can contribute further to the condensation and formation of PM2.5 particles. This interpretation is supported by the graph in Fig. 3A, which shows that PM2.5 concentrations are lowest during winter, when both temperature and humidity reach their lowest levels. According to the filtered data, the average winter temperature was 3.8 $^\circ$C. Further preprocessing showed that the linear correlations between the target variable and the predictors were not statistically significant. Kendall's correlation test, performed at the 5% significance level, indicated that the correlation co-

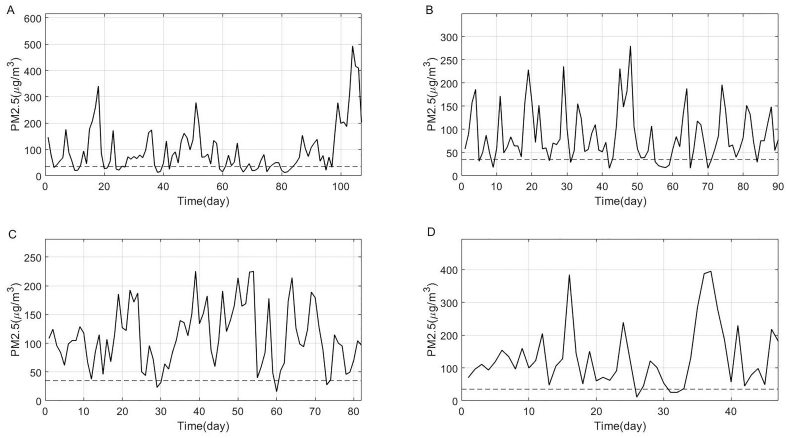


Fig. 3. Typical average daily PM2.5 particle concentrations: A-winter, B-spring, C-summer, D-autumn

efficients were negligible or borderline statistically significant, as presented in the table below.

Table 4. Correlations of PM2.5 particle concentrations with input variables

Parameters	**Correlation with PM2.5**
DEWP	0.303
TEMP	0.128
PRESS	-0.252
cwbd	0.268
Iws	-0.300
Is	-0.007
Ir	0.010

The data listed in the table above lead to the conclusion that the relationships between the variables in the system are nonlinear and complex, which is why it is necessary to use other methods to investigate them. Among the various methods, nonlinear regression is one of the best-known and widely used. Consequently, a nonlinear regression was performed between the target variable y and each predictor x separately, that is:

$$y = a_0 + \sum_{i=1}^{K} a_i x^i \tag{3}$$

where K is the highest degree of the polynomial in Eq. (3). If $K > 1$, it can be assumed that the system is nonlinear, i.e., for $K \gg 1$, the system is highly nonlinear. Fig. 4 shows the results. As can be seen, the degree of nonlinearity K is very high for all predictors, which means that nonlinear relationships prevail between the target variable and the predictors. These results are also consistent with the data in Table 4. In the cases of the DEWP and cwbd predictors, the Kendall coefficient values are borderline statistically significant, since the statistically significant values of the Kendall coefficient ρ occur when $|\rho| > 0.25$.

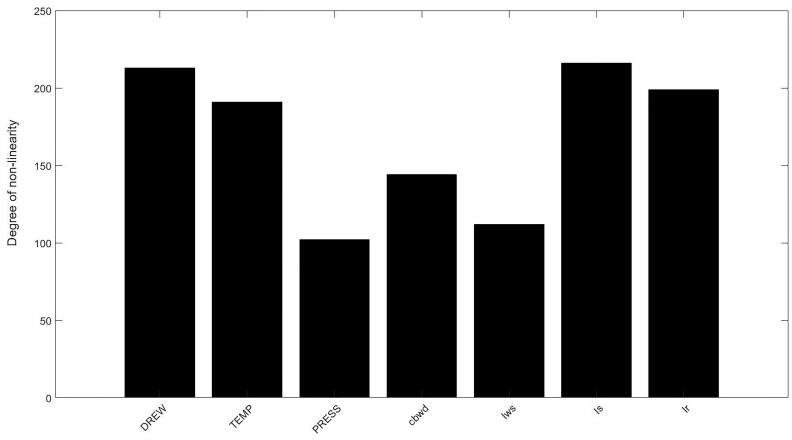


Fig. 4. The degree of nonlinearity between the target variable and the predictor

3.2. Performance criteria

The efficiency of ANNs in air pollution monitoring tasks is commonly evaluated using the following metrics: RMSE, MAE, and MAPE [3]. These metrics are defined by the following equations:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N}} \tag{4}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{5}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_i - y_i}{x_i}\right| \cdot 100 \tag{6}$$

3.3. Artificial neural networks

For air pollution prediction, LSTM networks and ANNs based on the LM algorithm were employed. LSTM networks, as a class of recurrent neural networks, are widely used in air pollution prediction problems [20]. They consist of conventional neural layers, including one or more characteristic memory layers, or cells. These layers differ in both structure and function, as they store short-term cell states and thereby enable the network to capture long-range dependencies in the data. This makes such networks particularly suitable for tasks that involve sequential data types. The structure of these cells is shown in Fig. 5. Four units with specific functions can be identified. The symbols σ and $\tanh$ represent the standard transfer functions sigmoid and hyperbolic tangent, and the operator $\odot$ denotes element-wise matrix multiplication. The operation of the memory layer can be described as follows: a sample vector X_t from the input dataset propagates through the characteristic units f_t, i_t, $\bar{C}_t$, and O_t, also called gates, while interacting with the previous states C_{t-1} and h_{t-1} of the layer (cell), resulting in updated states C_t and h_t. The first unit, f_t, known as the forget gate, is given by the formula:

