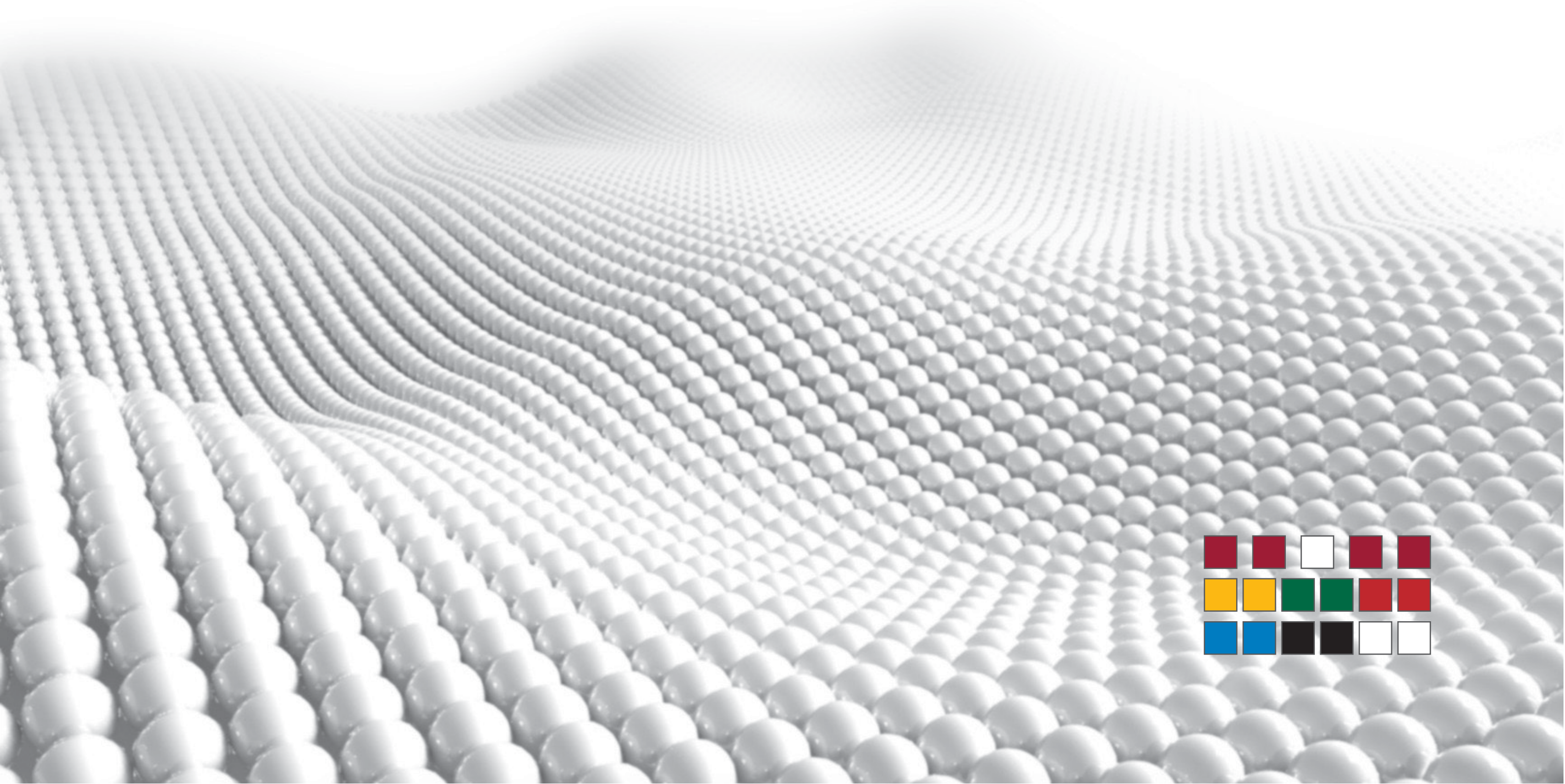
ISSN 2255-8950 (Online)
ISSN 2255-8942 (Print)

Volume 13(2025)
No. 2

Baltic Journal of Modern Computing



CO-PUBLISHERS

Vilnius
University

UNIVERSITY
OF LATVIA

Latvia University
of Life Sciences
and Technologies

Institute of Mathematics and Computer Science
University of Latvia

VIDZEME UNIVERSITY
OF APPLIED SCIENCES

EDITORIAL BOARD

Co-Editors-in-Chief

Managing Co-Editors

Editorial Board Members (in alphabetical order)

Prof. Dr.habil. **Hele-Mai Haav**, Tallinn University of Technology, Estonia
Dr. **Nissim Harel**, Holon Institute of Technology, Israel
Dr. **Delene Heukelman**, Durban University of Technology, South Africa
Prof. em. Dr. **Kazuo Iwama**, Kyoto University, Japan
Prof. Dr.sc.comp. **Anita Jansone**, Liepāja Academy at Riga Technical University, Latvia
PhD **Oskars Java**, Vidzeme University of Applied Sciences, Latvia
Prof. Dr.habil.sc.ing. **Igor Kabashkin**, Corresponding Member of Latvian Academy of Sciences,
Transport and Telecommunication Institute, Latvia
Prof. Dr. **Diana Kalibatienė**, Vilnius Gediminas Technical University, Lithuania
Prof. Dr.habil. sc.comp. **Audris Kalniņš**, Corresponding Member of Latvian Academy of Sciences, Institute
of Mathematics and Computer Science at University of Latvia, Latvia
Assoc. Prof. Dr.phys. **Atis Kapenieks**, Riga Technical University
Prof. Dr. **Egidijus Kazanavičius**, Kaunas University of Technology, Lithuania
Prof. Dr.sc.phys. and math. **Vladimir Kotov**, Belarusian State University, Belarus
Adj. Prof. Dr. **Dmitry Korzun**, Petrozavodsk State University, Russia
Prof. Dr.habil. **Algimantas Krisciukaitis**, Lithuanian University of Health Sciences, Lithuania
Assoc. Prof. Dr. **Olga Kurasova,** Vilnius University, Lithuania
Prof. Dr. **Ivan Laktionov**, Dnipro University of Technology, Dnipro, Ukraine
Assoc. Prof. Dr, **Audronė Lupeikienė,** Institute of Data Science and Digital Technologies,
Faculty of Mathematics and Informatics, Vilnius University
Prof. Dr. **Raimundas Matulevičius**, University of Tartu, Estonia
Prof. Dr.habil.sc.ing. **Yuri Merkuryev**, Full Member of Latvian Academy of Sciences,
Riga Technical University, Latvia
Prof. Dr.habil. **Jean Francis Michon**, retired from University of Rouen, France
Prof. Dr.habil. **Dalius Navakauskas**, Vilnius Gediminas Technical University, Lithuania
Prof. Dr.sc.comp. **Laila Niedrīte**, University of Latvia, Latvia
Assist. Prof. PhD **Anastasija Ņikiforova**, University of Tartu, Tartu, Estonia
Prof. Dr.sc.ing. **Oksana Nikiforova**, Riga Technical University, Latvia
Prof. Dr. **Vladimir A. Oleshchuk**, University of Agder, Norway
Prof. Dr.habil. **Jaan Penjam**, Tallinn University of Technology, Estonia
Assoc. Prof. PhD **Eduard Petlenkov**, Tallinn University of Technology, Estonia
Assoc. Prof. PhD **Ivan I. Piletski**, Belarussian State University of Informatics and Radioelectronics, Belarus
Prof. Dr.math. **Kārlis Podnieks**, University of Latvia, Latvia
Prof. Dr. **Boris Pozin**, Moscow State University of Economics, Statistics and Informatics (MESI),
Russian Federation
Prof. Dr. **Tarmo Robal**, Tallinn University of Technology, Estonia
Prof. Dr. **Andreja Samčović**, University of Belgrade, Serbia
Prof. Dr.sc.eng. **Egils Stalidzāns**, University of Latvia, Latvia
Prof. Dr.sc.comp. **Leo Seļavo**, University of Latvia, Latvia
Prof. Dr. **Janis Stirna**, Stockholm University, Sweden
Prof. Dr.phil. **Jurģis Šķilters**, University of Latvia, Latvia
Prof. Dr.sc.eng. **Uldis Sukovskis**, Riga Technical University, Latvia
Prof. Dr.sc.comp. **Darja Šmite**, Blekinge Institute of Technology, Sweden
Prof. Dr.sc.comp. **Juris Vīksna**, Full Member of Latvian Academy of Sciences,
Institute of Mathematics and Computer Science at University of Latvia, Latvia
Prof. Dr.sc.comp. **Māris Vītiņš**, University of Latvia
Prof. Dr.sc.ing. **Gatis Vītols**, Latvia University of Life Sciences and Technologies
Prof. Dr.art **Solvita Zariņa**, University of Latvia
Prof. em. Dr.rer.nat. habil. **Thomas Zeugmann**, Hokkaido University, Sapporo, Japan

Table of Content

Baltic J. Modern Computing, Vol. 13 (2025), No. 2, pp. 315–330
https://doi.org/10.22364/bjmc.2025.13.2.01

A Neural Network-Based Causal Model for Electricity Demand Estimation in Remote Areas: A Case Study in El Espino, Bolivia

Stefano SANFILIPPO[1] ⋆, José Juan HERNÁNDEZ-CABRERA[2] ⋆⋆, Christoph KÄNDLER[3] ⋆ ⋆ ⋆, José Juan HERNÁNDEZ-GÁLVEZ[2] †, José ÉVORA-GÓMEZ[2] ‡, Octavio RONCAL-ANDRÉS[2] §

[1] STAM S.r.l, Genoa, Italy
[2] Instituto Universitario de Sistemas Inteligentes y Aplicaciones Numéricas en Ingeniería - Universidad de Las Palmas de Gran Canaria, Las Palmas, Gran Canaria, Spain
[3] EIFER - Europäisches Institut für Energieforschung, Karlsruhe, Germany

Abstract. Designing microgrids in remote areas is challenging due to the lack of reliable and high-quality electricity demand data. These limitations arise from technological, economic, and logistical constraints, making it difficult to estimate demand—especially when daily demand curves cannot be fully constructed due to missing or incomplete data. Traditional methods, which rely on consistent and comprehensive datasets, often prove ineffective in such conditions.

To address this issue, this paper introduces a novel causal modelling approach, implemented using a neural network, to uncover the underlying relationships between key influencing factors—such as temperature, humidity, time of day, and seasonal variations—and electricity demand. Rather than requiring complete hourly demand curves as inputs, the proposed approach leverages available data to infer demand patterns more effectively.

We propose a neural network architecture that aims to capture causal dependencies in electricity demand by encoding input features into a high-dimensional latent space. Using an encoder-decoder structure, the encoder maps inputs to a latent space designed to preserve potential causal relationships, while the decoder generates the demand estimation. This approach hypothesizes that this configuration may help to get causal dependencies. To evaluate this, we compared our model against a simpler neural network architecture characterised by a triangular layer structure.

⋆ s.sanfilippo@stamtech.com ORCID: 0009-0001-0547-6222
⋆⋆ josejuan.hernandez@ulpgc.es ORCID: 0000-0003-2427-2441
⋆ ⋆ ⋆ christoph.kaendler@eifer.org ORCID: 0000-0002-0873-1137
† jose.galvez@ulpgc.es ORCID: 0009-0008-3626-7520
‡ jose.evora@ulpgc.es ORCID: 0000-0001-9348-7265
§ octavio.roncal@ulpgc.es ORCID: 0000-0003-3503-3833

Using real-world data from El Espino, Bolivia, our model achieved a Mean Squared Error (MSE) of 0.0511 with the Adam optimiser, representing a 61.8% improvement over the simpler neural network architecture.
A sensitivity analysis further confirmed the relevance of selected input variables, showing that excluding temporal-based features, such as the month of the year and weekend indicator, increased estimation error, with an 11.7% increase in MSE. These findings highlight the model's effectiveness in handling data limitations and its potential as a scalable solution for electricity demand estimation in remote areas.

Keywords: Microgrids, Remote areas, Electricity demand, Causal Model, Estimation, Neural Networks

1 Introduction

Microgrids are a key strategy for achieving electrification in remote areas (Nayanathara and Srilatha, 2018), as they integrate distributed energy resources to provide localised electricity generation, avoiding the challenges of transmission and distribution in inaccessible locations. Successfully designing such systems requires an accurate understanding of demand. Overestimating demand can inflate costs unnecessarily, while underestimating it may lead to undersized systems, which are unable to consistently meet energy needs (Sanfilippo and et al., 2023).

The term demand encompasses a range of needs, including electricity (Castillo et al., 2022), heat (Białek et al., 2022), and cooling (Abugabbara et al., 2022). In this study, the focus is specifically on electricity demand in remote areas, where reliable access to electricity not only improves local economic opportunities but also contributes to enhanced energy efficiency and integration of renewable energy (Stadler et al., 2016).

However, estimating electricity demand in remote areas is inherently challenging. In villages awaiting electrification, baseline demand data is often non-existent. Even in electrified areas, data can be sparse, unreliable, or of poor quality (Wassie and Ahlgren, 2023). Such data limitations hinder the ability to estimate demand accurately, making it difficult to design microgrids efficiently.

Addressing these constraints requires the development of robust, data-driven models tailored to conditions where high-quality, granular demand information is short. The case of El Espino, Bolivia, which offers an unusually large dataset compared to similar contexts, provides a good opportunity to apply artificial intelligence techniques for customised electricity demand estimation. The proposed approach aims to produce a model capable of delivering reliable estimations in remote areas facing persistent information gaps by leveraging this data-rich environment and integrating meta data (temperature and day time) to create a general representative model.

Given the challenges associated with electricity demand estimation in remote areas, various approaches have been explored to minimise the error and improve robustness. Existing methods generally fall into three categories: traditional statistical methods, computational intelligence methods, and hybrid approaches that combine both. This section provides an overview of these methodologies and their effectiveness in addressing the limitations of demand estimation in data-scarce environments.

2 Related Work

In recent years, the energy sector has encountered significant challenges in accurately estimating electricity demand. Various methodologies have been developed to address these challenges, which can generally be categorised into three main approaches: Traditional Statistical Methods, Computational Intelligence Methods, and Hybrid Methods. Each of these approaches offers distinct advantages and limitations, depending on the complexity of the problem, the availability of data, and the need for minimisation of error in forecasting.

Traditional statistical methods are grounded in mathematical principles and rely on historical data patterns, probability distributions, and regression techniques to predict electricity demand. Among these, time series analysis (Velasquez et al., 2022; Dilaver and Hunt, 2011) represents a fundamental approach, leveraging past consumption patterns to identify trends and seasonal variations that inform future demand projections. This method has been widely applied due to its interpretability and strong theoretical foundation. Additionally, econometric models (Dieudonné et al., 2022; Nasr et al., 2000; Gómez and Rodríguez, 2019; Zamanipour et al., 2023) extend the traditional statistical approach by incorporating relationships between electricity demand and macroeconomic indicators such as gross domestic product, economic growth, urbanization, and financial development. These models provide valuable insights into long-term dependencies and external factors affecting energy consumption patterns.

Further expanding on statistical modelling, probability-based approaches are employed to handle uncertainties in electricity demand estimation. Stochastic methods (Lombardi et al., 2019), for instance, integrate randomness into forecasting models, allowing for flexible predictions that account for variations in external influences such as weather fluctuations and human behaviour. Similarly, structural models (Michalik et al., 1997) adopt an engineering-based perspective, focusing on the physical and technical characteristics of the electricity system to determine demand based on infrastructure constraints and efficiency measures.

As the complexity of electricity demand forecasting has increased, computational intelligence methods have gained prominence. These methods, often associated with artificial intelligence and machine learning, are designed to handle non-linear relationships and large-scale datasets, surpassing the predictive capabilities of traditional statistical techniques. Neural networks, for example, have been widely used to model electricity demand by learning intricate consumption patterns through adaptive training mechanisms (Kandananond, 2011; Foldvik Eikeland et al., 2021). Similarly, Random Forest (Shin and Woo, 2022) has been applied to electricity consumption forecasting, offering enhanced accuracy and robustness in handling diverse input variables.

Beyond traditional statistical and computational intelligence approaches, hybrid methods (Shiraki et al., 2016) have emerged as a powerful alternative by integrating multiple forecasting techniques to improve accuracy and adaptability. By combining statistical models with AI-driven approaches, hybrid techniques mitigate the limitations of individual methods and leverage their strengths. For instance, scenario-based methods (Xia et al., 2022) often incorporate both econometric and machine learning components to assess how external variables such as climate change, economic fluctuations, and policy decisions influence electricity demand. Similarly, Geographic Information Systems methods (Torabi Moghadam et al., 2018) benefit from the integration of traditional spatial analysis with computational intelligence techniques to improve region-specific demand estimation, considering factors such as population density, land use, and climatic conditions.

Furthermore, agent-based models highlight another key area where hybridization of techniques proves beneficial. These models simulate the behaviour of individual consumers or groups, allowing for a more dynamic representation of energy demand influenced by social and behavioural patterns (Tian and Chang, 2020). When combined with stochastic and machine learning techniques, agent-based models become highly effective in capturing demand variability and providing more refined insights into consumption trends.

Overall, the landscape of electricity demand forecasting has evolved through the interplay of traditional statistical methods, computational intelligence approaches, and hybrid methodologies. While traditional statistical models offer well-established theoretical foundations and interpretability, computational intelligence techniques provide superior predictive power in handling complex, high-dimensional data. Hybrid methods, by integrating these diverse approaches, present a promising direction for enhancing accuracy and adaptability in demand estimation. The selection of an appropriate method depends on factors such as data availability, forecasting horizon, and the specific characteristics of the electricity market under analysis.

While these approaches have contributed significantly to demand estimation, they often rely on pattern recognition rather than explicitly modelling the causal factors driving electricity consumption. Traditional statistical methods assume stable demand patterns, while computational intelligence methods, such as neural networks, excel at pattern recognition. Hybrid approaches attempt to bridge this gap, yet they remain constrained by data limitations. To overcome these challenges, we propose a model that explicitly encodes causal relationships, enabling a more interpretable and robust framework for electricity demand estimation in remote areas.

3 Proposed Solution

Modelling electricity demand poses significant challenges due to its complex and nonlinear nature, as well as the intricate interdependencies that arise, particularly because user behaviour plays a central role (Lazzari et al., 2022). Traditional demand estimation methods rely on predefined demand curves, which assume stable and well-defined consumption patterns. However, these approaches fail to capture the evolving and dynamic nature of electricity demand, particularly in environments with incomplete or unreliable data. Moreover, they are based on correlations rather than identifying the causal mechanisms that drive demand variations.

Electricity demand is not merely the sum of independent factors but rather the result of dynamic interactions between environmental conditions, socio-economic factors, and technological adoption. These elements influence each other, creating causal dependencies that cannot be fully understood through conventional demand curve-based models alone. As a result, there is a need for an approach that moves beyond static demand profiles to one that models the underlying factors driving electricity consumption.

To address this, the proposed model is designed to describe the causal relationships between these factors, rather than relying on predefined demand curves or purely correlational patterns. The final proposal integrates variables—such as temperature, humidity, time of day, month, and whether it is a weekday or weekend—which were selected based on expert domain knowledge, ensuring that they are known to influence electricity demand. By incorporating expert, driven insights, the model is designed to capture the actual causal factors driving consumption, rather than relying solely on statistical correlations.

By structuring the model around cause-effect relationships, it aims to represent the way external conditions and user behaviour interact to shape electricity consumption. This approach allows for a more interpretable and robust demand estimation framework, capable of adapting to scenarios with incomplete or noisy data while maintaining a meaningful representation of the underlying processes driving electricity demand. The core innovation of this work lies in designing a neural network architecture specifically to capture these causal relationships, rather than merely identifying patterns, a fundamental shift from how neural networks are typically used.

Traditional neural networks excel at recognizing statistical dependencies in data but do not inherently distinguish between correlation and causation. A neural network is a computational model inspired by the structure and function of neurological systems, designed to represent complex, non-linear functions (Neervannan, 2018). It consists of interconnected layers of neurons, with each layer's arrangement determining its specific purpose or function. Optimised using methods such as backpropagation, neural networks effectively map input variables to output estimations, making them ideal for modelling systems like electricity demand, where relationships are non-linear and evolve over time.

Unlike traditional models, neural networks automatically identifying patterns and relations within multi variable datasets and utilize these insights of usually unseen couplings (Scarborough and Somers, 2006). Their ability to capture deep and complex dynamics allows them to represent emergent behaviours, surpassing the limitations of linear or polynomial models (Somers and Casal, 2009).

However, causality remains a challenge in neural networks, as they are traditionally designed to identify patterns rather than capture causal relationships. Our hypothesis is that by encoding inputs into a higher-dimensional space, it becomes possible to reveal underlying causal relationships that may not be apparent in the original input space. By doing so, the network moves beyond conventional pattern recognition, instead aiming to represent how different factors interact and influence electricity consumption in a structured manner.

In the context of electricity demand estimation, this means that the model could begin to identify meaningful dependencies that drive consumption, rather than merely recognizing statistical correlations. We posit that by leveraging the expressive power of high-dimensional representations, the model can uncover the true interactions that govern demand variations.

4 Data Collection, Preparation, and Analysis

During the data collection and analysis phase, available data were gathered, explored, and examined. Previous studies have aimed to establish common data platforms (Fioriti et al., 2023) that streamline this process. Such platforms would significantly reduce the effort required to obtain authentic electricity demand data, currently a task involving extensive literature reviews, contacting authors, interviewing stakeholders, and identifying specialised websites.

Although real measured data were identified for locations in Bolivia, Namibia, Mexico and Tanzania, many datasets contained missing measurements. After removing incomplete curves, the final dataset included a total of 869 days of measured electricity demand. Among these, El Espino in Bolivia provided the largest number of complete days, 578, which surpassed any other region under consideration. Consequently, El Espino was chosen for this study due to its comparatively abundant and reliable data.

The El Espino dataset, sourced from a GitHub repository (Balderrama Subieta, 2022), spans from 1 January 2016 to 31 July 2017, covering 578 days of recorded measurements. It encompasses data from 128 households, a hospital, a school, and street lighting systems and the wattage consumed at each timestamp. First, the data were preprocessed to address inconsistencies, remove outliers, and format it into a structured, consolidated, and normalised dataset.

Visual examinations were conducted to illustrate inherent variability in energy demand patterns. Figures 1 and 2 show the distribution of power usage by hour and by month, respectively, with outliers indicated by dots. Outliers were identified using the three-standard-deviation rule, where any data point x that deviates more than three times the standard deviation (σ) from the mean (μ) is considered an outlier. Mathematically, this is expressed as in Equation 1.

$$x < \mu - 3\sigma \quad \text{or} \quad x > \mu + 3\sigma. \tag{1}$$

By applying this approach, we mitigate the risk of extreme values disproportionately affecting the learning process. This helps preventing the exploding gradient problem, ensuring stable and efficient model training.

This method assumes a normal distribution of energy demand data and effectively detects extreme variations while maintaining robustness in identifying significant deviations. Figure 1 reveals a minimum at 8 a.m. and a peak at 8 p.m., while Figure 2 indicates that October experiences the highest demand.

Figure 3 illustrates the energy demand on weekends, highlighting significant patterns influenced by the hour of the day, the month of the year, and the effects seen during weekend. The heatmap reveals a distinct hourly pattern, with higher demand during specific hours, such as evenings. Additionally, a seasonal effect is evident, with increased energy usage during colder months, likely due to heating needs, and during warmer months, possibly due to cooling systems. The effects seen during weekend is also noticeable, as the patterns differ from those typically observed on weekdays, reflecting variations in social and economic activities. These insights are crucial for modelling energy demand and considering time-of-use factors in energy estimation.

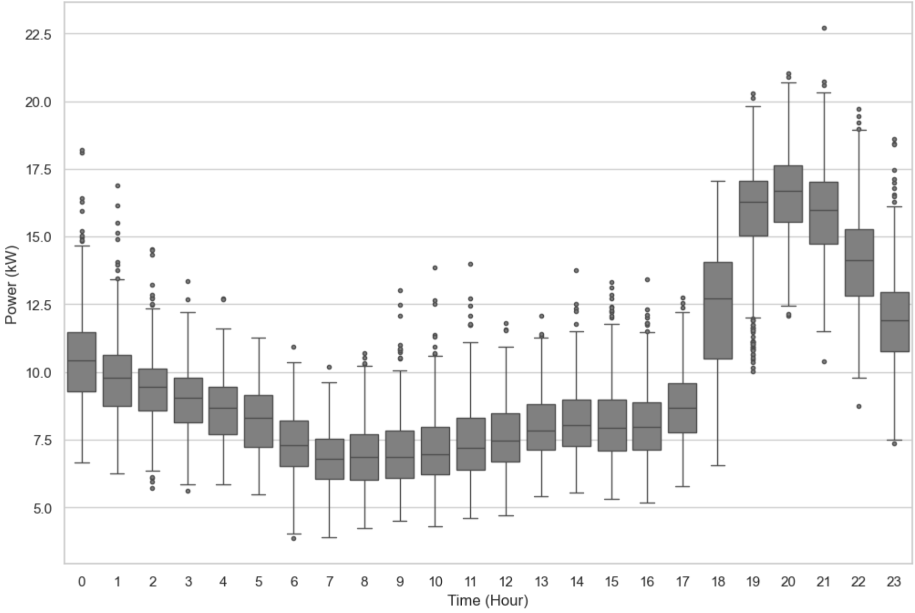


Fig. 1. Hourly Distribution of Electricity Demand Represented as a Boxplot with Outliers

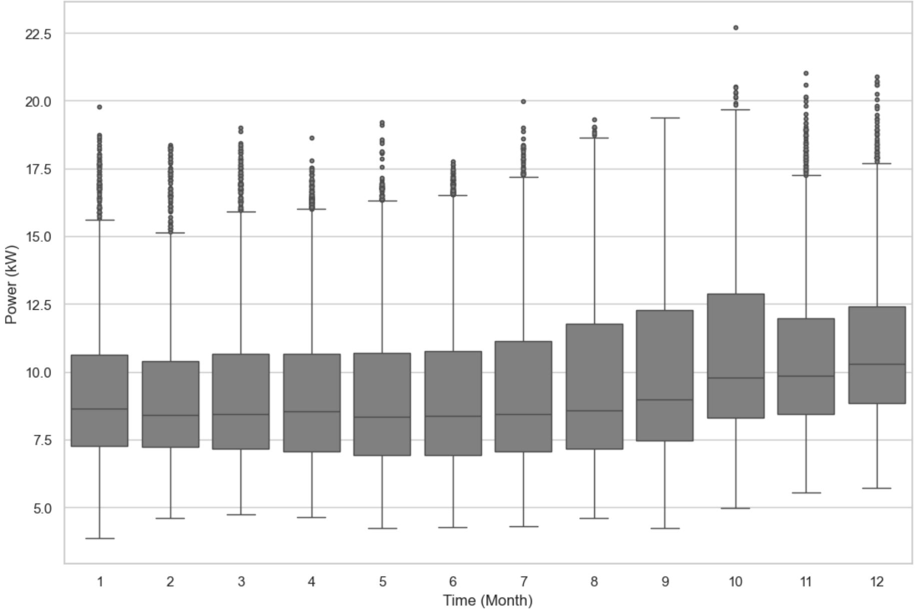
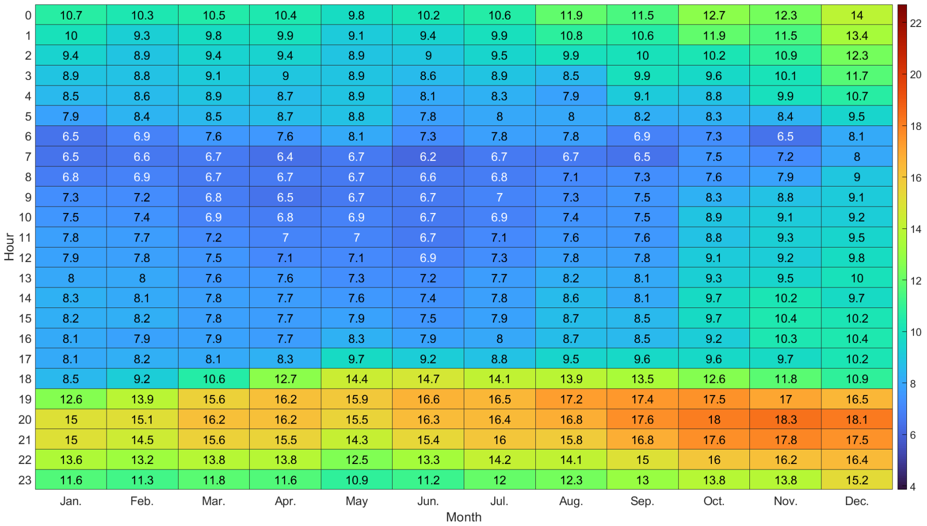


Fig. 2. Monthly Distribution of Electricity Demand Represented as a Boxplot with Outliers



Hour	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
0	10.7	10.3	10.5	10.4	9.8	10.2	10.6	11.9	11.5	12.7	12.3	14
1	10	9.3	9.8	9.9	9.1	9.4	9.9	10.8	10.6	11.9	11.5	13.4
2	9.4	8.9	9.4	9.4	8.9	9	9.5	9.9	10	10.2	10.9	12.3
3	8.9	8.8	9.1	9	8.9	8.6	8.9	8.5	9.9	9.6	10.1	11.7
4	8.5	8.6	8.9	8.7	8.9	8.1	8.3	7.9	9.1	8.8	9.9	10.7
5	7.9	8.4	8.5	8.7	8.8	7.8	8	8	8.2	8.3	8.4	9.5
6	6.5	6.9	7.6	7.6	8.1	7.3	7.8	7.8	6.9	7.3	6.5	8.1
7	6.5	6.6	6.7	6.4	6.7	6.2	6.7	6.7	6.5	7.5	7.2	8
8	6.8	6.9	6.7	6.7	6.7	6.6	6.8	7.1	7.3	7.6	7.9	9
9	7.3	7.2	6.8	6.5	6.7	6.7	7	7.3	7.5	8.3	8.8	9.1
10	7.5	7.4	6.9	6.8	6.9	6.7	6.9	7.4	7.5	8.9	9.1	9.2
11	7.8	7.7	7.2	7	7	6.7	7.1	7.6	7.6	8.8	9.3	9.5
12	7.9	7.8	7.5	7.1	7.1	6.9	7.3	7.8	7.8	9.1	9.2	9.8
13	8	8	7.6	7.6	7.3	7.2	7.7	8.2	8.1	9.3	9.5	10
14	8.3	8.1	7.8	7.7	7.6	7.4	7.8	8.6	8.1	9.7	10.2	9.7
15	8.2	8.2	7.8	7.7	7.9	7.5	7.9	8.7	8.5	9.7	10.4	10.2
16	8.1	7.9	7.9	7.7	8.3	7.9	8	8.7	8.5	9.2	10.3	10.4
17	8.1	8.2	8.1	8.3	9.7	9.2	8.8	9.5	9.6	9.6	9.7	10.2
18	8.5	9.2	10.6	12.7	14.4	14.7	14.1	13.9	13.5	12.6	11.8	10.9
19	12.6	13.9	15.6	16.2	15.9	16.6	16.5	17.2	17.4	17.5	17	16.5
20	15	15.1	16.2	16.2	15.5	16.3	16.4	16.8	17.6	18	18.3	18.1
21	15	14.5	15.6	15.5	14.3	15.4	16	15.8	16.8	17.6	17.8	17.5
22	13.6	13.2	13.8	13.8	12.5	13.3	14.2	14.1	15	16	16.2	16.4
23	11.6	11.3	11.8	11.6	10.9	11.2	12	12.3	13	13.8	13.8	15.2

Fig. 3. Heatmap of the mean energy demand on weekends, showing the variation across different hours of the day and months of the year.

Figure 4 presents the double standard deviation of Figure 3, which is significant lower than the single standard deviation plotted in Figures 1 and 2. The conclusion of this comparison is that the training of the neural network needs to be trained by including as meta data the hours, the days and if it is a weekend day or not. A decrease in the standard deviation of the metadata-sensitive demand profiles means less uncertainty due to unknown, influencing variables and minimising the error of the generated demand profiles. In principle, applying a two-input-based heat map might be an interesting approach to evaluate whether metadata affects the error and to assess the relative impact of the two inputs on each other.

The original measurements were recorded every 5 minutes. Since the objective is to achieve hourly estimations suitable for pre-design analysis, all 5-minute measurements within each hour were averaged. Additionally, two other components were integrated alongside hourly energy demand: a variable was included to denote *weekends (1)* versus *weekdays (0)* to account for potential differences in demand patterns arising from social or economic activities, and the *month* was included to capture potential seasonal variations in energy usage.

The inclusion of *temperature* and *humidity* variables was essential to complement the energy demand records due to the nature of the used appliances, such as refrigerators and other temperature-sensitive devices. These appliances exhibit energy demand patterns that are heavily influenced by ambient temperature and humidity. For instance, higher temperatures increase the cooling demand of refrigerators, while humidity levels can affect their efficiency and operation cycles. Incorporating these meteorological variables allows the model to account for external factors that significantly impact en-

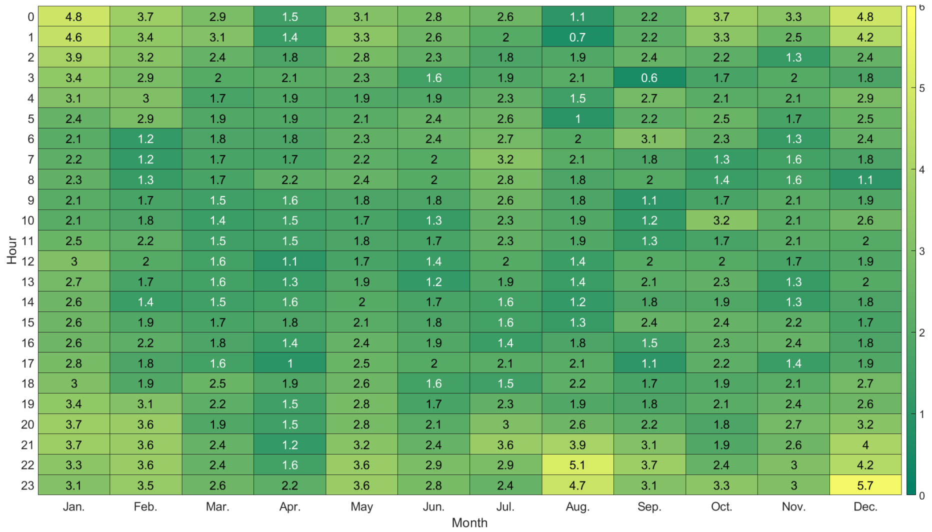


Fig. 4. Heatmap of double standard deviation energy demand on weekends, showing the variation across different hours of the day and months of the year.

ergy demand (Raza and Khosravi, 2015), thereby minimising the error and enhancing the reliability of the estimations. *Temperature* and *humidity* data were retrieved from (Weather Forecast API, 2023). These data were aligned by timestamps to ensure precise temporal synchronisation of all variables, producing a unified dataset for the El Espino case.

The resulting dataset provides the foundation for the modelling process described in the following section.

5 Architecture Definition

This work employs a neural network model designed to estimate hourly electricity demand using a set of metadata inputs: temperature (in degrees Celsius), humidity (in percent), hour of the day, month of the year, and a binary indicator for type of day (weekday or weekend). These variables were selected based on their broad availability in open-source datasets and their known influence on electricity demand patterns. The output node corresponds to the estimated hourly power in kW, meaning that for a given input (e.g., a specific hour and month along with the corresponding temperature and humidity), the model produces a single kilowatt value.

A key feature of the proposed approach is the use of a neural network architecture that transforms the input variables into a higher-dimensional space before making demand estimations. By expanding the input space, the model can capture complex dependencies and uncover latent causal structures that may not be evident in the original feature set. This higher-dimensional representation allows the network to move beyond simple correlations, enabling it to better model the intricate relationships that govern electricity demand.

The flexibility of neural network models in handling incomplete or irregular data has been highlighted in several studies (Owda et al., 2014; Hooshmand and Sharma, 2019). Unlike other methods that require complete daily data to generate estimations, this approach leverages any available measurement, allowing the construction of a dataset from incomplete or irregular records. It is possible to construct an entire demand curve for a selected time period by iterating this process across various hours.

The proposed architecture is represented in Figure 5. It incorporates a min-max normalization at the input stage, as shown in Equation 2, ensuring that all features are scaled to the $[0, 1]$ range. The minimum $(\min(x))$ and maximum $(\max(x))$ values for each feature are computed from the training dataset, ensuring consistency during inference. This normalization step is critical, as it prevents disproportionate influence from features with naturally larger magnitudes and generally improves model convergence and stability.

$$x_{\text{scaled}} = \frac{x - \min(x_{\text{train}})}{\max(x_{\text{train}}) - \min(x_{\text{train}})}. \tag{2}$$

The network consists of seven layers, structured to progressively increase the dimensionality of the feature space before refining the output. It begins with an input layer of 7 nodes, followed by hidden layers with 50, 250, and 750 nodes, capturing increasingly complex representations. The network then transitions through 300 and 150-node layers before reaching the single-node output layer. This progressive expansion in dimensionality enables the model to disentangle intricate dependencies, facilitating the capture of underlying causal relationships in the data.

Rectified Linear Unit (ReLU) activation functions are applied to each layer (unless otherwise specified) due to their effectiveness in modelling complex, non-linear relationships (Xu, 2015). Additionally, a batch normalization layer (indicated in orange) is incorporated after one of the hidden layers to stabilize and accelerate training. Finally, the red layer at the end represents a min-max descaling step, transforming the output from the normalised scale back to kilowatts, ensuring that estimations remain interpretable in domain-relevant units.

This hierarchical expansion and contraction of the feature space serves as a fundamental component of the model's ability to capture causal dependencies, as the higher-dimensional layers allow for richer representations before refining the output to a single predicted demand value.

The chosen architecture, normalisation strategies, and hyperparameters were guided by established best practices in neural network modelling and iterative empirical testing. Nonetheless, additional sensitivity analyses, alternative architectures (e.g., recurrent or attention-based networks), and more systematic hyperparameter optimisation could further enhance the model's estimation performance and transferability to different contexts.

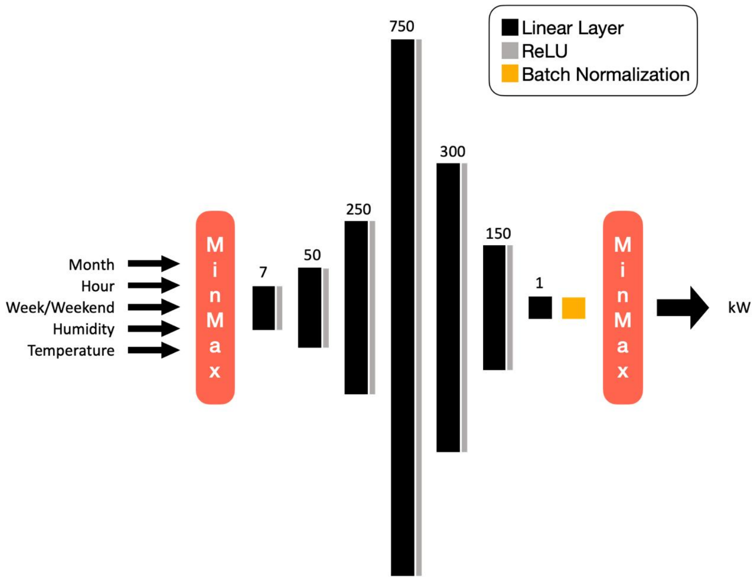


Fig. 5. Proposed neural network architecture, including input normalisation, hidden layers with ReLU activation, batch normalisation, and a final output descaling layer.

6 Model training

The dataset of 13,872 hourly measurements was randomly split into training (8,878 hours), validation (2,220 hours), and testing (2,774 hours) sets before training. This approximate 64%, 16%, 20% division is a standard practice aimed at ensuring robust model evaluation and preventing overfitting (Goodfellow et al., 2016). The training set was exclusively used to update model parameters, while the validation set was employed to monitor performance during training and prevent overfitting using early stopping techniques. The testing set was reserved for final evaluation, simulating real-world performance.

All weight parameters of the neural network were randomly initialised, a standard procedure in deep learning workflows. ReLU activation was uniformly applied to all nodes, as it avoids the vanishing gradient problem common in traditional sigmoid or tanh activations. MSE and MAE were used as the primary performance metric, given its widespread acceptance and straightforward interpretation (Hyndman and Koehler, 2006). Specifically, MAE quantifies the average magnitude of errors without considering their direction. In contrast, MSE assigns greater weight to larger errors due to its squared term, offering additional insight into the frequency and impact of significant discrepancies.

Four different optimisers were tested: Adaptive Moment Estimation (Adam), Adaptive Gradient (Adagrad), Adaptive Delta (Adadelta), and Stochastic Gradient Descent (SGD) (Tian et al., 2023), which are well-established and have demonstrated their applicability across a wide range of problems in neural network training. Each optimiser was applied with an initial learning rate of 0.01. Training was capped at a maximum of 20 epochs, and an early stopping mechanism was implemented to halt training once no further improvement on the validation set was observed, thus preventing overfitting.

7 Results

The primary objective of this study was to validate whether the selected metadata and input variables lead to an reliable approximation of electricity demand by capturing underlying causal relationships. By integrating key external factors—temperature, humidity, hour, month, and weekend indicator—the model aims to identify the causal influences that shape electricity demand patterns. Rather than relying on predefined demand curves or purely statistical correlations, this approach enhances interpretability and robustness in demand estimation.

To evaluate the model's performance, we analyse the impact of different optimisers on the error metrics. Table 1 presents the Mean Absolute Error (MAE) and Mean Squared Error (MSE) (Bhuyan et al., 2016) for each optimiser tested. The results indicate that the Adam optimiser achieves the lowest error values, with an MAE of 0.0536 and an MSE of 0.0511, making it the most effective optimiser for minimising errors.

To further validate the effectiveness of the proposed architecture, we compare its performance with a simpler model, the triangle-shaped architecture, as shown in Table 2. Across all tested optimisers, our approach consistently outperforms the simpler architecture in both MAE and MSE. Specifically, for the Adam optimiser, our architecture achieves an MAE of 0.0536 and an MSE of 0.0511, whereas the triangle-shaped architecture exhibits significantly higher errors, with an MAE of 0.1345 and an MSE of 0.1339. This results in a 60.2% reduction in MAE and a 61.8% reduction in MSE in our approach.

The performance gap is particularly evident across different optimisers. For Adagrad, our model's MAE is 0.0691, compared to 0.1818 in the triangle-shaped architecture, leading to a 62.0% improvement in MAE. Similarly, the MSE is reduced from 0.1744 in the simpler model to 0.0538 in our approach, which is a 69.2% reduction. Even for Adadelta, which produces the highest errors, our model achieves a lower MAE (0.1018 vs. 0.1364), resulting in a 25.3% improvement and MSE (0.1124 vs. 0.1962), which shows a 42.6% reduction, reinforcing the limitations of the simpler structure in capturing complex demand patterns. A similar trend is observed for the SGD optimiser.

These results confirm that our architecture enhances model precision and robustness, effectively capturing complex dependencies while minimising errors across different optimisation techniques. The substantial reduction in error compared to the triangle-shaped architecture supports the hypothesis that increasing model complexity, particularly through higher-dimensional encoding and carefully designed layers, leads to improved electricity demand forecasting.

Next, we conducted a sensitivity analysis using the Adam optimiser, as it performed the best with all the features, to determine the impact of excluding temporal features, such as the month of the year and weekend indicator, on the model's error. The results in Table 3 confirm that removing these variables leads to an increase in both MAE and

Optimiser	MAE	MSE
Adam	0.0536	0.0511
Adagrad	0.0691	0.0538
Adadelta	0.1018	0.1124
SGD	0.0904	0.1165

Table 1. Performance metrics of the proposed Neural Network Architecture.

Optimiser	MAE	MSE
Adam	0.1345	0.1339
Adagrad	0.1818	0.1744
Adadelta	0.1364	0.1962
SGD	0.1327	0.1403

Table 2. Performance metrics of the triangle-shaped architecture.

MSE, with a 10.7% increase in MAE and an 11.7% increase in MSE, further validating their relevance in electricity demand estimation.

8 Conclusions

Electricity demand estimation is crucial for effective microgrid design, particularly in remote areas where data availability is limited (Sanfilippo and et al., 2023). Unlike traditional methods that rely on predefined demand curves, the proposed approach introduces a causal model, implemented using a neural network, to uncover the underlying relationships driving electricity consumption. This study leveraged a dataset of 578 days from El Espino, Bolivia, integrating measured electricity demand with readily available metadata—such as temperature, humidity, hour, month, and whether the day is a weekday or weekend—to develop a novel estimation method. A key contribution of this work is the design of a neural network architecture specifically oriented toward capturing causality. Unlike conventional neural networks that primarily focus on pattern recognition, this architecture is structured to model causal relationships between input variables and electricity demand, ensuring that the learned representations reflect meaningful dependencies rather than surface-level correlations.