$$f_t = \sigma\big(W_f \cdot X_t + R_f \cdot h_{t-1} + b_f\big) \tag{7}$$

where W_f, R_f, and b_f denote, respectively, the corresponding neural and recurrent weights and the bias matrices. Passing the vector through these gates, i.e., by applying Eq. (7), the previous state of the cell is filtered, i.e., the unnecessary part of the previous state determined by the vector h_{t-1} is discarded or forgotten. Continuing further propagation through the input gate i_t, the cell state is prepared to be updated with new information. This is described by the following equation:

$$i_t = \sigma\big(W_i \cdot X_t + R_i \cdot h_{t-1} + b_i\big) \tag{8}$$

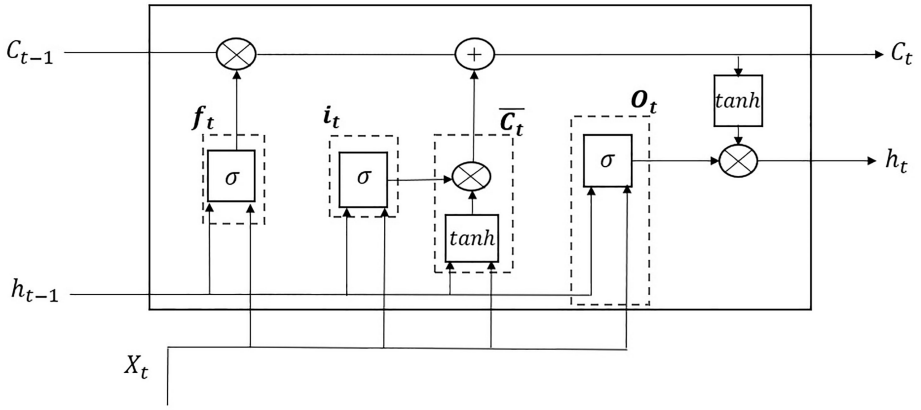


Fig. 5. Memory layer in LSTM networks

In the next phase, new information is added to the cell state:

$$C_t = f_t \odot C_{t-1} + i_t \odot \bar{C}_t \tag{9}$$

with a previously prepared candidate $\bar{C}_t$:

$$\bar{C}_t = \tanh\left(W_c \cdot X_t + R_C \cdot h_{t-1} + b_C\right) \tag{10}$$

Applying the output gate O_t then gives:

$$O_t = \sigma\big(W_o \cdot X_t + R_o \cdot h_{t-1} + b_o\big) \tag{11}$$

so that the new cell state vector h_t is finally obtained:

$$h_t = O_t \odot \tanh(C_t) \tag{12}$$

The structure of the LSTM network consisted of three layers: an input layer, an LSTM layer, and an output layer. Because this study considered real-time prediction of the benzene concentration based on the current values of the predictors, a single hidden LSTM layer was sufficient. The input layer, as the initial layer, had a number of neurons equal to the number of predictors, while the output layer had one neuron because the target variable was benzene concentration. The number of neurons in the LSTM layer was determined by optimization, which will be discussed in more detail later in the text.

The LM algorithm is one of the most efficient ANN algorithms for small and medium-sized samples [62, 55, 40]. Its modification with Bayesian regularization makes it particularly suitable for comparison with other algorithms, since this type of regularization introduces a special self-penalization of the model, equivalent to the principle of Occam's razor [34, 16]. This algorithm minimizes the following objective function:

$$F = \beta E_D(D \mid w, M) + \alpha E_w(w \mid M) \tag{13}$$

where $E_D = \frac{1}{M}\sum_{i=1}^{P}(y_i - y_{it})^2$ represents the standard objective function, while $E_w = \frac{1}{N}\sum_{i=1}^{N} w_i^2$ is the sum of the squares of the neural weights, and α and β are hyperparameters that need to be determined. The second term in Eq. (13) reflects the stochastic nature of the neural weight distribution and assumes that, for each point in the ANN hyperspace, multiple estimators may produce the same value of E_D. Therefore, by applying Bayes' theorem, the posterior probability $P(w \mid D, \alpha, \beta, M)$ is introduced:

$$P(w \mid D, \alpha, \beta, M) = \frac{P(D \mid w, \beta, M) \cdot P(w \mid \alpha, M)}{P(D \mid \alpha, \beta, M)} \tag{14}$$

where D is the input dataset, M is the estimator mathematical model, $P(w \mid \alpha, M)$ is the prior probability, $P(D \mid w, \beta, M)$ is the likelihood function and $P(D \mid \alpha, \beta, M)$ is the normalization factor. More specifically, application of Eq. (14) requires the determination of the maximum posterior probability so that, for a given distribution of neural weights, the most probable model M can be identified in each epoch. Globally, this corresponds to minimizing the objective function given by Eq. (13). The data distribution or likelihood function is given by the following expression [34]:

$$P(D \mid w, \beta, M) = \frac{e^{-\beta E_D(D|w,M)}}{Z_D} \quad ; \quad P(w \mid \alpha, M) = \frac{e^{-\alpha E_w(w|M)}}{Z_w} \tag{15}$$

where $Z_w = \left(\frac{2\pi}{\alpha}\right)^{N/2}$ and $Z_D = \left(\frac{2\pi}{\beta}\right)^{N/2}$. The normalization factor $P(D \mid \alpha, \beta, M)$ can be determined on the basis of Bayes' theorem:

$$P(\alpha, \beta \mid D, M) = \frac{P(D \mid \alpha, \beta, M) \cdot P(\alpha, \beta \mid M)}{P(D \mid M)} \tag{16}$$

Considering the general expression for the posterior probability:

$$P(w \mid D, \alpha, \beta, M) = \frac{e^{-F(w)}}{Z_F} \tag{17}$$

where $Z_F = \int d^3w\, e^{-F(w,\alpha,\beta)}$. Combining equations (14)–(17) yields the final posterior probability expression:

$$P(D \mid \alpha, \beta, M) = \frac{Z_F}{Z_D \cdot Z_w} \tag{18}$$

Determining the maximum posterior probability corresponds to minimizing the probability given in Eq. (18). The constants α and β are obtained from $A = \log\big(P(D \mid \alpha, \beta, M)\big)$ by solving:

$$\frac{dA}{dk} = 0, \quad k \in \{\alpha, \beta\} \tag{19}$$

Further details of the derivation are given in [34, 16], resulting in the following:

$$\alpha = \frac{\gamma}{2E_w(w_p)} \quad ; \quad \beta = \frac{N - \gamma}{2E_D(w_p)} \quad ; \quad \gamma = K - \mathrm{tr}(H^{-1}) \tag{20}$$

where $0 \leq \gamma \leq K$, and w_p denotes the most probable value of the neural weights obtained from the classical LM algorithm. The neural weights in the LM algorithm are updated in each epoch (iteration) according to the following formula:

$$W_{\text{new}} = W_{\text{old}} - \mu H^{-1} \cdot J \cdot e \tag{21}$$

In Eq. (20) and Eq. (21), the quantity $H \simeq 2\beta J^T \cdot J + 2\alpha I$ is the Hessian matrix, J is the Jacobian of the first derivatives of the network error e and tr is the trace operator.

A brief description of the LMBR algorithm can now be given. The procedure begins with loading the input data and the initial values of the constants α, β, and K. After the data have been propagated through the network, the neural weight matrices W are updated according to Eq. 21. Within the same epoch, the coefficients α, β, and γ are also updated, according to Eq. 20, and, at the end of the epoch, the convergence condition $|E_D| < \varepsilon \sim 10^{-4}, 10^{-5}$ is tested. If the condition is fulfilled, the entire cycle is terminated and the resulting model is saved. Otherwise, the cycle continues until convergence is achieved or the maximum number of epochs is reached.

3.4. Determining the optimal structure of artificial neural networks

Since the study involved two databases, special attention was paid to the fact that machine learning methods, including ANNs, are highly data-driven. Consequently, numerical experiments were performed separately to determine the optimal ANN structures for both

databases. The results showed that, for the Air Quality dataset, the optimal number of hidden neurons was 16 for the LMBR and LM networks and 11 for the LSTM network. In contrast, for the Beijing dataset, the optimal number of neurons in the hidden layer was 9 neurons for the LSTM network and 50 neurons for the LMBR and LM networks. In all cases, the input layer had a number of neurons equal to the number of predictors in the system, while the output layer had one neuron, since this was a regression task. To reduce the risk of overtraining, cross-validation was applied. The numerical experiments showed that 10-fold cross-validation provided the best results for the LSTM network on both databases, whereas 5-fold cross-validation provided the best results for the LMBR and LM networks. For all network configurations, the learning rate was set to the default value of 0.01, while the maximum number of epochs was 500. LM and LMBR operated in batch mode, while the LSTM network operated in semi-batch mode with a batch size of 80 samples.

4. Results and discussion

The results of the simulations are shown in Fig. 6 (Air Quality set) and Fig. 7 (Beijing PM2.5 set).

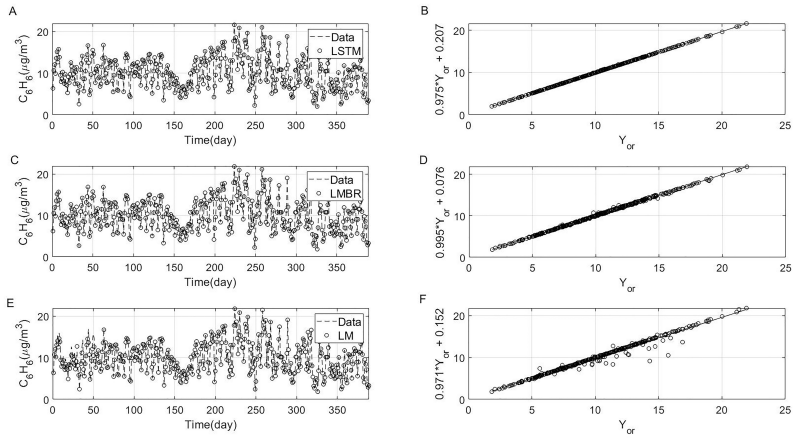


Fig. 6. Actual and simulated values of daily benzene concentrations for LSTM (A), LMBR (C) and LM (D) networks and regression plots: LSTM (B), LMBR (D) and LM (F)

To better visualize the differences in the simulation results, the corresponding values of the efficiency measures are listed in Table 5 and Table 6.

A visual inspection of the parameter values in Tables 5 and 6 shows that the LSTM networks perform efficiently in both cases, whereas LMBR and LM are competitive in the case of the Air Quality dataset. This can be explained by the presence of strong linear, i.e. monotonic, relationships between the benzene concentration and the predictors.