The study shows the superior performance of the causal-oriented neural network architecture over the triangle-shaped architecture in estimating hourly electricity demand. Our model consistently achieves lower errors, with a MAE of 0.0536 and an MSE of 0.0511, outperforming the triangle-shaped architecture by significant margins—up to 62% for MAE and 69% for MSE across various optimisers. The Adam optimiser, which outperforms Adagrad, Adadelta, and SGD in this context, enhances the model's ability to handle incomplete and noisy data. This approach not only offers a more reliable and

Feature Set	MAE	MSE
Full Feature Set	0.0536	0.0511
No Temporal Features	0.0593	0.0571

Table 3. Impact of Removing Temporal Features on Model Performance.

robust estimation but also provides greater interpretability, making it a valuable tool for complex demand pattern predictions. The results indicate that the neural network effectively captures causality between external factors and electricity demand. Additionally, the identified factors—such as temperature, humidity, time of day, month of the year, and weekend indicator-prove to be highly relevant in shaping electricity demand patterns. The sensitivity analysis, conducted using the Adam optimiser, further confirms that excluding these temporal-based features leads to an increase in error, demonstrating their importance in demand estimation. Specifically, removing the temporal variables results in a 10.7% increase in MAE and an 11.7% increase in MSE, underscoring their relevance. These results highlight that the selected variables not only contribute to minimising the error but also play a crucial role in capturing the causal relationships underlying electricity consumption, reinforcing the effectiveness of the proposed approach.

However, as datasets, methodologies, and computational tools continue to evolve, so too will the capacity to develop more scalable and widely applicable models. The integration of causality-driven neural networks into demand estimation holds significant potential for enhancing the reliability and efficiency of microgrid systems, particularly in underserved regions where precise energy planning is essential for sustainability and resilience.

While the chosen architecture—including layer sizes and the number of layers—was informed by preliminary experiments and computational constraints, future research could explore alternative neural network architectures to further improve the model's ability to capture causal relationships. Additionally, further work could investigate the impact of alternative input variables beyond those considered in this study. Incorporating additional environmental, socio-economic, or behavioural factors could enhance the model's ability to disentangle causal dependencies and improve demand estimation. Sensitivity analyses on a broader set of variables would help determine the most relevant causal factors for different microgrid contexts, ensuring the model remains adaptable and robust across diverse energy systems.

Although this approach was tested on a single region, its implications extend beyond El Espino. As data collection efforts improve—through expanded measurement campaigns, data-sharing platforms, or innovative sensing technologies—a similar methodology could be adapted for other remote areas, microgrids, renewable energy communities, and even individual buildings. Achieving broader applicability requires access to more extensive and diverse datasets, alongside ongoing methodological refinements. Future research should focus on validating the causal model across multiple sites and contexts to ensure robustness and transferability.

Acknowledgment

We acknowledge the European Institute for Energy Research (EIFER) for financing this internship, which contributed significantly to the development of this research.

References

Abugabbara, M., Javed, S., and Johansson, D. (2022). A simulation model for the design and analysis of district systems with simultaneous heating and cooling demands. *Energy*, 261:125245.

Balderrama Subieta, S. L. (2022). *Optimal design and deployment of isolated energy systems: The Bolivian pathway to 100 % rural electrification*. PhD thesis, ULiège - Université de Liège, Liège, Belgium.

Bhuyan, M. K., Mohapatra, D. P., and Sethi, S. (2016). Software reliability assessment using neural networks of computational intelligence based on software failure data. *Baltic J. Modern Computing*, 4(4):1016–1037.

Białek, J., Bujalski, W., Wojdan, K., Guzek, M., and Kurek, T. (2022). Dataset level explanation of heat demand forecasting ann with shap. *Energy*, 261:125075.

Castillo, V. Z., de Boer, H.-S., Muñoz, R. M., Gernaat, D. E., Benders, R., and van Vuuren, D. (2022). Future global electricity demand load curves. *Energy*, 258:124741.

Dieudonné, N. T., Armel, T. K. F., Vidal, A. K. C., and René, T. (2022). Prediction of electrical energy consumption in cameroon through econometric models. *Electric Power Systems Research*, 210:108102.

Dilaver, Z. and Hunt, L. C. (2011). Industrial electricity demand for turkey: A structural time series analysis. *Energy Economics*, 33(3):426–436.

Fioriti, D., Stevanato, N., Ducange, P., Marcelloni, F., Colombo, E., and Poli, D. (2023). Data platform guidelines and prototype for microgrids and energy access: Matching demand profiles and socio-economic data to foster project development. *IEEE Access*, 11:73218–73234.

Foldvik Eikeland, O., Bianchi, F. M., Apostoleris, H., Hansen, M., Chiou, Y.-C., and Chiesa, M. (2021). Predicting energy demand in semi-remote arctic locations. *Energies*, 14(4).

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gómez, M. and Rodríguez, J. C. (2019). Energy consumption and financial development in nafta countries, 1971–2015. *Applied Sciences*, 9(2).

Hooshmand, A. and Sharma, R. (2019). Energy predictive models with limited data using transfer learning. In *Proceedings of the tenth ACM international conference on future energy systems*, pages 12–16.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Kandananond, K. (2011). Forecasting electricity demand in thailand with an artificial neural network approach. *Energies*, 4(8):1246–1257.

Lazzari, F., Mor, G., Cipriano, J., Gabaldon, E., Grillone, B., Chemisana, D., and Solsona, F. (2022). User behaviour models to forecast electricity consumption of residential customers based on smart metering data. *Energy Reports*, 8:3680–3691.

Lombardi, F., Balderrama, S., Quoilin, S., and Colombo, E. (2019). Generating high-resolution multi-energy load profiles for remote areas with an open-source stochastic model. *Energy*, 177:433–444.

Michalik, G., Khan, M., Bonwick, W., and Mielczarski, W. (1997). Structural modelling of energy demand in the residential sector: 1. development of structural models. *Energy*, 22(10):937–947.

Nasr, G., Badr, E., and Dibeh, G. (2000). Econometric modeling of electricity consumption in post-war lebanon. *Energy Economics*, 22(6):627–640.

Nayanathara, C. and Srilatha, R. (2018). Electrifying villages using microgrids. In *2018 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, pages 300–305.

Neervannan, A. (2018). Evaluating the effectiveness of deep reinforcement learning algorithms in a walking environment. *Baltic J. Modern Computing*, 6(4):335–348.

Owda, H. M., Omoniwa, B., Shahid, A. R., and Ziauddin, S. (2014). Using artificial neural network techniques for prediction of electric energy consumption. *arXiv preprint arXiv:1412.2186*.

Raza, M. Q. and Khosravi, A. (2015). A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. *Renewable and Sustainable Energy Reviews*, 50:1352–1372.

Sanfilippo, S. and et al. (2023). Microgrid design optimization in benin within the leopard project: Evaluating the impact of inaccurate load profile estimation. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*.

Scarborough, D. and Somers, M. J. (2006). *Neural Networks in Organizational Research: Applying Pattern Recognition to the Analysis of Organizational Behavior*. American Psychological Association, Washington, DC.

Shin, S.-Y. and Woo, H.-G. (2022). Energy consumption forecasting in korea using machine learning algorithms. *Energies*, 15(13).

Shiraki, H., Nakamura, S., Ashina, S., and Honjo, K. (2016). Estimating the hourly electricity profile of japanese households – coupling of engineering and statistical methods. *Energy*, 114:478–491.

Somers, M. and Casal, J. (2009). Using artificial neural networks to model nonlinearity: The case of the job satisfaction–job performance relationship. *Organizational Research Methods - ORGAN RES METHODS*, 12:403–417.

Stadler, M., Cardoso, G., Mashayekh, S., Forget, T., DeForest, N., Agarwal, A., and Schönbein, A. (2016). Value streams in microgrids: A literature review. *Applied Energy*, 162:980–989.

Tian, S. and Chang, S. (2020). An agent-based model of household energy consumption. *Journal of Cleaner Production*, 242:118378.

Tian, Y., Zhang, Y., and Zhang, H. (2023). Recent advances in stochastic gradient descent in deep learning. *Mathematics*, 11(3).

Torabi Moghadam, S., Toniolo, J., Mutani, G., and Lombardi, P. (2018). A gis-statistical approach for assessing built environment energy use at urban scale. *Sustainable Cities and Society*, 37:70–84.

Velasquez, C. E., Zocatelli, M., Estanislau, F. B., and Castro, V. F. (2022). Analysis of time series models for brazilian electricity demand forecasting. *Energy*, 247:123483.

Wassie, Y. T. and Ahlgren, E. O. (2023). Determinants of electricity consumption from decentralized solar pv mini-grids in rural east africa: An econometric analysis. *Energy*, 274:127351.

Weather Forecast API (2023). Open-source weather api. https://open-meteo.com/.

Xia, Z., Ma, H., Saha, T. K., and Zhang, R. (2022). Consumption scenario-based probabilistic load forecasting of single household. *IEEE Transactions on Smart Grid*, 13(2):1075–1087.

Xu, B. (2015). Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*.

Zamanipour, B., Ghadaksaz, H., Keppo, I., and Saboohi, Y. (2023). Electricity supply and demand dynamics in iran considering climate change-induced stresses. *Energy*, 263:126118.

Received December 11, 2024 , revised February 21. 2025, accepted March 3, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 2, 331-356
https://doi.org/10.22364/bjmc.2025.13.2.02

Development and Validation of a Reference Architecture for the Smart Schoolhouse

Marge KUSMIN, Mart LAANPERE

Tallinn University, School of Digital Technologies, Narva mnt. 25, Tallinn, Estonia

margek@tlu.ee, martl@tlu.ee

ORCID 0000-0003-0243-0234, ORCID 0000-0002-9853-9965

Abstract. The last decades have shown a great many changes in the field of education, especially in the transition to a more learner-centred and technology-enhanced approach in the teaching and learning process. These changes have been brought about by both a need for it, but also by the opportunities that have risen from the rapid development of technology and the diversity of technological solutions. Technology can be used in the learning and teaching process to make learners' more engaged and allow them to take greater responsibility for shaping the their learning goals, process and environment, for instance, by accepting the BYOD model of technology use (Bring Your Own Device). The choice and ways of using the specific technologies should be defined by the purpose of the learning and teaching process itself and the requirements for the knowledge, skills, attitudes, and values of learners striving to become engaged members of both the knowledge society at large and also their chosen professional communities. Digital transformation of the European society creates demand for creative users of technology, i.e. is the need for innovative makers who can implement their knowledge in all kinds of different situations to create something new. This paper introduces a Smart Schoolhouse concept that has been designed to spark an interest in exploration among learners and to help them develop the proactive Maker mindset. As part of a four-year case study conducted in five Estonian schools, the possibilities of implementing the Smart Schoolhouse concept were analysed and evaluated. Suitable IoT technologies were tested, and the necessary support system was mapped. This process was supported by a comprehensive literature analysis, which provided an overview of the input collected during the case study for the creation and evaluation of the Smart Schoolhouse Reference Architecture. The article presents the development of RASS, which is based on the Industry 4.0 Reference Architecture Model (RAMI 4.0), created by the German Electrical and Electronic Manufacturers' Association (ZVEI). The conformity and completeness of RASS were validated using the RATE method to ensure that it meets stakeholders' expectations and requirements. The article concludes with an overview of the evaluated RASS and recommendations for its implementation.

Keywords: Smart Schoolhouse, reference architecture, reference architecture evaluation

1. Introduction

Global digitisation has brought about changed expectations for today's learners, i.e., members of society who will soon enter the workforce. In addition to their own chosen field they must be digitally competent and know how to solve interdisciplinary problems in cooperation with experts of other fields. Although for the last few decades schools have gone through big changes in their use of technology in the teaching and learning process, from simple content delivery on multimedia CDs to complex enterprise-level online learning systems (Moodle, EIS, eKool), the use of BYOD, cloud-based services, and IoT is pushing us towards another "*digital turn*" in schools: "*smart schoolhouse*".

The potential of IoT that links physical and virtual worlds and is used ever more in teaching and learning context has not yet been systematically researched, a coherent "*big picture*" is missing. Due to this the project "*Smart Schoolhouse by means of IoT*" was launched.

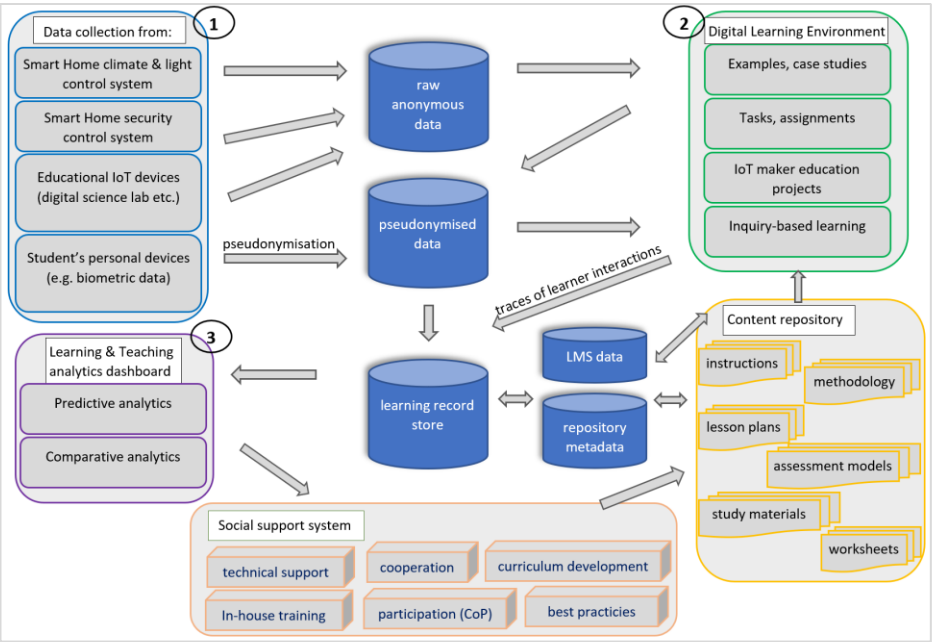


Figure 1. The data collection, handling, and use in a Smart Schoolhouse concept, and systems that support it.

This brought with it the opportunity to implement IoT devices and the data collected by them into the learning process, using various teaching methods and approaches (inquiry-based, problem-based learning, productive failure caused by ill-structured problems, learn-by-doing, Maker mindset and Maker movement, etc.). Also, the standardisation of their use in the everyday learning process. The goal was to provide a solution (shown in Figure 1) that 1) would make the background data collected by the Smart Home system, which is generally inaccessible to users, easily and conveniently available to learners for use in their studies; 2) would allow the use in combination of pseudonymised data from IoT devices and learners' digital footprints, along with data from the Smart Home system, to enhance STEM education (utilising this data in inquiry- and problem-based learning, research, etc.); 3) would allow the use of this data in a personalised form in learning analytics to gain a better understanding of the learning and teaching process, i.e. to identify the strengths and weaknesses of both the general and individual learning and teaching processes.

To generalise our research results, we decided to create a service oriented reference architecture that may be the best solution to describe the opportunities of Smart Schoolhouse to various stakeholders, e.g. in a context of preparing a large-scale software/hardware procurements and implementation in hundreds of schools simultaneously. Our paper outlines critical discussion points that aid in comprehending the probable technological, administrative and pedagogical solutions associated with Smart Schoolhouse. Furthermore, during the implementation phase, it promotes a more rapid, efficient, and seamless execution.

This research consists of two complementary parts: a theoretical part, involving a systematic literature review to establish a theoretical framework; and a practical part, in which this framework is applied to develop and evaluate the Reference Architecture for the Smart Schoolhouse (RASS). Given the complexity of the main goal, the study focused on answering the primary research question: How can a Reference Architecture for the Smart Schoolhouse be developed, validated, and implemented? This central question is further divided into five sub-questions (see Chapter 3.2). The results of the research questions are presented in both the theoretical see (Chapter 2), where initially the theoretical background is clarified, and the methodological (see Chapter 3), where it is demonstrated how theory is applied in practice, sections. In the final chapter, we summarise the topic and highlight limitations that future studies need to address.

2. Theoretical background

2.1. Related Research

To establish the context and foundation for our research and to address the research questions RQ1, RQ2, RQ3, and RQ6 (see section 3.2 below), we undertook a systematic literature review using a mapping technique (Kitchenham and Charters, 2007) and the guidelines for literature search, evaluation, and synthesis guidelines by vom Brocke et al. (2015).

Table 1. Articles selected for analysis in various fields, upon which this article and the research were based.

Key words	**Number of analysed papers**
Reference architecture	209
Service-oriented (reference) arhcitecture	84
Reference architecture in education	68
Reference architecture life cycle	27
Reference architecture evaluation	72

An iterative methodology was adopted, succinctly characterised as an iterative literature review approach geared towards an enhanced comprehension, wherein search, analysis, and synthesis are conducted concurrently and are interlinked. The search for relevant sources was progressively broadened in each iteration, both thematically and by authors, to delineate the lifecycle of the developed reference architecture—development-evaluation, usage, disposal—with a particular focus on its rigorous validation. It was also pivotal to ascertain the most recent information on architectures devised for the education sector to identify extant solutions and get confirmed the novelty of our design concept. While searching for relevant research literature, we used keywords "reference architecture", "reference architecture" + "education" (with alternatives replacing "educatio*": school, university, laboratory, classroom, teaching, learning, campus), + "lifecycle", + "evaluation" (with alternatives replacing "evaluat*": assess*, validate* AND method, framework), + "service oriented". Initially, a linear literature analysis process was utilised (sequential searching, analysing), reviewing articles from various repositories: Scopus, IEEE Xplore, ACM Digital Library and SpringerLink, with the search initially spanning a broader timeframe from 2010-2016. However, the approach soon became iterative, as the analysis of articles revealed the field's experts and most significant authors, leading to further iterations. Subsequent iterations primarily concentrated on IEEE Xplore to minimise duplicates and focused mostly recent four years.

The article selection process in these iterations was consistent: following the elimination of duplicates, we utilised analytical tools in MS Excel to identify key words from abstracts, which informed the selection of articles for in-depth review. The Table 1 below offers an overview of the research papers that were eventually shortlisted and closely examined, with the most significant ones being referenced in this article.

2.2. Reference Architecture

A reference architecture (RA) has been referred to as a blueprint or template for creating (software) systems, which, according to (Ünal, 2019; Abu-Matar and Mizouni, 2018), is claimed to offer a high-level structure and instructions for building applications in a specific context or domain, as by (Knodel and Naab, 2016) it helps to „*transform concerns within the problem space into decisions in the solution space*". Due to these (Hoel and Mason, 2018) states that it rather serves as a framework for designing a variety of systems. An explanation has also been used that RA is an accumulation of best practices (Ünal, 2019; Nakagawa et al., 2014), design patterns (Szwed et al., 2013), principles, and constraints (Cloutier et al., 2010; Weinreich and Buchgeher, 2014) over time within a specific application domain.

The RA „*provides, according to its objectives, discussion points for stakeholders*" (Ataei and Litchfield, 2022). It is used "*implicit knowledge and articulate it explicitly, facilitating the development of new products and product families*" (Cloutier et al., 2010). Reference architecture helps to "*understand the forms of likely solutions to certain domain problems*" (ISO/IEC/IEEE, 2019b), while providing "*a common (architectural) vision, lexicon and taxonomy*" (Cloutier et al., 2010) and guiding "*the development and deployment of applications of specific systems*" (Galster, 2015), "*implementation of new system architectures*" (Cloutier et al., 2010) or "*concrete architectures*" for specific instances of complex software systems (Angelov et al., 2009; Nakagawa et al., 2014). In general, majority of authors agree that a RA is used for development of concrete and standard architecture.

Concrete architecture, as implied by its name, is developed from an RA (Gidey et al., 2017) through the incorporation of specific software products and protocols (Angelov et al., 2009), and it is designed in a specific context and reflects specific objectives (Angelov et al., 2012). Considerably less abstract is the standard architecture, which is a specialisation of the RA within a specific organisation (Saay and Norta, 2016).

Angelov et al. (2012) conclude that architecture can only be an RA at higher levels of abstraction, reflecting the requirements of stakeholders and "*allowing its usage in differing contexts*": "*The RA is a generic architecture for a class of information systems, which is* *used as the basis for the development of concrete architectures*" (Angelov et al., 2009). While Cloutier et al. argue that the high level of abstraction in an RA makes understanding its role more complex, since several additional steps are required to create real software from it, the same article also explains the benefits it provides: RA (1) enables the reuse of good concepts and implementations in future projects, (2) may help control the complexity of an architecture, (3) provides a common understanding among stakeholders, and (4) helps mitigate risks. Therefore, the RA is deliberately maintained at an abstract level and designed with generality in mind (Kuppusamy & Suresh, 2020) to ensure its suitability for wide applicability (Guth et al., 2016).

2.3. Development of Reference Architectures

For the successful development of a reference architecture, it is imperative to comprehend the problem space and make design decisions within the solution space. According to the objectives, goals, and scope of reference architecture, it is primarily prescriptive (to recommend uniform solutions), descriptive (to create abstractions that simplify complexity), or predictive (to avoid reliance on trial and error). This, in turn, necessitates the selection of an appropriate style (e.g., Client-server, Component-based architecture, Data-driven architecture, Event-driven architecture, Layered architecture, Object-oriented architecture, Service-oriented architecture, etc.) and approach (such as Bottom-Up, Top-Down, Forward, Reverse, Zigzagging, etc.) suitable for its development (ISO/IEC/IEEE, 2019b).

Service-Oriented Architecture (SOA) is an architectural style that focuses on utilising services to meet software users' needs. By adhering to the service-oriented support architecture, organisations can design and implement software systems that are modular (Rabelo et al., 2015), flexible (Alsobhi et al., 2015), reusable (Alsobhi et al., 2015; Lopes et al., 2019; Kuppusamy and Suresh, 2020), and are capable of adapting to the changing needs. An essential aspect is the separation of these services from technology (Ataei and Litchfield, 2022).

Although there are several approaches to creating an RA, the most common is shown in Figure 2. These are top-down (Figure 2a) ("*comprehensive, technology-neutral coverage, often from the perspective of a particular Smart X application sector*") (Kearney and Asal, 2019) and bottom-up (Figure 2b) ("*user organisation creates first a standard architecture out of multiple concrete-architecture experience that matures into a RA*" (Norta et al., 2014).

The choice of which style or approach to use is largely determined by the objectives (starting points) of the RA being created, including the complexity of the system of interest, novelty, implementation mechanisms, and so on.

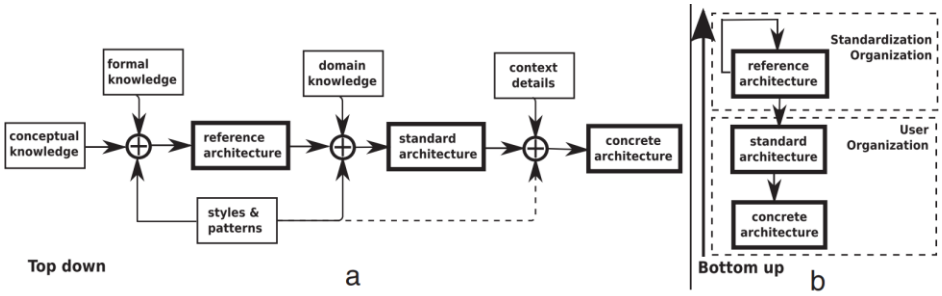


Figure 2. (a) Top-down architecture-framework process (b) Bottom-up practitioner process. (Norta et al., 2014)

Angelov et al. (2008) refers to the bottom-up approach as Practice-driven RA, as its creation is possible only when there is sufficient knowledge of the specific field to apply best practices. The top-down approach is described as Research-driven RA, also the "*greenfield*" approach (ISO/IEC/IEEE, 2019b), or Futuristic RA (Angelov et al., 2008; Abu-Matar and Mizouni, 2018) because these architectures are expected to become important sometime in the future (Angelov et al., 2008). Quite often, research centres are at the forefront of designing these preliminary architectures. These RA embody innovation and delineate the necessary components for systems implementing them. Angelov et al. argue that the origin of these RA lies predominantly in research-oriented environments that focus more on architectural innovation than on addressing the needs of domain stakeholders.

Therefore, Gidey et al. contend that the development of new RA requires attention to 1) architecturally significant requirements, and 2) the selection of appropriate architectural design decisions to implement these requirements (Gidey et al., 2017). To prolong the lifespan of the created RA, removing replaceable elements such as communication standards and protocols is necessary. Otherwise, the RA can quickly become obsolete. These architectures have to remain abstract, lacking specific technological implementations, standards, or protocols. The higher the level of abstraction at which the RA is presented, the longer its relevance endures (Angelov et al., 2009). However, an overly abstract RA can challenge stakeholders' understanding and may overwhelm them. Although it is often difficult to find the right level of abstraction (Cloutier et al., 2010), the RA must be abstract enough to allow for alternative decision-making while, at the same time, effectively ensuring the achievement of stakeholders' objectives (Galster, 2015).

The reference architecture life cycle comprises distinct phases or stages that an architecture undergoes, commencing with the recognition of the necessity for the architecture and concluding when it is deemed unnecessary or the architecture becomes obsolete (ISO/IEC/IEEE, 2019b). Multiple models, which can be characterised as structured frameworks comprising processes and activities arranged in sequential stages, have been put forward in (ISO and IEC, 2023) to enhance the management of the reference architecture lifecycle. Based on them, there are at least three main stages in the lifecycle of reference architecture: (1) development, (2) usage, and (3) discard. All of them may consist of multiply sub-stages. (ISO/IEC/IEEE, 2019b) presents sub-stages, that is grouped into three interacting processes (Conceptualisation, Elaboration, and Evaluation). These sub-stages are defining the problem, setting architecture goals, outlining its scope, and presenting potential solutions in a format suitable for stakeholders. Also, a mapping of quality indicators that enable the assessment of the value of the reference architecture, and finally, find a suitable evaluation method to assess the compliance of the reference architecture with the needs and concerns of stakeholders. Based on these steps, the final chapter of our paper proposes evidence-based recommendations for implementing the developed RASS to make it useful and meaningful to various stakeholders.

2.4. Evaluation of Reference Architectures

Architecture evaluation can be conducted at various stages of the system life cycle—from conceptual design to deployment and maintenance (ISO/IEC/IEEE, 2019a). However, the early evaluation of reference architectures (RAs) is particularly important to ensure alignment with stakeholder expectations and to mitigate risks related to quality, time, and budget (Knodel and Naab, 2016; Karlsson, 2016). While the primary aim of evaluation is to determine whether architectural objectives have been, or are likely to be, achieved, thereby supporting informed decision-making (Karlsson, 2016), the process also serves to validate architectural feasibility and minimise trial-and-error methods (Knodel and Naab, 2016), thus avoiding costly redesigns (Clements et al., 2010).

Although every software system is context-specific (Knodel and Naab, 2016), a wide range of evaluation methods has been developed. The literature review confirmed the dominance of scenario-based techniques, such as Architecture Level Modifiability Analysis (Garcés and Nakagawa, 2017; Batista et al., 2022; Fatima and Lago, 2023; Zbick, 2017; Ataei and Litchfield, 2020; Boyanov et al., 2020; Morkevicius et al., 2017), followed by experience-based approaches (e.g., focus groups) (Garcés and Nakagawa, 2017; Fatima and Lago, 2023; Zbick, 2017), prototyping (Ghantous and Gill, 2020; Palkar and Kamani, 2018), simulation (Garcés and Nakagawa, 2017; Baek et al., 2020; Li, et al., 2019; Fatima and Lago, 2023), model-based approaches (e.g., Architecture Description Languages) (Baek et al., 2020; Fatima and Lago, 2023; Nicolaescu and Lichter, 2016), and metric-based approaches (e.g., Software Productivity Metrics) (Fatima and Lago, 2023).

Angelov et al. (2008), applying the ATAM approach (ISO/IEC/IEEE, 2019a), argued that traditional evaluation techniques are frequently ill-suited to reference architectures due to their inherently high level of abstraction. This perspective is echoed by Ataei and Litchfield (2022), who point to the "*lack of dedicated evaluation methods for RAs.*" As a result, several frameworks have been developed as enhancements (Fatima and Lago, 2023; Boyanov et al., 2020; Knodel and Naab, 2016; Ehrlich et al., 2020), adaptations (Islam and Rokonuzzaman, 2009; de Oliveira Neves et al., 2018) or extensions (Fatima and Lago, 2023; Knodel and Naab, 2016) of existing methods. The choice of method depends on the type of architecture, the development stage, stakeholder interests, and specific evaluation objectives (ISO/IEC/IEEE, 2019a).

ATAM (Kazman et al., 2000), one of the most widely recognised evaluation methods, focuses on quality attributes such as modifiability, performance, and security. It promotes dialogue among stakeholders and supports informed architectural decision-making. This method was further developed from the Software Architecture Analysis Method (SAAM) (Kazman et al., 1994; Clements et al., 2010), which concentrated on modifiability (including portability, subset possibilities, and variability), and later incorporated analyses of performance, availability, and security (Kazman et al., 2000). Several notable developments and adaptations have emerged from ATAM.

Of particular relevance to this study is the Rapid ArchiTecture Evaluation (RATE) method, developed by Fraunhofer IESE. RATE is described as *“amalgamating best practices from existing methods and being adapted for pragmatic and rapid implementation in industrial contexts”* (Knodel and Naab, 2016). The method comprises five distinct checks: 1) DIC – verification of the integrity of stakeholder requirements, 2) SAC – assessment of the adequacy of the architectural solution, 3) DQC – scrutiny of the quality of architectural documentation, 4) ACC – conformance checks between implementation and architecture, and 5) CQC – general evaluation of code quality (Knodel and Naab, 2016).

The DIC check is critically important for aligning the concerns of various stakeholders and mapping these concerns onto evaluation criteria. Its purpose is to generate clearly structured problem descriptions based on stakeholder concerns, thereby ensuring that the architecture evaluation is both meaningful and effective. The primary aim of the SAC (Solution Adequacy Check) is to determine whether existing architectural solutions effectively address stakeholder concerns and whether there is sufficient confidence in their appropriateness. This assessment relies on a robust set of architectural drivers—typically formulated as scenarios—developed during the DIC process. Due to the abstract nature of architecture, evaluations rarely produce definitive outcomes; thus, it is essential to define the desired level of confidence and its implications early on. SAC supports early decision-making by validating architectural solutions before implementation resources are committed.

RATE was developed based on experiences where RAs lacked sufficient information to be evaluated under the ATAM framework. As a result, RATE incorporates several concessions compared to ATAM and demands fewer resources (Knodel and Naab, 2016). It is therefore well-suited for the evaluation of a RASS developed at the conceptual level.

2.5. Reference Architectures in the Field of Education

There exist not too many research papers that address the reference architectures created for the educational domain of our interest. The most common focus in such papers is the learning analytics or multimodal learning analytics enriched with IoT solutions (Drlik et al., 2018; Smith et al., 2018; Aleksieva-Petrova et al., 2020) or learning analytics in gamified eLearning (Maher et al., 2020). But there are also Assessment Analytics (Nouira et al., 2017), IoT curriculum (Abichandani et al., 2022), Robotics in Education (Kuppusamy and Joseph, 2020), Smart Education (Kuppusamy and Suresh, 2020), Smart Campus (Pandey et al., 2020), Context-aware Learning Environments (Lopes et al., 2019), and Tracking system for Online Laboratories (Zapata-Rivera and Petrie, 2018). These are only a few examples, but unfortunately none of them are suitable for implementation in our Smart Schoolhouse concept.

This underscores the necessity for a customised RA that addresses the unique requirements and objectives of the Smart Schoolhouse. In developing such an architecture, it is essential to consider various technological and pedagogical aspects that support learning and teaching in an innovative and effective manner. For instance, the architecture could integrate elements of IoT, data analytics, and gamification to create a dynamic and engaging learning environment that can respond to individual learners' needs and preferences. Furthermore, this architecture should promote flexibility and adaptability, enabling easy adjustments to align with developments in educational institutions and technology.

In the process of creating the RA, it is vital to involve a diverse array of stakeholders, including teachers, students, educational technologists, and administrators, to ensure that the final product meets the needs and expectations of all parties. Through further research and collaboration, a RA can be developed and implemented that not only meets current demands but is also sufficiently flexible to adapt to future educational and technological innovations.

3. Research Methodology

3.1. Research Aim

This research consists of two complementary parts: a theoretical part aimed at systematically mapping and analysing existing solutions for creating and evaluating reference architectures (RA) through an iterative literature review; and a practical part focused on applying these theoretical insights to develop and evaluate a reference architecture specifically tailored to the Smart Schoolhouse concept.

Over a period of four years, the practical component involved conducting a comprehensive case study across 19 Estonian schools. This included mapping (Kusmin et al., 2018), testing (Kusmin, 2019a, 2019b; Kusmin et al., 2019), and systematising (Kusmin and Laanpere, 2023) suitable IoT solutions for educational purposes, as well as identifying support system requirements (Kusmin and Laanpere, 2020). The findings of this case study contributed to the development and evaluation of a self-assessment model for the Smart Schoolhouse (SAMSS) (Kusmin and Laanpere, 2022; 2024).

Utilising these outcomes, the practical part of the present study aimed to create a robust RA to support the integration of physical and virtual learning environments, thereby facilitating the effective use of IoT-generated data and learning analytics within learner-centred, creative, and collaborative STEM education.

3.2. Research Questions

In this study, we sought to answer the question:

How to develop, validate, and implement RA for the Smart Schoolhouse?

To address the research question, two sub-studies were conducted: 1) a mapping of the literature (N=209), followed by the creation of the RASS, and 2) its evaluation. For the mapping and analysis of the literature, we established the following sub-questions:

RQ1: Which processes and phases constitute the life cycle of a RASS and how are they managed?

RQ2: Which existing reference architectures and their validation methods would be suitable or adaptable for our concept of Smart Schoolhouse?

RQ3: What methods are most commonly employed in the development of an RA?

During the evaluation of the Smart Schoolhouse reference architecture, we sought answers to the following sub-questions:

RQ4: To what extent does the RASS meet the expectations, requirements, and needs of the Smart Schoolhouse concept?

RQ5: Does the RASS meet the general criteria for reference architectures, i.e. is it adequately abstract and all-encompassing while still remaining understandable and executable?

We have already addressed the sub-questions RQ1, RQ2, and RQ3 in our literature review above. In the next chapters, we will describe the application of its results in design and validation of the RASS. To find answers to sub-questions RQ4 and RQ5, we conducted an evaluation of RASS based on the first two checks of the RATE method. At the end of this chapter, based on our findings, we provide guidelines for the implementation of RASS.

3.3. Research Design

We used a Design Science Research (DSR) to develop a RASS. DSR is defined as „*a problem-solving paradigm that seeks to enhance human knowledge via the creation of innovative artifacts*“ (vom Brocke et al., 2020). Hevner et al. clarify that an artefact can take the form of a construct, model, method, or instantiation (Hevner et al., 2004). According to Hevner, DSR is a fundamentally pragmatic approach, prioritising relevance and meaningful contributions to the application environment, but he adds that it is crucial to establish a harmonious balance between relevance and rigour in research studies (Hevner, 2007).

Our research design consists of the following steps: (a) conducting a literature mapping, (b) analysing existing architectures, (c) developing a reference architecture proposal based on requirements that were collected, analysed, grouped, and evaluated within the case study (SAMSS), and (d) evaluating the proposed architecture.

3.4. Development of the RASS

The analysis of the literature (in section 2) revealed that, although among the 137 relevant scientific articles examined in depth, including 68 articles that focused on the educational domain, none of these are suitable as the basis for the RASS. Many of them were too abstract, but the main issue was substantive – they were created for entirely different functionalities, such as developing e-learning environments, assessment analytics, curriculum development, etc. Therefore, we tried to find the most optimal solution, taking into account the experiences of others, to create the RASS as efficiently as possible.

Creating, evaluating, and maintaining a RA must be empirically justified to ensure their relevance and practical applicability. Building on Galster's interpretation of Karow et al., it is necessary to ensure: a) empirical foundation - the RA must be based on 1) real-life situations reflecting stakeholders' interests, 2) proven principles validated in practice, and 3) aspects reflected therein must be derived from the problem domain; b) empirical validity - evaluating the RA demonstrates its applicability and validity (Galster and Avgeriou, 2011).

The development of a SOA for the Smart Schoolhouse followed a six-stage framework of "*Empirically Grounded RA*" (EGRA) (Galster and Avgeriou, 2011), utilising a top-down research-driven approach (Angelov et al., 2008). Stakeholders’ concerns central to the Smart Schoolhouse concept were mapped out during a four-year project, "*Smart Schoolhouse by means of IoT*“, based on patterns of IoT tool selection and usage emerging from the learning process. These patterns were integrated into the Smart Schoolhouse Assessment Model (SAMSS). Subsequently, six personas and six scenarios were developed, utilised for both creating and evaluating the RA. In terms of structural design, we relied on the RAMI 4.0 model (Hankel and Rexroth, 2015).

The personas were as follows: 1) a 28-year-old engineer working as a smart home systems implementer in a technology company and conducting extracurricular activities at school; 2) two 15-year-old 9th-grade students, 3) a 47-year-old physics teacher with extensive professional experience but limited practical knowledge in ICT and IoT, 4) a 26-year-old young art and design teacher with no prior teaching experience but who has participated in two pedagogical internships, 5) a 34-year-old experienced ICT education specialist with experience in ICT who has been working as an educational technologist in schools for over ten years. As the concept of the Smart Schoolhouse is still under development, based on our previous experience in the project, we were able to create scenarios for the life cycle of the IoT devices used in the Smart Schoolhouse and their usage at three hierarchical levels (Disconnected, Online, Connected) (Kusmin and Laanpere, 2023) of IoT technology. Two higher levels of the hierarchy of IoT devices usage (Smart, Integrated) could not be mapped in the project, so they are theoretical and based on the SAMSS (Kusmin and Laanpere, 2022) validated by experts. Therefore, it is crucial to pay greater attention to them when evaluating the RA.

3.5. Evaluation of the RASS

To select a suitable evaluation method, a comprehensive literature analysis was conducted, examining scientific articles. Some of these articles provided brief overviews of RA evaluation, primarily focusing on RA development, while others offered an in-depth examination of evaluation processes.

In total, 72 scientific articles were analysed with the aim of identifying an evaluation method recommended by experts and empirically validated in practice. The objective was to ensure the empirical validity of the RASS evaluation, thereby guaranteeing that the selected method and measurement technique are of high quality and specifically designed to measure the required indicator. RATE consists of five critical checks (Knodel and Naab, 2016); however, only Driver Integrity Check (DIC) and Solution Adequacy Check (SAC) are relevant within the scope of this study. DIC identifies and explains ambiguous architectural drivers through specific scenarios, while SAC assesses the suitability of architectural solutions for these drivers, including confidence in their effectiveness. Thus, we employed two components of the RATE approach: DIC and SAC.

The evaluation of the RASS involved five experts with somewhat varying type of expertise in the fields of education, IoT, and engineering, including 4 males and 1 female as summarised in Table 2 below.

Table 2. Experts involved in the evaluation of the RASS, along with their age, gender, and experience in various fields.

age range	gen-der	experience as a software developer (in years)	experience with IoT devices	teaching experience	teaching areas or subjects
40-49	M	-	developed	more than 10	robotics, microcontroller programming, home automation, operating systems
40-49	M	-	configured	1-6	the use of produced IoT devices (in grades 5-7), assembling IoT devices and solutions on your own (in grade 7), guiding IoT devices UPT (usage, programming, and troubleshooting) (in grade 11)
30-39	M	7-10	configured	-	-
30-39	W	7-10	configured	1-6	software engineering
50-59	M	more than 10	-	more than 10	software development methodology, programming, etc.

Three of them had experience as software developers, with one having 7-10 years of experience and two having over 10 years. One expert had no exposure to IoT devices, while another had been involved in their development. In terms of teaching experience, one had none, while two had over 10 years of pedagogical experience, having taught subjects such as operating systems, software engineering, software development methodology, programming, robotics, microcontroller programming, home automation, the use and configuration of IoT devices, the creation of devices and solutions, and troubleshooting IoT devices.

The RATE approach to evaluation integrates several practices from both the software industry and academic research, focusing on five key checks: (1) evaluating the robustness of the architectural drivers, (2) assessing the adequacy of the architectural solution, (3) examining the quality of the architectural documentation, (4) verifying the alignment between the implementation and the architectural design, and (5) appraising the overall quality of the code (Knodel and Naab, 2016).

When evaluating a reference architecture using the RATE model, architectural drivers (typically formulated as scenarios) are employed. These are developed through the DIC process. Therefore, the DIC check plays a central role in transforming stakeholder concerns into clearly structured evaluation criteria.

In the first phase of the evaluation, we applied the DIC process according to the guidelines provided by Knodel and Naab (2016), creating six scenarios that reflected stakeholder concerns. These concerns had previously been collected, structured, and presented as an integrated whole within the Smart Schoolhouse Self-Assessment Model (Kusmin and Laanpere, 2022), which had been validated by experts using the Nominal Group Technique (Kusmin and Laanpere, 2024).

Through the DIC process, we identified the most critical aspects of the Smart Schoolhouse concept and, based on these, developed six scenarios. The first four scenarios (Life cycle of IoT device adoption, Disconnected, Online, Connected, Disconnected, Online, Connected) are drawn directly from real-world practice. In contrast, the final two scenarios (Smart, Integrated), while still conceptual due to their innovative nature, are grounded in data validated by both stakeholders and experts. Before proceeding to scenario analysis and their application in the second RATE check, the SAC, two domain-specific teachers contributed to evaluating and refining the clarity of the developed scenarios.

The evaluation's second stage, the SAC, took place in a Zoom session with experts. Incorporating external experts into the assessment of architecture adequacy using SAC techniques, as recommended by (Knodel and Naab, 2016), enables the attainment of solution reliability through comparison of detailed scenarios developed in the DIC phase with the RA. The online meeting adhered to the planned two-hour duration. An expert of considerable experience moderated the session, accompanied by an observer who undertook various roles: documenting the proceedings, sharing files (scenarios), presenting and sharing explanatory content in zoom's screen, gaining consent for video recording, managing the video recording process, and later transcribing the video to receive feedback for accuracy feedback from the experts on its accuracy.

The primary objective of the SAC was to assess the suitability of proposed architectural solutions relative to the identified architectural drivers and to ascertain the level of confidence in their appropriateness. The input for the SAC session consisted of six scenarios developed during the previous DIC session along with the three-dimensional and layered RASS.

The SAC encompassed five key steps: 1) an introductory session with Q&A, 2) the RASS evaluation, 3) a discussion, 4) a summary of pivotal observations and suggested amendments, ending with a vote, and 5) closing remarks, allowing experts to voice their final thoughts on the necessity and implementation of RASS.

Introduction aimed to ensure a shared understanding among all experts, facilitating effective collaboration. It covered the Smart Schoolhouse concept, its operational principles, data flow, security issues related to data collection and use with a focus on GDPR compliance, the three-dimensional RA (Figure 3), and its layered structure (Figure 4). With their questions answered, experts proceeded to validate the RASS against the scenarios.

The evaluation started with the scenarios describing the use of IoT devices with the lowest compatibility level (Disconnected), gradually moving towards better connectivity (Online, Connected, Smart, Integrated). Finally, the scenario describing the life cycle of IoT device deployment was evaluated. The evaluation process was conducted similarly across all scenarios. Initially, experts were given time to familiarise themselves individually with the scenario, shared via Google Drive. Subsequently, the moderator then led a discussion, querying the clarity of the scenario, the need for replenishment, and the compliance of the smart schoolhouse's three-dimensional and layered architecture with the described scenario. After collating expert opinions, a summary of key points and amendment recommendations for both the scenario and the RASS was compiled.

After the RASS assessment, which involved six similar evaluations based on scenarios, a feedback session was conducted. During this session, the proposed improvements and modifications for each scenario were discussed and prioritised according to their significance in order to identify all critical changes or potential enhancements. This information was subsequently used to refine the RASS. Subsequently, experts were asked to comment on the abstractness, comprehensiveness, understandability, and feasibility of the RASS. The aim was once again to determine whether the RASS corresponds to the described scenarios and is feasible in its proposed form.

3.6. Empirical Validity of the RASS Evaluation

The final stage involved the empirical validity assessment of the RASS evaluation. The discussion focused on three aspects of empirical validity: (1) construct validity, which examines whether the evaluation accurately measured what it was intended to measure; (2) external validity, which analyses the extent to which the results can be generalised; and (3) internal validity, which assesses the replicability of the experiment.

1) Construct validity assesses the extent to which the evaluation environment reflects its purpose with regard to dependent and independent variables (Galster et al., 2017). Within this framework, we highlight the following aspects that were confirmed during the discussion:

The utilisation of the RATE evaluation method: The RATE method, an advancement of ATAM, is designed to achieve objectives with optimised resources. The discussion confirmed that this evaluation method contributed effectively to fulfilling the assessment's objective.

Created and analysed scenarios: The scenarios employed in the evaluation were developed based on key aspects reflected in the SAMSS, which had previously been validated by experts using the Nominal Group Technique (NGT). The content and phrasing of the scenarios were coordinated with a broader target group prior to the evaluation. Therefore, it can be asserted that the scenarios employed effectively facilitated the evaluation of the specific criteria they were designed to assess. This was also corroborated by experts.

Preparation and management of the process: The evaluation was conducted online, adhering to the principles of the second control (SAC) of the RATE method. Following the pandemic, online meetings—particularly for IT experts—have become customary. Both the online meeting and file-sharing environments functioned flawlessly, providing all experts with the opportunity to offer both oral and written comments. The discussion confirmed that the evaluation process effectively supported the achievement of the objective.

Duration of evaluation: The pace of the RASS evaluation was measured and deliberate. The moderator guided the progression of the discussion in response to the level of expert engagement, introducing new questions or arguments as the feedback began to diminish. It was confirmed during the discussion that the time allocated and spent on the various stages of the evaluation was sufficient for all experts to explore the subject in depth and contribute as objectively as possible.

Interpreting visualised information: To avoid any issues, a thorough introduction to the topic was provided before the evaluation. Specifically, we explained the concept of the Smart Schoolhouse, including the principles of data collection, management, and use; the grouping of IoT devices identified based on usage patterns; their life cycle; and both the three-dimensional and layered RASS. Although the interpretation of visualised information largely depends on an individual's background and experience, experts confirmed that they were provided with a sufficient overview of the context and received answers to their questions before the evaluation.