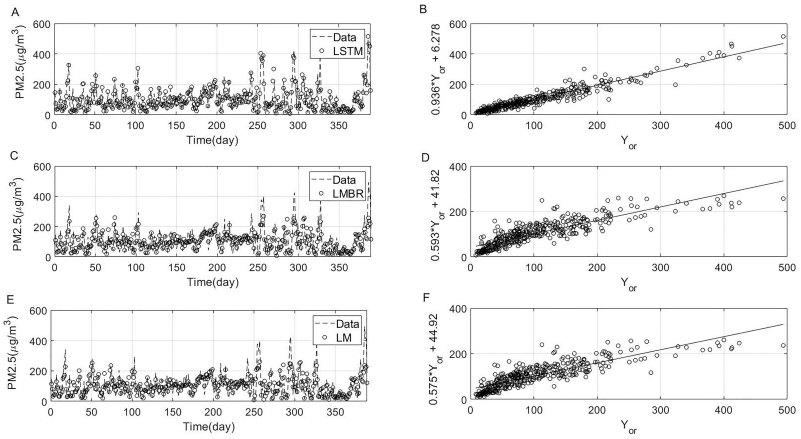


Fig. 7. Actual and simulated values of daily PM2.5 particle concentrations for LSTM (A), LMBR (C) and LM (D) networks and regression plots:LSTM (B), LMBR (D) and LM (F)

Table 5. Efficiency of ANNs in Air quality dataset

Model	**RMSE** (μg/m^3)	**MAE** (μg/m^3)	**MAPE** (%)
LSTM	0.11	0.09	1
LMBR	0.14	0.092	1
LM	0.57	0.2	2

Table 6. Efficiency of ANNs in Beijing PM2.5 dataset

Model	**RMSE** (μg/m^3)	**MAE** (μg/m^3)	**MAPE** (%)
LSTM	45.74	30.32	52
LMBR	64.94	42.83	72
LM	65.68	54.5	74.25

Based on the MAPE values reported in Tables 5 and 6, it can also be concluded that the LSTM networks are more robust in the case of the Beijing PM2.5 dataset, where monotonic relations are not dominant. The observed competitiveness of the LM/LMBR and LSTM networks across the two datasets provides an answer to research question RQ1. More importantly, the results highlight two key data characteristics that strongly influence model performance: sample size and monotonicity. This conclusion follows directly from the contrasting properties of the two datasets considered in this study.

Recent studies have examined the relationship between multilayer perceptron (MLP) networks, including LMBR/LM models and LSTM networks [66, 14, 50, 17]. The findings indicate that no universally optimal model exists, since performance depends largely on the structure of the dataset, the sequence length, and the processing of input attributes.

In general, LSTM networks tend to perform better in tasks involving sequential or temporal data, whereas MLP models can remain competitive when temporal dependencies are weak or when the dataset is relatively small ($<$10,000 samples). The results obtained in this study are consistent with these findings. Based on the RMSE, MAE, and MAPE values discussed previously, the performance of LMBR/LM networks is comparable to that of LSTM networks for the Air Quality dataset (9358 samples). However, for the Beijing dataset, where non-monotonicity relations dominate, LSTM networks achieve superior performance. Moreover, recent studies have shown that incorporating monotonic relationships between input variables can significantly improve the performance of ANNs, even when the available dataset is relatively small [63, 48].

The influence of monotonic relationships on the performance of ANNs is also indirectly confirmed by the Universal Approximation Theorem [23, 11]. According to this theorem, an ANN with a single hidden neural layer that uses a nonlinear transfer function can approximate any continuous function defined over a compact subset of R^n to arbitrary accuracy. In the mathematical literature, it has been shown that every non-decreasing (monotonic) function is pointwise and uniformly continuous and that, in general, monotonic functions have a finite and countable set of discontinuities, often suggesting a strong tendency toward continuity [45, 2].

The significant influence of monotonicity relationships has also been confirmed by other authors. Building on the monotonicity requirement in traditional generalized linear models, the authors showed that introducing a monotonicity constraint in ANNs leads to improved predictive accuracy [46]. They achieved this by introducing a new set of hidden-layer transfer functions. This set consisted of three new zero-centered transfer functions combined with the original ReLU transfer function. This approach led to a fully connected neural layer with constrained monotonicity that can control the concavity and convexity properties of its output function. Accordingly, ANNs formed from these layers are called constrained monotone neural networks. Furthermore, the influence of the monotonicity property was also considered in a study related to the distribution of biological species as a function of environmental and climatic conditions modeled by a range of different variables [21]. As this field was dominated by traditional regression analysis models, linearity was an important assumption that was often not satisfied, and attempts to ease this assumption using the structured additive regression model (STAR) did not yield satisfactory results. Therefore, the authors introduced the concept of probability of finding a species at a given point in space and time. This probability is defined as the logistic transformation of the regression function: $f = \mathbf{B}\boldsymbol{\beta}$, where $\mathbf{B}$ is a spline in the form of a matrix $n \times m$ (n – the number of samples and m – the order of the spline), while $\boldsymbol{\beta}$ represents the corresponding coefficients in the form of a matrix $m \times 1$. The regression function was decomposed into a global component that takes into account the effects of the input variables and an additional component related to non-stationary effects and spatial and spatiotemporal autocorrelations. The model is estimated by minimizing the least-squares criterion, with the monotonicity constraint controlled by an additional sum of squared residuals. The results clearly showed that the fit with the monotonicity constraint was superior to the fit without this constraint. Another interesting approach that confirms the advantage of the monotonicity property is the Deep Lattice Network. Using a variant of these networks (Deep Lattice Cross Network), researchers predicted aerodynamic-force values with high accuracy [73]. This multi-purpose machine learning model for predict-

ing lift and drag coefficients was trained on the basis of fluid-dynamics simulation results. Excellent agreement was observed between the results of these models.

Tables 5 and 6 also show that LMBR networks are more efficient than LM networks for both datasets. The advantages of the LMBR algorithm have also been observed in other studies. Of particular interest is the sharp decline in the performance of all networks for the Beijing for the, where nonlinear and non-monotonic relationships between the input variables dominate. Based on these results, it can be concluded that the LSTM networks are more robust to the nonlinearity of the input data, as indicated by the parameter values in the tables above. These results can be considered in the context of the answer to research question RQ1.