Expertise of evaluators: It is reasonable to assume that experts with more extensive experience in the analysis of RAs might have offered somewhat different responses, and their involvement could be considered in future evaluations. However, due to resource constraints, a purposive sample was employed, consisting of experts who are recognised and highly experienced in fields relevant to this study. Regarding whether the experts involved in the evaluation met the expectations placed upon them, the thoroughness of their input, along with the quality of their recommendations and proposals, confirms that they fulfilled the expected criteria.

2) Regarding external validity (the generalisability of results: whether the findings are of interest to others), we are confident that our results are generalisable. The mapping of stakeholders' requirements and concerns (i.e., the critical aspects of the Smart Schoolhouse concept) was conducted through focus group interviews and action research in 19 schools, which varied in size, location, and language of communication. This process was followed by a thorough analysis of the literature to corroborate the conclusions with scientific research. Although the experts stated during the evaluation that, following the implementation of their improvements and recommendations, the RASS is suitable for the next phase of the RA life cycle, namely, implementation, it is nonetheless advisable to conduct a new evaluation that also takes into account the specific requirements and concerns of the target group (whether at the school, municipal, or national level). This indicates that further evaluation will be necessary.

3) To ensure internal validity, we employed a reliable evaluation method for assessing the RASS and are confident in the results (particularly regarding how evaluation outcomes may depend on the experts' experience) because we adhered to the guidelines of the chosen method (RATE). The key assessment components of RATE, namely the first (DIC) and second (SAC) checks, are expert-driven activities which, due to the significant human factor involved, can be classified as qualitative in nature. As such, they are not considered the most reliable in the sense that identical results may not be replicated in a different context. Nevertheless, major discrepancies are unlikely to arise, as the participating experts were highly experienced and impartial. Furthermore, the RA evaluation was conducted using predefined metrics, specifically, scenarios developed in line with stakeholder requirements, which ensured that the experts assessed the RA from consistent and comparable perspectives.

The insights gathered from this SAC session with experts will be discussed in the next section.

3.7. Results of the RASS Evaluation

During the two-hour SAC session, a RASS evaluation took place. Immediately after the introductory part, numerous questions were posed to gain a better understanding of the developed RASS, and its three-dimensional and layered nature. Subsequently, during the presentation of the scenarios and the analysis of the resulting RASS, two major proposals were made, and seven recommendations were provided. In addition to these, the questions raised during the discussion about the Smart Schoolhouse concept provided food for thought and need consideration in the future implementation of the concept. The downside of implementing the SAC is that it is largely a manual task, requiring considerable effort, and primarily yields qualitative results (Knodel and Naab, 2016).

Suggestions and Discussion:

a) One expert suggested placing the presentation and business layers side by side rather than overlapping since they utilise the same data and services, but other experts did not consider it essential. Therefore, to gain a visually clearer overview, we postponed this suggestion for the time being.

b) Considering the Smart Schoolhouse concept, which involves data from the Smart House system, sensors inside and outside classrooms, learners' digital footprints, and, with learners' requests and parental permission, also from their personal wearable devices, it results in a large volume of diverse data that can be utilised in the learning process and learning analytics. In the initial RASS, the collected data was divided into four categories 1) raw, 2) processed, 3) pseudonymised, 4) Learning Record Store (LRS) data. During evaluation, a recommendation was proposed to include three additional groups in the data layer: 1) essential or objective data (such as fire or security data that should not be manipulated by students), 2) manipulative or experimental data (gathered/generated by students), and 3) simulation-based data (exclusively for modelling or educational purposes).

c) The third, and perhaps the most significant, supplement proposal was to add an integration layer or create integration capabilities for the data layer. It is crucial that devices which will need to be added or introduced to the created software system in the future can understand each other. These integration activities may be added to the documentation of the RASS, intended for the replacement and upgrade of existing IoT devices, or for acquiring entirely new functionality-providing IoT solutions. The idea was to add a recommendation that, in case the data layer does not support the standard of the offered IoT devices or solutions, manufacturers or providers should ensure integration capabilities by supplementing the IoT devices with adapter software.

d) Additionally, other smaller-scale suggestions and recommendations were made, such as the information that IoT devices suitable for home solutions may not always yield the best results when used in schools, but these are not reflected in the RASS.

In summary, it can be highlighted that the experts reached a consensus on two important additions: 1) additional grouping of data used in the learning process and 2) adding integration capabilities.

The development of the RASS was informed by the six-stage framework "Empirically Grounded RAs," (EGRA) (Galster and Avgeriou, 2011) and the design of RASS adhered to the principles of SOA. The stakeholders’ concerns, pivotal to the Smart Schoolhouse concept, provided essential input. In terms of structural design, we relied on the RAMI 4.0 model (Hankel and Rexroth, 2015). The architecture (Figure 3) manifests in three dimensions: 1) Life cycle of IoT devices or solutions adopted in the Smart Schoolhouse; 2) Hierarchy levels of IoT device usage, categorised as Disconnected, Online, Connected, Smart, and Integrated; and 3) an architectural framework comprising eight layers (shown in Figure 4): Devices, Data, Integration, Application, Business, Presentation, Support, Governance, and Security.

For evaluation purposes, we employed the first two checks of the RATE method, conducted with the expertise of five specialists. Regarding the validity of conclusions and recommendations, which are further explained in the discussion section, we must rely solely on trust in the expertise and contributions of experts, since, as Angelov et al. claim, "the progress and shortcomings of the RA can only be measured in a temporal perspective" (Angelov et al., 2012).

We do not assert that the developed RASS represents the best solution; however, it signifies an initial step within the context of the Smart Schoolhouse.

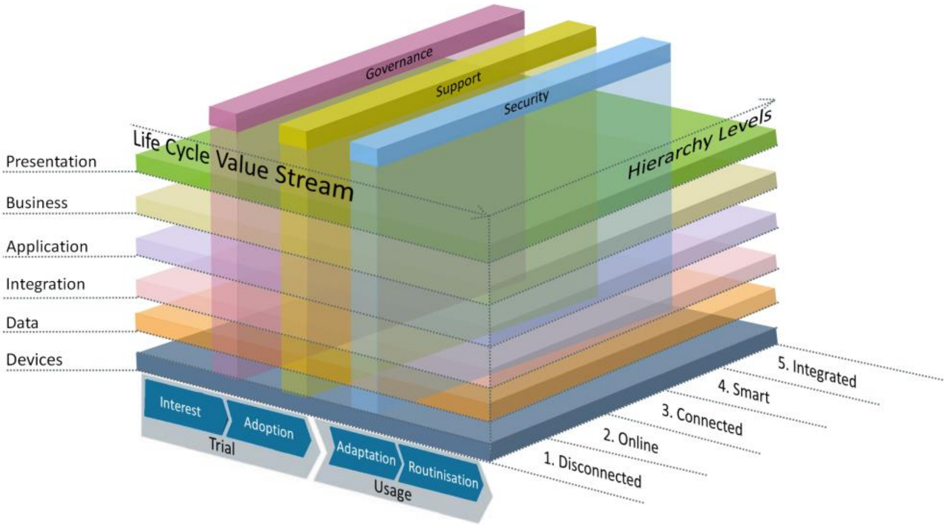


Figure 3. Three-Dimensional Reference Architecture of Smart Schoolhouse

3.8. Evaluated Reference Architecture of the Smart Schoolhouse

The Smart Schoolhouse reference architecture (RASS) is presented as a three-dimensional model in Figure 3, illustrating three distinct dimensions: 1) The lifecycle of IoT technology within the Smart Schoolhouse; 2) The hierarchical levels of IoT device usage, categorised based on usage patterns; 3) The layered architecture of RASS.

To enhance clarity, the service-oriented layered architecture is presented separately in Figure 4.

The presentation layer is responsible for providing essential technical and technological capabilities to modern learners. This includes applications, portals, and internal system user rights management. It also refers to the integration of diverse technologies into the learning process.

The primary function of the business layer is to manage various functionalities, such as user management, activities, tasks, communication, learning analytics etc., to establish a comprehensive overview of the learning experience and related activities. Although RASS is pedagogically neutral, the business layer outlines learning strategies and methods that must be considered at the next stage of RASS implementation.

The purpose of the application layer is to orchestrate data services, facilitating the collection, integration, and processing of data from various sources.

The integration layer was introduced during the evaluation process to enable the integration of IoT devices from different providers within the Smart Schoolhouse system, accommodating various standards and protocols.

The data layer is responsible for data management. Initially, it included raw data collected from various sources, processed data, and pseudonymised data to enable their use in the learning process. During the evaluation, three data categories were added: 1) Objective Data (e.g., safety or fire protection data), which should not be accessible for manipulation by students; 2) Manipulable Data, which students collect, create, and process; 3) Simulation-Based Data, intended solely for modelling or educational purposes.

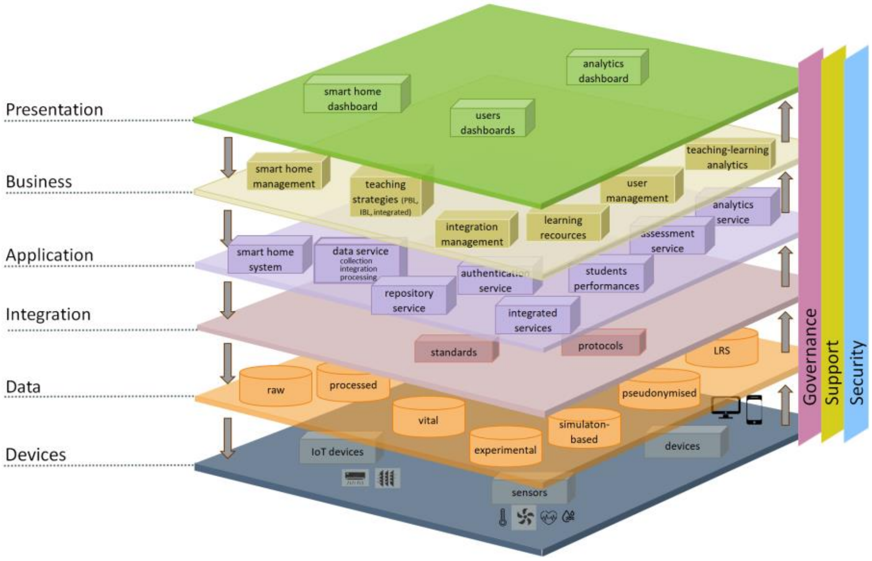


Figure 4. Layered Reference Architecture of Smart Schoolhouse

The device layer must facilitate the rapid and convenient addition and utilisation of IoT solutions, taking into account the hierarchical levels of IoT technology. To ensure technological neutrality, no specific standards are defined within RASS.

The support layer reflects the services required for education, including training programmes, guidelines, and the dissemination of knowledge.

The management layer ensures the identification, management, and dissemination of various standards, legal requirements, and regulatory obligations throughout the organisation.

The security layer must prioritise the accessibility and secure transmission of learners' personal data and data used within the learning process. Additionally, it is essential to ensure the system's reliability, confidentiality, integrity, and privacy.

3.9. Recommendations for the Implementation of the RASS

RASS can be applied in various future scenarios, such as: 1) formulating requirements for the development of a nationally commissioned learning environment that facilitates the rapid and convenient implementation of IoT technology; 2) integrating an IoT-enabled solution with an educational platform; 3) providing recommendations for the procurement of sustainable IoT technology, etc.

To effectively implement the concept of a Smart Schoolhouse within a specific purpose, the following recommendations are proposed to map the concerns of stakeholders. Based on the RASS:

R1: Identify the Components: Ascertain the key components and systems that constitute the smart schoolhouse architecture, tailored to the specific context. This may include sensors, devices, infrastructure, networks, databases, and learning analytics tools.

R2: Map the Data Flow: Visualise the flow of data within the architecture, commencing with the physical learning environment (data is collected via IoT sensors and smart devices from various sources) and concluding in sophisticated e-learning environments equipped with monitoring systems. It is essential to demonstrate how this data is transmitted and processed within the school infrastructure.

R3: Incorporate Digital Footprints: Integrate the digital footprints of learners, which may include data from personal devices (e.g. smartphones and tablets) and online platforms (including learning management systems, educational applications, and social media). Identify the connection points where this data is collected and synchronised with the smart schoolhouse architecture.

R4: Consider STEM Education Focus: Highlight any specific components or functionalities within the architecture that are pertinent to STEM education, such as integration with STEM-specific tools, virtual laboratories, or interactive learning resources.

R5: Design Data Integration: Analyse how data from both the physical learning environment and learners' digital footprints are integrated. This integration may involve processes of data harmonisation, aggregation, and transformation to ensure compatibility and consistency for learning analytics purposes.

R6: Include Learning Analytics: Demonstrate the components or tools responsible for conducting learning analytics. This may involve the use of algorithms, machine learning models, or dedicated analytics platforms that process the integrated data to generate insights and metrics related to learners’ performance, behaviour, or engagement.

R7: Address Privacy Concerns: Emphasise the measures implemented to protect the privacy of students and teachers while utilising the data. This may include techniques such as data anonymisation and pseudonymisation, encryption, access controls, and adherence to relevant privacy regulations and policies.

R8: Provide a Visual Legend and Explanations: Construct a legend or key that elucidates the symbols, labels, and connections used in the RA. Include explanatory notes or descriptions to clarify the purpose and functionality of each component.

R9: Engage with Relevant Stakeholders: Consult with relevant stakeholders, including educators, IT specialists, and privacy experts, throughout the process to ensure that the RA meets their concerns and adheres to best practices in data collection, integration, and privacy within the context of STEM education and Learning Analytics.

In light of the collected data and based on the RASS a concrete architecture (Gidey et al., 2017) can be developed by adhering to the steps delineated in the chosen framework (Cloutier, et al., 2010) for its construction. Concrete architecture is fashioned within a particular context, reflecting specific objectives (Angelov et al., 2012) and encompassing required functionalities (Angelov et al., 2009), domain knowledge (Saay and Norta, 2016), extant technologies (Kuppusamy and Suresh, 2020), pertinent standards, protocols, and other essential elements to ensure compatibility with existing software. Following the creation of a concrete architecture, it is imperative to assess its alignment with the expectations and concerns of stakeholders. Additionally, it is crucial to evaluate its quality requirements prior to the development of applications based upon it.

4. Research Results

The study comprises two complementary parts: a theoretical component, which focused on an iterative literature analysis to map and examine the possibilities for developing and evaluating a reference architecture (RA) for the Smart Schoolhouse (RASS); and a practical component, in which the knowledge derived from the theoretical part was applied to the development and evaluation of the RASS.

A total of 137 scientific articles were analysed in depth, including 68 directly related to reference architectures in the field of education. This analysis identified the key stages of the RA life cycle, provided recommendations for selecting an appropriate architectural style and approach, and offered essential input for defining quality indicators, selecting an evaluation method, and conducting the evaluation itself.

The insights gained from the theoretical component were applied to the creation of the RASS. The result is a three-dimensional, layered reference architecture, developed using a top-down approach grounded in service-oriented architecture (SOA) principles. Its substantive foundation derives from data collected over four years through case studies conducted in 19 Estonian schools. These data are represented in the form of the Smart Schoolhouse Self-Assessment Model (SAMSS) and may be characterised as stakeholder concerns. Using SAMSS as input and drawing on the first check (DIC) of the RATE model, six personas and scenarios were developed. These personas and scenarios were subsequently used in the evaluation of the RASS, corresponding to the second check (SAC) of the RATE model.

Given the abstract nature of architecture, SAC facilitates early, forward-looking decision-making by validating architectural solutions during the design phase, prior to significant implementation investments. Its primary purpose is to assess whether the current architectural solutions adequately address stakeholder concerns and instil sufficient confidence in their suitability.

The evaluation confirmed that the RASS broadly meets stakeholder expectations and is sufficiently flexible and applicable to support the integration of IoT solutions and the use of learning analytics data in STEM education. During the evaluation, experts highlighted the need to enhance the architecture with additional data categories and integration capabilities, which were incorporated into the final refinement of the RASS.

In conclusion, the study demonstrated that the developed RASS provides a strong foundation for the development of practical solutions, supporting the effective integration of learning environments and data-driven education.

5. Conclusion

The changed expectations for today's students, i.e. the members of tomorrow's society, have created a situation where schools have to be innovative to provide the education that society expects from them, despite lacking the necessary resources to do so.

To support schools in engaging learners with real-world problem-solving through various teaching methods, enabling the application of innovative technology and the analysis of collected data, we proposed the Smart Schoolhouse concept. To clarify this concept for stakeholders and to facilitate the development of an appropriate software solution for software developers, we introduced a reference architecture (RA) based on the Smart Schoolhouse concept in this article. To ensure that the developed RA would be both relevant and applicable, we grounded its development in empirically justified prior practice, drawing evidence from a comprehensive literature review. We aimed to answer the following research question: How can we develop, validate, and implement a reference architecture (RA) for the Smart Schoolhouse?

In the first part of the scientific literature analysis, we identified the various processes and stages of the RA life cycle, along with recommendations for their management. Additionally, we mapped out different methods that have already been applied in the creation of RAs. The development of RASS was based on the empirically grounded reference architecture framework (EGRA), utilising a top-down, evidence-based approach.

The created RASS is, similarly to RAMI 4.0, three-dimensional service-oriented architecture: 1) Life cycle of IoT devices or solutions to be adopted in the Smart Schoolhouse, 2) Hierarchy levels of IoT devices usage (Disconnected, Online, Connected, Smart, Integrated) and 3) architecture, comprising eight layers: Device Layer, Data Layer, Integration Layer, Application Layer, Business Layer, Presentation Layer, Support Layer, Governance Layer, and Security Layer. Since it is currently uncertain when the resources and readiness of schools will emerge for the implementation of the Smart Schoolhouse idea, we tried to create the RA abstract enough to ensure its longer life-cycle.

In the second part of the scientific literature analysis, we focused on RA evaluation methods to identify the most suitable one for RASS. To ensure the empirical validity of the assessment, confirming that the chosen method and measurement technique are of high quality and targeted towards what we intend to measure, we selected the RATE method, which is recommended by experts and validated in practice. The first two checks from the RATE method – DIC and SAC – were employed. The study included five experts in their respective fields, each possessing in-depth knowledge in at least two of the following fields: education, the Internet of Things (IoT), or software engineering. Six personas and six scenarios were utilised. This article presents the enhanced and refined RASS, which has been improved based on recommendations and suggestions from the evaluation process, in order to prevent the spread of misinformation.

Due to the constrained resources available during the evaluation of the RASS, this study is subject to several limitations, which present opportunities for improvement in future research and development endeavours. Among the most notable constraints, one may underscore 1) the geographical location of this study, 2) the novelty of the Smart Schoolhouse concept 3) the small size of the sample, 4) the experience of experts engaged in the RASS evaluation, 5) the level of abstraction of the RASS devised for its presentation and evaluation, and additionally, 6) the complexity of evaluating an abstract RA, 7) the restricted quantity of personas and scenarios fashioned for its evaluation.

1. Geographical limitation - The study encompassed a range of specialists; however, they all originated from the IT sector and hailed from a markedly homogeneous background with respect to both educational and living conditions. The inclusion of external experts would undeniably have furnished additional viewpoints and enhanced the evaluative procedure. The incorporation of international experts would necessitate broadening the linguistic spectrum, yet this may be contemplated in subsequent investigations.

2. The novelty of the concept of Smart Schoolhouse – The evaluation process may be biased due to the novelty of the concept of the Smart Schoolhouse. Since it is unknown when schools will have the necessary resources to implement Smart Schoolhouse concept, the idea was introduced and evaluated solely based on a RA. No additional resources were allocated for the development of a specific architecture or prototype, though this could be considered in future studies.

3. Small sample – As the current stage of development prioritised obtaining feedback on the alignment of the RA with stakeholder requirements and concerns, as well as its suitability for subsequent stages, minor flaws in the RA were of lesser importance. Therefore, a smaller sample sufficed for this study. In the future, as the Smart Schoolhouse concept is implemented, it is imperative to conduct a new mapping of stakeholder requirements. Based on the results, updating the RA is necessary, along with a comprehensive evaluation that should involve a larger number of experts.

4. Experts' Experiences – The efficacy of RASS assessment was significantly contingent upon the expertise and experience of the professionals engaged in the evaluation process. Although numerous software developers exist, software development frequently does not depend on RA, rendering it difficult to locate experts who possess simultaneous expertise across multiple domains (education, software development, the Internet of Things) with experience in software creation predicated on RA. In preparing for the following studies, it would be prudent to consider allocating resources in the budget to involve an expert with RA evaluation experience.

5. Abstraction of RA – Identifying the optimal level of abstraction frequently presents a considerable challenge. The RA requires a degree of abstraction that is sufficient to mitigate the risk of rapid obsolescence attributable to technological advancements, to facilitate alternative decision-making processes, and to guarantee its applicability for broad usage. Concurrently, it necessitates sufficient detail to offer a comprehensive overview of crucial elements, whilst efficiently securing the fulfilment of stakeholder objectives. In the assessment of the RASS, feedback from experts indicated that our selected approach is satisfactory.

6. The complexity of evaluating an abstract RA – Owing to its abstract nature, evaluating RA presents a considerable challenge, no universally applicable methods exist for undertaking such an evaluation. Consequently, it is advisable to customise an architecture evaluation framework. For the evaluation of RASS, the initial two checks of the RATE method were employed. The first check facilitated the development of scenarios rooted in stakeholder concerns, whilst the second aided in determining whether the solutions are satisfactory and meet the requirements. Since the second check predominantly relies on expert judgement, requiring substantial effort while producing only qualitative data, the outcomes are significantly influenced by the knowledge and experience of the experts.

7. The scope of use-case scenarios is restricted – In our evaluation of the RASS, a limited array of scenarios was employed. Our emphasis was placed on the technology innovation area pertinent to the self-assessment model, enriched by facets of change management and pedagogical innovation to enhance comprehension of the context. Our objective was to encompass the majority of descriptions outlined in the criteria. Subsequent research might gain from an expanded collection of succinct scenarios.

Our study undoubtedly has several limitations at different levels and related to various fields, but the ones mentioned above are those of which we are aware and recommend to be considered in future research.

In conclusion, a thorough literature review and input gathered from previous studies enabled the development of a reference architecture that supports the Smart Schoolhouse concept. This architecture was further refined through an evaluation involving five experts. This process confirms that the research question was effectively addressed and that, through the evaluation, the resulting RASS was validated as meeting the expectations of relevant stakeholders.

References

Abichandani, P., Sivakumar, V., Lobo, D., Iaboni, C., Shekhar, P. (2022). Internet-of-things curriculum, pedagogy, and assessment for stem education: A review of literature. IEEE Access 10, 38351-38369.

Abu-Matar, M., Mizouni, R. (2018). Variability modeling for smart city reference architectures. IEEE International Smart Cities Conference (ISC2) (p. 1-8). IEEE.

Aleksieva-Petrova, A., Gancheva, V., Petrov, M. (2020). Software Architecture for Adaptation and Recommendation of Course Content and Activities Based on Learning Analytics. International Conference on Mathematics and Computers in Science and Engineering (MACISE) (p. 16-19). IEEE.

Alsobhi, A., Khan, N., Rahanu, H. (2015). Dyslexia adaptive e-learning system based on multi-layer architecture. Science and Information Conference (SAI), (p. 776-780). IEEE.

Angelov, S., Grefen, P., Greefhorst, D. (2009). A classification of software reference architectures: analyzing their success and effectiveness. Joint Working IEEE/IFIP Conference on Software Architecture European Conference on Software Architecture (p. 141–150). IEEE.

Angelov, S., Grefen, P., Greefhorst, D. (2012). A framework for analysis and design of software reference architectures. Information and Software Technology, 54.4, 417-431.

Angelov, S., Trienekens, J., Grefe, P. (2008). Towards a method for the evaluation of reference architectures: Experiences from a case. Software Architecture: Second European Conference, ECSA 2008 (p. 225-240). Paphos, Cyprus: Springer Berlin Heidelberg.

Ataei, P., Litchfield, A. (2020). Big data reference architectures, a systematic literature review. ACIS 2020 Proceedings.

Ataei, P., Litchfield, A. (2022). The state of big data reference architectures: A systematic literature review. IEEE Access, 10, 223789-113807. doi:10.1109/ACCESS.2022.3217557

Baek, Y.-M., Mihret, Z., Shin, Y.-J., Bae, D.-H. (2020). A Modeling Method for Model-based Analysis and Design of a System-of-Systems. 27th Asia-Pacific Software Engineering Conference (APSEC) (p. 336-345). IEEE.

Batista, P., Rodrigues, C., Kassab, M. (2022). ARC-SoISE: Towards a Reference Architecture for Constituents of Educational Systems-of-Information Systems. 17th Annual System of Systems Engineering Conference (SOSE) (p. 142-147). IEEE.

Boyanov, L., Kisimov, V., Christov, Y. (2020). Evaluating IoT reference architecture. 2020 International Conference Automatics and Informatics (ICAI) (p. 1-5). IEEE.

Clements, P., Kazman, R., Klein, M. (2010). Evaluating software architectures. Pearson Education.

Cloutier, R., Muller, G., Verma, D., Nilchiani, R., Hole, E., Bone, M. (2010). The Concept of Reference Architectures. Systems Engineering, 13(1), 14-27. doi:10.1002/sys.20129

de Oliveira Neves, V., Bertolino, A., De Angelis, G., Garcés, L. (2018). Do we need new strategies for testing systems-of-systems? 6th International Workshop on Software Engineering for Systems-of-Systems, (p. 29-32).

Drlik, M., Skalka, J., Švec, P., Kapusta, J. (2018). Proposal of learning analytics architecture integration into university IT infrastructure. 12th International Conference on Application of Information and Communication Technologies (AICT) (p. 1-6). IEEE.

Ehrlich, M., Gergeleit, M., Trsek, H., Lukas, G. (2020). Towards automated security evaluation within the industrial reference architecture. 25th IEEE International conference on emerging technologies and factory automation (ETFA), 1, p. 1644-1651.

Fatima, I., Lago, P. (2023). A Review of Software Architecture Evaluation Methods for Sustainability Assessment. 20th International Conference on Software Architecture Companion (ICSA-C) (p. 191-194). IEEE.

Galster, M. (2015). Software Reference Architectures: Related Architectural Concepts and Challenges. Proceedings of the 1st International Workshop on Exploring Component-based Techniques for Constructing Reference Architectures (p. 5-8). CobRA .

Galster, M., Avgeriou, P. (2011). Empirically-grounded Reference Architectures: A Proposal. Proceedings of the joint ACM SIGSOFT conference--QoSA and ACM SIGSOFT symposium-

-ISARCS on Quality of software architectures--QoSA and architecting critical systems--ISARCS (p. 153-158). Boulder, Colorado, USA.

Garcés, L., Nakagawa, E. (2017). A Process to Establish, Model and Validate Missions of Systems-of-Systems in Reference Architectures. Proceedings of the Symposium on Applied Computing, (p. 1765-1772). doi:10.1145/3019612.3019799

Ghantous, G., Gill, A. (2020). The DevOps reference architecture Evaluation: A design Science research case study. IEEE International Conference on Smart Internet of Things (SmartIoT) (p. 295-299). IEEE.

Gidey, H., Marmsoler, D., Eckhardt, J. (2017). Grounded Architectures: Using Grounded Theory for the Design of Software Architectures. IEEE international conference on software architecture workshops (ICSAW) (p. 141-148). IEEE.

Guth, J., Breitenbücher, U., Falkentha, M., Leymann, F., Reinfurt, L. (2016). Comparison of IoT platform architectures: A field study based on a reference architecture. Cloudification of the Internet of Things (CIoT) (p. 1-6). IEEE.

Hankel, M., Rexroth, B. (2015). The reference architectural model industrie 4.0 (rami 4.0). p. 4-9.

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. Scandinavian Journal of Information Systems, 19(2), 87-92.

Hevner, A., March, S., Park, J., Ram, S. (03 2004. a.). Design Science in Information Systems Research. MIS Quarterly, 28(1), 75-106.

Hoel, T., Mason, J. (2018). Standards for smart education–towards a development framework. Smart Learning Environments, 5(1), 1-25.

Islam, S., Rokonuzzaman, M. (2009). Adaptation of ATAM SM to software architectural design practices for organically growing small software companies. 12th International Conference on Computers and Information Technology (p. 488-493). IEEE.

ISO and IEC. (2023). Systems and software engineering — System life cycle processes. International standard ISO/IEC/IEEE 15288, 1-15. International Organisation for Standardisation and International Electrotechnical Commission.

ISO/IEC/IEEE. (2019a). Software, Systems and Enterprise—Architecture Evaluation Framework. International standard ISO/IEC/IEEE 42030. ISO/IEC/IEEE.

ISO/IEC/IEEE. (2019b). Software, Systems and Enterprise—Architecture Processes. International standard ISO/IEC/IEEE 42020. ISO/IEC/IEEE.

Karlsson, E.-A. (2016). Architecture Evaluation - Threat or Opportunity? rmt: P. A. Abrahamsson, Product-Focused Software Process Improvement: 17th International Conference, PROFES 2016, Proceedings. Vol. 10027 (p. 765–766). Trondheim, Norway: Springer.

Kazman, R., Bass, L., Abowd, G., Webb, M. (1994). SAAM: A method for analyzing the properties of software architectures. Proceedings of 16th International Conference on Software Engineering (p. 81-90). IEEE.

Kazman, R., Klein, M., Clements, P. (2000). ATAM: Method for architecture evaluation. Pittsburgh PA 15213-3890: Carnegie Mellon University, Software Engineering Institute.

Kearney, P., Asal, R. (2019). ERAMIS: A Reference Architecture-based Methodology for IoT Systems. IEEE World Congress on Services (SERVICES) (p. 366-367). IEEE. doi:10.1109/SERVICES.2019.00106

Kitchenham, B., Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report EBSE-2007-01. UK: Software Engineering Group School of Computer Science and Mathematics Keele University and Department of Computer Science University of Durham.

Knodel, J., Naab, M. (2016). Pragmatic Evaluation of Software Architectures (Kd. 1). Springer International Publishing.

Kuppusamy, P., Joseph, K. (2020). Robotic Process Automation to Smart Education. International Journal of Creative Research Thoughts, 8(6), 3775-3784.

Kuppusamy, P., Suresh, J. (2020). Service-Oriented Reference Architecture to Smart Education. International Journal of Advance Scientific Research and Engineering Trends, 5(2).

Kusmin, M. (2019a). Co-designing the kits of IoT devices for inquiry-based learning in STEM. Technologies, 7(16). Allikas: https://www.mdpi.com/2227-7080/7/1/16

Kusmin, M. (2019b). Inquiry-Based Learning and Trialogical Knowledge-Creation Approach in Smart Schoolhouse Supported by IoT Devices. IEEE Global Engineering Education Conference (EDUCON) (p. 571-575). Dubai, United Arab Emirates: IEEE.

Kusmin, M., Laanpere, M. (2020). Supporting teachers for innovative learning in smart schools using Internet of things. IEEE Global Engineering Education Conference (EDUCON), (p. 1024-1030).

Kusmin, M., Laanpere, M. (2022). Design of the Smart Schoolhouse Self-assessment Model. 2022 IEEE Global Engineering Education Conference (EDUCON), (p. 526-531). Tunis, Tunesia. doi:10.1109/EDUCON52537.2022.9766535

Kusmin, M., Laanpere, M. (2023). The Implementation of the Smart Schoolhouse Concept. IEEE Global Engineering Education Conference (EDUCON) (p. 1-8). Salmiya, Kuwait: IEEE.

Kusmin, M., Laanpere, M. (2024). Validation of the Self-assessment model of the Smart Schoolhouse. In Press.

Kusmin, M., Kusmin, K.-L., Laanpere, M., Tomberg, V. (2019). Engaging Students in Co-designing Wearable Enhanced Learning Kit for Schools. Perspectives on Wearable Enhanced Learning (WELL) Current Trends, Research, and Practice, 97-120.

Kusmin, M., Saar, M., Laanpere, M. (2018). Smart schoolhouse — designing IoT study kits for project-based learning in STEM subjects. IEEE Global Engineering Education Conference (EDUCON) (p. 1514-1517). Santa Cruz de Tenerife, Spain: IEEE. doi:10.1109/EDUCON.2018.8363412

Li, J., Qiu, J., Dou, K., Liu, Y., Cheng, Y., Liu, S., . . . Wang, Q. (2019). A Reference Architecture and Evaluation Framework for Industrial Internet Platform. 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE) (p. 1290-1298). IEEE.

Lopes, Á., Barbosa, E., Vaccare, R. (2019). Service Oriented Reference Architecture for the Development of Context-awareLearning Environments. Frontiers in Education Conference (FIE) (p. 1-8). IEEE.

Maher, Y., Moussa, S., Khalif, M. (2020). Learners on focus: Visualizing analytics through an integrated model for learning analytics in adaptive gamified e-learning. Open Access Journal IEEEAccess, 197597-197616.

Morkevicius, A., Bisikirskiene, L., Bleakley, G. (2017). Using a systems of systems modeling approach for developing Industrial Internet of Things applications. 12th System of Systems Engineering Conference (SoSE) (p. 1-6). IEEE.

Nakagawa, E., Guessi, M., Maldonado, J., Feitosa, D., Oquendo, F. (2014). Consolidating a Process for the Design, Representation, and Evaluation of Reference Architectures. IEEE/IFIP Conference on Software Architecture, (p. 143-152). doi:10.1109/WICSA.2014.25

Nicolaescu, A., Lichter, H. (2016). Behavior-based architecture reconstruction and conformance checking. 13th Working IEEE/IFIP Conference on Software Architecture (WICSA) (p. 152-157). IEEE.

Norta, A., Grefen, P., Narendra, N. (2014). A reference architecture for managing dynamic inter-organizational business processes. Data & Knowledge Engineering, 52-89.

Nouira, A., Cheniti-Belcadhi, L., Brah, R. (2017). A Semantic Web Based Architecture for Assessment Analytics. International Conference on Tools with Artificial Intelligence (p. 1190). IEEE.

Palkar, S., Kamani, H. (2018). Web Based Tool for Traceability from Architecture Artifacts to ATAM. IEEE International Conference on Software Architecture Companion (ICSA-C) (p. 107-110). IEEE.

Pandey, J., Singh, A., Rana, A. (2020). Roadmap to smart campus based on IoT. 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (p. 909-9013). IEEE.

Rabelo, R., Noran, O., Bernus, P. (2015). Towards the Next Generation Service Oriented Enterprise Architecture. IEEE 19th International Enterprise Distributed Object Computing Workshop (p. 91-100). IEEE.

Saay, S., Norta, A. (2016). A reference architecture for a national e-Learning infrastructure. 9th International Conference on Utility and Cloud Computing (p. 404-409). IEEE/ACM: Shanghai, China.

Smith, B., Gallagher, P., Schatz, S., Vogel-Walcutt, J. (2018). Total Learning Architecture: Moving into the Future. Proceedings of the interservice/industry training, simulation, and education conference (I/ITSEC), (p. 1-11).

Szwed, P., Skrzynski, P., Rogus, G., Werewka, J. (2013). Ontology of architectural decisions supporting Ontology of architectural decisions supporting. Federated Conference on Computer Science and Information Systems, (p. 287-290).

Zapata-Rivera, L., Petrie, M. (2018). xAPI-based model for tracking on-line laboratory applications. EEE Frontiers in Education Conference (FIE) (p. 1-9). IEEE.

Zbick, J. (2017). A Web-based Reference Architecture for Mobile Learning: Its Quality Aspects and Evaluation. IEEE International Conference on Software Architecture Workshops (ICSAW) (p. 230-235). IEEE.

Weinreich, R., Buchgeher, G. (2014). Automatic Reference Architecture Conformance Checking for SOA-Based Software Systems. IEEE/IFIP Conference on Software Architecture (p. 95-104). Sydney, NSW, Australia: IEEE. doi:10.1109/WICSA.2014.22

vom Brocke, J., Hevner, A., Maedch, A. (2020). Introduction to design science research. Design Science Research. Cases, p. 1-13. doi:10.1007/978-3-030-46781-4_1

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R. (2015). Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. Communications of the association for information systems, 37.1(9), 205-224.

Ünal, P. (2019). Reference architectures and standards for the internet of things and big data in smart manufacturing. 7th international conference on future internet of things and cloud (FiCloud) (p. 243-250). IEEE.

Received September 20, 2024, revised March 31, 2025, accepted March 31, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 2, pp. 357–365
https://doi.org/10.22364/bjmc.2025.13.2.03

Using LLM-s for Zero-Shot NER for Morphologically Rich Less-Resourced Languages

Agris ŠOSTAKS[1], Sergejs RIKAČOVS[1], Artūrs SPROĢIS[1], Oskars MĒTRA[1], Uldis LAVRINOVIČS[2]

[1] Institute of Mathematics and Computer Science, University of Latvia
[2] LETA, Latvian Information Agency

agris.sostaks@lumii.lv, sergejs.rikacovs@lumii.lv,
arturs.sprogis@lumii.lv, oskars.metra@gmail.com,
uldis.lavrinovics@leta.lv

ORCID 0009-0003-5987-1644, ORCID 0009-0003-8989-0942, ORCID 0000-0002-2320-0887,
ORCID 0009-0002-3105-9059, ORCID 0009-0009-8378-2020

Abstract. Developing Named Entity Recognition (NER) solutions for morphologically rich but low-resource languages like Latvian is a complex task. Most state-of-the-art methods rely on deep learning models like BERT, which require substantial expertise in architectures, methods, and access to extensive computational resources and data. In this study, we explore the potential of using popular large language models (LLMs) in a zero-shot setting without additional training. We evaluate their performance on the publicly available Latvian dataset (Gruzitis, et.al., 2018) using the F1-score and find that their results are comparable to state-of-the-art methods. Moreover, LLMs offer a simpler, more resource-efficient alternative for NER tasks.

Keywords: LLM, zero-shot, NER

1 Introduction

Natural Language Processing (NLP) focuses on several key directions, including language understanding, generation, and transformation. Language understanding involves parsing, named entity recognition, and sentiment analysis, which aim to extract meaning and structure from text. Emerging trends include few-shot learning, large pre-trained models like transformers (BERT and LLM-s), and integrating multi-modal data, such as text and images, to enhance understanding and generation.

We focus on the named entity recognition (NER) task, which involves identifying and classifying entities in text into predefined categories, such as names of people, organizations, locations, dates, and more. For example, in the sentence "Barack Obama was born in Hawaii," an NER system would identify "Barack Obama" as a person and "Hawaii" as a location.

NER is crucial in several areas of NLP, including information extraction, where it helps convert unstructured text into structured data by identifying key entities like people, organizations, and locations. In search engines, NER improves the relevance of results by recognizing important entities in queries. It also enhances question-answering systems by identifying entities to provide more precise answers. In sentiment analysis, NER associates emotions or opinions with specific entities, such as brands or products. Additionally, NER supports machine translation by ensuring proper handling of named entities across languages, and it plays a role in text summarization by highlighting important entities to generate more informative summaries.

We focus on the automatic recognition of named entities, including people, organizations, and geographical locations. Our work involves extracting information from unstructured, low-quality texts, such as social media posts. Specifically, we analyze data in Latvian, a morphologically rich yet less-resourced language.

At LETA, Latvia's leading news and media monitoring agency, our practical work revolves around sentiment and propaganda analysis, where accurately identifying these named entities is essential. Since LETA has limited computational resources, we explore efficient alternatives to traditional, resource-intensive methods for tackling this task.

With the rise of LLMs, we explore their potential for our task in a zero-shot setting. The accessibility, versatility, and ease of integration of LLMs enable rapid development and incorporation into information extraction systems. Our research focuses on evaluating the quality of outputs from different LLMs in such settings. We designed a single prompt and tested it on several models, including Llama-3.1-405b, GPT-4o-mini, Gemma-2-9b-it, Llama-3.1-8b, and Chat-GPT-4o. These represent popular LLM families, such as open-source Llama models and commercial Chat-GPT systems. We compare models with large and smaller parameter sizes to provide insights into using LLMs both as external services and as locally deployed components. While we have worked with additional models (e.g., Chat-GPT-3.5, Gemini-1.5, LLAMA3-8b-chat), we excluded them due to poor initial results or obsolescence with newer versions.

We evaluated the models on the named entity annotation layer of the publicly available Latvian Multilayer Corpus (FullStack-LV dataset) (Gruzitis et al., 2018) to compare their performance with previous work. The evaluation used the F1-score, which balances precision and recall to measure accuracy. LLMs demonstrated F1-scores close to state-of-the-art, even in zero-shot settings. The best-performing model, Llama-3.1-405b, achieved an F1-score of 81.33, nearly matching the highest known score of 82.6 reported by (Znotiņš and Barzdins, 2020) on the same dataset. Smaller models showed lower performance, with GPT-4o-mini scoring 65.0 and Gemma-2-9b-it scoring 60.2.

The primary contribution of this paper is verifying a seemingly simple yet fundamental question: To what extent can LLMs replace task-specific NER tools in low-resource settings? We evaluate the popular and accessible LLMs for zero-shot Named Entity Recognition (NER) in a morphologically rich, low-resource language, specifically, Latvian. The study focuses on the most common named entity types: person (individual or group names), geopolitical entity and location (representing countries, cities, regions, and geographical places), and organization (including company and institution names). One might expect pre-trained, fine-tuned models to be significantly superior, yet our findings reveal that the performance gap between the best LLMs and state-of-the-art tools is surprisingly small. This unexpected result highlights an important message for the NLP research community working with less-resourced languages like Latvian. Just as deep learning and transformer models revolutionized NLP, LLMs are now reshaping the field once again—even in low-resource scenarios. Results show that LLMs achieve performance close to state-of-the-art while requiring significantly less effort and resources compared to traditional methods.

2 Related Work — Fast Changing State-of-the-Art

Numerous methods are available for performing NER, with deep learning approaches being the current mainstream. Most commonly used datasets are designed for large languages such as English, Chinese, and Arabic (Hu et al., 2024).

Early deep learning approaches used Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, to capture sequential information from text. Another important method involves Convolutional Neural Networks (CNNs), which are commonly used at the character level to capture morphological features such as prefixes and suffixes, improving entity recognition even in morphologically rich languages. CNNs are often combined with RNNs to improve performance, with CNNs processing local information at the character level and RNNs handling word-level sequences. The introduction of Transformers, particularly models like BERT (Bidirectional Encoder Representations from Transformers), revolutionized NER by enabling the model to capture the bidirectional context in text. Unlike RNNs, which process input sequentially, transformers can access all positions in a sequence simultaneously, allowing them to better understand the context. BERT-based models have achieved state-of-the-art performance on various NER tasks by fine-tuning them on labelled NER datasets. Moreover, Conditional Random Fields (CRFs) are often used on top of RNN or BERT architectures to model the dependencies between output labels, ensuring that the sequence labelling respects entity boundaries. However, building such models demands considerable effort and resources. LLMs have transformed NER by utilizing deep contextual understanding, enabling fine-tuning for specific tasks, supporting few-shot learning through prompting, and transferring knowledge across languages and domains. This makes LLMs powerful and versatile tools for modern NER systems.

Fine-tuning LLMs for NER adapts a pre-trained model to accurately recognize and classify entities within a specific dataset. This process begins with a base LLM pre-trained on large-scale textual data and fine-tunes it using a labeled NER dataset, where each token or word is annotated with its entity type. For example, SLIMER (Zamai et al., 2024) is a fine-tuned Llama-2-7b model designed specifically for NER tasks.

Few-shot learning for NER with LLMs leverages their ability to generalize from minimal labeled data, avoiding the need for extensive task-specific fine-tuning. This approach involves providing the LLM with a small set of labeled examples during inference, illustrating how text tokens correspond to specific entity types. The model uses its pre-trained contextual knowledge and these examples to identify and classify entities in unseen text. Prompt engineering plays a key role, with carefully designed prompts guiding the model's performance. Notable examples include methods to optimize prompt structure and example selection (Cheng et al., 2024), GPT-NER for labeling entities using few-shot learning (Wang et al., 2023), and PromptNER, which integrates entity definitions and examples directly in the prompt (Ashok et al., 2023). Few-shot learning is especially valuable in domain-specific or low-resource settings, as it reduces the need for extensive annotated datasets and computational power. Techniques like in-context learning (Jiang et al., 2024) and retrieval-augmented generation further enhance performance by improving the model's understanding of the NER task.

LLMs' ability to capture complex word dependencies and contextual relationships suggests they could perform well in zero-shot settings without specific pre-training or additional examples. Promising results in this area (Xie et al., 2023) have been achieved using techniques like syntactic prompting combined with tool augmentation. While such approaches have shown success for major languages, we aim to evaluate their effectiveness in zero-shot settings for less-resourced languages, specifically Latvian.

The first attempt to address NER for Latvian was made with the TildeNER toolkit (Pinnis, 2012), which uses a supervised conditional random field classifier enhanced with heuristic and statistical refinement methods. TildeNER achieved an F1-score of approximately 60 on a manually created dataset containing 881 named entities. The authors focused on three NER entity types: locations, persons, and organizations.

Vīksna and Skadina (2020) introduced a pre-trained BERT model trained on large Latvian corpora, achieving an F1-score of 81.91 across 9 NER types. Meanwhile, Znotins and Barzdins (2020) developed LVBERT, a BERT-based model fine-tuned specifically for Latvian to enhance performance on Latvian NLP tasks. LVBERT reached a state-of-the-art F1-score of 82.6 on the FullStack-LV dataset.