In most studies, the theoretical analysis of the impact of nonlinear relationships in the system on LSTM networks has been limited to numerical experiments and comparisons with classical methods, because an analytical approach is often too complex or unavailable. For example, using a special type of modified LSTM network with recurrent feedback (OR-LSTM), researchers have shown that classical nonlinear system dynamics can be successfully reconstructed through nonlinear transfer functions and memory mechanisms [9]. Van der Pol and Duffing oscillators, as well as other nonlinear mechanical systems, were examined using OR-LSTM networks, and their typical characteristic behaviors, such as chaotic regimes, were successfully reconstructed. These findings are consistent with the results of the present study. Similarly, it has been shown that LSTM networks can successfully reconstruct high-dimensional chaotic systems such as the Lorenz 96 model and the Kuramoto–Sivashinsky system [60]. In this manner, researchers also examined nonlinear stochastic dynamical systems with noise, such as stochastic Van der Pol and Mackey–Glas oscillators [67]. It was shown that LSTM networks can track the evolution of the states of these dynamical models, not deterministically, but instead by modeling the probabilities of transitions between system states. This approach has proven to be more robust than classical methods. Similar results have been reported in several other studies [6, 32]

The advantage of LMBR networks on smaller datasets has also been observed in other studies. Using air pollution data from the Putrajaya area in Malaysia, the authors evaluated three different methods for air pollution analysis and compared their performance [37]. These methods were the autoregression moving average (ARIMA) model and 40 ANNs based on the LMBR and conjugate gradient algorithms. They forecasted the API (Air Pollutant Index) as the target variable as a function of various pollutants (CO, O_3, PM_{10}) and showed that the LMBR algorithm outperformed the other two methods based on the mean squared error (MSE) and MAPE values. The lowest MSE values were 19.43 for LMBR, 19.6 for the conjugate gradient method, and 96.74 for ARIMA, while the corresponding MAPE values were 8.57 for LMBR, 8.82 for the conjugate gradient method, and 23.44 for ARIMA.

The results shown in Fig. 1 and Fig. 3, obtained by the IVS technique, represent the answer to the second research question RQ2. This technique is usually used in other studies to filter the most influential variables in order to improve the performance of ML methods. Although the most significant variables are selected for the ML method in this way, some important relationships between the input variables are also eliminated in the process. In addition, it remains an open question whether the same level of efficiency could be achieved with further parameter tuning in the hyperspace using all input vari-

ables. In this study, IVS techniques were applied to detect seasonal trends in the input data. In both databases, temperature, relative humidity, and absolute humidity are identified as highly influential variables. The influence of air humidity was confirmed in another study in which the authors examined the possibility of removing volatile organic compounds from the air using non-thermal plasma technology (NTP) [33]. They examined the properties of benzene oxidation during dielectric barrier discharge in a plasma reactor, which consisted of two coaxial quartz tubes. The inner tube contained a silver wire with a diameter of 8 mm, which served as the inner electrode and was connected to a voltage source. The outer stainless steel electrode was grounded and attached to the wall of the outer electrode. A gas mixture consisting of benzene, nitrogen, and oxygen was introduced into the plasma formed between these two electrodes to simulate real industrial conditions. With controlled introduction of water vapor into this system, it was shown that benzene decomposition increased with relative humidity until it reached 60%, after which it began to decrease. The efficiency of benzene decomposition initially increases because benzene is oxidized by OH radicals formed through collisions between electrons and water molecules. In addition, OH radicals have a substantially higher oxidation potential than oxygen and other oxidants present in NTP plasma. At relative humidity values greater than 60%, the efficiency of benzene decomposition begins to decline, indicating that water vapor plays a dual role in the process.

A similar result was reported in another study in which researchers measured the concentration of air pollutants in the Mali Lošinj area in Croatia [19]. Using multisorbent sample tubes and the DDA technique, they determined the concentration of non-methane hydrocarbons. They showed that during the autumn period, when the concentration of OH radicals is low, the concentration of hydrocarbons is high, which indirectly implies the interaction of OH radicals and benzene. The most abundant hydrocarbons in their measurements were benzene, toluene, propane, and ethyne. An additional interesting finding was reported in the same study. Specifically, it was found that hydrocarbons with five or more carbon atoms, such as benzene, toluene, etc., are positively correlated with each other. This was also the case in the present study, where high statistically significant correlations of benzene with other hydrocarbons were also confirmed ($\rho = 0.77$). The most likely explanation for this phenomenon is traffic as a common source of NMHC and benzene, as shown in Fig. 5. This may also explain why NMHC was selected as an input variable, since the mrMR algorithm selects variables that are weakly correlated with each other and highly correlated with the output variable.

5. Limitations

In this work, input data filtering techniques were applied to increase the generality of the analysis, as they are independent of the AI methods to which they pass the selected variables. It is important to note that these techniques were used to improve data preprocessing and detect seasonal trends, although they are more commonly used to select the optimal set of input variables. In such cases, special attention must be paid to the fact that not all IVS methods are independent of the AI method to which they pass the selected variables. This is particularly important because, in addition to the ANN types examined in this study, there are other machine learning methods that can be combined with various IVS techniques.

6. Conclusion

A comparative analysis of LSTM networks and neural networks based on the LM algorithm was performed for air pollution prediction tasks. The results showed that ANNs based on the LM algorithm were competitive with LSTM networks when there were strong linear correlations between the target variable and the predictors. When the relationships between the variables were non-monotonic and nonlinear, the performance of all networks decreased significantly, although the LSTM network showed greater robustness than the other models. More broadly, this issue should also be examined in forecasting tasks, where LSTM networks are expected to have a substantial advantage due to their memory capabilities. This remains an important direction for future research. The results indicate that the structure of the input data remains one of the most influential factors affecting the performance of machine learning methods.

Funding. This research did not receive external funding.

Conflicts of interest. The authors declare no conflict of interest.

References

1. Abirami, S., Chitra, P.: Regional air quality forecasting using spatiotemporal deep learning. Journal of Cleaner Production 283, 125341 (2021)
2. Apostol, T.M.: Mathematical Analysis. Addison-Wesley, Reading, MA, 2 edn. (1974)
3. Azan, A.N.A.M., Mototo, N.F.A.M.Z., Mah, P.J.W.: The comparison between arima and arfima model to forecast kijang emas (gold) prices in malaysia using mae, rmse and mape. Journal of Computing Research and Innovation 6(3), 22–33 (2021)
4. Batur, M., Babii, K.: The performance analysis of deep learning algorithms for modelling and forecasting the particulate matter (pm10) in the eastern part of turkey. In: IOP Conference Series: Earth and Environmental Science. vol. 1348, p. 012046. IOP Publishing (2024)
5. Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G., Di Carlo, P.: Recursive neural network model for analysis and forecast of pm10 and pm2. 5. Atmospheric Pollution Research 8(4), 652–659 (2017)
6. Bonassi, F., La Bella, A., Panzani, G., Farina, M., Scattolini, R.: Deep long-short term memory networks: Stability properties and experimental validation. In: 2023 European Control Conference (ECC). pp. 1–6. IEEE (2023)
7. Cateni, S., Colla, V., Vannucci, M.: Improving the stability of the variable selection with small datasets in classification and regression tasks. Neural Processing Letters 55(5), 5331–5356 (2023)
8. Chao, B., Guang Qiu, H.: Air pollution concentration fuzzy evaluation based on evidence theory and the k-nearest neighbor algorithm. Frontiers in Environmental Science 12, 1243962 (2024)
9. Chen, R., Jin, X., Laima, S., Huang, Y., Li, H.: Intelligent modeling of nonlinear dynamical systems by machine learning. International Journal of Non-Linear Mechanics 142, 103984 (2022)
10. Chen, S.: Beijing pm2.5 [dataset]. https://doi.org/10.24432/C5JS49 (2015), uCI Machine Learning Repository
11. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems 2(4), 303–314 (1989)

12. Dai, H., Huang, G., Wang, J., Zeng, H., Zhou, F.: Prediction of air pollutant concentration based on one-dimensional multi-scale cnn-lstm considering spatial-temporal characteristics: A case study of xi'an, china. Atmosphere 12(12), 1626 (2021)
13. Donnelly, A., Misstear, B., Broderick, B.: Real time air quality forecasting using integrated parametric and non-parametric regression techniques. Atmospheric Environment 103, 53–65 (2015)
14. Fei, X., Ye, M., Du, Z., Miao, H.: A comparative study of mlp and lstm neural networks for shale gas production prediction based on numerical simulation data. PLoS One 20(11), e0336782 (2025)
15. Gulati, S., Bansal, A., Pal, A., Mittal, N., Sharma, A., Gared, F.: Estimating pm2. 5 utilizing multiple linear regression and ann techniques. Scientific Reports 13(1), 22578 (2023)
16. Gull, S.F.: Developments in maximum entropy data analysis. In: Maximum Entropy and Bayesian Methods: Cambridge, England, 1988, pp. 53–71. Springer (1989)
17. Guo, Q., He, Z., Wang, Z.: Assessing the effectiveness of long short-term memory and artificial neural network in predicting daily ozone concentrations in liaocheng city. Scientific Reports 15, 6798 (2025)
18. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research 3(Mar), 1157–1182 (2003)
19. Herjavić, G., Matasović, B., Arh, G., Kovač-Andrić, E.: Investigation of non-methane hydrocarbons at a central adriatic marine site mali lošinj, croatia. Atmosphere 11(6), 651 (2020)
20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
21. Hofner, B., Müller, J., Hothorn, T.: Monotonicity-constrained species distribution models. Ecology 92(10), 1895–1901 (2011)
22. Hong, N., Liu, A., Zhu, P., Zhao, X., Guan, Y., Yang, M., Wang, H.: Modelling benzene series pollutants (btex) build-up loads on urban roads and their human health risks: Implications for stormwater reuse safety. Ecotoxicology and Environmental Safety 164, 234–242 (2018)
23. Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural networks 2(5), 359–366 (1989)
24. Huang, C.J., Kuo, P.H.: A deep cnn-lstm model for particulate matter (pm2. 5) forecasting in smart cities. Sensors 18(7), 2220 (2018)
25. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial intelligence 97(1-2), 273–324 (1997)
26. Kovacs, B., Rusu-Both, R.: Neural networks based real-time indoor air quality monitoring and prediction iot system. In: 2025 29th International Conference on System Theory, Control and Computing (ICSTCC). pp. 812–817. IEEE (2025)
27. Lee, M.C., Chang, J.W., Hung, J.C., Chen, B.L.: Exploring the effectiveness of deep neural networks with technical analysis applied to stock market prediction. Computer Science and Information Systems 18(2), 401–418 (2021)
28. Li, T., Zhang, Q., Peng, Y., Guan, X., Li, L., Mu, J., Wang, X., Yin, X., Wang, Q.: Contributions of various driving factors to air pollution events: Interpretability analysis from machine learning perspective. Environment International 173, 107861 (2023)
29. Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T.: Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. Environmental pollution 231, 997–1004 (2017)
30. Liu, Y., Wen, L., Lin, Z., Xu, C., Chen, Y., Li, Y.: Air quality historical correlation model based on time series. Scientific Reports 14(1), 22791 (2024)
31. Llamelo, C., Medina, R., Fajardo, A.: Performance of enhanced cnn-lstm prediction model for vehicle-mounted air quality monitoring. In: 2024 15th International Conference on Information and Communication Technology Convergence (ICTC). pp. 66–71. IEEE (2024)
32. Llerena Cana, J.P., Garcia Herrero, J., Molina Lopez, J.M.: Forecasting nonlinear systems with lstm: analysis and comparison with ekf. Sensors 21(5), 1805 (2021)