Next, Vīksna and Skadina (2022) investigated the performance of various multilingual NER models within the state-of-the-art natural language processing framework, Flair. They found that for Latvian, the more specialized LitLat BERT model achieved the best F1-score of 81.97 on the FullStack-LV dataset. Therefore, BERT-based fine-tuned models currently deliver the best results for morphologically rich, less-resourced languages like Latvian. However, creating such models is a complex and resource-intensive process.

3 Prompt Engineering

We use LLMs to address NER tasks. While LLMs can be fine-tuned for specific tasks using deep learning methods, this requires significant resources, including large datasets. As an alternative, zero-shot prompt-based methods are used to guide the model without the need for extensive fine-tuning. These prompts direct LLMs to perform specific tasks by providing clear, structured instructions within the input. Instead of retraining the model, prompts leverage its existing knowledge by specifying the task, input format, and output requirements.

For our task, we create prompts that instruct the LLM to extract mentions of different named entity types from Latvian text. We design a separate prompt for each entity type, as adjusting a single prompt for all entity types is challenging due to differing errors. Since the structure of the prompts is similar for each type, we illustrate the case of named entities for persons. Here is a step-by-step breakdown of the prompt:

Task definition: LLM needs to act as an NLP expert and apply NER techniques to identify all mentions of individuals (persons) in the provided text. This reduces ambiguity and ensures the model works within the intended scope of NER.

```
Act as a NLP researcher performing Named Entity Recognition (NER).
Analyze the following text fragment labeled TextToAnalyze.
From that fragment, extract a list with named entity mentions that represent persons (named individuals).
```

Clarification of the task: LLM has to exclude generic terms. The prompt emphasizes that only specific individuals should be listed. Generic roles or titles like ”teacher,” ”president,” or ”doctor” should be excluded. This ensures that the list only includes names of actual people and not their job positions or generic designations. This aligns the model’s attention with the task and minimizes false positives.

```
Ensure that named entities representing persons refer to specific individuals by excluding generic terms such as titles or roles.
If a named entity refers to a role or position, exclude it from the list of people.
Before giving the answer analyze the list you created and exclude from this list items that are not referring to named individuals.
```

Output definition: the extracted person entities must be returned in two forms: a) the original form as it appears in the provided text; b) the name of the person converted to the nominative case (which is the default grammatical case for the subject in Latvian, like ”John” instead of ”John’s”). This ensures consistency in the representation of named entities, even if they appear in different grammatical forms in the text. It enhances the usability of the output by providing the proper ”base” form of names, which is crucial for downstream tasks like database matching or reference alignment. The final output must be a JSON object with the key *persons*, storing an array of the extracted named person entities. If no person entities are found, the JSON should return an empty list. The instruction explicitly states, ”Do not give any additional explanation,” forcing the LLM to stick to the task at hand and focus on generating the output in the desired format without unnecessary verbosity or commentary, improving the response’s efficiency and clarity.

```
Return JSON object. This object should have field persons, containing extracted person mentions.
For each item in this list provide both latvian text labeled as lv, and same text but in nominative case labeled lv_nc.
If there are no entity mentions to return - return empty list.
Do not give any additional explanation.
Ensure that you are returning valid JSON.
```

Input:

```
TextToAnalyze:
"Gleznas attēlo , kā Jānis Bērziņš paraksta laulību līgumu ar Annu Kalniņu."
```

Output Example:

```
{
  "persons": [
    {
      "lv": "Jānis Bērziņš",
      "lv_nc": "Jānis Bērziņš"
    },
    {
      "lv": "Annu Kalniņu",
      "lv_nc": "Anna Kalniņa"
    }
  ]
}
```

4 Experiment

We use the FullStack-LV dataset (Gruzitis et al., 2018) to evaluate LLMs. This Latvian corpus is designed for broad applications, including natural language understanding (NLU), abstractive text summarization, and knowledge base population. It features hierarchical named entity annotations with both outer and inner (nested) entities. The dataset includes 3947 paragraphs of text, containing 9697 outer entities and 944 inner entities, categorized into nine types: geopolitical entities (GPE), person, time, location, product, organization, money, event, and a general ”entity” category. It adopts a simplified CoNLL-2003 format with BIO (Begin, Inside, Outside) labeling.

The FullStack-LV dataset is commonly used to train and evaluate NER models, including multilingual transformers and other machine learning approaches. Experiments with multilingual transformers have shown strong results, though F1-scores vary depending on the model and training parameters. The dataset's hierarchical structure presents an additional challenge, particularly for models not optimized for nested entity recognition.

We focus on three named entity categories: persons, locations and GPEs, and organizations. The subset of the dataset used for evaluation includes 3,104 person entities, 2,031 GPEs and locations, and 1,847 organization entities, making up 72% of all outer entities in the dataset. This subset is sufficient for evaluation since results vary similarly across entity types, as shown in previous research (Pinnis, 2012), (Vīksna, 2020), and in practical applications, prompts would need to be tailored for each type separately.

It is important to note that our evaluation differs from previous research on Latvian NER. Traditionally, tokenization is performed first; for example, LVBERT uses LVTagger (Paikens et al., 2013) for sentence tokenization. In such cases, the model classifies tokens directly, labeling each token individually. This simplifies F1-score calculation, as results are straightforward to interpret by comparing gold labels with classified labels to identify true and false positives.

In our approach, the process is more complex. Extracted entities are compared to gold-standard entities and their labels. The gold data includes two forms: the exact string from the text and its nominative (base) form. Models are asked to extract both forms, and we perform cross-comparison. If either extracted form matches the gold data, it is counted as a true positive; otherwise, it is a false positive. Additionally, false negatives—entities missed by the models—must be accounted for.

This method introduces room for errors. For instance, quotation marks in organization names can cause mismatches when models omit them. Simple data cleansing, such as removing quotation marks and trimming leading or trailing whitespace from the gold data, significantly improves results.

We evaluated five models: Llama-3.1-405b, chat-gpt-4o, gpt-4o-mini, gemma-2-9b-it, and Llama-3.1-8b. Access to these models was provided via online services using their respective APIs. The aggregate F1-scores for all evaluated NER types on the FullStack-LV dataset (Gruzitis et.al., 2018) are presented in Table 1. The first row presents the baseline result achieved by LVBERT (Znotins and Barzdins, 2020). The second result, shown in parentheses (81.3), was obtained after applying cleansing procedures to the gold data.

Table 1. The results of LLM tests for the NER task.

Model	F1-Score
LVBERT (Baseline)	82.6
Llama_3.1_405b	76.6 (81.3)*
gpt_4o	71.9
gpt_4o_mini	65.0
gemma-2-9b-it	60.2
Llama_3.1_8b	16.6

Larger models, such as Llama-3.1-405b and chat-gpt-4o, achieve performance close to state-of-the-art, while smaller models like gpt-4o-mini and gemma-2-9b-it perform worse, with Llama-3.1-8b showing even significantly lower results.

It is important to note that LLMs exhibit considerable variation in performance across different NER types. For instance, Llama-3.1-405b achieves a high F1-score of 91.0 for person entities but only 50.0 for organizations, which improves to 69.0 after applying data cleansing. Similar disparities are observed with other LLMs, aligning with discrepancies noted in previous research (Pinnis, 2012), (Vīksna, 2020).

5 Conclusion

LLMs represent a transformative technology for NER tasks, even for morphologically rich and less-resourced languages like Latvian. Despite not being specifically trained for Latvian—e.g., Llama 3 contains only 5% non-English data—their ability to process texts in Latvian with near state-of-the-art quality is remarkable. What truly sets LLMs apart is their out-of-the-box usability, eliminating the need for pre-training or fine-tuning. This greatly simplifies and broadens the application of NER technology for less-resourced languages.

However, challenges remain in using LLMs for private data that cannot be processed via third-party servers or transmitted over the Internet. Smaller LLMs, which demand fewer computational resources, currently lack the required quality for effective NER tasks. Conversely, deploying larger LLMs requires significant investment unless sufficient computational infrastructure is available.

Looking ahead, we anticipate ongoing advancements in LLM quality for NER as models continue to evolve. Techniques like prompt engineering could further improve the extraction of specific NER types, while fine-tuning and few-shot learning are likely to enhance overall performance even further. We expect LLMs to soon surpass state-of-the-art NER methods. However, a more in-depth performance and error analysis is necessary, as F1-scores vary significantly across different NER types. This analysis will provide a clearer understanding of the limitations and best-use scenarios for LLMs in NER tasks.

Acknowledgements

The research leading to these results has received funding from the research project ”Competence Centre of Information and Communication Technologies” of the EU Structural funds, contract No.5.1.1.2.i.0/1/22/A/CFLA/008 signed between IT Competence Centre and Central Finance and Contracting Agency, Research No. 2.6 “Applications of Large Language Models in Analyzing Large Text Corpora”.

References

Ashok, D., and Lipton Z. (2023) ”PromptNER: Prompting for Named Entity Recognition.” arXiv preprint arXiv:2305.15444.

Cheng, Q., Liqiong, C., Zhixing, H., Juan, T., Qiang, X., Binbin, N. (2024) ”A Novel Prompting Method for Few-Shot NER via LLMs.” Natural Language Processing Journal, Volume 8, 2024, 100099. ISSN 2949-7191. https://doi.org/10.1016/j.nlp.2024.100099.

Gruzitis, N., Pretkalnina, L., Saulite, B., Rituma, L., Nespore-Berzkalne, G., Znotins, A., Paikens, P. (2018) ”Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU.” Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC).

Hu, Z., Hou, W., Liu, X. (2024) ”Deep Learning for Named Entity Recognition: A Survey.” Neural Comput & Applic 36, 8995–9022. https://doi.org/10.1007/s00521-024-09646-6.

Guochao, J., Zepeng, D., Yuchen, S., Deqing, Y. (2024) ”P-ICL: Point In-Context Learning for Named Entity Recognition with Large Language Models.” arXiv preprint arXiv:2405.04960.

Paikens, P., Rituma, L., Pretkalnina, L. (2013) ”Morphological Analysis with Limited Resources: Latvian Example.” In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), pp. 267–277.

Pinnis, M. (2012) ”Latvian and Lithuanian Named Entity Recognition with TildeNER.” Seed 40: 37.

Vīksna, R., and Skadiņa, I. (2020) ”Large Language Models for Latvian Named Entity Recognition.” Human Language Technologies–The Baltic Perspective. IOS Press, pp. 62-69.

Vīksna, R., and Skadiņa, I. (2022) ”Multilingual Transformers for Named Entity Recognition.” Baltic Journal of Modern Computing, Volume 10, Issue 3.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Wang, G. (2023) ”GPT-NER: Named Entity Recognition via Large Language Models.” arXiv preprint arXiv:2304.10428.

Xie, T., Li, Q., Zhang, J., Zhang, Y., Liu, Z., Wang, H. (2023) ”Empirical Study of Zero-Shot NER with ChatGPT.” In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 7935–7956.

Zamai, A., Zugarini, A., Rigutini, L., Ernandes, M., Maggini, M. (2024) ”Show Less, Instruct More: Enriching Prompts with Definitions and Guidelines for Zero-Shot NER.” arXiv preprint arXiv:2407.01272.

Znotiņš, A., and Barzdiņš, G. (2020) ”LVBERT: Transformer-Based Model for Latvian Language Understanding.” Human Language Technologies–The Baltic Perspective. IOS Press, pp. 111-115.

Received January 13, 2025 , revised March 11, 2025, accepted April 7, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 2, 366-388
https://doi.org/10.22364/bjmc.2025.13.2.04

Statistics-based Quality Control of Patterns

Andrejs NEIMANIS[1], Liva PURINA[1], Zane BICEVSKA[1],
Edgars DIEBELIS[1], Janis BICEVSKIS[2]

[1]DIVI Grupa Ltd, Fridriha Candera str. 1, Riga, LV-1046, Latvia
[2] University of Latvia, Raina Blvd. 19, LV-1586, Riga, Latvia

Andrejs.Neimanis@di.lv, Liva.Purina@patterns.di.lv,
Zane.Bicevska@di.lv, Edgars.Diebelis@di.lv, Janis.Bicevskis@lu.lv

ORCID 0000-0002-3320-8209, ORCID 0009-0008-0339-5956, ORCID 0000-0002-5252-7336,
ORCID 0000-0002-5950-9915, ORCID 0000-0001-5298-9859

Abstract. This study addresses a statistics-based testing procedure aimed at optimizing quality assurance for custom-made and made-to-measure (CM/M2M) clothing patterns. The procedure focuses on identifying machine-detectable issues in patterns before they progress to the expensive and time-intensive manual tailoring phase, thereby streamlining the testing cycle and reducing both time and costs. By analysing patterns for individuals with similar body measurements and applying statistical methods, the study identified potential design errors in pattern pieces based on measurable properties such as perimeters, areas, and contour-defining lines. Advanced statistical techniques, including residual analysis, Cook's distance, and Mahalanobis distance, were employed to detect outliers and pinpoint potential construction errors. Further analysis of line properties using predictive models—such as linear regression, random forests, generalized additive models (GAM), and rpart decision trees—revealed that a high frequency of outliers often correlates with construction anomalies. This research demonstrates that predictive modelling and outlier detection are effective tools for identifying errors in CM/M2M pattern construction, contributing to improved garment accuracy and production efficiency.

Keywords: Custom-made Clothing, Made-to-measure, Garment accuracy, Statistical testing, Outlier detection, Predictive modelling

1. Introduction

Personalized or custom-tailored clothing patterns are created in various contexts — both in traditional custom clothing production (the atelier model, where individual measurements are taken from the client and the garment is constructed according to these specific measurements), as well as in fashion houses (for small-scale clothing collections and individually crafted, high-quality designer garments). Additionally, specialized companies offer clients the option to purchase custom-designed patterns, often in digital formats.

According to a 2024 report (Sneha, 2024) by *Cognitive Market Research*, the global market for custom-tailored clothing was valued at approximately $50 billion, with a projected average annual growth rate (CAGR) of 10%. Researchers believe that, with the rise in digital accessibility, industrially manufactured custom-sized garments will gradually replace the current dominance of mass-produced standard-size clothing. This shift will significantly alter consumer purchasing habits, clothing production methods, and associated global supply chains. People will no longer need to buy off-the-rack clothing that only partially meets their preferences or fits their body size; instead, they will be able to order any desired garment (in a design of their choice, possibly even self-designed) tailored precisely to their measurements and delivered directly from the manufacturer.

The primary challenge in advancing the custom-made/made-to-measure (CM/M2M) clothing segment (McKee, 2024) lies in the objective difficulty of ensuring that garments sewn from automatically generated (algorithm-based) clothing patterns really fit the individuals for whom they are intended. In traditional custom tailoring, fit verification occurs manually — the garment is created, tried on by the wearer, and then altered (sometimes in multiple iterations) until it fits perfectly. However, if a pattern is generated automatically for a person located at a geographical distance, the fitting and alteration phase becomes not only costly (fabric, time) but may also be impractical (if the person is unavailable for fittings).

A solution must be found to ensure that the algorithm used to generate custom clothing patterns can, with a high probability, produce a pattern that fits well. Complete verification of the algorithm's accuracy could theoretically be achieved by sewing a garment for every person on Earth, but this is, of course, an unrealistic task.

The study describes how to accelerate the testing/improvement cycle for creating personalized products (clothing patterns). In today's fashion industry, developing such products involves a highly manual and iterative quality improvement process (sewing/fitting). The testing method described in the study, which is based on statistical methods, has the potential to speed up and thereby reduce the cost of the testing/improvement cycle for such products, as it allows manual fitting to be deferred until automatically identifiable issues have already been resolved.

The statistical methods described in the study can utilize variables that are either already available or can be easily obtained during the product construction process. The study concludes that, in practical application, the number of indicators to be tested will be significant and that these must be interpretable within the context of their application. Therefore, the main challenge of the outlined testing method is to quickly identify and exclude from further testing those indicators that are not useful for identifying issues in the tested pattern within the context of statistical methods.

The variables used in this testing method can be applied with any personalized garment construction system that produces results in vector graphics form. The description and its implementation can also be adapted for other CAD systems that create two-dimensional construction images from interrelated input data.

1.1. Problem Identification

Compared to a typical situation in software testing, quality assurance of a CM/M2M product under development faces several specific challenges — both in terms of input data and in evaluating the results achieved.

The input data for CM/M2M products (garments) consists of body measurements of individuals to be clothed where some of these measurements might be specific to the construction method used. Since a CM/M2M product is generally developed with the intent to be suitable for a large part of the population (e.g. all adult women), there is a need to provide data that covers the potential market of the product's potential buyers regarding the relevant body measurements.

Regardless of whether the CM/M2M product is offered to the customer as a finished garment or merely as a CM/M2M pattern, its quality is ultimately assessed by how well the resulting garment fits the individual. In addition to the physical process of tailoring/fitting, there should also be considered that the quality of fit is only partially quantifiable.

We will discuss these specific aspects of quality assurance for CM/M2M products in detail in the following sections and derive an accelerated quality assurance procedure from this analysis.

1.2. Main Idea of Solution

The procedure presented in the study aims to detect machine-identifiable issues within the CM/M2M product before the costly tailoring/fitting process begins. The concept is based on the assumption that the CM/M2M patterns for two individuals with similar body measurements must also be similar. Assuming it's possible to generate automated quality measurement indicators of "similar" or "different" for patterns, statistical methods are then used to identify test cases that deviate from the expected results.

Although this approach would not allow for assessing the overall quality of the pattern in terms of fitting, it holds significant potential to speed up the testing/improvement cycle for CM/M2M products. This approach would postpone the expensive and entirely manual fitting checks until after addressing the more apparent quality and measurement data issues in the patterns, which can be identified automatically.

2. Quality Control of Patterns

2.1. CM/M2M patterns used in the study

The procedure for statistics-based, automated problem identification presented in this study (hereafter referred to as the Testing procedure) is derived step by step. Initially, the simplest possible variant is applied to a straightforward yet practically relevant example: a basic skirt pattern. Even from this simple application, prerequisites for the use of statistical methods in quality assurance can be deduced.

While the first application case utilizes only two characteristics that define the size of the patterns, the second application case, involving the basic bodice pattern, also takes the shape of the patterns into account, significantly increasing the number of characteristics considered.

To avoid the implicit dependency of the Testing procedure on the construction method used by the authors for pattern creation, the procedure is also applied to similar patterns from a publicly available CM/M2M product study (Harwood et al., 2020), specifically the basic skirt pattern and the basic sweatshirt pattern (including two sleeve variations) provided with this solution. For this purpose, the .dxf files generated by this system were first converted into .svg files; the details of this process can be found in (Neimanis et al., 2025) – refer dxftosvgconverter.zip.

All the aforementioned examples used in this study are base constructions (patterns that serve as base to develop more complex garments) consisting only of one or few pattern pieces. They were selected with the goal to derive the Testing procedure but not to overload it. The practical application of the Testing procedure in real-life CM/M2M product development is foreseen in upcoming research.

2.2. Input Data used for the study

In the development of CM/M2M products, body measurements pertaining to the person to be clothed serve as input data. Unfortunately, CM/M2M products lack a standardized system for determining body measurements — each manufacturer uses its own, distinct set of measurements and interpretations (Januszkiewicz, 2021).

However, we can reasonably assume that each CM/M2M product manufacturer has access to their own set of body measurement data and is aware of the quality of their test data. Furthermore, it can be assumed that they are able to obtain measurement data from new individuals involved in the process (for example, new clients purchasing patterns or ordering custom garments online).

In another study (Bicevskis et al., 2024), the authors of this study analyzed the possibilities to extract test data from CM/M2M product development data. As result of that study a test database with 469 female body profiles was accessible, each profile having 37 measurements. As discussed in details of that study the data covers a wide range of female body types, but the quality of that data is not homogenous. In Section 2.4, we will demonstrate what impact low-quality input data can have on the Testing procedure described here.

In order to use the second CM/M2M pattern-generating system (Harwood et al., 2020) discussed above (refer Section 2.1), some of the measurement data had to been transferred matching the definitions of that second system.

2.3. Determining the Quality of Patterns

Determining the quality of patterns as a product is quite complicated at its core, as this intermediate product's quality can ultimately only be assessed by a person after trying on the clothing made from the pattern (a person visually confirms whether the garment fits the specific individual or not). Therefore, the quality of the pattern can only be partially determined by quantitatively expressible and measurable indicators, and subjective factors play a significant role.

Garment fitting has been extensively studied, analyzing both its subjective factors (Fan et al., 2004), (Hernández, 2018), (Zhang et al., 2011), (Brownbridge et al., 2013) and attempting to create objectively verifiable criteria (Sayem et al., 2017). Although there have been initial steps toward the digitalization of fitting (WEB (a) (WEB (b)), (WEB (c)), there is currently no convincing alternative to the iterative sew/fit approach (Keefe et al., 2017), (Lagė et al., 2020), which is a significant barrier to the development of CM/M2M products (new clothing models) and a crucial cost factor in the overall development of such clothing design products.

The decision that a garment is "sufficiently well-fitting" is made during a fitting process conducted by a person, often under variable conditions (such as the fabric used, lighting, available time, etc.). Even if the "sufficiently well-fitting" decision is made by a qualified specialist, it will never be entirely free of subjectivity. Given this subjectivity in fitting practices, there has yet to be a universally adopted, quantitatively expressible, and practically measurable indicator that could serve as the basis for making a "sufficiently well-fitting" decision.

Attempts have been made to overcome subjective factors and evaluate quality with quantitative indicators (De Silva et al., 2024), (Wang et al., 2006), for example, by determining the pressure exerted by the garment or measuring the space between the body and the garment. However, the practical application of these indicators is challenging — their assessment cannot be conducted without additional labor and/or technical equipment.

In the absence of digital alternatives, it is advisable to apply the manual, costly tailoring/fitting process only after automatically detectable issues have been resolved. In the following, we will demonstrate how it is possible to identify potential construction deficiencies in the CM/M2M product using purely computer-based methods by analyzing the properties of the generated patterns.

2.4. Fault Diagnosis on Piece Level

To identify potential design errors based on the created patterns, it is first necessary to express the dissimilarity of two patterns generated for two individuals.

Since, in a computer-aided process, we are naturally unable to apply the subjective criteria discussed in Section 2.3, we require quantitatively expressible comparison metrics.

Viewing patterns as sets of two-dimensional images makes comparing them a complex task. It is immediately clear that this comparison problem can be reduced to examining each pattern piece individually.

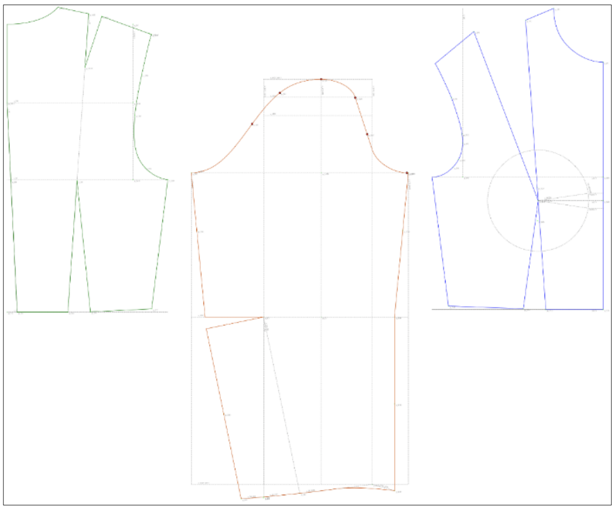


Figure 1. Samples of pattern pieces and their outer contours

From a purely mathematical perspective, there exist numerous methods to compare two graphics in a two-dimensional space; however, we will limit ourselves to those that are both professionally justifiable and practically computable.

Each piece, in the context that interests us in terms of its relevance and applicability from a professional perspective, consists of a line describing its outer contour, which represents the boundary of the fabric to be cut. The other lines and symbols visible in the pattern image merely provide supplementary information for the pattern user and are irrelevant in this context.

Defining the notion of "identical" in terms of pattern use as an abstraction for cut fabric is straightforward: two patterns are identical if the shape and size of their outer contour lines are the same.

Note: By default, we assume that the orientation of comparable pieces on the fabric (i.e., the grainline direction) will always be the same across different input data sets, meaning that comparable pieces will not be rotated.

However, if we want to determine how "different" two pieces are, we quickly realize that quantifying their difference with a single number is not feasible. Even a simple piece like a rectangular belt demonstrates this; the shape and size of the belt can only be described using at least two numbers (e.g., length and width). This simple rectangular example highlights that it's impossible to define a single universally applicable value to characterize the difference between two pattern pieces. Therefore, we will examine differences based on various indicators, guided by both the application context and the practical consideration that these indicators must also be computable.

When searching for parameters that can characterize a piece, we intuitively consider its Perimeter (the length of the outer contour) and the Area (surface) it encompasses. These two parameters are advantageous, both because they are clearly suitable in this context (as the size of a person increases, more fabric is required, leading to larger Perimeters and Areas) and because they can be easily obtained — either directly from the CM/M2M pattern-generating system or, in any case, from vector data found in the file representing the pattern.

This leads to the following research question:

Research question 1: How and under which requirements potential errors in the construction of a CM/M2M pattern can be identified using quantitatively measurable properties of pattern pieces?

To initially focus on the approach, we chose a pattern-making program of medium complexity (base pattern for skirts) as the first application example, whose quality is good based on practical experience. The same data, which was already used in the validation of input data (refer section 2.2), was used as input data here. Thus, patterns were generated for 469 datasets of different quality, and the length of the Perimeter and the Area for the two pieces FRONT and BACK were calculated.

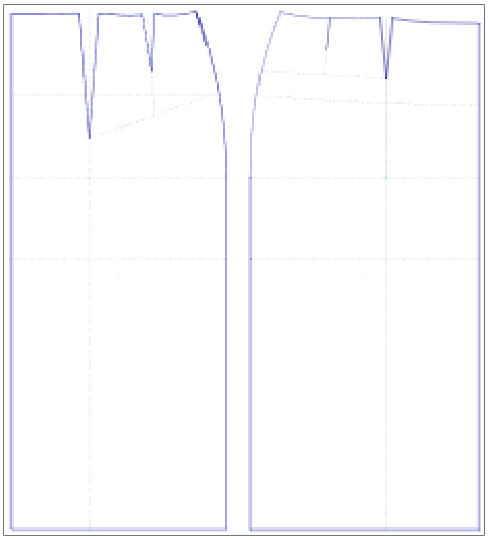


Figure 2. Base skirt pattern sample

The results were analyzed using the statistical programming language R (R Core Team, 2024). It was calculated how much the Perimeter and Area of each piece correlate with the measurements used to generate them. The Area showed a higher correlation than the Perimeter. The Pearson correlation coefficient (R Core Team, 2024) for the BACK piece Area was 0.955, while for the FRONT piece it was 0.956. Lower results were found for the correlation of measurements with the Perimeter of the pieces: FRONT - 0.614, BACK – 0.865.

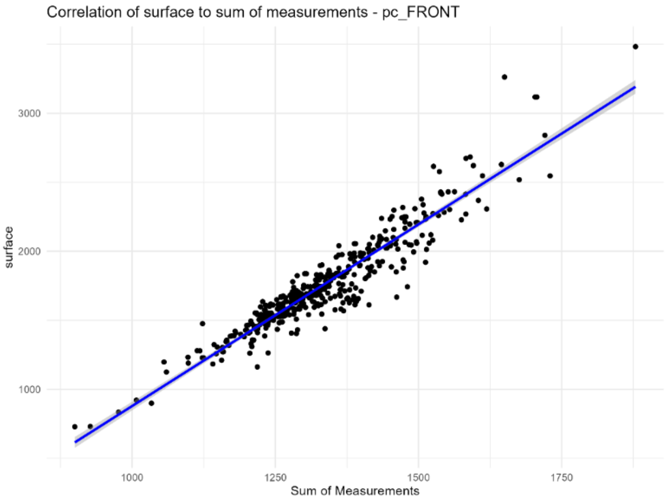


Figure 3. Correlation of the Area of the skirt's piece FRONT with the measurements

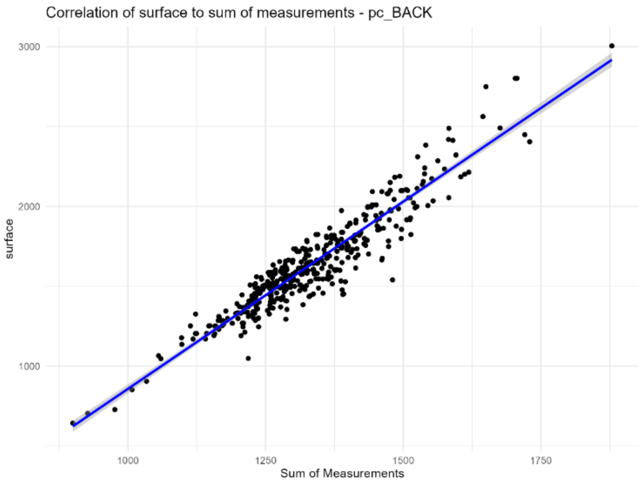


Figure 4. Correlation of the Area of the skirt's piece BACK with the measurements

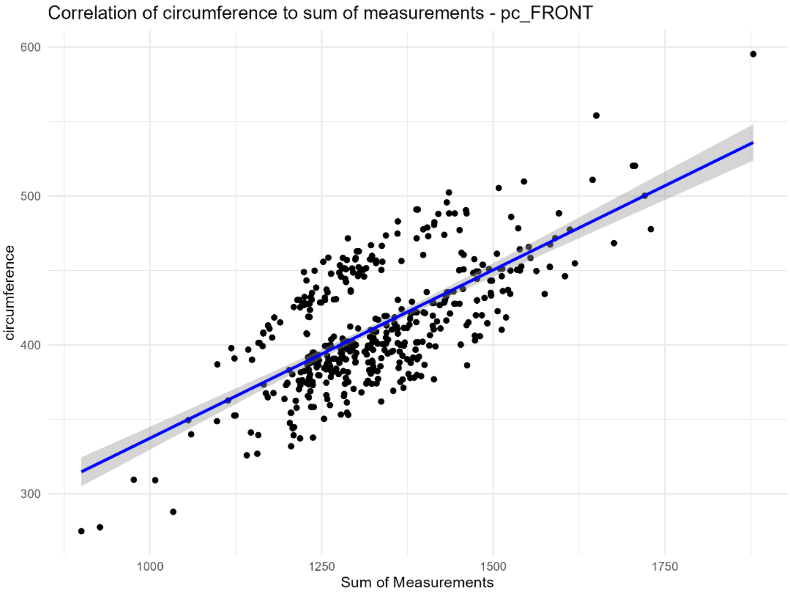


Figure 5. Correlation of the Perimeter of the skirt's piece FRONT with the measurements

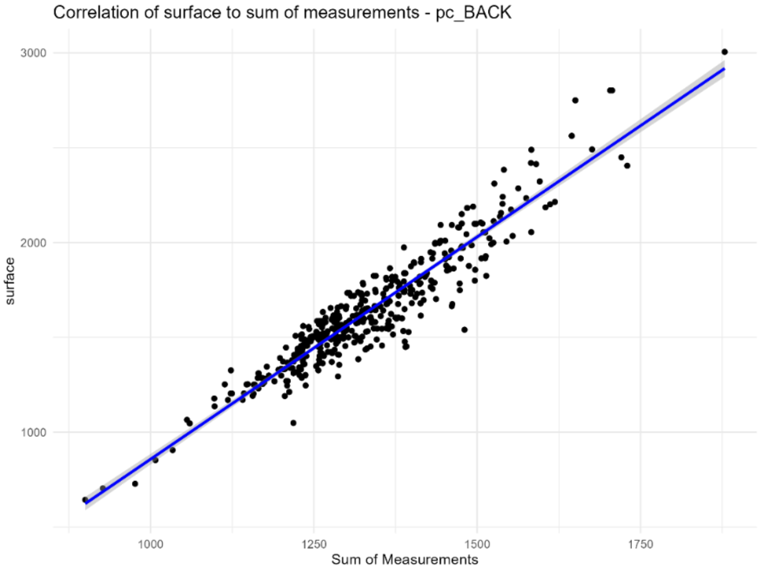
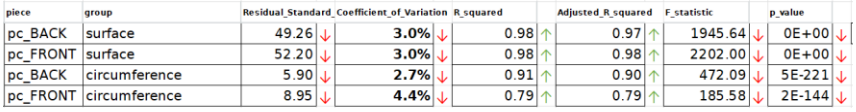


Figure 6. Correlation of the Perimeter of the skirt's piece BACK with the measurements

A statistical model was repeatedly created and trained for each piece and dependent variable (Perimeter or Area).

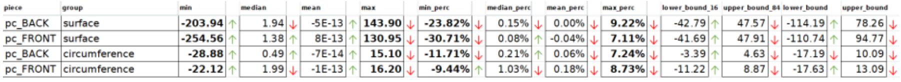
Initially using a linear regression approach (the lm() function (R Core Team, 2024)), followed by use of predictive models with resampling and hyperparameter tuning (the train() function (Kuhn, 2008)) the predicted results were compared with the actual ones and iteratively improved.

The summary indicators of each model (Residual Standard Error, Coefficient of Variation, R², Adjusted R², F-statistic, p-value) were also recorded, and the maximum, minimum, and average deviations from the regression line were calculated.



piece	group	Residual_Standard_	Coefficient_of_Variation	R_squared	Adjusted_R_squared	F_statistic	p_value
pc_BACK	surface	49.26 ↓	**3.0%** ↓	0.98 ↑	0.97 ↑	1945.64 ↓	0E+00 ↓
pc_FRONT	surface	52.20 ↓	**3.0%** ↓	0.98 ↑	0.98 ↑	2202.00 ↓	0E+00 ↓
pc_BACK	circumference	5.90 ↓	**2.7%** ↓	0.91 ↑	0.90 ↑	472.09 ↓	5E-221 ↓
pc_FRONT	circumference	8.95 ↓	**4.4%** ↓	0.79 ↑	0.79 ↑	185.58 ↓	2E-144 ↓

Figure 7. Main statistical indicators of the Area and Perimeter regression models for the FRONT and BACK pieces



piece	group	min	median	mean	max	min_perc	median_perc	mean_perc	max_perc	lower_bound_16	upper_bound_84	lower_bound	upper_bound
pc_BACK	surface	**-203.94** ↑	1.94 ↓	-5E-13 ↑	**143.90** ↓	**-23.82%** ↓	0.15% ↓	0.00% ↓	**9.22%** ↓	-42.79 ↑	47.57 ↓	-114.19 ↑	78.26 ↓
pc_FRONT	surface	**-254.56** ↑	1.38 ↑	8E-13 ↑	**130.95** ↓	**-30.71%** ↓	0.08% ↑	-0.04% ↓	**7.11%** ↓	-41.69 ↑	47.91 ↓	-110.74 ↑	94.77 ↓
pc_BACK	circumference	**-28.88** ↑	0.49 ↑	-7E-14 ↑	**15.10** ↓	**-11.71%** ↓	0.21% ↑	0.06% ↓	**7.24%** ↓	-3.39 ↑	4.63 ↓	-17.19 ↓	10.09 ↓
pc_FRONT	circumference	**-22.12** ↑	1.99 ↓	-1E-13 ↑	**16.20** ↓	**-9.44%** ↑	1.03% ↓	0.18% ↓	**8.73%** ↓	-11.22 ↑	8.87 ↓	-17.63 ↑	13.09 ↓

Figure 8. Deviation indicators from Area and Perimeter regression lines for the FRONT and BACK pieces

The finally reached indicators are visible in the above table *(Remark: the red and green arrow icons indicate the model's performance compared to the previous iteration)*. The according R script is available in (Neimanis et al., 2025) – refer R scripts.zip.

Various approaches to identifying outliers were tested:

(i) **Residuals** (WEB (d)) – differences between predicted and observed values. The highest 2% were marked as outliers. The results were visualized with a **Q-Q plot**(Wilk and Gnanadesikan, 1968), and points (profiles) far from the blue line can be considered outliers.

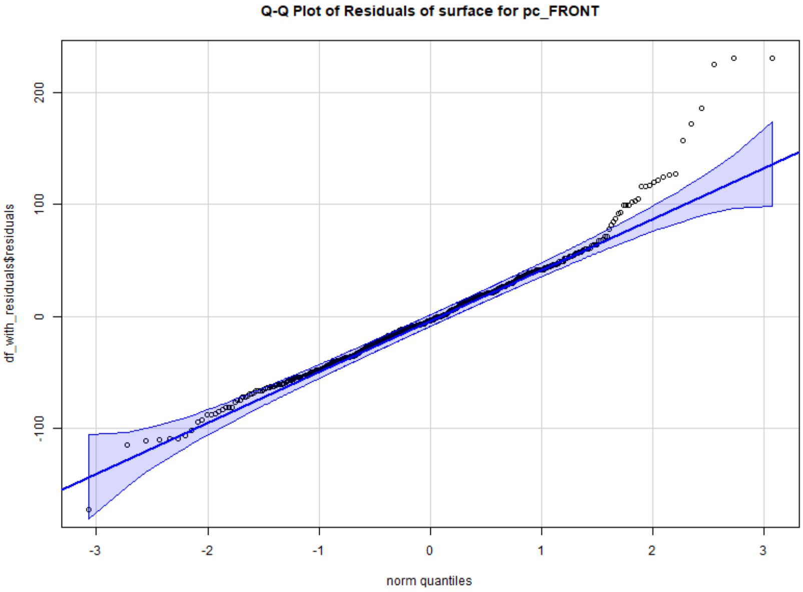


Figure 9. Visualization of the residuals for the Area of FRONT piece with a Q-Q plot

(ii) **Cook's distance**(Cook, 1977) - helps to identify data points that could significantly influence the results of the regression model. Outliers are determined using the formula 4 / (n - k - 1), where the number 4 is used as an empirical constant often applied to identify influential data points in a regression model; n is the number of observations in the dataset; k is the number of independent variables in the model (number of regression coefficients excluding the intercept). In this case, data points whose Cook's distance exceeds the calculated threshold value are considered influential. In figure 10, the threshold is illustrated with a red dashed line. There are quite a few outliers but there exist some distinctly influential data points between them.

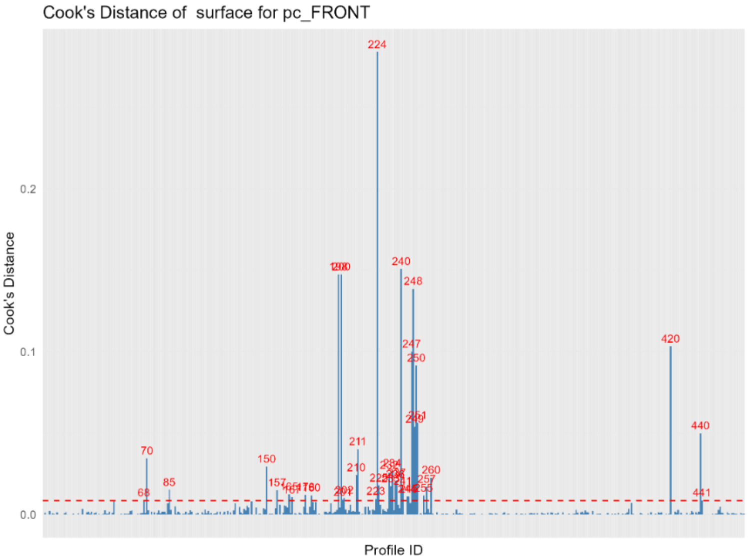


Figure 10. Visualization of Cook's distance for the Area of FRONT piece

(iii) **Mahalanobis distance** (Mahalanobis, 1936) - used to identify outliers in multidimensional data. It is calculated in R using the mahalanobis() function (R Core Team, 2024) which uses data matrix, the mean vector, and the covariance matrix of the residuals. The highest 2% are marked as outliers.

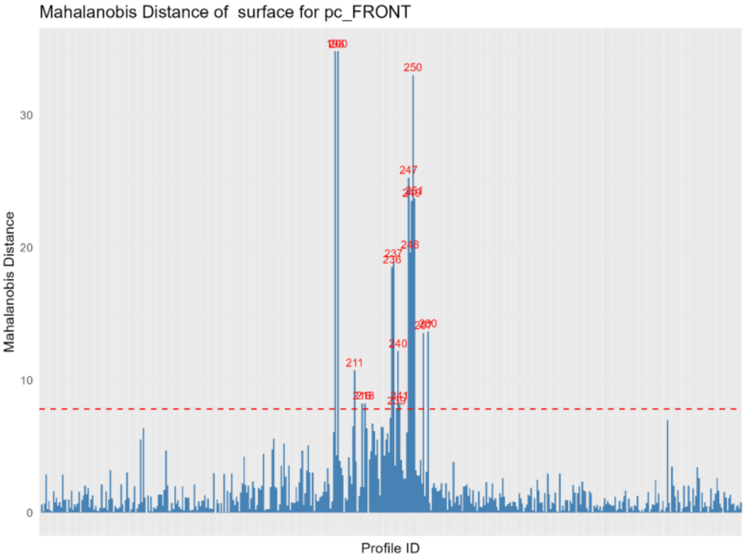


Figure 11. Visualization of Mahalanobis distance for the Area of FRONT piece

Outliers identified by Cook’s and Mahalanobis distances were also recorded in a scatter plot, which visualizes how consistent the data is for training and predicting with the linear regression model.

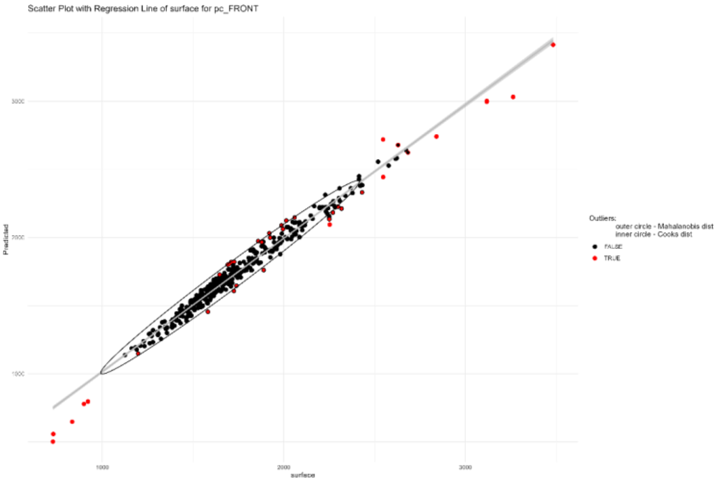


Figure 12. Scatter plot of Cook’s and Mahalanobis distances for the Area of FRONT piece

The output of all outlier detection methods were combined and summarized in EXCEL tables, which are available in (Neimanis et al., 2025) - refer Perimeter and Area attachments.zip.

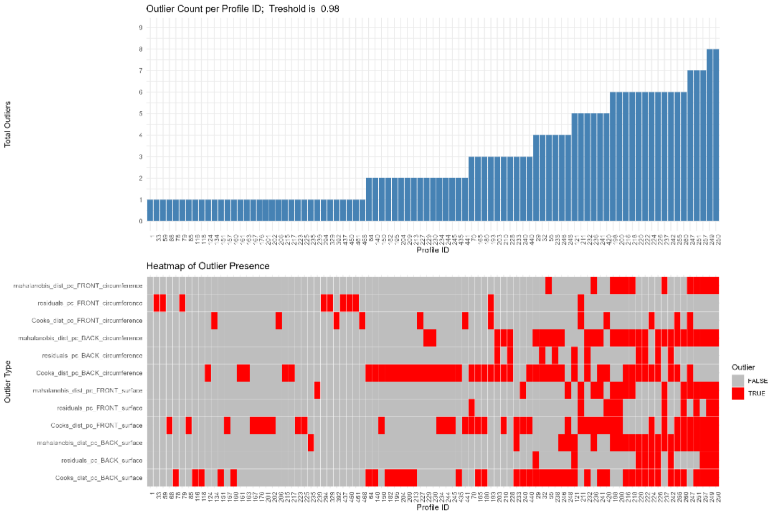


Figure 13. Identified profiles with the highest error probability for the Perimeter and Area of the FRONT piece

2.5. Fault Diagnosis on Line Level

The previously analyzed variables Perimeter and Area can only be used as indicators of large structural deviations, as neither of these quantities considers the possible shape differences of any two pieces. Therefore, in further analyses, we will focus on how to also consider shape.

Research question (2.5): How and under which requirements potential errors in the construction of a CM/M2M pattern can be identified using properties characterizing the pattern pieces’ shape?

To identify shape differences, one can leverage the fact that each CM/M2M piece is constructed using specific types of lines. The characteristics of these lines can then be analyzed to detect differences.

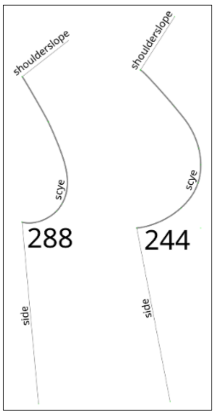


Figure 14. Sample set of comparable lines for patterns of profiles 288 and 244

From a purely mathematical perspective, there are many conceivable variants for defining the indicators used to compare individual lines. However, considering the aspect of applicability, the crucial question becomes how we interpret the location of the line within the common piece. Therefore, we determine that each line’s location is characterized relative to the mass center of the respective piece. By aligning the two comparable pieces at their mass centers (without changing their directions defined in the CM/M2M system), we are able to compare the pairs of lines forming the construction of these two pieces.

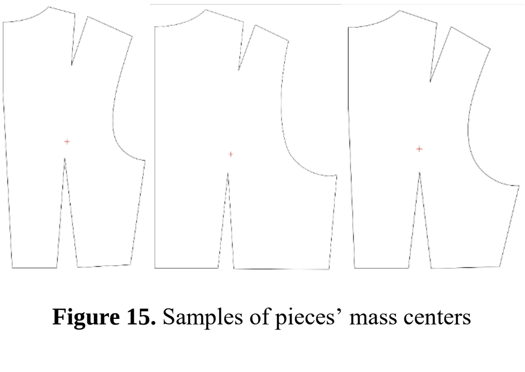


Figure 15. Samples of pieces’ mass centers

The following quantities can be used for comparison, all of which (including the required mass centers in this context) are derived from the CM/M2M system or vector data contained in the file representing the pattern:

i. length of the line;
ii. vector (represented as x/y-coordinates) of the starting point from the piece's mass center;
iii. vector (represented as x/y-coordinates) of the endpoint from the piece's mass center;
iv. angle of the line at its starting point;
v. angle of the line at its endpoint.