33. Ma, T., Zhao, Q., Liu, J., Zhong, F.: Study of humidity effect on benzene decomposition by the dielectric barrier discharge nonthermal plasma reactor. Plasma Science and Technology 18(6), 686 (2016)
34. MacKay, D.J.: Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. Network: computation in neural systems 6(3), 469 (1995)
35. Manisalidis, I., Stavropoulou, E., Stavropoulos, A., Bezirtzoglou, E.: Environmental and health impacts of air pollution: a review. Frontiers in public health 8, 14 (2020)
36. Méndez, M., Merayo, M.G., Núñez, M.: Long-term traffic flow forecasting using a hybrid cnn-bilstm model. Engineering Applications of Artificial Intelligence 121, 106041 (2023)
37. Mun, C.K., Abd Rahman, N.H., Ilias, I.C.: Performance of levenberg-marquardt neural network algorithm in air quality forecasting. Sains Malaysiana 51(8), 2645–2654 (2022)
38. Năstase, G., Șerban, A., Năstase, A.F., Dragomir, G., Brezeanu, A.I.: Air quality, primary air pollutants and ambient concentrations inventory for romania. Atmospheric Environment 184, 292–303 (2018)
39. Neo, E.X., Hasikin, K., Lai, K.W., Mokhtar, M.I., Azizan, M.M., Hizaddin, H.F., Razak, S.A., et al.: Artificial intelligence-assisted air quality monitoring for smart city management. PeerJ Computer Science 9, e1306 (2023)
40. Ngia, L.S., Sjoberg, J.: Efficient training of neural nets for nonlinear adaptive filtering using a recursive levenberg-marquardt algorithm. IEEE Transactions on Signal Processing 48(7), 1915–1927 (2002)
41. Prado-Rujas, I.I., Garcia-Dopico, A., Serrano, E., Córdoba, M.L., Pérez, M.S.: A multivariable sensor-agnostic framework for spatio-temporal air quality forecasting based on deep learning. Engineering Applications of Artificial Intelligence 127, 107271 (2024)
42. Radovic, M., Ghalwash, M., Filipovic, N., Obradovic, Z.: Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC bioinformatics 18(1), 9 (2017)
43. Ramírez-Gallego, S., Lastra, I., Martínez-Rego, D., Bolón-Canedo, V., Benítez, J.M., Herrera, F., Alonso-Betanzos, A.: Fast-mrmr: Fast minimum redundancy maximum relevance algorithm for high-dimensional big data. International Journal of Intelligent Systems 32(2), 134–152 (2017)
44. Ren, Q.: Air quality prediction based on lstm algorithm. In: Sixth International Conference on Electromechanical Control Technology and Transportation (ICECTT 2021). vol. 12081, pp. 1119–1127. SPIE (2022)
45. Rudin, W.: Principles of mathematical analysis. 3rd ed. (1976)
46. Runje, D., Shankaranarayana, S.M.: Constrained monotonic neural networks. In: International Conference on Machine Learning. pp. 29338–29353. PMLR (2023)
47. Sarigiannis, D.A., Karakitsios, S.P., Gotti, A., Papaloukas, C.L., Kassomenos, P.A., Pilidis, G.A.: Bayesian algorithm implementation in a real time exposure assessment model on benzene with calculation of associated cancer risks. Sensors 9(02), 731–755 (2009)
48. Schlieper, P., Dombrowski, M., Nguyen, A., Zanca, D., Eskofier, B.: Data-centric benchmarking of neural network architectures for the univariate time series forecasting task. Forecasting 6(3), 718–747 (2024)
49. Schürholz, D., Kubler, S., Zaslavsky, A.: Artificial intelligence-enabled context-aware air quality prediction for smart cities. Journal of Cleaner Production 271, 121941 (2020)
50. Seo, B., Yoon, Y., Lee, K.H., Cho, S.: Comparative analysis of ann and lstm prediction accuracy and cooling energy savings through ahu-dat control in an office building. Buildings 13(6), 1434 (2023)
51. Singh, S., Gupta, P.: Comparative study id3, cart and c4. 5 decision tree algorithm: a survey. International Journal of Advanced Information Science and Technology (IJAIST) 27(27), 97–103 (2014)