As an application example for quality analysis using the properties of lines, we use a base bodice construction. This construction, like the base skirt construction used above, has already proven itself in practice. Using the above-mentioned line parameters the quality of a garment construction can be further analyzed applying several statistical methods. To detect potential outliers in the line properties and find out which profile's patterns should be checked for quality control, a similar approach as described earlier in piece analysis (refer section 2.4) was used. First, data was scaled and centered. For each piece, several predictive models were trained using the lines as factors. Several models were applied and the results compared:

- Linear regression (R Core Team, 2024),
- Random forest (Lang et al., 2019), (Wright and Ziegler, 2017),
- GAM (Wood, 2011)
- rpart decision tree (Therneau and Atkinson, 2023).

Their results (using the metrics MAE, MAPE, and R2) were compared to choose the ones with the best performance in each category of predictors:

Model for Length –

type: linear regression model (R Core Team, 2024),

predictors: all measurements used to create the pattern; circumference and surface of the piece where the line is located.

Model for Coordinates –

type: random forest (Lang et al., 2019), (Wright and Ziegler, 2017),

predictors: all measurements used to create the pattern; length of the line.

Model for Angles –

type: rpart decision tree (Therneau and Atkinson, 2023),

predictors: all measurements used to create the pattern; length of the line; start and end coordinates of the line.

These models were used to make predictions on the same data they were trained on, allowing us to compare actual versus predicted values. High differences between these values are candidates for anomalies in the geometry of the lines.

Outliers were detected using Mahalanobis distance and residuals applying the same thresholds as in the piece level analysis (refer Section 2.4) - top 2% of residuals; Mahalanobis distance was assessed using chi-squared distribution (Kahle, 2017), (R Core Team, 2024) to determine if an observation was unusually far from the mean of a distribution.

If the distance or the residual were greater than the threshold the respective outlier was marked. All marked values were summed up for every single line and for every measurement profile. The goal was to detect profiles having the highest outlier count indicating that their measurement data combination has led to some unforeseen geometry.

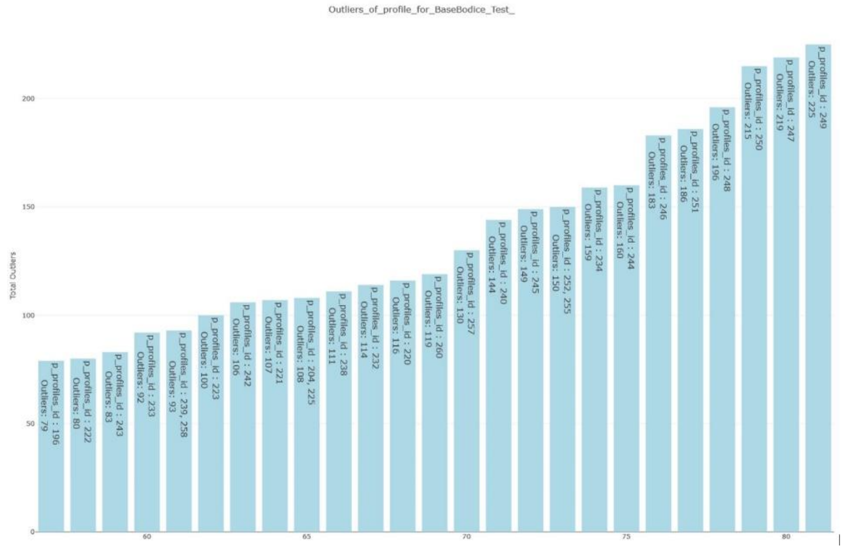


Figure 16. Profiles with highest count of outliers for base bodice pattern

To illustrate the application of such results we can take a closer look onto the profile with the id 249 having the highest number of outliers (225). In the following image of the base bodice pattern for that profile the lines having at least one outlier in one of their parameters are colored red. As a counterpart we added the pattern for the profile with id 254 which has quite similar measurements as profile with id 249 but way less outliers (64).

The comparison between the two patterns illustrates that a high outlier count helps to detect construction anomalies, e.g., the shape of the scye is obviously uneven for profile 249, but smooth for profile 254.

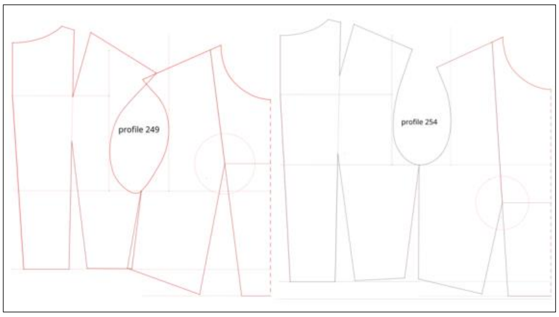


Figure 17. Base bodice patterns for two profiles.

Looking at the list of lines with the highest number of outlier profiles (refer Figure 18), the cl_ScBc line (back scye) appears at the top. This suggests that the scye in the base bodice construction may require improvement.

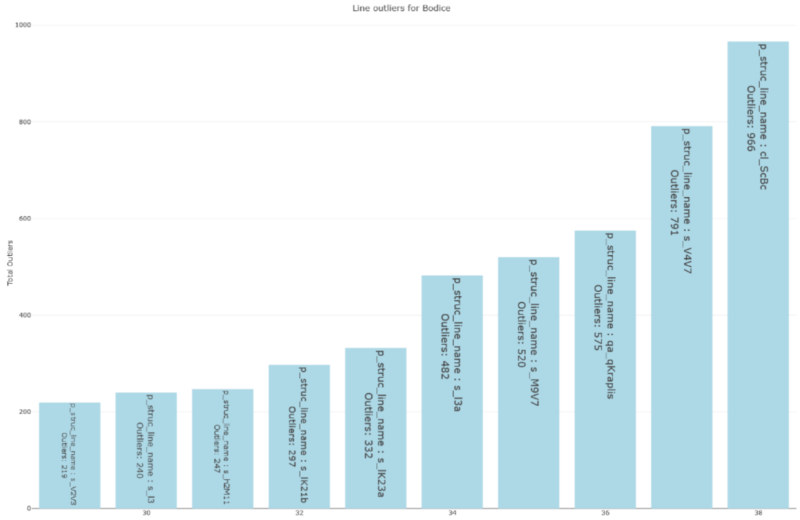


Figure 18. Lines with the highest number of outliers

But it should be considered that scye lines are long and curved. The shape of a scye largely depends on the pattern-making method used to create the bodice, as these methods employ different approaches to combining body measurements. The parameters used so far for the analysis (length, location and angles of start/end) cannot fully characterize the shape of such a line. The two samples from the executed test dataset illustrate that scye lines with nearly identical start/end-angles can have a substantially different form.

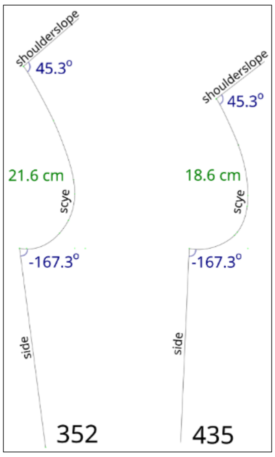


Figure 19. Samples of scye with similar start-/end-angle of profiles 352 and 435

2.6. Fault Diagnosis for .dxf patterns

The samples discussed thus far were created using a CM/M2M construction system that outputs result in a structured format. Every pattern produced by this system comprises the same set of lines arranged in a consistent sequence. Regardless of the profiles for which the program is executed, it generates patterns that are structurally identical. The only specifics are that a line might exist as a 1-point-line (a line having a fixed location but no length). Since that construction method's implementation uses the .svg standard, all these lines can be characterized by a fixed number of parameters, resulting in a structurally fixed set of output parameters The use of statistical methods within the Testing procedure is based on that fixed structure premises.

But not all CM/M2M-construction systems support the .svg standard. Many systems produce their results in other formats, and one of the most popular such standards is .dxf (WEB (e)).

Although .dxf as a standard allows to implement an analogous fixed structure approach, in real life another approach is more popular. That approach is to implement all the pattern's lines as straights connecting many closely placed vertexes as illustrated in the samples below (further on called Vertex-polygon-type).

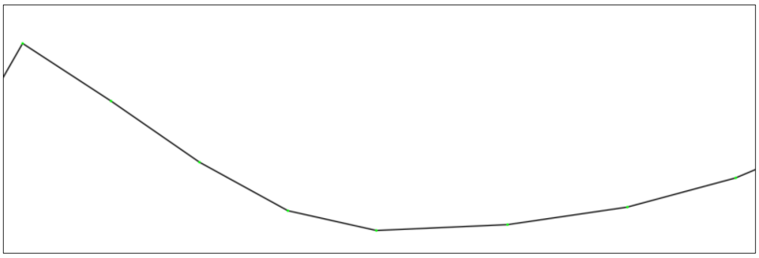


Figure 20. Sample of a Vertex-polygon-type line

In such kind of patterns implementation, the count of vertexes is not fixed but depends on the length and the curvature of the patterns' lines. Applying the statistical approach used so far is not possible since the number of output parameters is not fixed. However, the authors of this study applied a conversion approach (details to be found in (Neimanis et al., 2025) – refer dxftosvgconverter.zip) allowing to approximate the .dxf polygons with a combination of straights and Bezier curves. The resulting error (e.g., the maximum distance between the lines) is irrelevant to the subject of investigation.

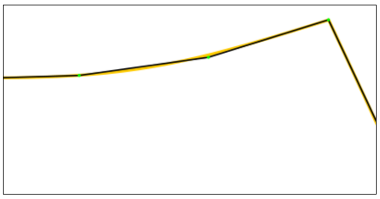


Figure 21. Conversion sample (green/black: vertex-polygon-type, orange: approximation with .svg)

The approach was applied on the output created by an open-source CM/M2M-system (Harwood et al., 2020) which provides the results in that Vertex/polygon-type format. That system is published with several executable pattern samples. We selected similar patterns as discussed above – a base skirt and a sweatshirt. By providing (the partly transferred – refer section 2.2) input data from the test data we created 469 .dxf patterns and applied the above-mentioned conversion to .svg creating the prerequisites to apply the Testing procedure.

Thanks to the Testing procedure, before reviewing any pattern we identified immediately 5 profiles for whom the pattern creation algorithm had created a senseless .dxf result, simply by comparing the number of created lines during the conversion.

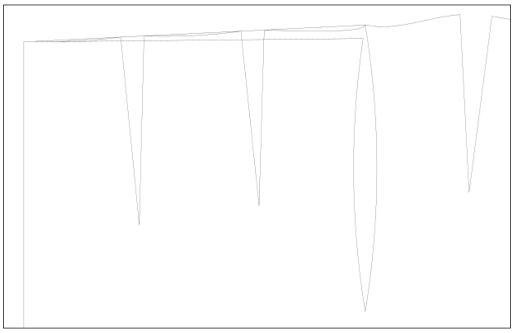


Figure 22. One of five identified senseless results for the skirt in .dxf format

For the remaining 464 profiles results in the same form as described in the previous sections were reached.

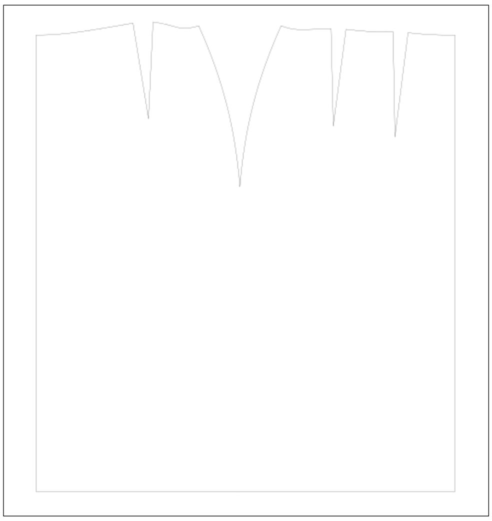


Figure 23. Skirt sample (converted to .svg format)

The analogue execution of the sweatshirt sample, consisting of front, back and two sleeve versions resulted in .svg-files with identical line structure, hence the provided outlier data could be used for detection of potential construction problems.

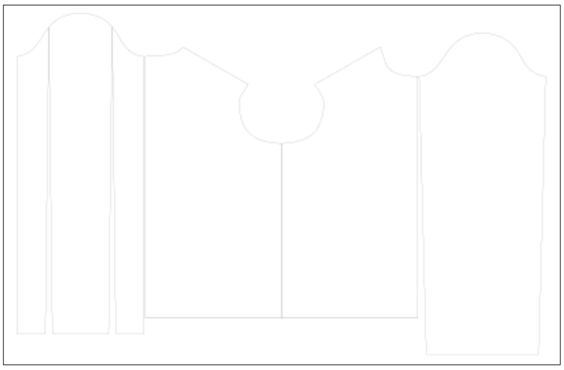


Figure 24. Sweatshirt sample (converted to .svg format)

3. Results

3.1. Piece Level

As discussed in Section 2.4, the correlation coefficients for the piece-level parameters Area and Perimeter differed significantly, even though practical experience suggests that the basic skirt pattern examined here should not exhibit anomalies in its construction.

To investigate the reason for these differences, we closely examined the pattern's algorithm for the waist darts. We found that the number, length, and placement of the darts are not defined continuously but rather in discrete steps. As a result, the prediction of the Perimeter using linear regression is less accurate compared to that of the Area, which is less influenced by the waist darts.

When investigating the reason for the inhomogeneous distribution of Cook's distance (refer Figure 10), we found that a small number of outliers, which are influential data points, also appear to have low credibility. This suggests that the identified outliers are more likely to indicate low-quality body measurements rather than construction errors. Overall, we identified 12 profiles with potentially incorrect measurements, which adversely affect the generated skirt base patterns.

Answer on research question 1: Potential errors in CM/M2M pattern construction can be identified by analyzing quantitatively measurable properties like the area and enclosed surface of pattern pieces' outer contour. Preconditions for a senseful use of the according statistical methods are on the one hand qualitative input data and on the other hand a consistent implementation of the pattern construction for the whole variety of input data. By using the measurements as input for predictive models (e.g., linear regression), and analyzing the deviations between observed and predicted values with Mahalanobis distance, Cook's distance, and residual analysis, the method can detect major outliers and irregularities, highlighting issues such as scaling errors or miscalculations in the pattern construction process.

3.2. Line Level

Similar to the skirt base pattern discussed in the context of potential piece-level failures, the bodice pattern used for line-level failure detection has been well-tested in practice. When investigating the reasons for the numerous outliers in profile 249, discussed at the end of Section 2.5, we found that this profile had already been classified as having "low" credibility in the study presented in Section 2.2. Similar to the piece-level analysis, the line-level analysis is more likely to indicate low-quality input data rather than actual construction errors.

Answer to Research Question 2: Under the same preconditions outlined in the answer to Research Question 1, potential errors in CM/M2M pattern construction can be identified by analyzing the properties of the lines forming the outer contour of pattern pieces. By using body measurements as input for predictive models (e.g., linear regression) and analyzing the deviations between observed and predicted values through residual analysis, this method can detect major outliers and irregularities. It can highlight issues such as scaling errors or miscalculations in the pattern construction process.

4. Conclusions

4.1. Summary

Our investigations indicate that statistically based methods for quality assurance of CM/M2M patterns have great potential for avoiding unnecessary costs in the product development of CM/M2M products. By using the Testing procedure described above, it is possible to determine, even during the development phase, which tested profile’s patterns are not fitting into the remaining test data result spectra. Targeted identification of risk candidates accelerates the testing process because, when CM/M2M patterns are programmed, it is important to consider as many possible and realistic measurement combinations as possible. However, due to human factors and time constraints, certain specific cases may remain unprocessed. The goal is to minimize such cases, and outlier analysis serves as a tool to achieve this. By identifying standout profiles, pattern programmers can visually inspect these specific patterns for potential defects without having to review all generated patterns.

The most significant benefit of the Testing procedure, however, lies in uncovering technologically detectable errors before the CM/M2M product is validated through physical tailoring and fittings.

4.2. Usage Hints

The results discussed in Sections 3.1 and 3.2 show that the Testing Procedure loses significant value if the measurement data used as input is of low quality. The general statistical principle "Garbage In, Garbage Out" fully applies to this procedure as well. Therefore, the Testing Procedure requires high-quality body measurement data that adequately covers the entire range of the target buyer group for the tested pattern.

By nature, the Testing Procedure is only applicable if the line structure of the tested pattern's pieces remains identical across all applied body measurement datasets. Furthermore, the poor Perimeter results discussed in Section 3.1 indicate that not every quantitatively measurable property of pattern pieces is inherently suitable for a statistically based testing procedure. To use parameters such as Perimeter, one must either (i) Implement an appropriate segmentation of the test data, or (ii) remove the parameter from the test altogether.

The Testing Procedure delivers meaningful results only for the tested items. The discussion regarding the scye line (see end of Section 2.5 and Section 3.2) suggests that potential issues cannot always be identified by the specific set of tested parameters used in this approach. For highly complex lines that depend on multiple body measurements, it is advisable to split them into equidistant segments before applying the Testing Procedure.

4.3. Practical Application

This publication presents only the results of applying the developed method to base patterns. The authors have begun applying the procedure to much more complex use cases in ongoing CM/M2M product developments. A major challenge will be to avoid being overwhelmed by a flood of data, as the number of lines in a typical CM/M2M product is in the high double digits, and the number of analyzable parameters is in the triple digits. The task is therefore to derive a practical and effective application methodology for the procedure.

Furthermore, in addition to the previously considered .svg and .dxf variants, it is our intention to support additional output formats to expand the range of supported garment construction systems.

Acknowledgments

This work has been conducted within the research project "Competence Centre of Information and Communication Technologies" of The Recovery and Resilience Facility, contract No. 5.1.1.2.i.0/1/22/A/CFLA/008 signed between IT Competence Centre and Central Finance and Contracting Agency, Research No. 1.7 “Technology for quality assessment of graphical objects described using polygons"

References

Bicevskis, J., Bicevska, Z., Diebelis, E., Purina L. (2024). Quality Control of Body Measurement Data Using Linear Regression Methods, *Annals of Computer Science and Intelligence Systems, Open Access, Issue 2024,* pp. 289 – 300, 19th Conference on Computer Science and Intelligence Systems, FedCSIS 2024, Belgrade, 8 - 11 September. Available at https://doi.org/10.15439/2024F6463

Brownbridge K., Gill S., Ashdown S. (2013). Effectiveness of 3D Scanning in Establishing Sideseam Placement for Pattern Design, in *Proc. of 4th Int. Conf. on 3D Body Scanning Technologies*, Long Beach CA, USA, pp. 41-49. Available at http://dx.doi.org/10.15221/13.041.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**(1), 15–18. DOI: 10.1080/00401706.1977.10489493.

De Silva R. K. J., Navodhya P. U., Gill S. (2024). Analysis of body-to-pattern relationship using traditional pattern drafting techniques: implications for automated digital prototyping, *International Journal of Fashion Design, Technology and Education*, **17:1**, 37-49, available at https://doi.org/10.1080/17543266.2023.2230270.

Fan J., Yu W., Hunter L. (2004). Clothing appearance and fit: Science and technology, *Woodhead Publishing Limited and CRC Press LLC*, p. 31-42, available at http://182.160.97.198:8080/xmlui/bitstream/handle/123456789/1397/Introductory%20pages.pdf?sequence=1&isAllowed=y.

Hernández, N. (2018). *Does it really fit? : improve, find and evaluate garment fit*. PhD dissertation, University of Borås, Faculty of Textiles, Engineering and Business. Available at https://urn.kb.se/resolve?urn=urn:nbn:se:hb:diva-14321.

Harwood A., Gill J., Gill S. (2020). JBlockCreator: An open source, pattern drafting framework to facilitate the automated manufacture of made-to-measure, *SoftwareX Volume 11, January–June 2020, 100365*. Available at https://www.sciencedirect.com/science/article/pii/S2352711018302528

Januszkiewicz, M. (2021). The Human Factors of 3D Body Scanning (3DBS) *Data Presentation and Service Interaction*, available at https://pure.manchester.ac.uk/ws/portalfiles/portal/194692958/FULL_TEXT.PDF.

Kahle D. (2017). *The Chi Distribution.* Available at https://cran.r-project.org/web/packages/chi/chi.pdf.

Keefe A., Kuang J., Daanen H. (2017). NATO Research Task Group: 3D Scanning for Clothing Fit and Logistics, in *Proc. of 3DBODY.TECH 2017 - 8th Int. Conf. and Exh. on 3D Body Scanning and Processing Technologies*, Montreal QC, Canada, 11-12 Oct. 2017, pp. 201-209, available at http://dx.doi.org/10.15221/17.201.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, **28**(5), 1–26. Available at https://doi.org/10.18637/jss.v028.i05.

Lagė A., Ancutienė K., Pukienė R., Lapkovska E., Dāboliņa I. (2020). Comparative Study of Real and Virtual Garments Appearance and Distance Ease, in *Vol. 26 No. 2 (2020): Materials Science*, available at https://doi.org/10.5755/j01.ms.26.2.22162.

Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, Au Q, Casalicchio G, Kotthoff L, Bischl B (2019). mlr3: A modern object-oriented_machine learning framework in R. _*Journal of Open Source Software*_. Available at https://doi.org/10.21105/joss.01903.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, **2**(1), 49–55. Available at http://bayes.acs.unt.edu:8083/BayesContent/class/Jon/MiscDocs/1936_Mahalanobis.pdf.

Neimanis, A., Purina, L., Bicevska, Z., Diebelis, E., Bicevskis, J. (2025). Supplementary Material for the Publication "Statistics-based Quality Control of Patterns". *Zenodo*. Available at https://zenodo.org/records/14607759

McKee, M. (2024). *Buying a suit: Know the difference between made to measure, custom, and bespoke*, available at https://www.themanual.com/fashion/made-to-measure-vs-bespoke-vs-custom-suits/.

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.

Sayem M., Sadat A., Bednall A. (2017). A Novel Approach to Fit Analysis of Virtual Fashion Clothing. In: 19th edition of the *International Foundation of Fashion Technology Institutes* conference (iffti 2017), 28 March 2017 - 30 March 2017, The Amsterdam Fashion Institute (AMFI), Amsterdam. Available at https://e-space.mmu.ac.uk/618256/3/Revised_Full%20Paper_%20IFFTI%202017_Fashion%20disruptive%20technology_Abu%20Sayem_MFI%20(1).pdf.

Sneha, M. (2024). *Custom Clothing Market Report 2024 (Global Edition),* available at https://www.cognitivemarketresearch.com/custom-clothing-market-report.

Therneau T, Atkinson B (2023). *_rpart: Recursive Partitioning and Regression Trees*. R package version 4.1.23, available at https://CRAN.R-project.org/package=rpart.

Wang, Z., Newton, E., Ng, R., Zhang, W. (2006). Ease distribution in relation to the X-line style jacket. Part 1: Development of a mathematical model. *The Journal of The Textile Institute*, *97*(3), 247–256. Available at https://doi.org/10.1533/joti.2005.0239.

WEB (a). *Fashion design software*, https://tailornova.com/.

WEB (b). *Pattern Design Software*, https://www.lectra.com/en/fashionWEB

WEB (c). *Pattern Design Software*, https://optitex.com/products/2d-and-3d-cad-software/.

WEB (d). *Errors and residuals* https://en.wikipedia.org/wiki/Errors_and_residuals.

WEB (e). *AutoCAD DXF*, https://en.wikipedia.org/wiki/AutoCAD_DXF.

Wilk, M. B., Gnanadesikan, R. (1968). Probability plotting methods for the analysis for data. *Biometrika*, **55**(1), 1–17. Available at https://doi.org/10.1093/biomet/55.1.1

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society* Volume 73, Issue 1, 3-36, January 201, available at https://doi.org/10.1111/j.1467-9868.2010.00749.x.

Wright, M. N., Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, **77**(1), 1–17._Available at https://doi.org/10.18637/jss.v077.i01.

Zhang, L., Zhang, W., Xiao, H. (2011). Subjective Assessment of Women's Pants' Fit Analysis Using Live and 3D Models. In: Chen, R. (eds) *Intelligent Computing and Information Science. ICICIS 2011*. Communications in Computer and Information Science, vol 135. Springer, Berlin, Heidelberg. Available at https://doi.org/10.1007/978-3-642-18134-4_109.

Received January 7, 2025, revised March 10, 2025, accepted April 7, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 2, 389-406
https://doi.org/10.22364/bjmc.2025.13.2.05

An Innovative Data Hiding Scheme for Color Images

Kanza FATIMA[1], Nan-I WU[2], Chi-Shiang CHAN[3],
Min-Shiang HWANG[1,4,*]

[1]Department of Computer Science & Information Engineering,Asia University,
Taichung 41354, Taiwan
[2]Department of Information Management, Lunghwa University of Science and
Technology,Taoyuan 33306, Taiwan.
[3]Department of M-Commerce and Multimedia Applications, Asia University,
Taichung 41354, Taiwan
[4]Department of Medical Research, China Medical University Hospital, China Medical
University,Taichung 41404, Taiwan

fatimakanza@outlook.com,naiwu@gm.lhu.edu.tw,CSChan@asia.edu.tw,
mshwang@asia.edu.tw

ORCID 0009-0008-8738-6458, ORCID 0009-0004-7758-5200, ORCID 0000-0001-7324-8320,
ORCID 0000-0001-5502-8033

*Corresponding author

Abstract. Steganography is the method of obfuscating the fact that a message is being delivered by enclosing it within non-secret cover media, such as pictures, audio videos, or text files. Steganography seeks to make the sheer existence of communication challenging to discover, in contrast to cryptography, which scrambles the contents of a message to make it unreadable. This research presents a model technique to hide and retrieve message data within RGB images without losing the original data. The mapping technique is used to conceal the secret message by using PYTHON. A message of different sizes is generated using a random number generator, and different image resolutions are used as cover images, such as 256x256, 512x512, and 1024x1024. The results show the best PSNR, which is 48.90, 56.31, and 62.45 dB, respectively. The MSE and SSIM are 0.815, 0.152, 0.036, and 0.998, 0.999, 0.999 respectively.

Keywords: Human Visual System, Data hiding, Information security, Image Steganography

1. Introduction

Internet has become the most practical and effective form of communication. The challenges of protecting people’s privacy get increasingly complex as electronic transmission technology advances and increases in computer power and storage. Due to advancements in cyber intelligence and an enormous increase in social media applications availability in handheld devices majorly contributed to the exponential increase in data transmission rates, data protection has recently received more attention (Hwang et al., 2024; Song et al., 2025). Data security involves preventing illegal users from accessing, altering, sharing, or simply viewing data and enabling only certified users to do so. Secure transmission is crucial in the modern world when there are thousands of daily malware attacks (Li and Li, 2023; Yang, 2023).

Due to the high dependency on electronic media for information exchange and sending different digital messages, information must be delivered to its intended recipients without allowing illegal individuals to view, alter, and delete their contents. Electronic information security is the top priority of the governments at that time. Secure data and copyright transfers are demanded by official users, hence many methods have been developed to meet this need. Although steganography, Cryptography, and watermarking operate private data in distinct ways, they all aim to send data safely (Bawaneh et al., 2021). Before the invention of electronic images, secret communications were conducted using manuscript letters referring to a list of numbers that may refer to anything else, such as a book or a set of street numbers (Mahdi, 2021). The security of sensitive data is under the purview of numerous security technology fields; the two most significant are cryptography and steganography. Cryptography, which deals with various data encryption techniques, is one of the most fascinating subfields in the data analysis discipline. It aims to transfer confidential data into an encrypted form using various encryption algorithms (Kadhim et al., 2019). In cryptography, data is not sent directly from the sender to the recipient; instead, data is first converted from plain text into coded text using an encryption algorithm and then sent coded text is to the recipient, and the recipient decrypts the coded text into plain text to read the real data. However, keeping the message's contents private may not be successful. It uses two types of keys such as asymmetric and symmetric keys (Kilic and Evrensevdi, 2022). In contrast, the second technique is steganography. Steganography is the procedure of hiding information in plain sight, whether it be in an image, video, audio, or file (Gupta and Bhagat, 2019). The fundamental goal of a digital steganographic procedure is to encode personal or confidential information inside cover media covertly (Kilic and Evrensevdi, 2022). The data hidden in the file is known as the stego file, and the original file is called the cover image file.

While providing secure end-to-end information transmission is the same goal as cryptography and steganography, their methods differ (Kilic and Evrensevdi, 2022). This research focuses on embedding more data using the color. Wavelength property and protecting data security using bit matching and substitution steganography techniques. Steganography is the art of writing a secret message without knowing the third person other than the sender and receiver. The word "Steganography" comes from the Greek term. The word "Steganography" is divided into two words: "Stegano" which means concealed, and "graphy" which means represented. Although there are many definitions of steganography, the concept is hiding secret information inside the cover media; the secret information may be in binary bits, images, audio, or text data (Kadhim et al., 2019). The main objective of steganography is to send confidential information without the knowledge of attackers. Attackers were not aware that the information was hidden in the sender file. However, if doubt is increased, the goal of steganography is all in vain. There are four main elements of steganography.

- **Cover message:** The cover message is any medium in which data is hidden. It is a media file, i.e. (image, audio, video, or text).
- **Secret message:** The secret message is data about to be hidden.

- **Embedding algorithm:** The procedure through which data is embedded in the cover media.
- **Extracting algorithm:** The procedure for extracting data from the cover media.

In the steganography framework, before hiding the secret message, the sender must select the proper medium for the data, such as an image, audio, video, or text document. A proper embedding and extracting algorithm must be used to embed and extract the data with less chance of damage. The steganography structure is shown in Figure 1.

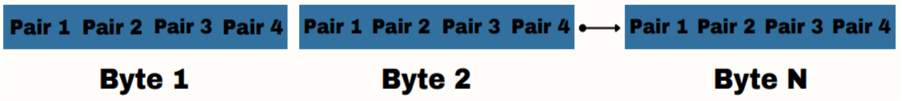

Pair 1 Pair 2 Pair 3 Pair 4
Pair 1 Pair 2 Pair 3 Pair 4
Pair 1 Pair 2 Pair 3 Pair 4
Byte 1
Byte 2
Byte N

Figure 1. Steganographic Structure.

2. Literature review

This section depicts a few past examinations that utilized the LSB procedure to work on the security of the installed secret message or to work on the limit of the cover picture. In paper (Dumre and Dave, 2021) researcher proposed image steganography by altering the order of RGB planes while embedding the concealed message by combining the least significant bits with AES-128 encryption for data protection. Another study of (Jiao and Feng, 2021) used distinctive try to merge information technology with optical illusions by using a greyscale picture background integrated with color image pixels for visual appearance in image steganography. Spatial domain steganography is applied by using three criteria (the size of the Gaussian filter, a low threshold, and a high threshold) specified for canny edge detection to hide concealed messages into the LSB of the blue color channel of the host image (Almazaydeh, 2020).

Many steganographic strategies have been proposed. To bury the image into another picture is another famous method of image steganography. For data encryption, a K-LSB method is proposed in (Elharrouss et al., 2020). Edge detection operation is used on the decoding time to know the hidden image blocks. An image quality enhancement algorithm is used to boost the quality of an image after the extraction process. There are different types of image formats which are used for image steganography.

The study of comparison between different file types based on their performance is presented in (Ansari et al., 2019), which shows the BITMAP images as the best PSNR value and high capacity requirement.

A variety of methods has been introduced for image steganography. These are used according to the requirement of security level. The authors Darwis and Pamungkas, (2021) presented the comparative study of different types of image steganography methods, such as Least significant bit, Picture value difference, and modulus function, and their results showed that LSB is considered more suitable as compared to others in the sense of image quality. For better embedding capacity PVD algorithms are best. Authors Sally and Maisa (2021) introduced a new technique by integrating the least and most significant bits. LSB values have been replaced with concealed message bits based on MSB values. The result of this proposed approach has no change in the stego image. The secret key is executed using a point curve between sender and receiver (Mahdi, 2021). The challenges of image distortion and inadequate transmission security within image information-hiding techniques are tackled through the introduction of several novel algorithms. To mitigate image distortion, particularly in gray-scale images, an adaptive enhancement algorithm is proposed. Furthermore, a reversible information-hiding approach is developed by Li (2024), leveraging a fixed tone plane for data embedding, ensuring reversibility. To increase transmission security, a new algorithm based on Most Significant Bit (MSB) prediction and error embedding is presented. This combined approach aims to improve both the visual quality of the stego-image and the confidentiality of the embedded information. Author Xu and Zhang (2024) propose a new image encryption algorithm that uses dynamic transformation matrices to enhance security and randomness in image encryption. It aims to overcome the limitations of traditional methods that rely on fixed transformation matrices. Meng and Wu (2024) present a novel color image encryption algorithm aimed at strengthening information security. Their approach combines the ZigZag transform, DNA coding techniques, and a fractional-order five-dimensional hyper-chaotic system (F5DHS). The integration of the F5DHS ensures a significantly large key space, thereby enhancing the overall robustness and security of the encryption process.

Pramanik et al. (2020) combined cryptography and steganography to provide multi-layer security. The encrypted data is hidden in the LSBs of the host image by mapping function; the RSA algorithm of cryptography is used for encryption.

The study of Gupta and Bhagat (2019), proposed less significant bits substitution method is used for picture steganography by using pre-shared password. Pixels can be modified and split into three channels and then camouflage the confidential data in it. (Adiyan et al.,2018; Mulyono et al.,2019) merged steganography with cryptography to conceal confidential information in JPEG image and audio file format by using Vigen`ere cipher.

Safitri and Ahmed (2021) introduced a novel approach to three-layer image steganography using the salesman travel problem model. The results showed the best PSNR average of size 512×512 images is 61.56. In colorful images, human eyes have high sensitivity compared to monochrome images. In the study by Astuti et al. (2020), the purpose of the approach is to use the bit flipping method to three-channel images with the embedding capacity of 1 bit per pixel and having an increase in PSNR up to 13db as compared to black and white images. Moran et al. (2018) proposed a novel strategy that used optimization techniques for colorful combined an iterated search algorithm with a Greedy Randomized adaptive search procedure to determine the substitution matrix, which is further used to hide the data in a channel picture.

Abdel (2021) presented a unique steganography approach based on human visual properties. Firstly, different numbers of bits are used for every color channel according to human eye sensitivity and then circularly embedding confidential data starting from edges toward the center of the picture. All in all, distributed research on steganography has targeted different procedures for improving the restricted information in cover media. However, only a few consider an image’s visual appearance. Therefore, the proposed study inspired with (Abdel, 2021) targets the visual appearance of three channel colors by embedding different numbers of bits according to their intensity.

3. The Proposed Scheme

The proposed methodology is described in this section. The proposed framework diagram is given in Figure 2.

1. **Data Information of Cover Image:** In this experiment multiple cover images are used with multiple resolutions. The cover image can be shared to the receiver through public channels. The cover images were obtained from the USC-SIPI dataset (Araghi and Megías, 2024). Each of these RGB image pixels contains 3 bytes, one for each channel, as shown in Figure 2. The Green byte in a pixel (p) has two pairs of bits $G1_p$ and $G2_p$ used to match the secret message pair. In Red byte $R1_p$ and $R2_p$ pairs are matched lastly in Blue byte $B1_p$ and $B2_p$ bit pairs are matched to message bits. The selection of bit pairs in each channel is based on visual properties described in (Abdel, 2021).
2. **Secret message:** For confidential message data, we employ a random number function that generates a set of values in a range of 0–255 by keeping in view that RGB images are used in experiments. In our experiment we use different message sizes such as 1, 2, 4, 8, 16, 32, 64, 128 and 256 kilobytes. The secret message will be hidden in the cover image is used as a stream of bytes. Each pixel byte is further divided into four pairs of two bits as shown in Figure 2 where each pair in the message is labeled as M_i.
3. **Embedding procedure:** The embedding process starts by accessing pixel data from the cover image and secret message as bytes. Each pixel is of three bytes for green, red, and blue channels. Each channel has two pairs of bits, as mentioned earlier. Each secret message byte is processed sequentially and divided into four pairs of two non-overlapping consecutive bits. The step-by-step procedure for embedding data is as follows:
   - **Step 1:** For each pixel pin Cover Image, initially green channel has two-bit pairs $G1_p$ and $G2_p$, which are matched with message pairs sequentially and their results are stored in Steganography Image respective pixel bits $C1_p$ and $C2_p$.
   - **Step 2:** In the second stage, the red channel has two-bit pairs $R1_p$ and R2p, which are matched with message bits consecutively. Their results are stored in the Steganography image, respectively, as bits $C3_p$ and $C4_p$.
   - **Step 3:** In the third stage, the blue channel has two-bit pairs B1p and B2p, which are matched with message bits consecutively. Their results are stored in the Steganography image, respectively, as bits $C5_p$ and $C6_p$.
   - **Step 4:** The same procedure can be repeated until all the message bits are successfully embedded in the cover image. The arrangement of color channels is based on the perspective of the human eye mentioned in (Abdel, 2021).

For example, consider embedding the message "69" into the cover image. First, convert the message into binary bits and divide them into pairs, denoted as M_1, M_2, M_3, ..., M_n, as illustrated in Figure 3. Next, for each pixel in the RGB cover image, convert the pixel values into 8-bit binary format and then split them into pairs. In our research, we prioritize the green channel for embedding. The green channel values are divided into two pairs, G1p and G2p, which are compared against the message pairs. Compare $G1_p$

with the first message pair ($\mathbf{M_1}$). If they match, set $\mathbf{C1_p = 1}$; otherwise, set it to **0**. Next, compare $\mathbf{G2_p}$ with $\mathbf{M_2}$. If they match, set $\mathbf{C2_p = 1}$; otherwise, set it to **0**.If $\mathbf{M_2}$ does not match **R1p**, set **C3p = 0**.Compare **M2** with $\mathbf{R2_p}$; if they match, set $\mathbf{C4_p = 1}$.The message pair remains unchanged until it no longer matches the pixel values. Compare $\mathbf{M_3}$ with $\mathbf{B1_p}$. If they match, set $\mathbf{C5_p = 1}$. Compare **M4** with $\mathbf{B2_p}$. If they match, set $\mathbf{C6_p = 1}$

.

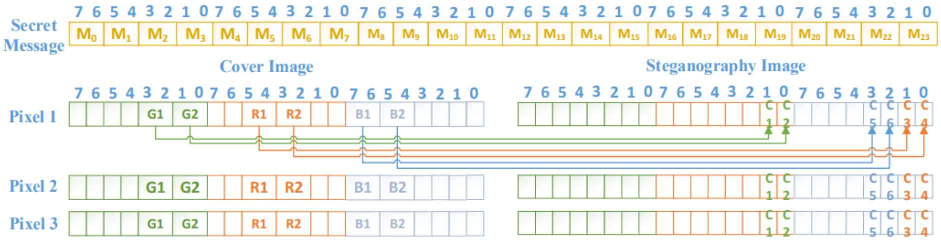


Figure 2. Structure of Proposed Scheme

M_1 is the first message pair, and it is compared with the first green channel pair ($G1_p$). Since it finds a match immediately, the process moves on to the next message pair.M_2 is first compared with the second green channel pair ($G2_p$). If it matches, the bit is set accordingly. If M_2 does not match $G2_p$, it is compared with the first red channel pair ($R1_p$).If there is still no match, M_2 is further compared with the second red channel pair ($R2_p$). This means that M_2 remains the active message pair until it finds a match or exhausts the predefined matching conditions. Unlike M_1, which immediately finds a match with $G1_p$, M_2 requires multiple attempts across different channels before a match is confirmed.

This mechanism effectively embeds the message while maintaining imperceptibility in the cover image. Using the same approach, the process continues iteratively for subsequent message pairs (M_3, M_4, ...).

This process continues iteratively until all message pairs are embedded into the cover image. The detailed steps of this embedding process are outlined in **Algorithm 1**.

4. **Extraction Algorithm:** The next stage is to extract a secret message from the Steganography image. The sender encodes the secret message into the cover image C, creating the stego image. The receiver then obtains the stego image through a communication channel and decodes it to extract the hidden secret message file. It starts by checking S bits in the Steganography image and copying respective data into a secret message container. The extraction process is explained in steps:
    - **Step 1:** First apply the check bit function to the Red pixel to check the bit at index $C1_p$. If the bit is set to 1, get two bits from green $G1_p$ as a message pair M_1.
    - **Step 2:** Now apply the function on the Red pixel to check the bit at index $C2_p$. If the bit is set to 1, get two bits from green $G2_p$ as a second message pair M_2.

- **Step 3:** Same as above apply check bit function to check indexes $C3_p$, $C4_p$, $C5_p$, $C6_p$ and get the two bits of message pairs from $R1_p$, $R2_p$, $B1_p$ and $B2_p$ respectively.

The extraction algorithm is also explained as shown in Algorithm 2.

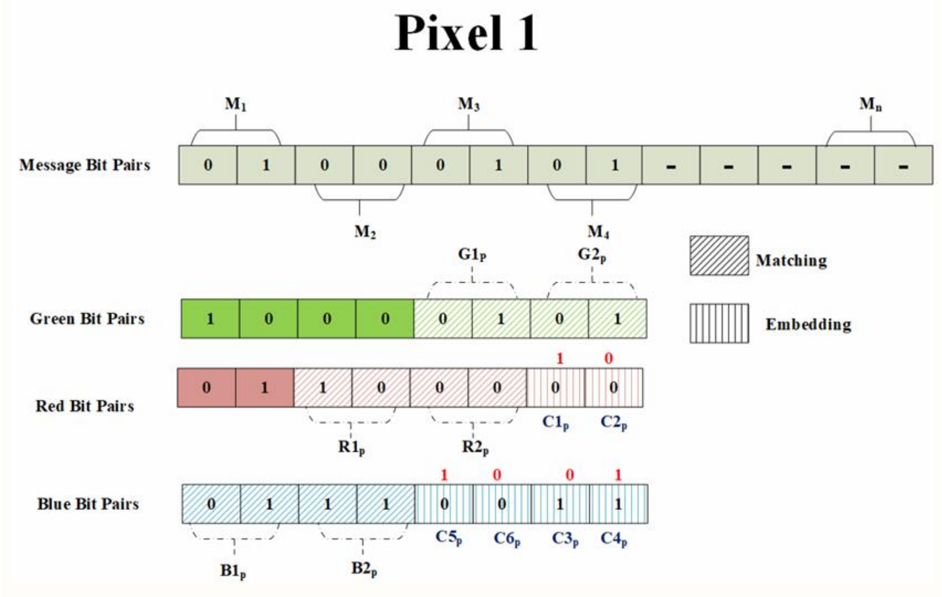


Figure 3. Example of Proposed method

Algorithm 1 Embedding Algorithm

```
Input=Secret Message(M)and Cover
Media(C)
Output=Steganographic Media(S)
i ← 0
For each pixel p in C do
if $G1_p==M_i$ then
        $S1_p$ ← 1
i ← i+1
else
        $S1_p$ ← 0
end if
if $G2_p==M_i$ then
        $S2_p$ ← 1
i ← i+1
else
        $S2_p$ ← 0
```

```
end if
if $R1_p==M_i$ then
     $S3_p \leftarrow 1$
$i \leftarrow i+1$
else
     $S3_p \leftarrow 0$
end if
if $R2_p==M_i$ then
     $S4_p \leftarrow 1$
$i \leftarrow i+1$
else
     $S4_p \leftarrow 0$
end if
if $B1_p==M_i$ then
     $S5_p \leftarrow 1$
$i \leftarrow i+1$
else
     $S5_p \leftarrow 0$
end if
if $B2_p==M_i$ then
     $S6_p \leftarrow 1$
$i \leftarrow i+1$
else
     $S6_p \leftarrow 0$
end if
endfor
```

Algorithm 2 Extraction Algorithm

```
Input = Steganographic Media (S)
Output = Secret Message (M)
$i \leftarrow 0$
For each pixel p in C do
  if $S1_p==1$ then
$M_i \leftarrow G1_P$
$i \leftarrow i+1$
end if
  if $S2_p==1$ then
$M_i \leftarrow G2_P$
$i \leftarrow i+1$
end if
  if $S3_p==1$ then
$M_i \leftarrow R1_P$
$i \leftarrow i+1$
end if
  if $S4_p==1$ then
$M_i \leftarrow R2_P$
```

```
i ← i+1
end if
  if $S5_p$==1 then
$M_i$ ← $B1_P$
i ← i+1
end if
  if $S6_p$==1 then
$M_i$ ← $B2_P$
i ← i+1
     end if
end for
```

4. Quality Measures of Performance

The following measures are used to assess how well the suggested model performs:

1) **Payload capacity:** The amount of pixels embedded in the cover image is represented by payload capacity (Abdel, 2021). It is represented by bits per pixel. The payload capacity C is given as:

$$\mathrm{C}=\frac{\textbf{Number of message bits}}{\textbf{Cover image pixels}} \quad (1)$$

2) **PSNR:** It is a performance metric that mostly can be used to evaluate the image quality. The PSNR formula (Kadhim et al., 2019) can be derived from Mean square error where E and F are dimensions; and c, d are the pixels of the targeted image, and x is the cover or x' is the stego image (Abdel, 2021). With PSNR and SSIM, higher values show considerable picture similarity, while with MSE; large values specify lower image similarity :

$$\mathbf{MSE}=\frac{\mathbf{1}}{\mathbf{ExF}}\sum_{\mathbf{c}}^{\mathbf{E}}\sum_{\mathbf{d}}^{\mathbf{F}}[\mathbf{x}(\mathbf{c},\mathbf{d})-\mathbf{x`}(\mathbf{c},\mathbf{d})]^{\mathbf{2}} \quad (2)$$

&

$$\textbf{PSNR}=\mathbf{10log_{10}}\left[\frac{\mathbf{255^2}}{\boldsymbol{MSE}}\right]\mathbf{db} \quad (3)$$

3) **Structural Similarity Index Measure (SSIM):**

SSIM (Kadhim et al., 2019) is a more current estimation device that is planned in light of three elements, for example, luminance, differentiation, and design, to more readily suit the functions of the human visual framework (Setiadi, 2021):

$$\textbf{SSIM (u,v)}=\frac{(2\mu_u\mu_v+c_1)(2\sigma_{uv}+c_2)}{(\mu_u{}^2+\mu_v{}^2+c^1)(\sigma_u{}^2+\sigma_v{}^2+c^2)} \quad (4)$$

$$c_1=(O_1P)^2$$

$$c_2=(O_2P)^2$$

where μ_u and μ_v are the mean intensity values of images u and v. σ_u^2 is the variance of u, σ_v^2 is the variance of v and σ_{uv}^2 is the covariance of u and v. c_1 and c_2 are the two stabilizing parameters, P is the dynamic range of pixel values (2#bitsperpixel−1) and the contents O_1=0.01 and O_2=0.03.

5. Experimental Dataset

The dataset in this research for cover images is from the USC-SIPI image database (Araghi and Megías, 2024). The USC-SIPI image database is a collection of digitized images. It is created to support research in image processing, image analysis, and machine vision. The primary edition of the USC-SIPI image database was distributed in 1977, and many fresh images have been added since then.

The database is divided into volumes based on the basic quality of the pictures. Images in each volume are of various sizes, such as 256×256 pixels, 512×512 pixels, or 1024×1024[1] pixels.

All images are 8 bits/pixel for black and white images and 24 bits/pixel for color images. The designed approach is implemented in the PYTHON 3.10 module.

The proposed method has experimental images of 256×256, 512×512, 1024 × 1024 resolutions some of which are shown in Fig 4, Fig 5, and Fig 6 respectively. All the images are of tiff extensions. Other than these images we use Lady.tiff, Girl.tiff, Candies.tiff and Benties.tiff for 256x256.For 512x512 and 1024x1024, Splash.tiff, F16.tiff, House.tiff, Pepper.tiff or Grass.tiff, Mountian.tiff, River.tiff, Birds.tiff respectively.

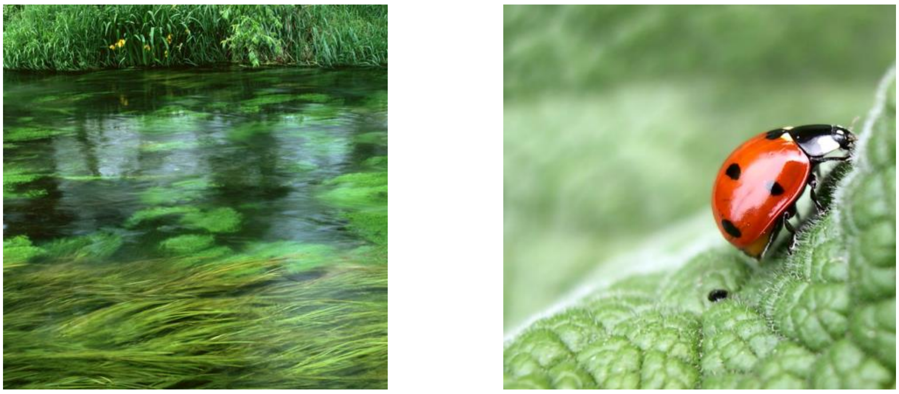
(a) Zelda.tiff **(b)Couple.tiff**

Figure 4. Cover images used with resolution 256 × 256 pixels.

[1] https://www.mathworks.com/matlabcentral/answers/307261-from-where-i-get-the-image-set-of-size-1024-1024-or-more-for-image-processing-in-matlab

(a) House (b) Lena

Figure 5. Cover images used with resolution 512x512 pixels.



(a) Grass (b) Bug

Figure 6. Cover images used with resolution 1024x1024 pixels.

6. Results

6.1. Capacity and quality analysis

Table 1 presents the results of different embedding message capacities in kilobytes in different images of 256×256×3,512×512×3and 1024×1024×3shown in Figures 4, 5 and 6 respectively. The images obtained a maximum PSNR of 49.016 and a minimum of 47.891 for the Couple.tiff image after embedding a message of 1 kilobyte. A minimum PSNR of 34.902 is obtained for Candies.tiff. For 512x512 images the results have shown that this algorithm PSNR for a minimum message size of 64 kilobytes is 36.131. Due to the increase in cover image size, algorithm PSNR for 1-kilobyte message is raised to a maximum of 56.312. The result of the proposed methodology by using different message sizes in kilobytes in images 1024×1024 has shown that this algorithm PSNR for a maximum message size of 256 kilobytes is 37.651 and minimum is 36.752. Due to the increase in cover image size, algorithm PSNR for 1 kilobyte message is raised to a maximum of 62.453. Figure 7 shows a histogram of Couple.tiff from Figure 4. It represents the effects of multiple sizes of secret messages on a Couple.tiff. It can be observed that up to 8kB secret message histogram remains similar.

We compare our proposed technique with Naveen and Jayaraghavi (2024) LSB (Least Significant Bit) method, demonstrating improved results. The image sizes used in our evaluation are RGB 256×256, 512×512, and 1024×1024.Table 2 below presents the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) values for different cover images. The results indicate that our proposed method achieves higher PSNR values compared to both the traditional method, signifying better image quality. Additionally, the SSIM values remain consistently high, ensuring minimal perceptual distortion.

6.2. Security analysis

Steganography is the technique of hiding sensitive information, while steganalysis involves breaking the steganographic algorithm to detect and extract the concealed data. Steganalysis is categorized into active and passive approaches. Active steganalysis not only detects hidden information but also deciphers or alters it, whereas passive steganalysis focuses solely on identifying its presence. Various steganalysis methods such as RS analysis are applied to the proposed technique to assess the security of the hidden data.

The most popular and reliable statistical steganalysis to determine whether hidden information is present in the stego-image for LSB-based approaches is the RS analysis. To determine if a steganography algorithm is secure, Fridrich et al. (2001) devised the steganalysis technique known as RS analysis. S analysis, or Residual-based Steganalysis, is a technique used in steganography to detect the presence of hidden information in digital media, such as images or audio files.

The basic idea behind RS analysis is to examine the residuals, or differences, between the original signal and the modified signal that contains the hidden information. The residual is the difference between the two signals, and in steganography, it represents the hidden data. In RS analysis, statistical features are extracted from the residual signal to detect the presence of hidden information. These features can include measures such as variance, skewness, and kurtosis, which are then used to train a machine learning algorithm to classify the residual as either containing hidden information or not.

Using three distinct cover images, Table 2 evaluates the effectiveness of a suggested steganography technique against the Naveen and Jayaraghavi (2024) Least Significant Bit (LSB) method. The comparison uses two measures of picture quality: the Structural Similarity Index (SSIM) and the Peak Signal-to-Noise Ratio (PSNR). While SSIM values nearer 1 imply greater similarity between the original and changed images, higher PSNR values suggest superior image quality. According to the results, the suggested approach outperforms the LSB methods in terms of image quality and robustness for data hiding within images, achieving higher PSNR and equivalent or marginally enhanced SSIM values. However, Table 3 details the computational efficiency of the proposed technique, showing the execution time in seconds for both the embedding and extraction stages.

Table 1. PSNR of proposed embedding algorithm

Msg Size	**1 KB**	**2KB**	**4KB**	**8KB**	**16KB**	**32KB**	**64KB**	**128KB**	**256KB**
256x256									
Max	49.016	46.872	44.855	41.063	38.803				
Min	47.891	45.058	40.734	38.219	34.903				
512x512									
Max	56.312	52.443	49.351	46.341	43.257	40.206	37.120		
Min	53.846	50.992	41.348	41.348	36.344	39.010	36.131		
1024x1024									
Max	62.453	59.345	56.349	53.356	50.407	47.355	44.267	40.801	37.651
Min	60.399	57.088	53.887	50.965	48.240	45.237	42.208	39.871	36.752

Table 2. Comparison of PSNR and SSIM values measured in LSB methods and proposed technique.

Cover Image	PSNR		SSIM	
	Naveen's method	Proposed method	Naveen's method	Proposed method
Peppers	40.7581	54.685	0.998	0.999
Lena	40.7825	54.486	0.9985	0.999
Tulips	40.8356	53.846	0.9931	0.999

Table 3. The execution time of our proposed technique.

Phase	**Execution Time(s)**
Embedding Phase	0.3992
Extraction Phase	0.6500

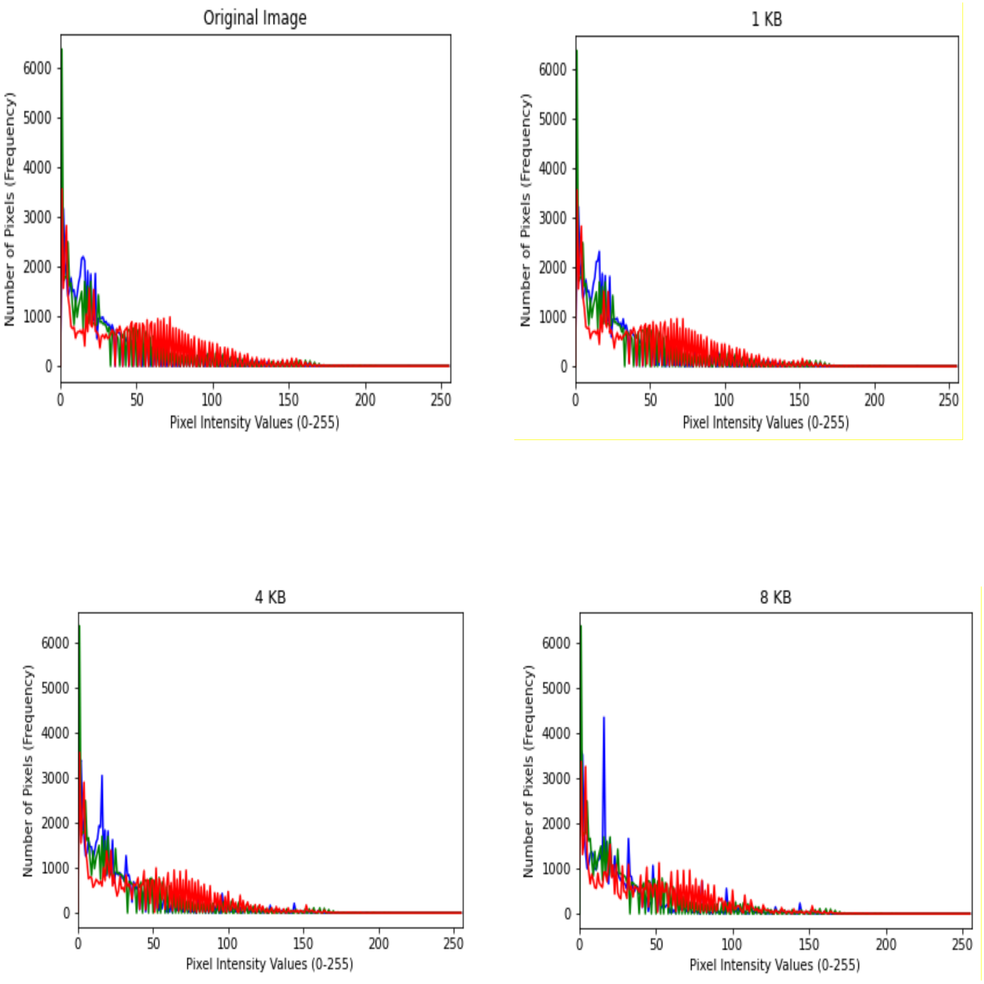


Figure 7. Histogram comparison of Couple.tiff

Figure 7 shows a histogram of Couple.tiff with different sizes of data in KB from Figure 4. It represents the effects of multiple sizes of secret messages on Couple.tiff. It can be observed that up to 8kB secret message histogram remain similar. Red, Blue and Green lines on graphs illustrate the amount of pixel at specific number.

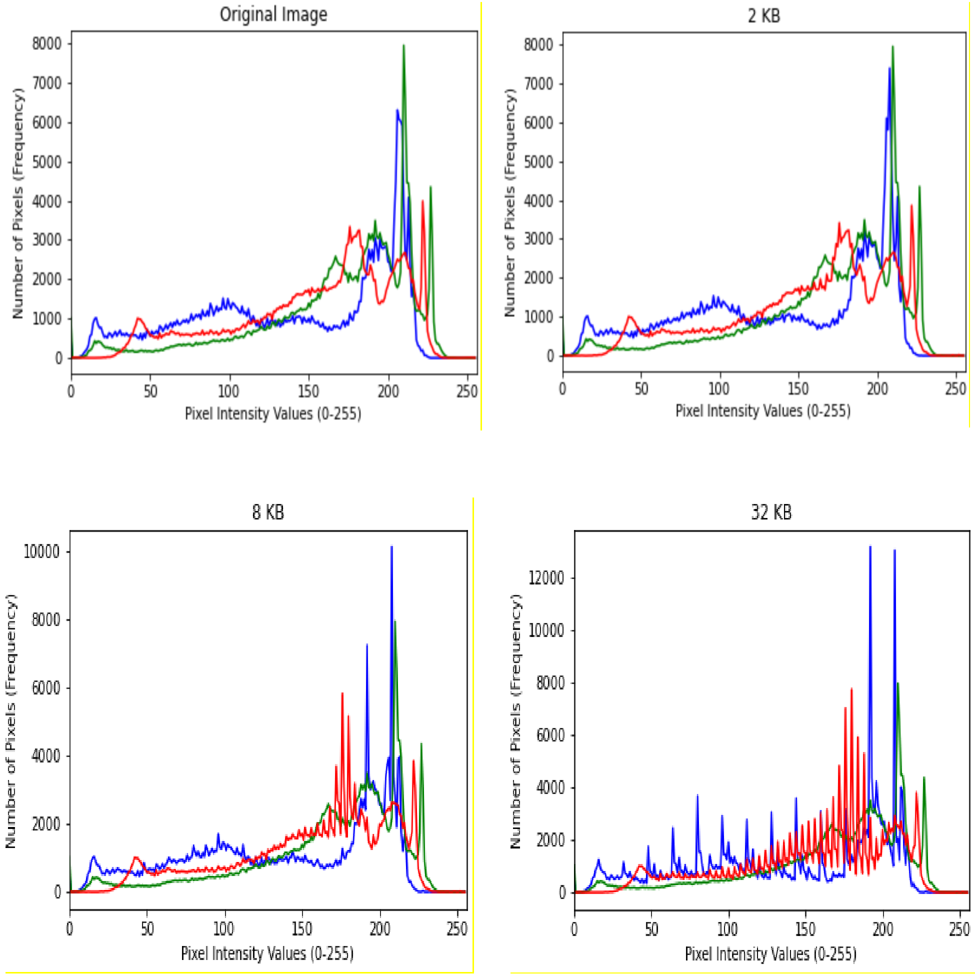


Figure 8. Histogram comparison of House.tiff

Figure 8 shows a histogram of House.tiff from Figure 5 with different sizes of data in KB from Figure 5. Red, Blue and Green lines on graphs illustrate the amount of pixel at specific number.

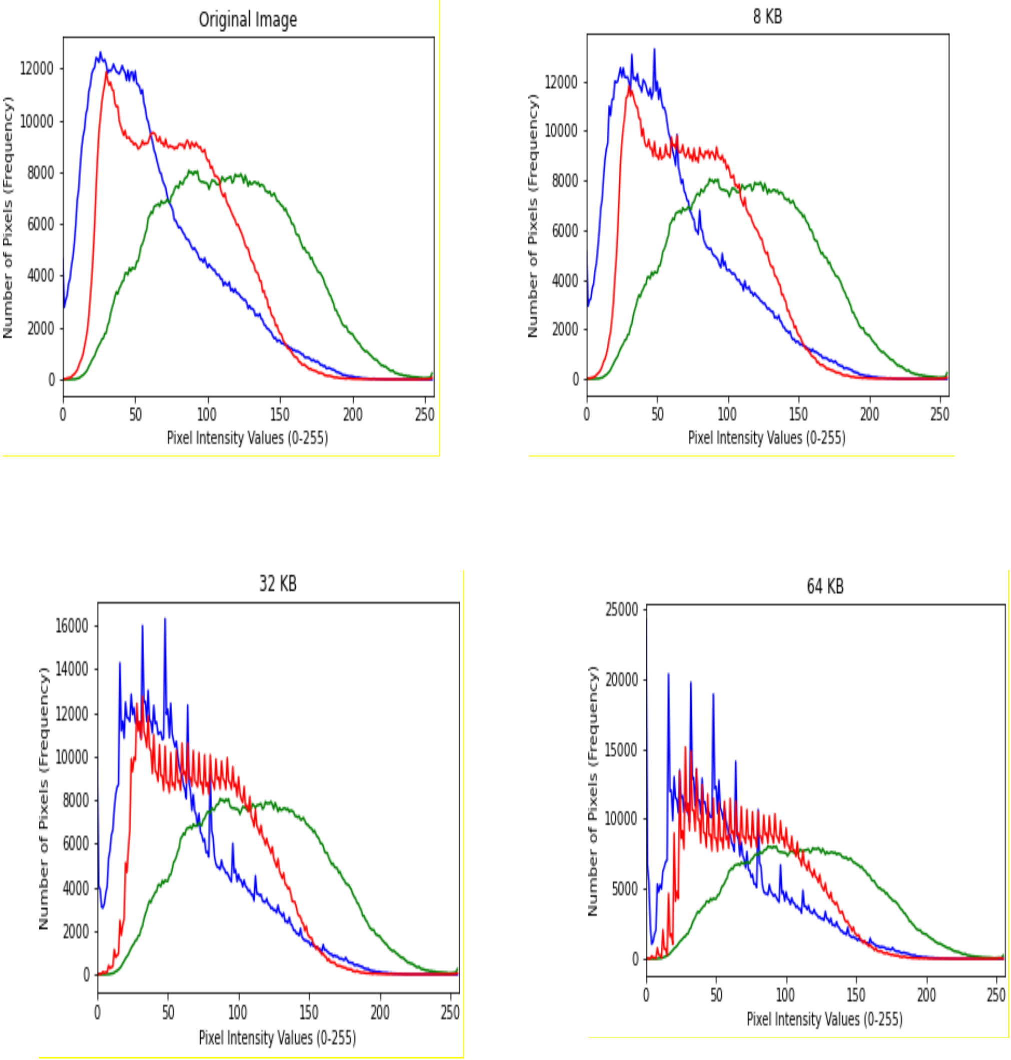


Figure 9. Histogram comparison of Grass.tiff

Figure 9 shows a histogram of Grass.tiff from Figure 6. It has also shown similar behaviour as in Couple.tiff in Figure 7.

7. Conclusion and future work

The approach used in this research proposed a security enhancement methodology to hide the confidential information that is steganographically concealed within the cover images. The proposed scheme has the following objectives:

1) Embedding is done using human visual properties and wavelengths of colors.
2) More data is embedded with less change in the cover image.

The experimental work was carried out on different sizes of 256x256x3, 512x512x3, 1024x1024x3 images, in which secret data of different kilobytes, such as 1, 2, 4, 8, 16, 32, 64, 128, 512, and so on, was embedded. The obtained result reveals that the embedding is performed by using a wavelength of color properties that are 99% similar image to the original image, having the capacity to embed $\frac{1}{2}$bits and $\frac{1}{4}$bits can be modified by the proposed methodology. The highest PSNR achieved by using this scheme with the image sizes 256x256, 512x512, and 1024x1024 are 49.01, 56.31, and 62.45, respectively, and the average PSNR with the same image sizes is 34.90, 36.13, and 36.75, respectively. For future work, we try to improve our technique and make it useable in the field of practical applications in secure communication, digital watermarking, and medical image security**,** maximizing the method’s real-world impact.

References

Abdel-Hafeez, S., Alqunaysi, A. THRESHOLD ANALYSIS on 2-bit LSB STEGANOGRAPHY (R, G, B) COLOR IMAGE. Researchgate.net

AbdelRaouf, A. (2021). A new data hiding approach for image steganography based on visual color sensitivity. *Multimedia Tools and Applications*, *80*(15), 23393-23417.

Almazaydeh, L. (2020). Secure RGB image steganography based on modified LSB substitution. *International Journal of Embedded Systems*, *12*(4), 453-457.

Amahdi, S., 2021. An improved method for combine (LSB and MSB) based on color image RGB. Engineering and Technology Journal, 39(1B), pp.231-242.

Ansari, A. S., Mohammadi, M. S., Parvez, M. T. (2019). A comparative study of recent steganography techniques for multiple image formats. *International Journal of Computer Network and Information Security*, *11*(1), 11-25.

Araghi, T. K., Megías, D. (2024). Analysis and effectiveness of deeper levels of SVD on performance of hybrid DWT and SVD watermarking. *Multimedia Tools and Applications*, *83*(2), 3895-3916.

Astuti, E. Z., Setiadi, D. R. I. M., Rachmawanto, E. H., Sari, C. A., Sarker, M. K. (2020). LSB-based bit flipping methods for color image steganography. In *Journal of Physics: Conference Series* (Vol. 1501, No. 1, p. 012019). IOP Publishing.

Bawaneh, M.J., Al-Shalabi, E.F., Al-Hazaimeh, O.M., 2021. A novel RGB image steganography using simulated annealing and LCG via LSB. *International Journal of Computer Science & Network Security*, 21(1), pp.143-151.

Danny Adiyan, Z., Purboyo, T.W. Nugrahaeni, R.A., 2018. Implementation of secure steganography on jpeg image using LSB method. *International Journal of Applied Engineering Research*, 13(1), pp.442-448.

Darwis, D., Pamungkas, N. B. (2021). Comparison of Least Significant Bit, Pixel Value Differencing, and Modulus Function on Steganography to Measure Image Quality, Storage Capacity, and Robustness. In *Journal of Physics: Conference Series* (Vol. 1751, No. 1, p. 012039). IOP Publishing.

Dumre, R., Dave, A. (2021). Exploring lsb steganography possibilities in rgb images. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-7). IEEE.

Elharrouss, O., Almaadeed, N., Al-Maadeed, S. (2020, February). An image steganography approach based on k-least significant bits (k-LSB). In *2020 IEEE international conference on informatics, IoT, and enabling technologies (ICIoT)* (pp. 131-135). IEEE.

Fridrich, J., Goljan, M., Du, R. (2001). Detecting LSB steganography in color, and gray-scale images. *IEEE MultiMedia*, 8(4), 22-28.

Gupta, P., Bhagat, J. (2019). Image steganography using LSB substitution facilitated by shared password. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 1* (pp. 369-376). Springer Singapore.

Hwang, M.S., Chang, Y.L., Lin, K.Y., Yang, C.Y., Lin, I.C. (2024). Research on Data Security and Privacy of Smart Grids. *International Journal of Network Security*, 26(5), 901-910.

Jiao, S., Feng, J. (2021). Image steganography with visual illusion. *Optics Express*, *29*(10), 14282-14292.

Kadhim, I.J., Premaratne, P., Vial, P.J., Halloran, B. (2019). Comprehensive survey of image steganography: Techniques, Evaluations, and trends in future research. *Neurocomputing*, 335, pp.299-326.

Kılıç, E., Evrensevdi, B. A. (2022). Review on the Different Types of Steganography.

Li, Y.P., Li, Y.C. (2023). IoT Malware Threat Hunting Method Based on Improved Transformer. *International Journal of Network Security,* 25(2), 267-276.

Li, Z. (2024). Image Enhancement and Cloud Secure Transmission Based on Reversible Image Information Hiding Technology. *International Journal of Network Security*, 26(4), 703-712.

Meng, F., Wu, G. (2024). Color Image Encryption Algorithm with ZigZag Transform and DNA Coding Based on Fractional Order 5D Hyperchaotic System. *International Journal of Network Security*, 26(2), 244-251.

Moran, M. B., Ochi, L. S., Conci, A., Araujc, A. S., Muchaluat-Saade, D. C. (2018). Iterated local search for RGB image steganography. In *2018 25th International conference on systems, signals and image processing (IWSSIP)* (pp. 1-5). IEEE.

Mulyono, I.U.W., Susanto, A., Anggraeny, T. Sari, C.A. (2019). Encryption of Text Message on Audio Steganography Using Combination Vigenere Cipher and LSB (Least Significant Bit). *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp.63-74.

Naveen, P., Jayaraghavi, R. (2024). Image Steganography Method for Securing Multiple Images using LSB–GA. *Wireless Personal Communications*, 135(1), 1-19.

Pramanik, S., Samanta, D., Dutta, S., Ghosh, R., Ghonge, M., Pandey, D. (2020). Steganography using improved LSB approach and asymmetric cryptography. In *2020 IEEE international conference on advent trends in multidisciplinaryresearch and innovation (ICATMRI)* (pp. 1-5). IEEE.

Safitri, P.H., Ahmad, T. (2021). December. Developing RGB Image Steganography using Travel Salesman Problem Modelling. In *2021 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)* (pp. 209-213). IEEE.

Setiadi, D. R. I. M. (2021). PSNR vs SSIM: imperceptibility quality assessment for image steganography. *Multimedia Tools and Applications*, *80*(6), 8423-8444.

Song, X., Yang, H., Cao, W. (2025). Research on Data Security in Supply Chain Financial Business from the Perspective of Blockchain Technology. *International Journal of Network Security*, 27(1), 46-50.

Xu, C., Zhang, Y. (2024). Color Image Encryption Based on Chaotic Systems and Dynamic Transformation Matrices. *International Journal of Network Security*, 26(5), 867-873.

Yang, D. (2023). A Study on Detection of Malware Attacks Using Machine Learning Techniques. *International Journal of Network Security*, 25(6), 1042-1047.

Received December 6, 2024, revised March 19, 2025, accepted March 19, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 2, 407-418
https://doi.org/10.22364/bjmc.2025.13.2.06

Acoustic Cues of Prosody Phrasing in Lithuanian at Different Speech Rates

Asta KAZLAUSKIENĖ, Gailius RAŠKINIS,
Eidmantė KALAŠINSKAITĖ-ZAVIŠIENĖ

Vytautas Magnus University, V. Putvinskio str. 23-204, LT-44243 Kaunas, Lithuania

asta.kazlauskiene@vdu.lt, gailius.raskinis@vdu.lt,
eidmante.kalasinskaite-zavisiene@vdu.lt

ORCID 0000-0002-6135-1635, 0000-0001-5793-868X, 0009-0003-6007-6621

Abstract. The aim of this research is to determine whether pauses, decreases in intensity, pitch lowering, lengthening of the final sound, and certain voice quality features are equally significant in marking phrase boundaries at different reading rates. The research material consists of 9 recordings of Aesop's fable "The North Wind and the Sun". This text was read by three male actors 3 times: at a natural reading rate, at a fast reading rate, and at a slow reading rate. The presence and duration of pauses, the intensity and F0 of the entire phrase and its final syllable, as well as the duration, intensity, F0, jitter, shimmer, and HNR of the final vowel at the end and in phrase-internal position, were analysed.

The results indicate that all parameters can serve as indicators of phrase boundaries, particularly for intonational phrases. Intonational phrases can be differentiated from intermediate phrases based on the level of intensity decrease, jitter, and to some extent, pauses and F0 lowering. However, among these parameters, only duration-related features, such as pauses and the lengthening of the final sound, show a more prominent sensitivity to changes in reading rate.

Keywords: Lithuanian, prosodic phrasing, phrase, pause, intensity, fundamental frequency, duration, voice quality, speech rate

1. Introduction

Prosodic phrasing involves the grouping of words into larger units, known as intermediate phrases (ip) and intonation phrases (IP). This grouping is influenced by a syntactic and prosodic structure of a language, as well as by semantic and pragmatic factors, and individual speaker characteristics (e.g. speech rate, emotions, and experience in speaking aloud). While phonetic features of phrase boundaries are likely universal, their significance may vary across languages. In many studies, the most important indicators of phrase boundaries include pitch and intensity changes, pausing, segmental lengthening (Peters, 2003; Michelas and D'Imperio, 2010; Petrone et al., 2017; Brugos et al., 2018; Żygis et al., 2019; Harrington Stack and Watson, 2023; Liu et al. 2023; de Souza, 2023; Steffman et al. 2024; Volín et al., 2024), and non-modal phonation

(Kohler, 2000; Crowhurst, 2018). However, as syntactic structure varies between languages, phrasing patterns in one language cannot simply be transferred to another.

Prosodic phrasing affects the perception and production of language phenomena. For appropriate clause and sentence processing, prosodic and syntactic boundaries must align (for research on this relationship, see Watson, Gibson, 2004; Himmelmann, 2022).

Research on prosodic phrasing based on acoustic cues is important for improving the performance of human language technologies such as automatic speech recognition (ASR), speech synthesis, and natural language understanding (NLU) systems. Acoustic cues, including pitch variation, duration, and pauses, are fundamental in signalling phrase boundaries in spoken language. These cues help disambiguate sentence structure, clarify meaning, and improve the flow of communication. By developing algorithms that can accurately detect phrase boundaries using these acoustic signals, ASR systems can produce more accurate transcriptions, particularly in real-time applications where syntactic structure impacts comprehension.

In speech synthesis, accurately predicting pause durations based on speech tempo can lead to more natural and expressive prosody, significantly improving the overall quality of generated speech. For NLU systems, precise identification of phrase boundaries allows for better parsing and understanding of input, leading to improvements in tasks like machine translation and dialog systems. Ultimately, this research is crucial for refining the ability of language technologies to handle the complexities of spoken language, making them more effective in diverse applications such as virtual assistants, automated transcription services, and language learning tools.

Given this context, it is crucial to identify the basic features that signal intonation phrase and intermediate phrase boundaries and to determine which of these features are consistent. Research on the Lithuanian language in this area is still in its early stages. Dereškevičiūtė and Kazlauskienė (2022), Kazlauskienė, Dereškevičiūtė, and Sabonytė (2023) have shown that phrase endings are signalled by a decrease in intensity, F0 lowering, lengthening of the final sound, and an increase in amplitude perturbations. While pauses are a possible indicator of phrase boundaries, they are not mandatory.

2. Aim, Material and Methods

The research presented in this paper represents one stage of a broader study. The aim of this stage is to determine whether pauses, decreases in intensity, pitch lowering, lengthening of the final sound, and certain voice quality features are equally significant in marking phrase boundaries at different reading rates. For this purpose, a controlled-read speech analysis was conducted.

To carry out the experiment, three native male speakers (professional actors) read Aesop's fable *The North Wind and the Sun*. A translation of this fable is usually chosen to illustrate the International Phonetic Alphabet, as these illustrations provide accounts of the sound structure of different languages. The text comprises seven sentences, 68 lexemes, and 405 sounds, and it was read three times: at a natural rate, a fast rate, and a slow rate. Breaks were provided between the recordings. The speakers were instructed to read the text clearly, paying attention to the semantic and grammatical connections between words. They were allowed to choose the phrasing of the sentences freely, and the phrasing did not necessarily have to be identical across all recordings. Recordings were made in a professional sound recording studio.

Until further research is conducted, intonation phrases in read Lithuanian are identified with sentences. In this study, an intonation phrase is defined as a phrase consisting of one intermediate phrase or the final intermediate phrase within a sentence. The number of intonation phrases in each recording was 7.

Intermediate phrases are identified based on acoustic, syntactic, and semantic criteria. The number of intermediate phrases per recording varied across speakers and reading rates. Including the last intermediate phrase of an intonation phrase, as well as intonation phrases consisting of only one intermediate phrase, speaker S1 produced 24 intermediate phrases during fast reading, 22 during natural reading, and 28 during slow reading. Speaker S2 produced 23 intermediate phrases in the fast reading, 27 in the natural reading, and 29 in the slow reading. Speaker S3 produced 23 intermediate phrases in the fast reading, 24 in the natural reading, and 25 in the slow reading.

To investigate whether **intensity** and **fundamental frequency** (F0) signal phrase boundaries, the intensity and F0 of entire intonation and intermediate phrases, as well as the nucleus of the final syllable in each phrase, were compared. F0 values were extracted from audio recordings using an autocorrelation technique with sinc interpolation to enhance the precision of lag and peak height estimates (Boersma, 1993).

To assess whether the final sound in a phrase is lengthened, the **durations** of the vowels [eː] and [ɑː] were compared. These vowels occur both at the end of a word within a phrase and at the end of intonation or intermediate phrases, making them suitable for such comparisons.

Additionally, to explore whether aperiodicity, stability, and noisiness in the voice signal serve as indicators of phrase boundaries, several voice quality features of phrase-final vowels were compared to those in phrase-internal positions. The analysed features included:

- **Jitter:** The period-to-period variability in pitch, estimated using waveform-matching, where the duration of each period is determined via a cross-correlation technique (Boersma, 2009).
- **Shimmer:** The average absolute difference between the amplitudes of consecutive periods, normalized by the average amplitude.
- **Harmonics-to-Noise Ratio (HNR):** The ratio of the relative power of the periodic (harmonic) component to the noise component. HNR values were calculated using the technique described by Boersma (1993)

All the above-mentioned acoustic features were extracted manually using the Praat software tool (Boersma and Weenink, 2018). As this study focuses on acoustic parameter ratios for individual speakers, data normalization was deemed unnecessary.

The ratios in this study are calculated either by dividing one analysed quantity by another or by converting the difference in decibels to a ratio. For instance, if the average fundamental frequency (F0) of the nucleus of the final syllable in a phrase is 83 Hz, and the average F0 of the voiced units in the entire intermediate phrase is 124 Hz, the resulting ratio is 0.67 (83/124 = 0.67). Similarly, if the intensity at the end of the phrase is 55.7 dB and the average intensity of the entire intonation phrase is 61.1 dB, the resulting ratio is calculated as $10^{(55.7-61.1)/10} = 0.288$

3. Results

3.1 Duration of Pauses and Pausing

Pauses can occur for a variety of reasons, including psychological, physiological, rhythmic, and other factors. One of their primary linguistics functions is to divide speech into distinct intonation units. Consequently, pauses (or prosodic breaks) are likely indicators of phrase boundaries.

Table 1 presents the pause duration ratios at different reading rates. The natural reading rate was used as a baseline for calculating ratios, with the pause duration ratio determined by dividing the mean pause duration during slow or fast reading by the mean duration for natural reading. All speakers tended to shorten pauses between intonation phrases during fast reading and lengthen those between intermediate phrases during slow reading.

Table 1. Duration ratios of pauses. S1, S2, and S3 denote the speakers; IP refers to an intonation phrase or the last intermediate phrase of the intonation phrase; ip represents an intermediate phrase; N, F, and S correspond to natural, fast, and slow reading rates, respectively.

Researched unit	**S1**	**S2**	**S3**
ip (F:N:S)	0.7:1:1.2	0.3:1:1.5	0.7:1:2
IP (F:N:S)	0.4:1:1.1	0.2:1:1.3	0.6:1:1.4

When comparing pauses between intermediate phrases in fast and natural reading, it was observed that two speakers shortened these pauses by one-third during fast reading, while one speaker shortened them by nearly two-thirds. These shortenings were statistically significant. In this study, statistical significance was assessed using the *t-test*, with a significance level of 0.05.

Conversely, during slow reading, the speakers extended the pauses from 1.2 to 2 times, with these differences also being statistically significant.

Pauses between intonation phrases were shortened by two to four-fifths during fast reading. During slow reading, pauses after intonation phrases were extended slightly less than those between intermediate phrases, ranging from 1.1 to 1.4 times. Regardless of whether the reading rate increased or decreased, the differences in pause durations remained statistically significant.

Pauses can not only be shortened but may also be entirely omitted when the reading rate is increased. Conversely, additional pauses may be introduced when the rate is slowed. An analysis of the number of pauses revealed that during fast reading, the number of pauses decreased by half to one-tenth compared to natural reading. The difference between natural and slow reading was considerably smaller: on average, slow reading introduced only about 1.5 times more pauses than natural reading. Importantly, only pauses between intermediate phrases were omitted or added. Pauses between intonation phrases were consistently present after each sentence, with one exception in the fast-rate recordings where such a pause was omitted.

It is essential to assess the proportion of total recording time occupied by pauses at different reading rates. For fast reading, pauses accounted for only 4% to 14% of the total duration (S1—13%, S2—4%, S3—14%). In contrast, at a natural reading rate, pauses constituted over 20% of the total duration (S1—22%, S2—28%, S3—26%). During slow reading, pauses occupied even more time, ranging from 27% to 38% (S1—27%, S2—35%, S3—38%).

An analysis of the duration and number of pauses, as well as the total time occupied by pauses in the recordings, indicates that pauses consistently mark the boundaries of intonation phrases, irrespective of the reading rate. However, pauses between intermediate phrases are treated differently. During fast reading, these pauses may be completely omitted, rendering them unreliable markers for identifying intermediate phrases. Conversely, in slow reading, additional pauses may occasionally be inserted within a phrase, creating intermediate phrases not based on syntax or semantics, and thus necessitating caution when identifying phrase boundaries based solely on pauses. Nonetheless, a pause reliably separates the two phrases.

3.2 Intensity

The results of the research reveal certain tendencies. As predicted, intensity decreases at the end of any phrase (see Table 2). However, the range of the decrease varies significantly: for intermediate phrases, it ranges from 0.16 to 0.60 times, and for intonation phrases, it ranges from 0.01 to 0.13 times, with the last syllable exhibiting lower intensity than the entire phrase.

The data in this study do not indicate a strong correlation between intensity decreases and reading rate for intermediate phrases. During natural reading, the intensity decreases quite consistently for all speakers, averaging a decrease of 0.27 times. In fast reading, speaker S1 shows a significantly smaller intensity decrease at the end of the phrase compared to natural reading. Conversely, speakers S2 and S3 demonstrate a slightly greater intensity decrease during fast reading than during natural reading, though the differences are minimal. When reading slowly, speakers S2 and S3 exhibit even greater intensity decreases, whereas speaker S1 shows the opposite trend: although the decrease in intensity is more prominent than during fast reading, it remains less prominent than during natural reading.

Table 2. Intensity ratios of phrases and their final syllables. σ marks the final syllable of a corresponding phrase

Researched unit	S1	S2	S3
ip_N:σ(ip_N)	1:0.29	1:0.27	1:0.24
ip_F:σ(ip_F)	1:0.60	1:0.21	1:0.22
ip_S:σ(ip_S)	1:0.34	1:0.16	1:0.19
IP_N:σ(IP_N)	1:0.03	1:0.13	1:0.03
IP_F:σ(IP_F)	1:0.03	1:0.12	1:0.01
IP_S:σ(IP_S)	1:0.03	1:0.05	1:0.03

At the end of an intonation phrase, the intensity decreases more than at the end of an intermediate phrase. In 5 out of 9 cases, the intensity of the last syllable is 0.03 times lower than that of the entire intonation phrase. It can be argued that the decrease in intensity at the end of an intonation phrase is even less influenced by reading rate than the decrease in intensity at the end of an intermediate phrase. This conclusion is supported by the data from speakers S1 and S3, where the decrease in intensity is equal. Speaker S2 maintains a similar intensity difference during natural and fast reading, while slow reading results in a greater intensity decrease.

To claim that a decrease in intensity at the end of a phrase can reliably serve as an indicator of a phrase boundary, it is necessary to determine whether this decrease is not merely characteristic of word-final positions in general. Only the vowels [ɑː] and [eː] were used at the end of both phrase types and at the end of a word in a phrase-internal position. Therefore, a comparison was conducted using the data of these vowels.

The results confirm the conclusion drawn from comparing the results for the entire phrase and its last syllable. In all cases, the intensity at the end of a phrase is lower than that of the last syllable of a word not at the end of a phrase (see Table 3). Furthermore, the last vowel of an intonational phrase (V_{IP}) exhibits lower intensity than that of an intermediate phrase (V_{ip}), and even more so than the last vowel of a word within a phrase ($V_{\#}$).

The decrease in intensity at the end of an intonation phrase is not influenced by reading rate, as speakers consistently decrease intensity in this position regardless of the reading rate. However, the data for intermediate phrases shows greater variability. Speaker S1 demonstrates the greatest decrease in intensity during natural reading, a moderate decrease in slow reading, and the least decrease in fast reading. Speaker S2 shows the largest decrease in intensity during fast reading, the smallest during slow reading, and a moderate decrease during natural reading. Speaker S3 exhibits minimal variation in intensity decrease, with the largest decrease during fast reading, slightly less during natural reading, and the least during slow reading.

Table 3. Intensity ratios of vowels in the final syllable. V marks vowel

Researched unit	**S1**	**S2**	**S3**
$(V_{\#}:V_{ip}:V_{IP})_N$	1:0.18:0.02	1:0.24:0.06	1:0.16:0.01
$(V_{\#}:V_{ip}:V_{IP})_F$	1:0.63:0.02	1:0.15:0.03	1:0.12:0.004
$(V_{\#}:V_{ip}:V_{IP})_S$	1:0.43:0.04	1:0.19:0.02	1:0.18:0.02

The difference between the intensity of the entire phrase and that of the final syllable is statistically significant in all cases. This suggests that the decrease in the intensity of the final syllable serves as a reliable indicator of phrase boundaries.

The present study did not identify consistent patterns related to reading rates, likely due to the limited amount of material examined.

3.3 Fundamental Frequency

Changes in F0 at the end of a phrase can either fall or rise. A decline in F0 is typical of utterances that express a complete thought, while an increase in F0 often signals a question or an incomplete thought. The research material did not include intonation phrases ending with rising F0, although 21% of the intermediate phrases ended with rising F0. Only phrases with a falling F0 were selected for this study, as those with rising F0 require semantic and syntactic analysis and were therefore excluded at this stage of the research.

F0 (as well as voice quality features) were undefined at the end of 45% of the intonation phrases due to creaky voice or devoicing. This also indicates the boundary of the phrase, as all three speech production systems—the respiratory, phonatory, and articulatory—are returning to a neutral position. These cases were also excluded from this analysis.

The data obtained from our material indicate that the F0 of the final syllable is, on average, one-fifth lower than the F0 of the entire intermediate and intonation phrase (see Table 4). Speaker S2 lowers F0 consistently at the end of both types of phrases, with minimal dependence on the reading rate. In contrast, the other two speakers show variability in F0 lowering.

At the end of intermediate phrases, speaker S3 lowers F0 equally (i.e. one-fifth) in natural and slow reading but reduces it by nearly two-fifths in fast reading. Speaker S1 halves the F0 in fast reading but lowers it similarly in slow and natural reading (nearly one-fifth).

At the end of intonation phrases, speaker S3 halves the F0, and this is independent of the reading rate. Whereas for speaker S1, no significant decrease is observed. Statistically significant differences are marked with an asterisk in Table 4.

Table 4. F0 ratios of phrases and their final syllables

Researched unit	**S1**	**S2**	**S3**
ip_N:σ(ip_N)	1:0.8*	1:0.8*	1:0.8
ip_F:σ(ip_F)	1:0.5*	1:0.8	1:0.6*
ip_S:σ(ip_S)	1:0.9	1:0.8*	1:0.8
IP_N:σ(IP_N)	1:0.9	1:0.8*	1:0.5*
IP_F:σ(IP_F)	1:1	1:0.8*	1:0.5*
IP_S:σ(IP_S)	1:0.9	1:0.9	1:0.5*

Similar tendencies are also shown by the analysis of the last vowels F0 ratios in the middle and at the end of the phrase (see Table 5). In the S1 speaker recordings, the F0 of the last vowels at the end of the phrases is one tenth lower than in the middle of the phrases, and this lowering is not dependent on the type of phrase or the reading rate. Although, as can be seen from Table 4, analysing F0 for all (not only identical) last syllable centres and the entire intermediate phrase more significant differences were observed, especially in the case of fast reading.

Speaker S2 vowel data shows more variety than the data for the whole phrase and all the last syllables. There is no F0 lowering of the last vowel in slow reading, and much more (almost 30% and 40%) lowering of the last vowels of the phrase in fast reading.

Speaker S3 shows very similar results when comparing the last vowels with each other and with the F0 of the whole phrase only his speech failed to identify F0 at the end of the intonation phrase in the fast reading.

Table 5. F0 ratios of vowels in the final syllable.

Researched unit	S1	S2	S3
$(V_{\#}:V_{ip}:V_{IP})_N$	1:0.9:0.9	1:0.9:0.8	1:0.7:0.6
$(V_{\#}:V_{ip}:V_{IP})_F$	1:0.9:0.9	1:0.7:0.6	1:0.6:–
$(V_{\#}:V_{ip}:V_{IP})_S$	1:0.9:0.9	1:1:1	1:0.7:0.5

Overall, F0 is consistently lowered at the end of a phrase, but there were no clear patterns in the F0 changes at the end of intonation phrases that were related to reading rate. In the intermediate phrases F0 was lowered more in fast reading than in natural and slow reading, while F0 lowering in slow and natural reading was very similar.

3.4 Duration of the Final Sound

The analysed text is relatively short, and there are few same sounds used in the positions under investigation. Nevertheless, this feature was included in the current study to analyse the duration of [eː] and [ɑː] vowels. It should be noted that the duration of long vowels at the end of a word is particularly complex in Lithuanian due to frequent shortening and reduction. However, since the speakers involved in the study were professional actors, their vowels retained the duration and tension characteristic of Standard Lithuanian. Otherwise, vowel duration results would need to be interpreted with caution.