52. Stockwell, D.R., Peterson, A.T.: Effects of sample size on accuracy of species distribution models. Ecological modelling 148(1), 1–13 (2002)
53. Sun, W., Huang, C.: A hybrid air pollutant concentration prediction model combining secondary decomposition and sequence reconstruction. Environmental Pollution 266, 115216 (2020)
54. Tao, H., Jawad, A.H., Shather, A., Al-Khafaji, Z., Rashid, T.A., Ali, M., Al-Ansari, N., Marhoon, H.A., Shahid, S., Yaseen, Z.M.: Machine learning algorithms for high-resolution prediction of spatiotemporal distribution of air pollution from meteorological and soil parameters. Environment international 175, 107931 (2023)
55. Tijani, M.A., Nwiabu, N.D., Bennet, E.O.: An artificial neural network model for small and medium size data analysis using levenberg-marquardt optimization algorithm. Journal of Scientific and Engineering Studies (JSES) 10(6), 29–39 (2024)
56. Uwimana, E., Zhou, Y., Sall, N.M.: A short-term load demand forecasting: levenberg–marquardt (lm), bayesian regularization (br), and scaled conjugate gradient (scg) optimization algorithm analysis. The Journal of Supercomputing 81(1), 55 (2025)
57. Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J.: Machine learning algorithm validation with a limited sample size. PloS one 14(11), e0224365 (2019)
58. Ventura, P., Khodamoradi, M., Costa, R., Figueiras, P., Jardim-Gonçalves, R.: Real-time environmental monitoring in smart buildings using federated learning. Procedia Computer Science 263, 680–687 (2025)
59. Vito, S.: Air Quality [dataset] (2008), https://doi.org/10.24432/C59K5F, uCI Machine Learning Repository
60. Vlachas, P.R., Byeon, W., Wan, Z.Y., Sapsis, T.P., Koumoutsakos, P.: Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 474(2213) (2018)
61. Wang, W., Mao, W., Tong, X., Xu, G.: A novel recursive model based on a convolutional long short-term memory neural network for air pollution prediction. Remote Sensing 13(7), 1284 (2021)
62. Waseem, M., Lin, Z., Yang, L.: Data-driven load forecasting of air conditioners for demand response using levenberg–marquardt algorithm-based ann. Big Data and Cognitive Computing 3(3), 36 (2019)
63. Wu, X., Fu, D., Li, Z., Yang, T.: Graph-enhanced & monotonic embeddings: A novel approach to tabular data representation. Neurocomputing p. 132108 (2025)
64. Xayasouk, T., Lee, H., Lee, G.: Air pollution prediction using long short-term memory (lstm) and deep autoencoder (dae) models. Sustainability 12(6), 2570 (2020)
65. Xiao, F., Yang, M., Fan, H., Fan, G., Al-Qaness, M.A.: An improved deep learning model for predicting daily pm2. 5 concentration. Scientific reports 10(1), 20988 (2020)
66. Yahia, A.B., Kadir, I., Abdallaoui, A., Elazhari, K.: Architectural optimization of a multilayer perceptron (mlp) neural network enhanced by the levenberg–marquardt algorithm for predicting relative humidity: application to tangier, morocco. Water SA 51(3), 279–287 (2025)
67. Yeo, K., Melnyk, I.: Deep learning algorithm for data-driven simulation of noisy dynamical system. Journal of Computational Physics 376, 1212–1231 (2019)
68. Yonar, A., Yonar, H.: Modeling air pollution by integrating anfis and metaheuristic algorithms. Modeling Earth Systems and Environment 9(2), 1621–1631 (2023)
69. Yuan, H., Xu, G., Lv, T., Ao, X., Zhang, Y.: Pm2. 5 forecast based on a multiple attention long short-term memory (mat-lstm) neural networks. Analytical Letters 54(6), 935–946 (2021)
70. Zhang, B., Zhang, Y., Zhang, K., Zhang, Y., Ji, Y., Zhu, B., Liang, Z., Wang, H., Ge, X.: Machine learning assesses drivers of pm2. 5 air pollution trend in the tibetan plateau from 2015 to 2022. Science of the Total Environment 878, 163189 (2023)
71. Zhang, L., Liu, P., Zhao, L., Wang, G., Zhang, W., Liu, J.: Air quality predictions with a semi-supervised bidirectional lstm neural network. Atmospheric Pollution Research 12(1), 328–339 (2021)

72. Zhang, Y.L., Cao, F.: Fine particulate matter (pm2. 5) in china at a city level. Scientific reports 5(1), 14884 (2015)
73. Zhao, J., Zeng, L., Lin, A., Shao, X.: Deep learning prediction method for aerodynamic forces on morphing aircraft considering physical monotonicity. Advances in Aerodynamics 7(1), 7 (2025)
74. Zhou, Y., Chang, F.J., Chang, L.C., Kao, I.F., Wang, Y.S.: Explore a deep learning multi-output neural network for regional multi-step-ahead air quality forecasts. Journal of cleaner production 209, 134–145 (2019)

Goran Keković is an Associate Professor at Faculty of Information Technology, Alfa BK University. He received his PhD in 2008 from the School of Electrical Engineering, University of Belgrade. His early research focused on condensed matter theory, after which he shifted his interests toward biosignal analysis (EEG) and medical data processing, employing statistical methods and machine learning techniques. His current research interests include data processing and artificial intelligence.

Rade Božović received his M.Sc. degree in 2007, and Ph.D. degree in 2019 in Electrical Engineering from the School of Electrical Engineering, University of Belgrade. In 2020, he joined the ALFA BK University as an Assistant Professor. His research interests include wireless communication systems, cognitive radio, signal processing. Also, he works as technical manager at company CRONY, Belgrade, Serbia, specialized for wireless communication.

Sonja Ketin is a Professor at Academy of Technical and Art Applied Studies Belgrade (Serbia), Dpt. Railway College of Applied Sciences. She graduated from the Faculty of Technology and Metallurgy, University of Belgrade, in the field of chemical engineering. At the Faculty of Technical Sciences in Novi Sad, she enrolled in post-graduate studies in the field of Environmental Engineering (master's degree and doctoral dissertation).

Vladimir Mikić is an Assistant Professor at the Faculty of Information Technology, Alfa BK University. He received his Ph.D. in Information Technology from the same institution. His research interests include technology-enhanced learning, personalized e-learning systems, learning analytics, and artificial intelligence in education.

Miloš Ilić is an Assistant Professor at the Faculty of Information Technology, Alfa BK University. He earned his Ph.D. in Information Technology from the same faculty. His research focuses on intelligent tutoring systems, artificial intelligence, data-driven e-learning, and programming.

Boban Vesin is a Professor of IT and Information Systems at the University of South-Eastern Norway. His research focuses on engineering learning technologies using artificial intelligence, learning analytics, and personalized learning, with recent work extending toward explainable AI and its application in medical education and clinical decision support.

Received: November 1, 2025; Accepted: May 20, 2026.