Although it was anticipated that the longest vowels would occur at the end of intonation phrases, the results do not support this hypothesis. The data are ambiguous but certain tendencies can be seen (see Table 6). All speakers lengthen the final vowels of intonation phrases by an average of one-third during natural reading. Speakers S2 and S3 also lengthen the final vowels of intermediate phrases, and these vowels are lengthened more than the final vowels of intonation phrases.

During fast reading, none of the speakers significantly lengthen the final vowels of intermediate phrases, and only speaker S2 exhibits lengthening at the end of intonation phrases.

In slow reading, all speakers, on average, lengthen the final vowels of intermediate phrases by one-third. However, the final vowels of intonation phrases are lengthened by only one-tenth.

Statistically significant differences were observed in three cases, which are marked with an asterisk in Table 6. This can be attributed to the large variance in the length of vowels at the end of phrases, as indicated by the high standard deviation.

Table 6. Duration ratios of vowels in the final syllables

Researched unit	**S1**	**S2**	**S3**
$(V_{\#}:V_{ip}:V_{IP})_N$	1:1:1.4*	1:1.5*:1.3*	1:1.3:1.2
$(V_{\#}:V_{ip}:V_{IP})_F$	1:1.1:1	1:0.9:1.3	1:1:0.9
$(V_{\#}:V_{ip}:V_{IP})_S$	1:1.3:1.1	1:1.2:1.1	1:1.4:1.1

Therefore, while vowel lengthening at the end of a word can reliably signal the end of a phrase, the absence of vowel lengthening does not reliably indicate that it is not the end of a phrase. Furthermore, only one consistent pattern was observed regarding reading rate: speakers do not prolong final vowels during fast reading.

3.5 Voice quality features

Higher jitter values indicate greater frequency perturbations, which are expected to be highest at the end of an intonation phrase, lower at the end of an intermediate phrase, and lowest at the end of a word in a phrase-internal position.

The findings of this research reveal that jitter values at the end of intonation phrases are higher compared to those in the middle of a phrase or at the end of an intermediate phrase, and jitter values at the end of intermediate phrases are higher compared to those in phrase-internal positions. However, due to large variation in the data, only one-third of the differences are statistically significant (9 out of 27; these are indicated with a dash in Table 7). Furthermore, in 11 out of 36 cases, jitter values were undefined at the end of intonation phrases due to devoicing, which is particularly characteristic of speaker S3. These cases were excluded from the calculation of jitter and shimmer ratios and statistical significance. As mentioned earlier, devoicing clearly signals the end of a phrase. No consistent patterns of jitter changes related to reading rate were observed.

Shimmer measurements, which indicate increased amplitude perturbations, are expected to be higher at the end of larger prosodic units. However, our analysis does not fully support this hypothesis. While shimmer values at the end of intonation phrases are higher than those in phrase-internal positions, the intermediate phrase data are contradictory. In only half of the cases (5 out of 9) are shimmer values higher at the end of intermediate phrases than in phrase-internal positions. Moreover, in 7 out of 9 cases, shimmer values at the end of intermediate phrases are lower than those at the end of intonation phrases.

Only three cases (S1 and S3 natural reading, S2 slow reading) align with the general pattern: shimmer values are highest at the end of intonation phrases, lower at the end of intermediate phrases, and lowest in phrase-internal positions. Unfortunately, none of these differences, nor other differences observed, are statistically significant.

As with jitter, no clear correlation was found between shimmer values and reading rate.

The HNR data show that vowels at the end of an intonation phrase tend to be produced with less periodicity, regardless of the reading rate. The HNR of the final vowel in an intonation phrase is statistically significantly lower than that of vowels in phrase-internal positions (except for speaker S1 during natural and fast reading, where this parameter is equal).

The data for intermediate phrases are less consistent. In 4 out of 9 cases, the HNR at the end of an intermediate phrase is lower than in phrase-internal positions, while in 5 cases, it is higher. Thus, it can be argued that HNR cannot reliably differentiate between intonation and intermediate phrases.

Table 7. Jitter, shimmer, and HNR ratios of vowels in the final syllables

Researched unit	**S1**	**S2**	**S3**
jitter			
$(V_{\#}:V_{ip}:V_{IP})_N$	1:2.3*:3.5*	1:1.1:2.5*	1:1.8:2
$(V_{\#}:V_{ip}:V_{IP})_F$	1:1.3:1.9*	1:1.1:4.2*	1:1.4:–
$(V_{\#}:V_{ip}:V_{IP})_S$	1:1.6:1.9*	1:1.8*:2.1*	1:1.8:8*
shimmer			
$(V_{\#}:V_{ip}:V_{IP})_N$	1:1.3:2.2	1:0.9:1.7	1:1.2:2.6
$(V_{\#}:V_{ip}:V_{IP})_F$	1:0.4:3	1:0.7:0.9	1:2.1:–
$(V_{\#}:V_{ip}:V_{IP})_S$	1:1:1.2	1:1.6:2	1:1.4:1.2
HNR			
$(V_{\#}:V_{ip}:V_{IP})_N$	1:0.7:0.3	1:2.1:1	1:0.6:0.2
$(V_{\#}:V_{ip}:V_{IP})_F$	1:1.1:0.2	1:1:1	1:0.5:–
$(V_{\#}:V_{ip}:V_{IP})_S$	1:1.8:0.4	1:0.5:0.3	1:1.5:0.2

To summarise, jitter can be used to identify not only the end of a phrase but also the type of phrase. Shimmer and HNR can indicate the end of an intonation phrase; however, their effectiveness in marking the end of an intermediate phrase should be considered with caution. This study suggests that the reading rate does not have a significant effect on voice quality at the end of a phrase.

4. Conclusions

The results of this stage of the research lead to the following preliminary conclusions:

- **Pauses** consistently mark the boundaries of intonation phrases but are not always necessary for intermediate phrases. An increase in reading rate reliably shortens pauses, while a decrease extends them for both intonation and intermediate phrases. However, pauses between intermediate phrases are less shortened during fast reading and are more extended during slow reading compared to pauses between intonation phrases. There may also be fewer pauses between intermediate phrases in fast reading and more in slow reading.
- **Intensity decreases** are a reliable indicator of phrase boundaries. Based on the level of intensity decrease, it is possible to identify the type of phrase. For intonation phrases, the difference in intensity between the last syllable and the whole phrase is nearly twice as great as for intermediate phrases. Reading rate does not appear to influence the level of intensity decrease at the end of intonation phrases. However, it is unclear whether reading rate affects intensity decrease at the end of intermediate phrases.

- **F0 lowering** consistently occurs at the end of phrases, but its relationship to phrase type and reading rate is not always clear. In fast reading, F0 lowering at the end of intermediate phrases may be more pronounced. Additionally, devoicing of the final syllable reliably signals the end of an intonation phrase.
- **Lengthening of sounds** at the end of a phrase serves as an indicator of phrase boundaries. The data show one clear tendency related to reading rate: speakers accelerate reading by reducing or omitting lengthening of the final vowels of a phrase. This suggests that speakers speed up their reading rate not only by shortening or omitting pauses but also by avoiding the prolongation of final vowels.
- **Jitter** results support the assumption that an increase in jitter is a reliable indicator of phrase boundaries. Jitter can also help distinguish phrase types, as larger intonation units exhibit higher jitter values, indicating a characteristic frequency perturbation.
- **Shimmer** results did not confirm predicted patterns related to phrase type. While shimmer values are higher at the end of intonation phrases compared to phrase-internal positions, intermediate phrase data are inconsistent. This is likely due to the limited scope of the material examined.
- **HNR** data show that vowels at the end of intonation phrases tend to be produced with less periodicity. However, HNR does not consistently signal the end of intermediate phrases.
- The study found **no significant effect of reading rate on voice quality**, as parameters like jitter, shimmer, and HNR remained relatively stable across different reading rates.

All the analysed features signal the end of a phrase to varying degrees. However, according to the data from this study, only the duration-related features—namely, the existence of pauses, their duration, and the lengthening of the final sound—respond to changes in reading rate, while the other features remain relatively constant.

The next stage of the study aims to verify the tendencies identified here using a larger data sample, which will include longer and more varied texts, as well as recordings of spontaneous speech. Additionally, the qualitative characteristics of final sounds should be investigated. Phrase onset data in Lithuanian has not yet been thoroughly analysed, except in perceptual experiments (Kazlauskienė et al., 2023, pp. 87–155). Therefore, these phenomena should be incorporated into future research on prosodic phrasing. The effect of discovered pausing duration patterns on the naturalness of synthetic speech at varying reading rates will be examined using our experimental speech synthesis system.

References

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, Institute of Phonetic Sciences, University of Amsterdam, *Proceedings* 17, p. 97-110, available at https://www.researchgate.net/publication/2326829_Accurate_Short-Term_Analysis_Of_The_ Fundamental_Frequency_And_The_Harmonics-To-Noise_Ratio_Of_A_Sampled_Sound

Boersma, P. (2009). Should *jitter* be measured by peak picking or by waveform matching?, *Folia Phoniatrica et Logopaedica* 61: p. 305–308. doi: 10.1159/000245159.

Boersma, P., Weenink, D. (2018). *Praat: doing phonetics by computer, version 6.035*. www.praat.org [computer program], available at https://www.fon.hum.uva.nl/praat/

Brugos, A., Breen, M., Veilleux, N., Barnes, J., Shattuck-Hufnagel, S. (2018). Cue-based annotation and analysis of prosodic boundary events, *Proceedings of the 9th international conference on Speech Prosody 2018*, p. 245-249. doi: 10.21437/SpeechProsody.2018-50.

Crowhurst, M. J. (2018). The joint influence of vowel duration and creak on the perception of internal phrase boundaries, *Journal of Acoustic Society of America*, 143(3): EL147. doi: 10.1121/1.5025325.

Dereškevičiūtė, S., Kazlauskienė, A. (2022). Prosodic phrasing in Lithuanian: preparatory study, *Baltic J. Modern Computing*, 10(3), 317–325. doi:10.22364/bjmc.2022.10.3.05.

Harrington Stack, C., Watson, D. G. (2023). Pauses and Parsing: Testing the Role of Prosodic Chunking in Sentence Processing. *Languages*, 8, no. 3, 157, https://doi.org/10.3390/languages8030157.

Kazlauskienė, A., Dereškevičiūtė, S., Sabonytė, R. (2023). *Standard Lithuanian intonation: phrase centre, boundaries, and marking* (in Lithuanian), https://doi.org/10.7220/9786094675782.

Kohler, K. J. (2000). Linguistic and paralinguistic functions of non-modal voice in connected speech, *Proceedings of the 5th Seminar on Speech Production: Models and Data*, 89-92, available at https://www.ipds.uni-kiel.de/kjk/pub_exx/kk2000_2/51.pdf

Liu, R., Liu, B., Li, H. (2023). Emotion-Aware Prosodic Phrasing for Expressive Text-to-Speech. *arXiv preprint arXiv:2309.11724*, https://doi.org/10.48550/arXiv.2309.11724.

Michelas, A., D'Imperio, M. (2010). Durational cues and prosodic phrasing in French: Evidence for the intermediate phrase. *Speech Prosody*, 2010-196. doi: 10.21437/SpeechProsody.2010-196.

Napoleão de Souza, R. (2023). Segmental cues to IP-Initial boundaries: Data from English, Spanish, and Portuguese. *Prosodic boundary phenomena*, p. 35-86. Language Science Press. doi: 10.5281/zenodo.777752.

Peters, B. (2003). Multiple cues for phonetic phrase boundaries in German spontaneous speech, *Proceedings 15th ICPhS*, Barcelona, p. 1795-1798, available at: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_1795.pdf

Petrone, C., Truckenbrodt, H., Wellmann, C., Holzgrefe-Lang, J., Wartenburger, I., Höhle, B. (2017). Prosodic boundary cues in German: Evidence from the production and perception of bracketed lists, *Journal of Phonetics*, 61, p. 71-92, https://doi.org/10.1016/j.wocn.2017.01.002.

Steffman, J., Kim, S., Cho, T., Jun, S. (2024). Speech rate and prosodic phrasing interact in Korean listeners' perception of temporal cues. *Speech Prosody 2024,* p. 1090–1094, https://doi.org/10.21437/speechprosody.2024-220.

Volin, J., Šturm, P., Skarnitzl, R., Bořil, T. (2024). *Prosodic phrase in spoken Czech*. Prague: Charles University, Karolinum Press.

Żygis, M., Tomlinson, J., Petrone, C., Pfütze, D. (2019). Acoustic cues of prosodic boundaries in German at different speech rate, *Proceedings of ICPhS 2019*, p. 999-1003, Melbourne, Australia. ⟨hal-02097707⟩

Received January 21, 2025, accepted April 16, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 2, 419-435
https://doi.org/10.22364/bjmc.2025.13.2.07

Optical Character Recognition Tool for Georgian Handwritten Text Recognition Based on YOLOv8

Ana CHIKASHUA[1], Maia ARCHUADZE[2], Magda TSINTSADZE[2], Manana KHACHIDZE[2]

[1] Industrial Analytics IA GmbH, Berlin, Germany
[2] Department of Computer Sciences, Iv. Javakhishvili Tbilisi State University, Georgia

ana.chikashua5118@ens.tsu.edu.ge, maia.archuadze@tsu.ge, magda.tsintsadze@tsu.ge, manana.khachidze@tsu.ge

ORCID 0009-0008-9930-9249, ORCID 0000-0002-9484-1016, ORCID 0000-0001-9316-3295, ORCID 0000-0003-4545-2679

Abstract. Optical character recognition is one of the important tasks in the fields of artificial intelligence, computer vision, natural language processing tasks, and machine learning. It remains a major challenge for less common languages such as Georgian. In addition, we face the complexity and variety of writing forms characteristic of this language, caused by the integration of three different scripts implemented over time.

Discussed in the presented work are three models: convolutional neural networks (CNN), ResNet, and Yolov8, through which the OCR system was created. The methodology presented in this paper is the first attempt to create for the Georgian language an OCR system based on the Yolov8 platform. The paper analyzes the results obtained by all three methods, based on which we conclude that the accuracy of YOLOv8 compared to CNN and ResNet was improved.

After training and evaluating the aforementioned models on the training dataset, it was determined that the YOLOv8 model yielded the best results. The CNN model, incorporating convolutional, pooling, and dropout layers, achieved an accuracy of 0.9230, whereas the ResNet model achieved an accuracy of 0.9302. In contrast, YOLOv8 achieved an accuracy of 0.98848, with a loss value of 0.04479. These results indicate that YOLOv8 performed exceptionally well during live testing. Furthermore, it is noteworthy that during the testing phase, YOLOv8 demonstrated remarkable speed in generating results.

Keywords: OCR, YOLOv8, CNN, ResNet

1. Introduction

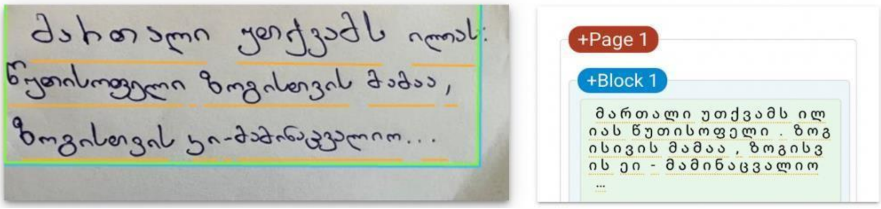
In our view, the low accuracy of recognizing Georgian texts by OCR systems is accounted for by several factors. One of them is the uniqueness of the language itself, as its script is a combination of three different scripts: **Asomtavruli**, **Nuskhuri,** and **Mkhedruli**, each with a graphic style of its own. This leads to challenges with variations in handwritten styles. Due to this peculiarity, the same symbol can be tilted, rotated, or presented in different forms. Some letters may have similar outlines in handwritten text, making it difficult to distinguish them. For instance, "ღ" and "დ," "ღ" and "ლ," "უ" and "ყ," "ვ" and "გ." Additionally, the position of symbols is crucial, as in Georgian, symbols are written above and below the line, each with different meanings. Variations in different fonts further affect the accuracy of the recognition system.

2. Experimental

In addition to notable commercial OCR systems such as Azure OCR, Amazon Textract, and Google OCR, various open-source OCR tools are available today. Whereas this technology is the latest, there are still no OCR products that can recognize all types of text with 100% accuracy.

Pytesseract is a Python package that simplifies the use of Tesseract. It can read all image types supported by the Pillow and Leptonica visualization libraries, including jpeg, png, gif, bmp, tiff, and other libraries (Shubham 2020). EasyOCR can currently recognize text in more than 80 languages, including English, German, Hindi, Russian, and others. (Rosebrock 2020). It is worth noting that most open-source OCR systems do not support the Georgian language at all, (for example, ABBYY FineReader/ABBYY Cloud OCR SDK and OnlineOCR.net.). Based on our experiments, we can conclude that those systems that do support the Georgian language have a low recognition accuracy.

Below are the results of an experiment on a Georgian text for those OCRs that support the Georgian language. Google Cloud Vision's output for a specific example in the Georgian language demonstrates some issues, such as missing characters or incorrect character recognition.

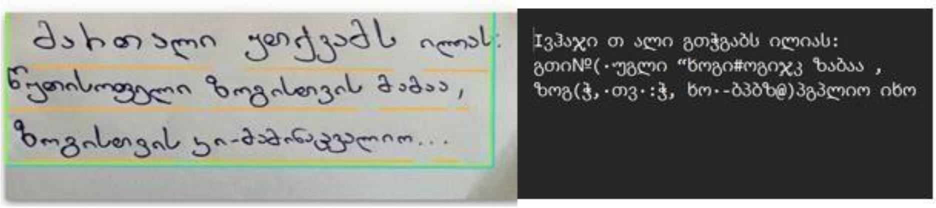


მართალი უთქვამს ილიას: წუთისოფელი ზოგისთვის მამაა, ზოგისთვის კი-მამინაცვალიო...

მართალი უთქვამს ილ
იას წუთისოფელი . ზოგ
ისივის მამაა , ზოგისვ
ის ეი - მამინაცვალიო
...

Figure 1. Example of Google Cloud Vision

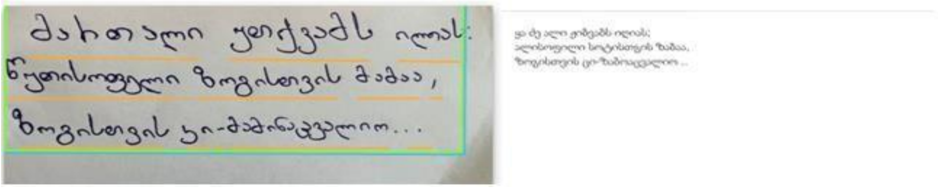
Tesseract OCR: In this case, the structure of the text and summary are unclear.



```
Iვჰაჯი თ ალი გთჭგაბს ილიას:
გთიი№(·უგლი “ხოგი#ოგიჯკ ზაბაა ,
ზოგ(ჭ,·თვ·:ჭ, ხო·-ბპბზ@)პგპლიო იხო
```

Figure 2. Example of Tesseract OCR

The result of using i2OCR in this case contains several errors.



```
ყა ძე ალი გიზგაბს ილიას;
ალისოფელი სოგისთვის ზაბაა,
ზოგისთვის კი-ზაზინაცვალი ...
```

Figure 3. Example of i2OCR

Georgian language is also supported by quite a popular software Nebo (Microsoft 2023) which is a whiteboard-type application available on Android, iOS, and web platforms. It offers remarkable speed and accuracy in calculations. However, Nebo requires that the entries be created directly on the screen of your device. It is important to note that Nebo uses ICR (Intelligent Character Recognition), not OCR (Optical Character Recognition), which typically converts existing text entries into an editable format.

In addition, we ought to mention that scientific studies have implementation regarding ~~to~~ this task for the Georgian language using a convolutional neural network VGG 16 and GoogleNet architecture trained with 200,000 data (Soselia, et al. 2018). In our research, we use CNN, ResNet, and YOLOv8. YOLOv8 represents the latest version, released in 2023, which has provided us with quite good results. The limitation of the previous research is that it only works on individual characters and doesn't consider word recognition.

3. Methodology

Study Limitations and the Process Description:

It is also important to note that the successful operation of the system is contingent upon several key assumptions outlined in this paper.

The input image is a full, straight, (fairly) standard-sized sheet of paper with visible borders;

The text is written horizontally with non-intersecting lines;

Distance between words should be taken into account;

We should also pay attention to the fact that the symbols stand separately and not linked with one another.

The project is divided into two parts. In the first scenario, the model is directly trained, which comprises the following stages:

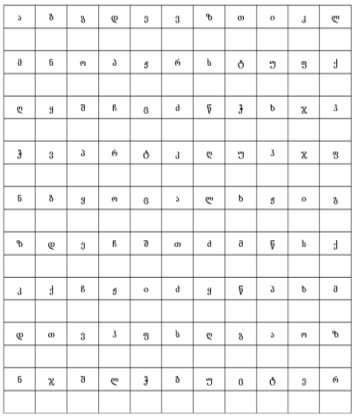
- Data gathering and labelling;
- Data processing (to format images for the model's input);
- Data augmentation;
- Data partitioning (into training, validation, and testing datasets);
- Model training, validation, and testing;
- Results analysis;
- Model deployment.

This process results in a model capable of working with various inputs, specifically numbers, symbols, and words.

In the second scenario, the focus shifts to segmenting text into individual symbols. Despite the complexity of each step, as elaborated below, the objective is to exclusively provide our trained model with characters it can accurately interpret. Below we consider each of the steps in detail.

3.1. Data Gathering and Labeling

We started the research process by collecting data. We create template that consist of Georgian characters and asked colleges, students, friends to fill this and then create database based on this data. The main idea was to have already labelled characters as many as we can. Working with this model required a database of Georgian characters. This database is a certain table filled by different people. The base consists of 48625 characters.



ა	ბ	გ	დ	ე	ვ	ზ	თ	ი	კ	ლ
მ	ნ	ო	პ	ჟ	რ	ს	ტ	უ	ფ	ქ
ღ	ყ	შ	ჩ	ც	ძ	წ	ჭ	ხ	ჯ	ჰ
ჭ	ვ	პ	რ	ტ	კ	ღ	უ	ჰ	ჯ	ფ
ნ	ბ	ყ	ო	ც	ა	ლ	ს	ჟ	ი	გ
ზ	დ	ე	ჩ	შ	თ	ძ	მ	წ	ს	ქ
კ	ქ	ჩ	ჟ	ი	ძ	ყ	წ	პ	ხ	მ
ღ	თ	ვ	ჰ	ფ	ს	ღ	გ	ა	ო	ზ
ნ	ჯ	შ	ლ	ჭ	ბ	უ	ც	ტ	ე	რ

Figure 4. Data collection table

The MNIST (Modified National Institute of Standards and Technology) database, which consists of more than 70,000 scrambled images of numbers, was used to recognize the numbers.

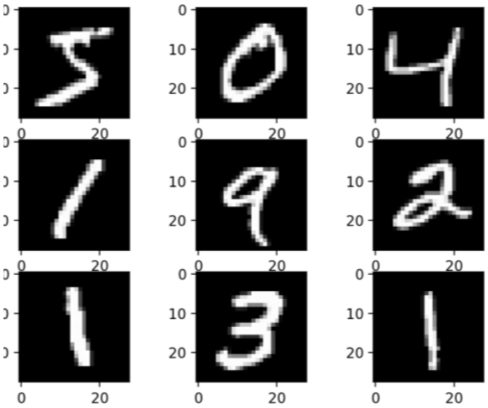


Figure 5. The MNIST database

Data for two models (ResNet and CNN) were taken in the following ratio: 80% for training, 10% for validation, and 10% for testing.

3.2. Data processing

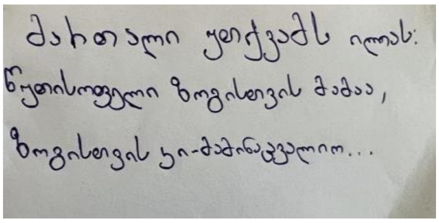
The above-mentioned pre-processing system takes an image of a complete handwritten page and returns images of words. Images of individual words are crucial because they can be fed into Handwritten Text Recognition (HTR) systems.

Preprocessing consists of the following stages:

1. Initial image processing;
2. Edge Removing;
3. Line Removing;
4. Gray-scaling and blurring;
5. Segmentation Edges(Ganny);
6. Filter Connected Components;
7. Segmentation (GMM-based text lines detection);
8. Selection of clusters; 10. Segmentation into words.

Let's consider each of them in detail.

3.2.1. Initial image processing



მართალი უთქვამს ილიას:
წუთისოფელი ზოგისთვის მამაა,
ზოგისთვის კი-მამინაცვალიო...

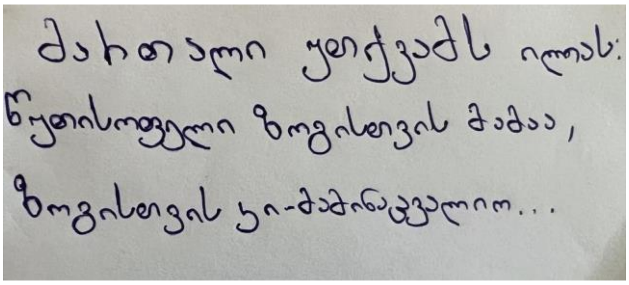
Figure 6 . Original

Ideally, the image should encompass the entire page with all page borders visible. The page should also be as straight as possible and well-lit. 'Well-lit' means that the image is neither too dark nor too bright.

Initially, image processing is required, in order to get the character of a text which would be an input for a model. It is more like an image processing and is the same for each language. This step which includes tasks such as edge detection, grayscale conversion, noise reduction, and the retention of only the components necessary for recognition. These steps are crucial for enhancing the quality of the input data and improving the accuracy of subsequent text recognition. Here’s a detailed approach tailored for Georgian handwritten text.

3.2.2. Edge Removing

The first step in preprocessing is to find the page contour, outline the page (leaving only the page). This can be done quite easily by using cv2.findContours (Rosebrock 2020)] and cv2.approxPolyDP to find the largest contour whose approximate shape is a square. cv2.approxPolyDP is a function of the OpenCV library used to smooth the contour by reducing the number of points while preserving the overall shape of the contour. This reduces computational complexity and improves efficiency. The purpose of approxPolyDP is to replace a polygon with another simpler (less squiggly) polygon.

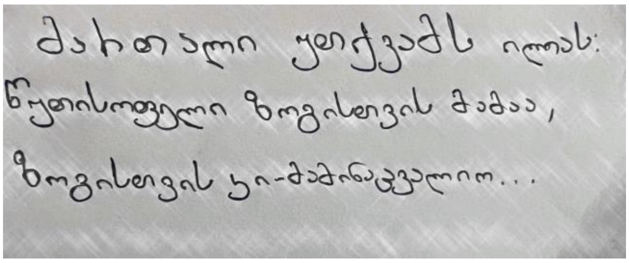


მართალი უთქვამს ილიას:
წუთისოფელი ზოგისთვის მამაა,
ზოგისთვის კი-მამინაცვალიო...

Figure 7. Edge removed

3.2.3. Line Removing

Removing lines on a drawn page is crucial because these lines can meet each other. In related components (these components are discussed below)) they cannot be filtered off. The mentioned methodology uses hough lines and Fourier transform to remove the lines. Doing this several times improves the result of cv2.canny. Therefore, it is necessary to remove them at the beginning. This method does not completely remove the lines, but significantly cuts the line components of the page. (Although in the example we discussed the line sheet is not used, it is still useful, because it gives us quite good results for removing noise).



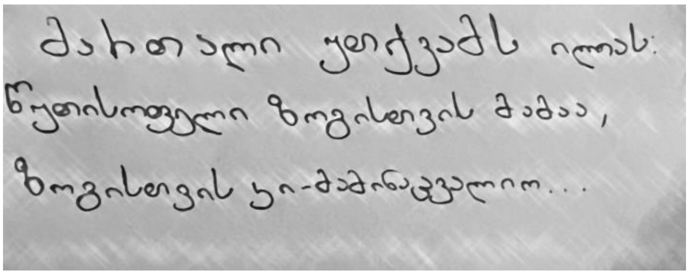
მართალი უთქვამს ილიას:
წუთისოფელი ზოგისთვის დედაა,
ზოგისთვის კი-დედინაცვალიო...

Figure 8. Line removing

3.2.4. Gray-scaling and blurring

The image then becomes gray and blurry for cv2.canny. For blurring we used Gaussian blurring (cv2.GaussianBlur) which in terms of image processing, smoothed out any sharp edges in images and excessive blurring is minimized. In (Microsoft 2023) Gaussian smoothing is used to remove noise that approximately follows a Gaussian distribution.

The end result is a less blurry image, but more "naturally blurry" than any other image that we could get by using other methods.



მართალი უთქვამს ილიას:
წუთისოფელი ზოგისთვის დედაა,
ზოგისთვის კი-დედინაცვალიო...

Figure 9. Gray and Blurred

3.2.5. Segmentation Edges (Canny)

This process allows us to localize the text, and we get sharp contours around the text. To get edges of characters we used cv2.canny. This is a good way to detect text in an image.

Invariant parameters are used here to detect all text - any text that is not detected will not make it to the final result, so detecting everything, even noise, is fine. As a result:

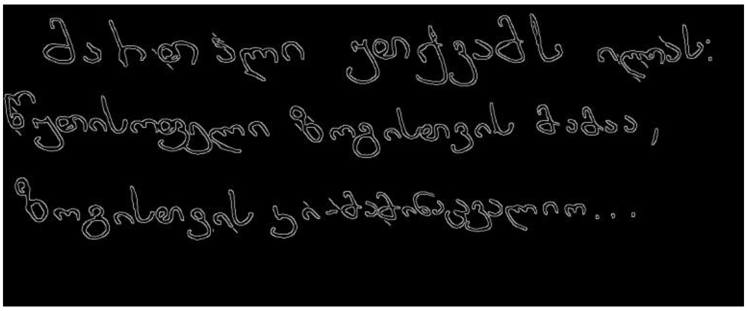


Figure 10. Segmentation Edges

3.2.6. Filter Connected Components

In image processing, connected components refer to distinct regions or sets of pixels in a binary image that are related to each other based on some criterion. These connected components can represent individual objects, regions, or parts of an image.

To find the components, we used the OpenCV function connected Components With Stats, the ConnectComponentsWithStats function being a part of OpenCV, a popular computer vision library.

Area filtering: Area filtering is a common technique used to filter connected components (regions) in an image based on their area, which refers to the number of pixels in each component. The idea is to keep components that meet a specific size criterion and remove those that are too small or too large.

Filters are based on the height, width, and area of the resulting images. Filtering is performed according to different components;

Filtering the bounds of connected components is based on their dimensions. After this type of filtering, only the connected components that are useful for defining or separating text lines remain;

Filtering of stand-alone 'stray' components: This process once again filters connected components, this time removing 'stray' components along the y-axis in addition to removing lines.

Although simple and unlikely to remove many components, this filtering is crucial for detecting text lines.

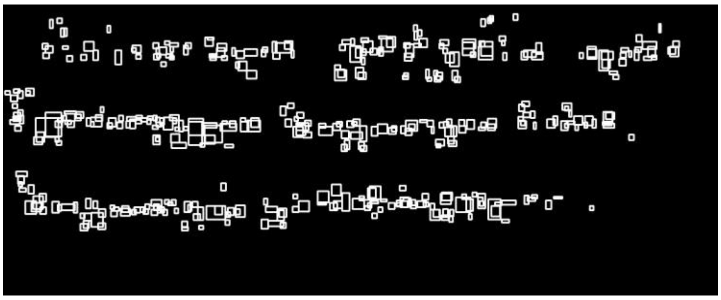


Figure 11. Components with bounding boxes

3.2.7. Segmentation (GMM-based text lines detection)

We used the sklearn.mixture.GaussianMixture function to see how many horizontal lines of the text are there and where there are - this is the assumption of "Horizontal lines". Horizontal lines are not part of the textual content and can be ignored during the recognition process. This helps improve the accuracy of OCR results and prevents horizontal lines from being mistakenly converted to text. Through using it, text lines were grouped according to their y values.

The Gaussian Mixture Model (GMM) is a probabilistic model used to estimate clustering and distribution. It is a generative model which assumes that the data points are generated from several Gaussian distributions. Each Gaussian distribution represents such as mean and covariance, that best explain the data observed.

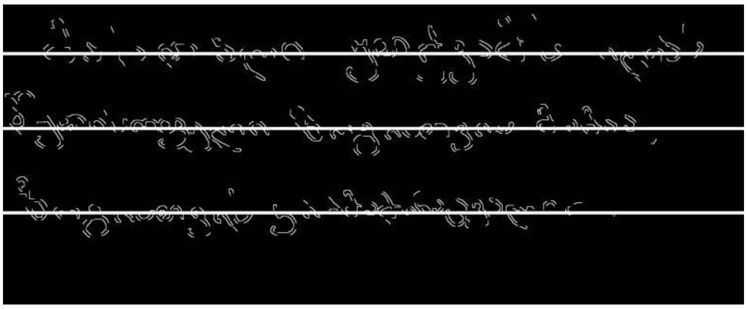


Figure 12. "Horizontal Lines"

3.2.8. Selection of Clusters

To select the number of clusters, we used an approach similar to the elbow method with a lower bound (error) reduction. It is a technique used to determine the optimal number of clusters in a clustering algorithm. The idea of the elbow method is to plot the within-cluster sum of squares (WCSS) for different values of k (number of clusters) and identify the "elbow point" on the graph.

In Gaussian Mixture Model (GMM) clustering, the lower bound is the log-likelihood lower bound (also known as the evidence lower bound (ELBO) in variational inference). This is a value that:

1. Measures how well the model fits the data given a certain number of components.
2. Is unitless, but it's in the log-probability space, typically representing the log-likelihood of the observed data under the GMM model.
   X-axis: Number of GMM components (clusters).
   Y-axis: Lower bound (log-likelihood).

A higher (less negative) lower bound means a better fit. The sharp increase from 1 to - 3 components shows that the model fits the data much better as more clusters are added. After 3 or 4 components, improvements are marginal, so adding more clusters doesn't significantly improve the model.

We determined the optimal number of clusters, 3 in our case.

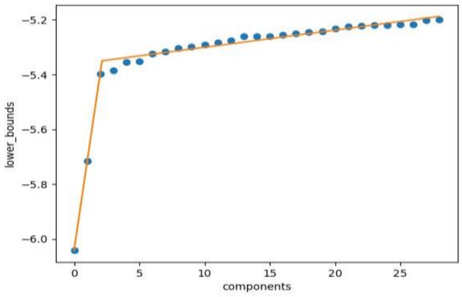


Figure 13. Selection of Clusters

3.2.9. Segmentation into words

One of the main challenges was segmentation into words. After dividing the document into lines, it must be further divided into words. To accomplish this, we find the minimum spacing interval between two words in proportion to the width of the page and the words themselves. This function estimates the minimum spacing threshold that can be used to distinguish individual words in a line of text. It analyzes the horizontal gaps between characters or elements on a line and compares them relative to the page width. By calculating the expected number of words per line based on line width proportions and filtering out lines with insufficient data, it derives an average of the largest spacing values—interpreted as spaces between words. This threshold is crucial for accurately segmenting text into words during document layout analysis. Based on this value, we segment the sentence into individual words. This process entails two main challenges: determining the locations of the separating spaces and obtaining the final word images that can serve as input for Handwriting Text Recognition (HTR).

3.3. Data Augmentation

During the training, we lacked data, which naturally had a significant impact on the accuracy of the model. CNN, ResNet50, required data augmentation using the following techniques:

Morphological changes, its implementation was necessary to correct text lines. Adding noise, which involves removing black pixels or adding white pixels to the image. In the process of research, various methods were used, and in the case of a specific method, we got the following results: RandomRain with a black drop greatly damages the image and makes it more difficult to distinguish. So we tried to add less of this type of augmentation. RandomShadow will blur text with lines of varying intensity. PixelDropout turns random pixels black.

Rotate- ShiftScaleRotate: This parameter is quite problematic. We tried to avoid cutting the text from the initial form. During this process both shape changes and rotation takes place. We paid attention to the fact that the image obtained as a result of rotation corresponds to the predetermined dimensions, which the input image of the model is supposed to have.

Image blurring - to lighten the dark pixels in the image.

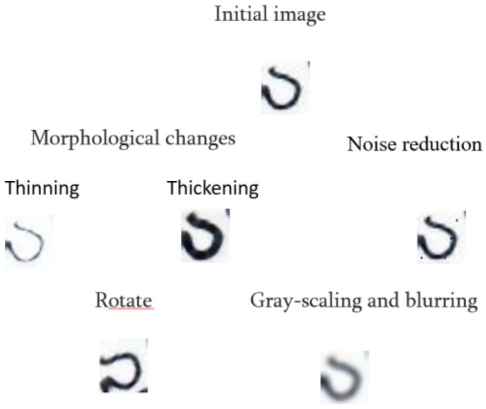


Figure 14. Data Augmentation

As for YOLOv8, we needed no augmentation process because it performs this process by itself and uses Mosaic Data Augmentation for data augmentation. Mosaic Data Augmentation is a simple data augmentation technique in which four images are combined and transferred to the model. This forces the model to learn real objects from different positions.

It is important to note that different models in this study were trained using their respective standard data augmentation strategies. The CNN and ResNet50 models were trained on synthetically augmented datasets, while YOLOv8 utilized its built-in Mosaic Data Augmentation. Although the augmentation methods differ, our aim was not to perform a strictly controlled comparison of architectures under identical input conditions but rather to evaluate each model in a practical, real-world context using its native training pipeline. YOLOv8 is designed to work seamlessly with Mosaic augmentation, which is considered part of its core strength. Therefore, we argue that the variation in augmentation does not invalidate the comparison but rather reflects realistic usage scenarios where each model is used as intended by its designers. The performance differences observed are still meaningful, as they showcase how each model performs with optimal or standard preprocessing strategies.

3.4. Model Training/ Validate/ Testing

Our research aimed to develop an OCR system for the Georgian language that could provide improved results compared to the previously discussed systems. To achieve this, we employed a new model, distinct from the existing ones, marking the first attempt to create such a system not only for Georgian but also for other languages. Consequently, the research, conducted concurrently with other models, was also interesting and applicable in assessing the model's effectiveness in similar tasks.

We trained three models: CNN, ResNet, and YOLOv8.

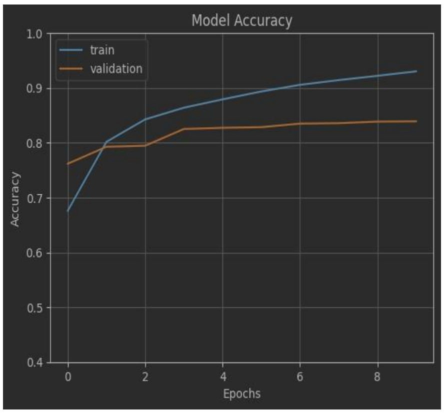
The convolutional neural network (CNN) is constructed from several convolutional layers, dropout, and max-pooling (Nielsen 2019). We utilized sparse_categorical_crossentropy for the loss function and Adam for optimization.

Additionally, we experimented with ResNet50, which is a 50-layer neural network (Kaiming, et al. 2015), and applied it to the aforementioned Georgian manuscript data. After 30 iterations, we obtained the following results for both models:

Table 1. Results for CNN and ResNet

Evaluation/Models	CNN	ResNet
Loss:	0.2507	0.2456
Accuracy:	0.923	0.9302
Val_loss:	0.9264	0.6916
Val_accuracy:	0.8018	0.8391

YOLOv8 is the latest version of the YOLO models, officially released on January 10, 2023 (Shubham, et al. 2022), (Cheng 2020), (Krishnakumar 2023), (Jocher 2024) These models demonstrate high performance compared to their predecessors. Notably, YOLOv8 has shown remarkable success in aerial image detection for Unmanned Aerial Vehicles (UAVs) (Yiting, et al. 2023), (Redmon and Farhadi 2017).



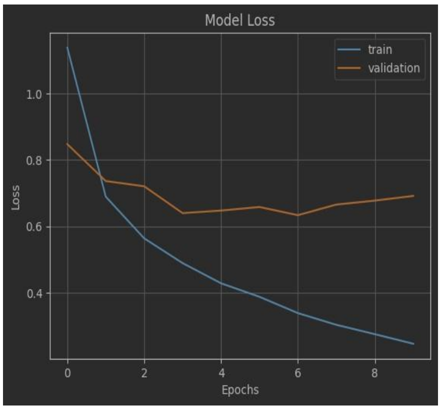
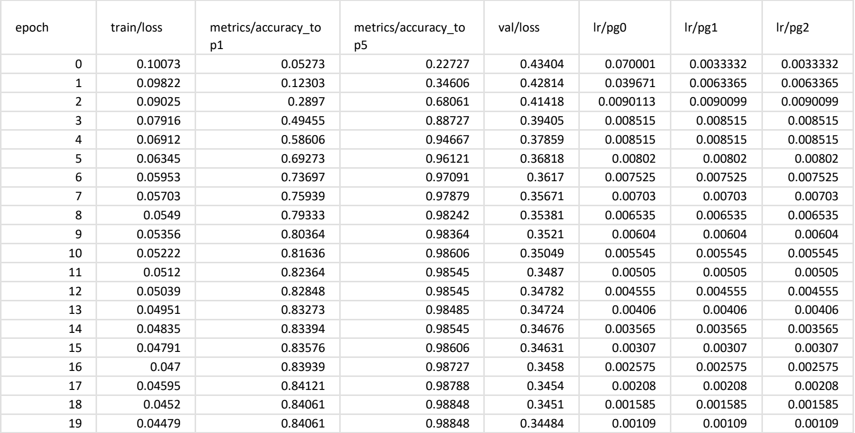


Figure 15. Epochs vs Accuracy (ResNet)

Picture 16. Model Loss (ResNet)

Previous YOLO models have been used for OCR in various languages, such as English, German, French, Latin, and more (Chaudhuri, et al. 2016), (Alghyaline 2022), (Subramanian, et al. 2022), (Farhadi 2018). However, our project marks the first attempt to implement the YOLOv8 model for an OCR system. In our project we employed the 'YOLOv8n-cls' model to analyze the above mentioned data, focusing on solving a classification problem. The model includes an integrated augmentation process, eliminating the need for additional adjustments. After multiple iterations, we obtained model weights, and the best-performing weight can be selected by the user. Additionally, our project generated a CSV file with the following format:

Table 2. Model Weights (csv file)



epoch	train/loss	metrics/accuracy_top1	metrics/accuracy_top5	val/loss	lr/pg0	lr/pg1	lr/pg2
0	0.10073	0.05273	0.22727	0.43404	0.070001	0.0033332	0.0033332
1	0.09822	0.12303	0.34606	0.42814	0.039671	0.0063365	0.0063365
2	0.09025	0.2897	0.68061	0.41418	0.0090113	0.0090099	0.0090099
3	0.07916	0.49455	0.88727	0.39405	0.008515	0.008515	0.008515
4	0.06912	0.58606	0.94667	0.37859	0.008515	0.008515	0.008515
5	0.06345	0.69273	0.96121	0.36818	0.00802	0.00802	0.00802
6	0.05953	0.73697	0.97091	0.3617	0.007525	0.007525	0.007525
7	0.05703	0.75939	0.97879	0.35671	0.00703	0.00703	0.00703
8	0.0549	0.79333	0.98242	0.35381	0.006535	0.006535	0.006535
9	0.05356	0.80364	0.98364	0.3521	0.00604	0.00604	0.00604
10	0.05222	0.81636	0.98606	0.35049	0.005545	0.005545	0.005545
11	0.0512	0.82364	0.98545	0.3487	0.00505	0.00505	0.00505
12	0.05039	0.82848	0.98545	0.34782	0.004555	0.004555	0.004555
13	0.04951	0.83273	0.98485	0.34724	0.00406	0.00406	0.00406
14	0.04835	0.83394	0.98545	0.34676	0.003565	0.003565	0.003565
15	0.04791	0.83576	0.98606	0.34631	0.00307	0.00307	0.00307
16	0.047	0.83939	0.98727	0.3458	0.002575	0.002575	0.002575
17	0.04595	0.84121	0.98788	0.3454	0.00208	0.00208	0.00208
18	0.0452	0.84061	0.98848	0.3451	0.001585	0.001585	0.001585
19	0.04479	0.84061	0.98848	0.34484	0.00109	0.00109	0.00109

As a result of the analysis, we can show the results of model training using the Python library matplotlib:

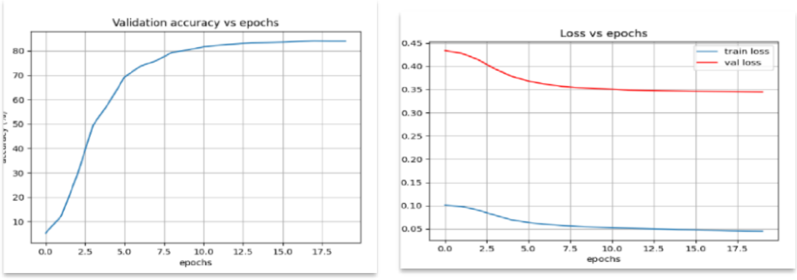


Figure 17. Results of Model Training

While the figure highlights the Top-1 accuracy, it is worth noting that the Top-5 accuracy also yields significant insights. It reflects the model’s ability to identify the correct class among its top predictions, offering a more nuanced perspective on overall performance - particularly in cases where multiple classes may appear visually similar. Top-1 accuracy considers only the model’s most confident prediction, marking it correct only if it exactly matches the ground truth label. In contrast, Top-5 accuracy is more lenient, deeming the prediction correct if the true label appears within the model’s top five highest-probability outputs. This distinction is especially important in complex classification tasks with many categories, where the correct class may not always be the top-ranked prediction but is still among the most plausible candidates.

3.5. Results and Analysis

For the evaluation of the OCR system, accuracy, precision, recall, and the F-score were calculated based on the Confusion Matrix for all three models. The results are as follows.

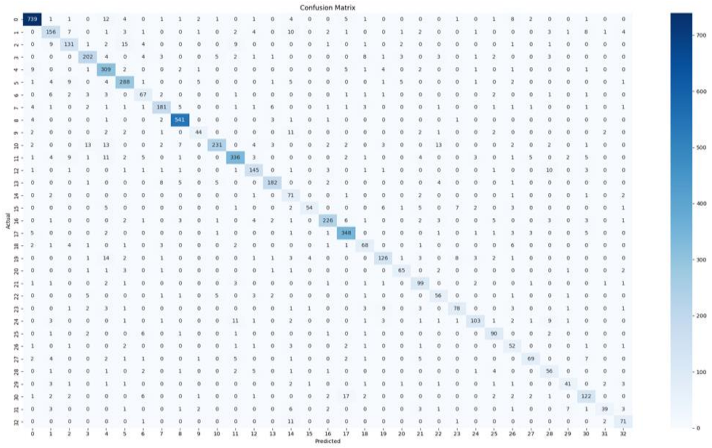


Figure 18. Confusion Matrixes for ResNet

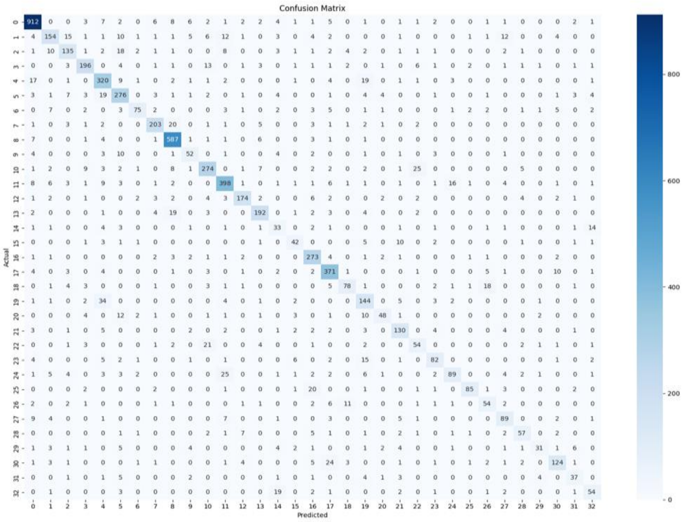


Figure 19. Confusion Matrixes for CNN

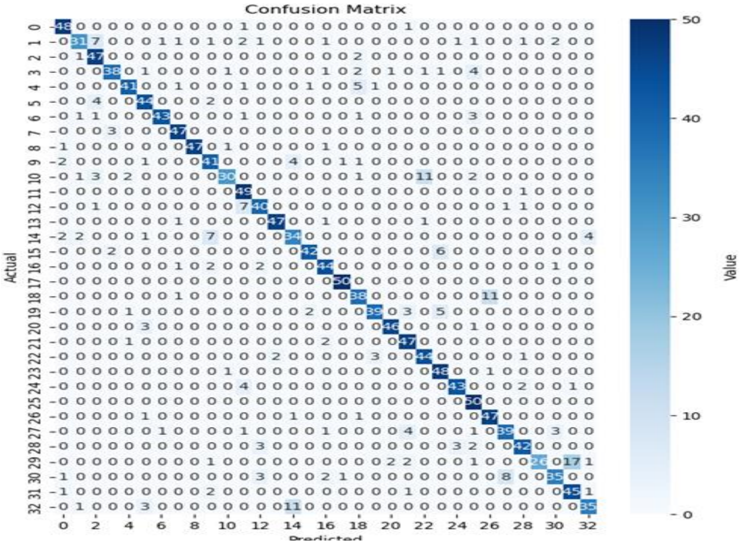
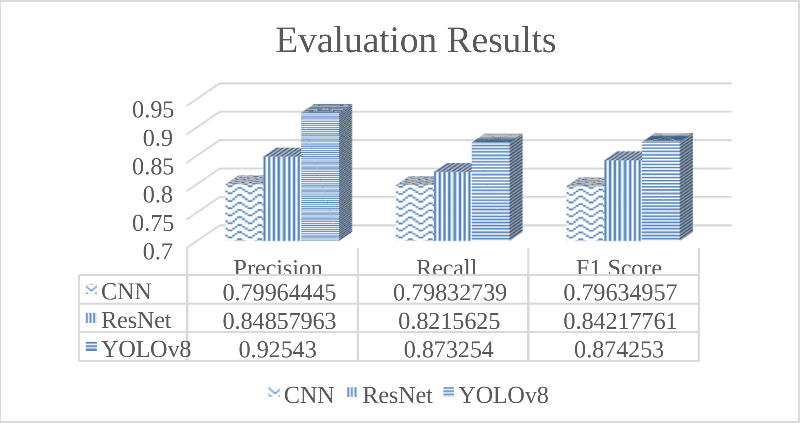


Figure 20. Confusion Matrixes for YOLOv8



	Precision	Recall	F1 Score
CNN	0.79964445	0.79832739	0.79634957
ResNet	0.84857963	0.8215625	0.84217761
YOLOv8	0.92543	0.873254	0.874253

Figure 21. Precision, Recall, F1 Score

As seen from the results, the YOLOv8 model outperformed the other two models on the same data. Advance YOLOv8 model showed some improvements mentioned above, it works better with small letters and major differences that Georgian language characters have. That is why accuracy is much higher and model performance is also faster.

3.6. Model deployment

In conclusion, the deployment of the Georgian language OCR system through Flask, designed to emulate a whiteboard system for uploading whole images and returning typed text, represents a significant step towards bridging the gap between handwritten and digital content in the Georgian language. The deployment details for this system can be found in our GitHub repository (Chikashua 2023). This innovative solution not only leverages advanced optical character recognition techniques tailored to the unique characteristics of Georgian script but also offers a user-friendly experience, making it accessible for a wide range of users. By enabling an easier transformation of handwritten text into digital format, this system plays a crucial role in preserving and digitizing historical documents, improving data accessibility, and promoting the utilization of the Georgian language in the digital age. As technology continues to evolve, this Georgian OCR system serves as a valuable tool in the fields of linguistics, cultural heritage preservation, and document digitization, opening up new possibilities for research, education, and archival work.

4. Conclusion

While this technology represents the latest advancements, there are still no OCR products capable of recognizing all types of text with a 100% accuracy. In conclusion, the goal of this project was to explore one of the intriguing aspects of computer vision related to OCR and the Georgian language.

Tools like whiteboards, with fewer artifacts, distinct colors, and other advantages, tend to yield good results with the mentioned model. However, it also exhibits a reasonable level of accuracy when processing text written on paper. Progress in the realm of whole document analysis is crucial, although it still requires further development.

Despite the existing advancements in this field, finding a one-size-fits-all solution that consistently delivers accurate results remains challenging. Thus, this paper represents a step towards addressing the problem under study.

References

Alghyaline, Salah. 2022. " A Printed Arabic Optical Character Recognition System using Deep Learning." *Journal of Computer Science,* vol. 18, pp. 1038-1050.

Chaudhuri, Arindam, Krupa Mandaviya, Pratixa Badelia, and Soumya K Ghosh. 2016. "Character Recognition Systems for Different Languages with Soft Computing." In *Studies in Fuzziness and Soft Computing book series*. Springer.

Cheng, Richeng. 2020. "A survey: Comparison between Convolutional Neural Network and YOLO in image identification." *Journal of Physics: Conference Series,* vol. 1453, no. 012139.

Chikashua, Ana. 2023. "Georgian Language OCR with TensorFlow." *GitHub.* https://github.com/AnaChikashua/geo-alphabet.

Farhadi, Joseph Redmon and Ali. 2018. "YOLOv3: An Incremental Improvement." *arXiv preprint arXiv:* 1804.02767.

Jocher, Glenn. 2024. *Ultralytics YOLOv8 Docs.* November 12. https://docs.ultralytics.com/datasets/classify/imagenet.

Kaiming, He, Zhang Xiangyu, Ren Shaoqing, and Sun Jian. 2015. *Deep Residual Learning for Image Recognition.* Microsoft Research.

Krishnakumar, Mukilan. 2023. "Weight & Biases." *Yolov8 locally on a non-GPU machine.* June 26. https://pchenlab.wordpress.com/2023/06/26/yolov8-locally-on-a-non-gpu-machine/.

Microsoft. 2023. *NEBO.* https://apps.microsoft.com.

Nielsen, M. 2019. *Neural Networks and Deep Learning.*

Redmon, Joseph, and Ali Farhadi. 2017. "YOLO9000: Better, Faster, Stronger." *Computer Vision and Pattern Recognition (CVPR),*. USA: Honolulu. pp. 1-23.

Rosebrock, A. 2020. *Getting started with EasyOCR for Optical Character Recognition.* September 14. https://machinelearningknowledge.ai/easyocr-python-tutorial-with-examples/.

Shubham, Prasad. 2020. "https://www.topcoder.com/thrive/articles/python-for-character-recognition-tesseract. ." *https://www.topcoder.com/how-it-works.* December 10.

Shubham, Srivastava, S, Verma Ajay, and Sharma Shekhar. 2022. "Optical Character Recognition Techniques:A Review." *IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS).* Bhopal, India,.

Soselia, D, M Tsintsadze, L Shugliashvili, I Koberidze, S Amashukeli, and S Jijavadze. 2018. "Georgian Handwritten Character Recognition." *Elsevier.*

Subramanian, Srividya, Vineet Kekatpure, Gladina Raymond, Kapil Parab, Shashikant Dugad, and rchana Shirke. 2022. "“TEYSuR - Text Extraction with YOLO and Super Resolution." *in International Conference for Advancement in Technology (ICONAT),*. Goa.

Yiting, Li, Fan Qingsong, Haisong Huang, Zhenggong Han, and Gu Qiang. 2023. "A Modified YOLOv8 Detection Network for UAV AerialImage Recognition." *Drones* vol. 7, no. 304, pp. 2-26.

Received November 19, 2024, revised May 5, 2025, accepted May 6, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 2, 436-452
https://doi.org/10.22364/bjmc.2025.13.2.08

Predictive Mathematical and Computer Model for Determining Harmful Effects of Dust Pollution on the Environment and Workers

Yevhenii LASHKO[1], Serhii SUKACH[1], Ivan LAKTIONOV[2], Olha CHENCHEVA[1], Dmytro RIEZNIK[1], Olena KORTSOVA[3,4]

[1]Institute of Education and Science in Mechanical Engineering, Transport and Natural Sciences, Department of Civil and Labour Safety, Geodesy and Land Management, Kremenchuk Mykhailo Ostrohradskyi National University, 20 Universytetska str., Kremenchuk, 39600, Ukraine
[2]Faculty of Information Technologies, Department of Computer Systems Software, Dnipro University of Technology, 19 Dmytra Yavornytskoho av., Dnipro, 49005, Ukraine
[3]Institute of Education and Science in Mechanical Engineering, Transport and Natural Sciences, Department of Ecology and Biotechnologies, Kremenchuk Mykhailo Ostrohradskyi National University, 20 Universytetska str., Kremenchuk, 39600, Ukraine
[4]Promecologia Scientific Technological Centre, Ltd, 2/7 Likaria Bohaievskoho str., Kremenhcuk, 39600, Ukraine

elashko@kdu.edu.ua, ssukach@kdu.edu.ua, Laktionov.I.S@nmu.one,ochencheva@kdu.edu.ua, dreznik@kdu.edu.ua, lenakortsova@yahoo.com

ORCID 0000-0001-9691-4648, ORCID 0000-0002-6834-0197, ORCID 0000-0001-7857-6382, ORCID 0000-0002-5691-7884, ORCID 0000-0003-1258-6136, ORCID 0000-0002-8101-322X

Abstract. The paper is devoted to mathematical modeling and computer simulation of the process of movement of dust particles in the working area of quarries and adjacent territories in order to protect workers of mining enterprises. The object of the research is a three-dimensional digital model of the relief of a quarry that is currently under development. The subject of the research is simulation the trajectory of dust particles and determining the zones of their maximum permissible concentration, taking into account the prevailing meteorological characteristics of the area. The scientific and practical value of the research results of this article lies in the fact that for the first time the regularities of aerodynamic processes occurring in the working area of quarries and adjacent territories, taking into account the exact shape of the contour of the quarry itself and the corresponding profile of its geological section, were determined by mathematical modeling and computer simulation methods.

Keywords: forecasting, simulation, emissions of harmful substances, atmospheric air, deposits, dust particles.

1. Introduction

In the modern mining industry, up to 80% of production is extracted by open pit mining. The creation of open pits causes changes in the microclimate, primary terrain and hydrography of the area. At the same time, open-pit mining is being improved by intensifying all industrial processes (from drilling and primary crushing, mainly through blasting, to loading and transporting rock mass), using more and more high-performance machines (drilling rigs, excavators, loaders), as well as rail, road and conveyor transport. In addition, current technologies related to the use of crushers, screens, etc. are being introduced in quarries during the development of rocks [Honcharevskyi et al., 2013].

This development of open-pit mining is accompanied by a sharp increase in the amount of emissions of various harmful substances into the atmosphere: dust, toxic gases, and other chemicals hazardous to human health, which causes various disorders in the animal and plant life. The intensity of dust and gas pollution in and around quarries depends on a number of natural, technological and technical factors. Their ratio determines the level of concentration of dust and other harmful substances in the atmosphere of open-pit mines. At the same time, the number of cases of exceeding the maximum permissible concentrations and creating hazardous situations is increasing.

The type of deposit to be developed, its geological and hydrogeological characteristics, shape and conditions of mineral occurrence, chemical composition and physical and mechanical properties of the mineral are determined by analyzing the mining and geological conditions of the deposit. The need for such an analysis is due to the fact that at the design stage and during operation, depending on geological and mining data, an efficient development technology and appropriate equipment are selected, classified into groups based on the intensity of the emission of harmful mixtures and dust during its operation. Based on the mineralogical composition of mining operations, dust content standards are set and preventive measures are developed to prevent the increased content of harmful mixtures and normalize the atmosphere in the quarry space and beyond.

For example, the most unfavorable process for the atmosphere of quarries is blasting, which dramatically worsens the composition of the atmosphere in terms of dust and gas criteria. The hazard of blasting is exacerbated by the fact that harmful mixtures are released from the extracted rock mass not only during the explosion, but also afterwards, during loading and transportation.

When assessing the impact of open pit mining on the atmosphere, it is also necessary to take into account the order of operations, the relative position and concentration of equipment in relation to the direction of air flows, the location and composition of rock mass along the contours of pits and openings, as well as the wind pattern and speed, climatic conditions of the area and the microclimate of pits. Taking into account the above, it is necessary to characterize the atmosphere of production spaces in order to properly select methods of intensifying natural and organizing artificial air exchange during open-pit mining.

Therefore, the task is to build a predictive mathematical and computer model of the spread of dust pollution in quarries during their planned activities. Modern methods of mathematical modeling and computer simulation make it possible to research the movement of multifractional particles and determine the stable patterns that arise in this process, as well as to analyze not only the parameters of individual particles, but also the overall parameters of aerodynamic systems.

This research considers mathematical modeling and computer simulation of the process of movement of dust particles in the working area of quarries and adjacent territories in order to protect workers of mining enterprises. The object of the research is a three-dimensional digital model of the relief of a quarry that is currently under development. The subject of the research is the simulation the trajectory of dust particles and the identification of the zones of their maximum permissible concentration, taking into account the prevailing meteorological characteristics of the area. The main tasks that arise during such simulation are obtaining profiles of concentrations of harmful substances, determining distances and dangerous wind speeds that correspond to the formation of maximum concentrations of pollutants from sources. The article contains a mathematical description of the determination of dust and gas emissions from massive (volley) explosions, the design of a three-dimensional parametric model of a quarry, and computer simulation of dust pollution.

The scientific and practical value of the research results of this article lies in the fact that for the first time the regularities of aerodynamic processes occurring in the working area of quarries and adjacent territories, taking into account the exact shape of the contour of the quarry itself and the corresponding profile of its geological section, were determined by mathematical modeling and computer simulation methods, which will make it possible to implement them at the national level both in the field of labor protection and in the field of environmental protection technologies.

2. Methods, tools and approaches to research

2.1. Generalized research approaches

The research of dust sources, its movement in the atmosphere, its impact on human health, and methods for simulation these processes covers a wide range of approaches considered in the modern scientific literature.

For example, (Liu et al. 2019) researched the impact of a dust removal system during tunnel construction using CFD simulation technology. This method allowed for a detailed analysis of the effectiveness of the dust removal system in the context of mechanized tunnel construction, which helped to reduce the level of dust pollution. Its advantage is the high accuracy of simulation the distribution of dust in real conditions, but this research requires significant computing resources, which limits its application on a full scale.

The authors (Xiu et al., 2020) used numerical simulation to research the character of dust pollution in coal mines and determine the optimal airflow parameters for dust control. The use of CFD methods for simulation emissions and ventilation allows for accurate pollution predictions, but this accuracy depends on the model settings and the realism of the input data. In addition, the method is highly dependent to changes in ventilation conditions and the mineralogical composition of the extracted material.

A research (Zhou et al., 2022) examined the dynamics of dust pollution in mines using the DPM-DEM model. This approach allows for a more accurate assessment of the movement of dust particles and their interaction with the air flow, which makes it possible to accurately identify areas of high pollution. However, the model has limitations during simulation complex ventilation conditions where a large number of turbulent flows occur, and requires large computing power for full implementation.

The literature review also includes researches that focus on other aspects of pollution, such as the use of more complex numerical models to assess the risks and effectiveness of dust emission reduction technologies.

In (Liaskoni et al. 2023), a highly detailed simulation of wind dust emissions in Europe was carried out using the WRF-Chem system. The method allowed to take into account the complex interactions between dust particles and the atmospheric condition, which ensures high accuracy of the forecast of PM_{10} and $PM_{2,5}$ concentrations. The main advantages of the approach are the spatial and temporal resolution and complexity of data processing, but the method is very resource-intensive and dependent to the quality of input parameters.

In another research (Tong et al., 2018), the Monte Carlo method was used to assess the health risk of workers at construction sites. The approach provides flexibility in simulation uncertainties, allowing assessing variations in dust exposure under different scenarios. However, the effectiveness of the method largely depends on the reliability of the input data and the number of iterations.

The authors (Askarova et al., 2021) have developed three-dimensional CFD models of solid fuel combustion processes to reduce harmful emissions. The advantage of the approach is the realistic consideration of heat and mass transfer and chemical reactions in the furnace. However, the results require careful calibration and validation, which increases the complexity of using such CFD models in practice.

Statistical methods, such as multivariate analysis of variation (MANOVA), have been used (Giancristofaro et al., 2015) to analyze regional differences in odor perception. Although statistical tests are an effective tool for assessing differences between groups, the results are dependent to subjective factors and data selectivity.

An important contribution to the research was made by researchers (Konglok et al., 2016) who used the fractional step method to numerically solve the problems of pollution spreading under different classes of atmospheric stability. The method provides increased computational stability and allows for variable meteorological conditions, although its effectiveness decreases in cases of highly turbulent flows.

The authors (Kumar et al., 2020) conducted multiphase CFD simulation and laboratory testing of the Vortecone device for air dust removal. The combination of numerical simulation and experimental data allowed for highly accurate results, but multiphase CFD calculations require significant computational resources.

In (Oyjinda et al., 2017), simplified numerical diffusion models were used to simulate pollution in the vicinity of industrial areas. The advantage of this approach is the speed of calculations and the ability to quickly adapt the model, but the simplification of physical processes limits the accuracy of predictions in complex conditions.

Thus, a review of the available literature shows that modern research actively uses both complex physical and chemical models (CFD, atmospheric models) and statistical and stochastic methods (Monte Carlo, MANOVA) to analyze dust pollution and its consequences. The choice of methodology largely depends on the objectives of the research: from predicting concentrations on a large scale to local risk assessment or optimization of engineering solutions.

The research of this article is a logical sequential continuation of our own theoretical and experimental researcher in the field of mathematical modeling and computer simulation of the aerodynamic process of movement and removal of dust particles in the working area, which are reflected in scientific publications (Chencheva et al., 2023; Lashko et al., 2024).

We improved the known dependencies of dust pollution dispersion, which for the first time take into account local meteorological characteristics, as well as particle size and quarry topography, which allowed creating a mathematical description for further computer simulation. A limitation of the research is that it does not take into account emission factors for different dust sources, such as drilling and crushing, which could provide more accurate and general predictions in the future.

It is also worth noting that calculating air pollution using the «EOL+» software version 5.3.8, which implements the Methodology for Calculating Air Concentrations of Harmful Substances Contained in the Emissions of GRD-86 Enterprises approved at the state level in Ukraine at (Order of the Ministry of Ecology and Natural Resources of Ukraine, 2021), does not provide correct results for volley emissions, taking into account the influence of terrain. At the same time, it is important to improve this methodology in order to bring it closer to European environmental requirements (Directive (EU) 2024/2881 of the European Parliament and of the Council of 23 October 2024 on ambient air quality and cleaner air for Europe (recast), 2024), which is relevant to the topic of the research.

2.2. Structural and algorithmic description of a computer-oriented model

The COMSOL Multiphysics® version 5.6 is designed for simulation three-dimensional fluid and gas flows in technical and natural objects, as well as visualizing these flows using computer graphics (CFD Modeling and Simulation. COMSOL Multiphysics, 2020). The simulation flows include stationary and unsteady, compressed, incompressible and uncompressed liquid and gas flows. The use of various turbulence models and an adaptive computational grid allows simulating complex fluid motions, including flows with strong swirling, combustion, and free-surface flows.

This software is based on the finite volume method of solving fluid dynamics equations and uses a rectangular adaptive mesh with local refinement. This technology allows importing geometry from CAD systems and exchanging information with finite element analysis systems. The use of this technology allows solving the problem of automatic mesh generation – to generate a mesh, it is enough to set only a few parameters, after which the mesh is automatically generated for the design area with geometry of any degree of complexity.

2.3. Mathematical description of the system under study

The most powerful source of instantaneous dust emission and formation of dust and gas clouds in the atmosphere of quarries is massive (volley) explosions (Fig. 1).

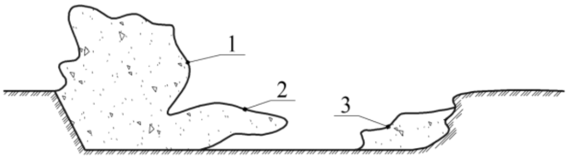


Figure 1. Scheme of dust and gas clouds formation during massive explosions in a quarry: 1 – primary cloud, 2 – secondary cloud, 3 – cloud formed due to shock wave and seismic oscillations

The mineralogical composition of dust is usually close to the mineralogical composition of blasted rocks.

The chemical composition of dust is also close to that of rocks, but some particles may be introduced from other sources. Poisonous dust contains lead, mercury, chromium, manganese and other toxic elements. Non-poisonous dust includes quartz dust (dangerous due to silicosis) and coal dust.

The dispersion composition of dust is determined by natural and technical and technological factors and differs in the size of dust particles that accumulate in the atmosphere of open pit mining. The shape of the dust particles determines the rate of their deposition and depends on the method of rock destruction. The true integrity of a single dust particle is equal to the original one.

The height of rise of dust and gas aerosols is calculated by the formula:

$$h_0 = \frac{\Delta t}{(\gamma_a - \gamma) - t_c / \left[g(cb)^2 R_1 - 4{,}3\right]}, \tag{1}$$

where *Δt* – the temperature difference of the ambient explosion products at a height of 10–15 m from the surface to be blown up, C°; *γ* – the vertical temperature gradient, C°/100 m, m; *g* – the free fall acceleration, m/s$^{(2)}$; *c*, *b* – the dimensional experimental constants (*c*=11,5; *b*=0,2); R_1 is the primary radius of the dust and gas cloud, m.

The primary radius of a dust and gas cloud is calculated by the formula:

$$R_1 = \sqrt[3]{\frac{3V}{4\pi}}, \tag{2}$$

where *V* – the volume of gases released during the explosion of an explosive, m^3.

The volume of gases released during the explosion of an explosive is calculated by the formula:

$$V = mAV_0, \tag{3}$$

where *m*=0,6÷0,75 – a coefficient that takes into account the actual amount of gases that enter the atmosphere (some gases remain in the exploded rock mass), 1/kg; V_0 – the volume of gases formed during decomposition; 1 kg of explosive (V_0=0,6÷1,1) m^3; *A* – the mass of explosive, kg.

A secondary dust and gas cloud occurs after blasting in the dust and gas throwing zone at the foot of the ledge within the radius of rock fragments:

$$R_2 = k\sqrt{A} / r\sqrt{p_i} / n\left(S / d^2\right), \tag{4}$$

where *k* – a coefficient that takes into account the type and design of the explosive charge; *p* – the density of the rock to be blown up, kg/m$^{(3)}$); *r* – the advancement of the face, m; *S* – the cross-sectional area of the face to be blown up, m^2; *D* – the diameter of the well, m.

The above mathematical dependencies were used as the basis for computer simulation of dust pollution spread. The initial data for the calculation are grouped in Table 1. The components of these data include the characteristics of the source of pollutant emissions, namely blasting, multifractional particles models in the range from 1,4 to 50 μm and meteorological characteristics that determine the conditions for dispersion of pollutants in the air, created by directional airflow computer model.

Table 1: Initial data of preprocessing

Parameters of the dust and air mixture	
Volume, m^3/s	0,01
Emission capacity determined	
Calculation, g/s	470
Calculation, tones per year	5,64
Average annual wind rose, %.	
N	10,8
NE	8,5
S	10,1
SE	11,9
S	12,9
SW	14,2
W	19,9
NW	11,7
Wind speed (based on average long-term data), the recurrence of exceeding which is 5%, m/s	9–10

Additionally, the topography of the open pit and surrounding areas was taken into account, as shown in Figure 2.

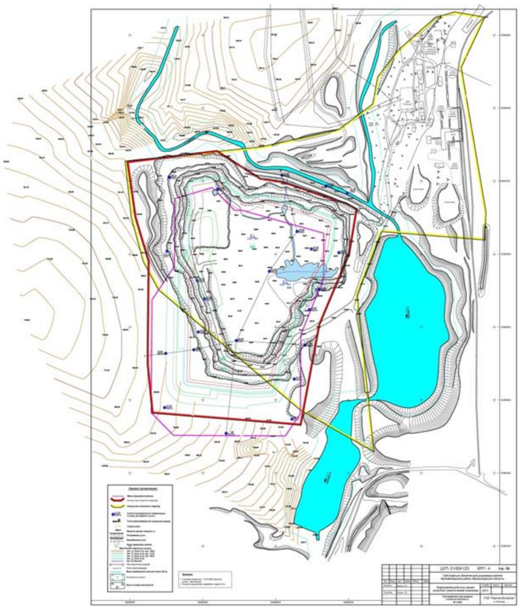


Figure 2. Surface plan from the field development and reclamation project

On this topographic plan, the deposit is marked in purple, the boundaries of the license area are marked in red, and the boundaries of the land allotment are marked in yellow. All of these data were used to build three-dimensional parametric models as a calculation space for computer simulation of the movement of dust particles in it.

3. Research results

3.1. Computer model

A three-dimensional model of the studied quarry as an initial structural element is shown in section in Fig. 3. It was built using SolidWorks computer-aided design software.

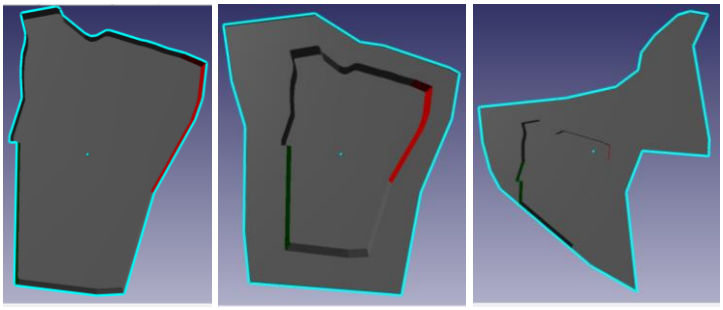


Figure 3. Three-dimensional models of the studied quarry in the section with indication of boundary restrictions

To verify the theoretical positions, computer simulation was carried out using COMSOL Multiphysics® version 5.6, the results of which are shown in Figs. 4–6. The results show the movement of particles directly in the open pit and adjacent areas in the form of vectors, where red indicates maximum values and blue indicates minimum or infinitesimal values.

The main purpose of the prediction model is to assess the possible environmental response to the direct or indirect impact of planned activities, and to address the challenges of future rational use of natural resources in relation to expected environmental conditions. At the same time, it is also important to protect employees at workplaces from dust pollution that can lead to respiratory diseases.

Management of production processes includes current (operational) and perspective (long-term) aspects. In this case, strategic management is used for operational management and forecasting of the mining enterprise, i.e., identification of prospects and development of fundamental solutions to prevent possible negative impact on environmental components and health of employees.

For long-term forecasting, computational (analytical, approximation) models based on the solution of the turbulent diffusion equation are most often used. These are plume models, finite-difference models, etc., which form the basis of the Methodology for Calculating Air Concentrations of Harmful Substances Contained in Industrial Emissions (GRD-86) (Order of the Ministry of Ecology and Natural Resources of Ukraine, 2021).

The practical effectiveness of short-term air pollution forecasts is clearly demonstrated when the sources are known and measures can be taken to reduce emissions during periods of unfavorable weather conditions.

For operational forecasting, statistical models of linear and nonlinear regression are widely used. For operational forecasting of air pollution during emergency salvo emissions, it is necessary to use computational (analytical) models – «tangle» models – which are used to predict the spread of impurities from instantaneous point sources.

The development of forecasting methods begins with the identification of periods with significant air pollution. Then, correlations are established between the degrees of air pollution observed during these periods and some meteorological variables or a certain combination of them, which are considered as predictors.

The long-term forecasting of environmental pollution from quarry emission sources was carried out using the method of software (electronic) calculation and simulation, which is based on the use of software products approved for use at the state level, which, in turn, are based on the algorithms of existing model calculation approaches. In particular, during the environmental impact assessment, the dispersion was calculated using the «EOL+» software package.

The forecasting results in the production of profiles of concentrations of harmful substances, determination of distances and dangerous wind speeds corresponding to the formation of maximum concentrations of pollutants from enterprise sources.

The most interesting is the forecasting of air pollution during salvo emissions. In these cases, the pollution forecast is performed using the expected change in emissions, taking into account specific meteorological conditions (predictors).

The choice of predictors is usually based on general physical ideas about the possible causes of changes in impurity concentrations, such as changes in wind direction or speed, atmospheric stability, leaching or transformation of impurities, etc.

In addition to the previously mathematically calculated data, the topographic plan of the existing quarry surface was used as the initial data for the long-term forecast, and meteorological characteristics that determine the conditions for dispersing pollutants in the air (average annual wind rose) were used as predictors.

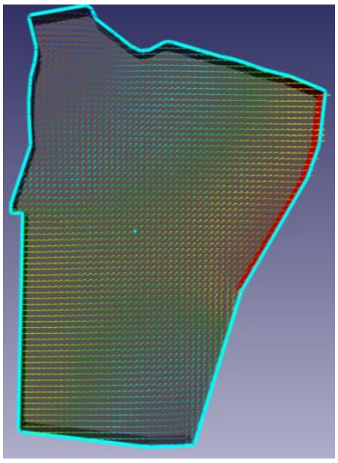


Figure 4. Velocity of particles in the air flow, displayed in vector form in the range from 0 to 10 m/s, directly in the deposit

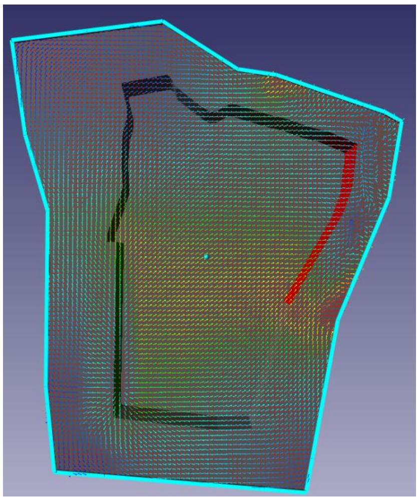


Figure 5. Velocity of particles in the air flow, displayed in vector form in the range from 0 to 10 m/s, within the license area

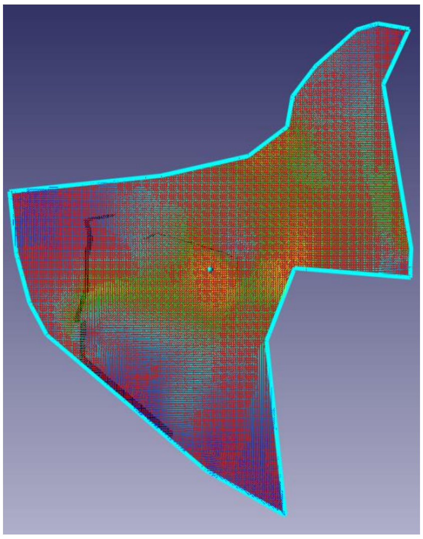


Figure 6. Velocity of particles in the air flow, displayed in vector form in the range from 0 to 10 m/s, within the land allotment

By means listing of software air flow velocities exceeding the critical ones, which affect the ability to blow dust from work space, were determined (Fig. 7). At the same time, by means formula (3) blasting operations accounted during predicting the overall balance of harmful substances in the quarry space (Fig. 8).

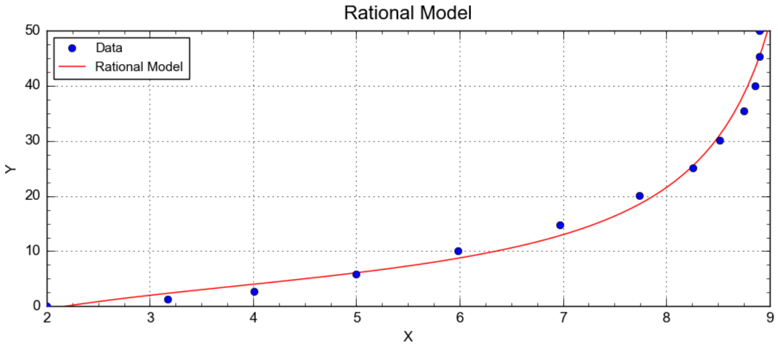


Figure 7. Dependence of specific dust blowing *c*, m/cm^2 (*y*-axis) on air flow velocity *v*, m/s (*x*-axis)

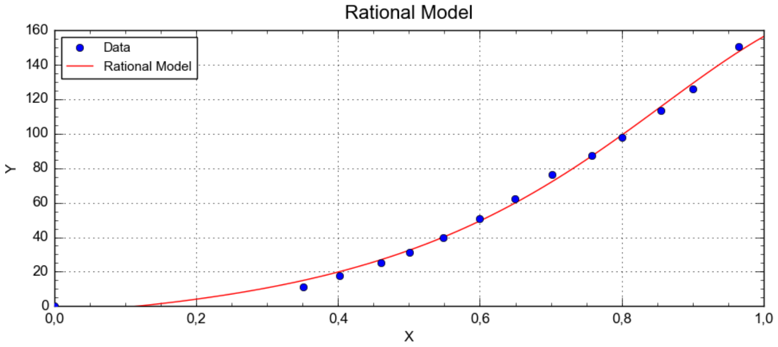


Figure 8. Dependence of the amount of dust generated during the explosion g_n, g/m^3 (*y*-axis) on the specific consumption of explosive A, kg/m^3 (*x*-axis)

As a result of computer simulation, it was established that the movement of dust particles according to the prevailing meteorological characteristics is directed to the territory of economic structures while maintaining its intensity within the land allotment. The nature of the distribution of dust movement vectors is somewhat different from the wind rose and indicates a significant impact on its annual distribution not only of the transforming wind directions, but also of its speed and the shape of the quarry contour. Therefore, it can be concluded that workers are exposed to dust from the quarry during its development, especially during volley emissions from blasting operations.

3.2. Experimental validation of the model

Considering that the concentration of harmful particles in the air changes over time, it is necessary to additionally perform several iterations of calculations, namely at the moment of the explosion flare and the subsequent phase of rarefaction. It is during the rarefaction phase that dust particles begin to settle, being directly influenced by the direction and speed of air movement.

One of the most common measures during the construction and operation of industrial facilities in the quarry area is the wind rose, which reflects the frequency of winds blowing from different directions. However, this is often not enough to assess the direction of distribution and concentration of industrial emissions, as the concentration of harmful substances in the dust of a gas cloud or a stream plume is significantly affected by wind speed and atmospheric turbulence. Accordingly, the quarry space is identified as a linear source, accumulated along its longer or shorter axis according to the dominant wind directions. For intermediate directions, the source is projected on a long axis directed normal to the wind direction. The corresponding parameter is used to assess the impact of emissions:

$$K = \sum_{i=1}^{m} C_x p_i, \tag{5}$$

where C_x – the concentration of hazardous mixtures along the plume axis at a distance x at a given wind speed and direction, g/m^3; p – the probability or frequency of wind repetition of a given wind speed and direction. The calculation is carried out in different directions, based on the results of which a plan of the zone adjacent to the quarry is drawn up with perimeter isolines K.

The concentration of the mixtures at different distances from the pit contour as a linear source can be determined by the Setton formula:

$$C_x = \frac{K_0 M}{u} e^{\frac{y^2}{c^2 x^{2-n}}}, \tag{6}$$

where K_0 – the specific concentration equal to the concentration of harmful substances from a pollution source of 1 g/s at a wind speed of 1 m/s:

$$K_0 = \frac{2000}{\frac{2-n}{3} - \frac{h^2}{c^2 x^{2-n}}}, \tag{7}$$

where c – the scattering coefficient (c=0,05 at n=0); n – a coefficient that depends on the temperature gradient of the atmosphere, surface roughness and the surface to be washed (n=0 under average meteorological conditions); x – the distance along the calculated wind direction from the source to the line perpendicular to the wind direction and passing through the point of determining the concentration of mixtures, m; h – the conditional height of the linear source emission, m; M – the intensity of 1 m of the source length, g/s; u – the calculated wind speed, m/s; y_2 – the normal distance from the calculated point to the line passing through the center of the linear source perpendicular to the wind direction.

According to the formula:

$$e - \frac{y^y}{c^2 x^{2-n}}, \tag{8}$$

calculate the decrease in concentration along the width (at n=0 it is equal to $e - \frac{y^y}{c^2 x^{2-n}}$).

Therefore, the next step is to use this methodology in an operating quarry, where the calculations of the concentration of harmful substances in the explosion products can be used to reflect the expected distribution of dust particle movement vectors in the quarry itself and adjacent areas.

The above mathematical dependencies were used as the basis for computer simulation of the determination of dust concentrations (Fig. 9, 10).

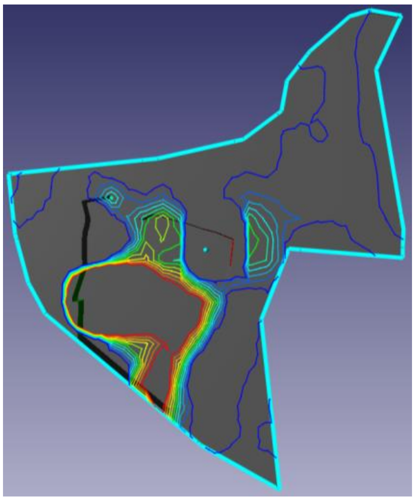


Figure 9. Concentration of particles in the air flow at the moment of the explosion flare, displayed in isolines form in the range from 0 to 5 mg/m^3, within the land allotment

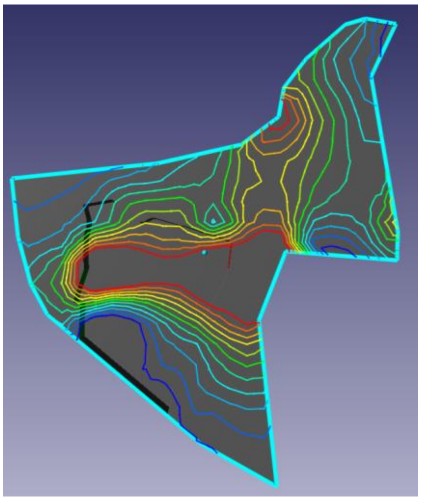


Figure 10. Concentration of particles in the air flow at the phase of rarefaction, displayed in isolines form in the range from 0 to 5 mg/m^3, within the land allotment

The results were validated by conducting field studies directly on the site (Table 2).

Table 2: Change in the dispersion composition of settling dust, depending on the distance from the explosion site

Distance from the explosion site, m	Dispersion composition, % by dust fractions, µm				
	1,4	1,4-4	4-15	15-50	50
40	63,09	23,46	9,03	1,12	1,30
60	68,79	23,13	6,76	0,92	0,40
90	65,74	22,69	9,89	1,66	0,02
120	70,21	19,90	8,62	1,24	0,03
200	74,31	17,52	7,33	0,80	0,04
300	75,11	19,50	4,80	0,57	0,02
600	79,87	15,77	3,70	0,50	0,16

Comparing the results of computer simulation and field measurement data, it is worth noting their convergence. The maximum concentration of different fractional dust particles is observed at the time of formation of the primary dust-gas cloud (explosion flare) directly in the deposit space. Further dispersion of dust particles is determined by the direction and speed of movement of air masses at the walls of the quarry and adjacent territories. Dust particles of larger fractions settle faster, and smaller ones continue their movement over long distances, where can be workers.

4. Discussion and suggestions for future research

The generalized results of the research are important in several aspects, which should be highlighted. Firstly, the mathematical description of aerodynamic processes occurring in the quarry and adjacent areas made it possible to establish quantitative indicators of key parameters of the entire system functioning, taking into account the shape of the quarry contour and the profile of the geological section. Secondly, computer simulation allowed us to determine the speed and trajectories of dust particles and set the zones of their maximum permissible concentration, taking into account meteorological characteristics. It is worth noting that these data are fully consistent with those obtained as a result of field studies at the facility, which indicates the sufficient efficiency of the proposed predictive model. Thirdly, the maximum concentration of multifractional dust particles is observed at the time of formation of the primary dust-gas cloud and further dispersion of dust particles is determined by the direction and speed of movement of air masses at the walls of the quarry and adjacent territories. The presented research complements the known data on the movement of particles in a two-phase flow in terms of a better understanding of the aerodynamics of the process. Thus, modern software tools allow obtaining results that demonstrate full agreement with field measurements, making it possible to implement them at the national level in the field of labor protection and environmental protection technologies.

Prospects for further research should be related to the assessment of the impact of all harmful substances entering the recirculation zone of the quarry from internal and external sources, including loading and unloading and transport operations, which will allow deriving the equation of their overall balance. Prospects for further research also include the use of air curtains in working premises, which will prevent air from entering the premises from the outside, while creating an air and dust barrier to protect workers (Naserzadeh et al., 2017).

5. Conclusions.

As a result of the comprehensive research, the following can be noted:

1) computer simulation of dust particle movement in the working area of quarries and adjacent territories, based on the obtained mathematical dependencies, showed that the maximum concentration of different fractional dust particles (more than 5 mg/m^3) is observed at the time of formation of the primary dust-gas cloud directly in the deposit space. Further dispersion of dust particles is determined by the direction and speed of movement of air masses at the walls of the quarry and adjacent territories. Dust particles of larger fractions settle faster and smaller ones continue their movement over long distances;

2) the most intense release of multifractional dust from the primary cloud is observed at a distance of 40–600 m from the explosion site;

3) the distribution of dust particle motion vectors differs from the wind rose and indicates a significant impact on its annual distribution not only of the transforming wind directions, but also of its speed and the shape of the quarry contour;

4) in terms of a better understanding of the aerodynamics of the process, air flow rates that exceed the critical ones (from 2 m/s) that affect the ability to blow dust from work surfaces were determined;

5) during predicting the overall balance of harmful substances in the quarry space, it is necessary to take into account the consumption of explosives.

All of the above leads to the general conclusion that the results of the predictive computer model can be effectively applied, which can be recommended for implementation in the mining sector in order to prevent the negative impact of pollutant emissions on environmental components and the health of workers.

6. Acknowledgments

The research was carried out under the contract for research work commissioned by Promecologia Scientific Technological Centre, Ltd dated 04.12.2024 No. 597/24-CSGL, which was used at the national level by the Ministry of Ecology and Natural Resources of Ukraine for public discussion of the quarry development project to establish compliance with the requirements for environmental protection and employees of the enterprise. Some of the above research results are publicly available on the Internet in accordance with the Law of Ukraine «On Access to Public Information» dated 08.10.2023 No. 2939-VI.

References

Askarova, A., Bolegenova, S., Maximov, V., Bolegenova, S., Askarov, N., Nugymanova, A. (2021). Computer Technologies of 3D Modeling by Combustion Processes to Create Effective Methods of Burning Solid Fuel and Reduce Harmful Dust and Gas Emissions into the Atmosphere. *Energies*, **14**(5), 1236. https://doi.org/10.3390/en14051236.

Chencheva, O., Lashko, Ye., Rieznik, D., Cheberyachko, Yu., Petrenko, I. (2023). Research of the aerodynamic process of carbon dust removal from the working zone. *Municipal Economy of Cities*, **1**(175), 208–220. https://doi.org/10.33042/2522-1809-2023-1-175-208-220.

CFD Modeling and Simulation. COMSOL Multiphysics. URL : https://www.comsol.com/cfd-module.

Directive (EU) 2024/2881 of the European Parliament and of the Council of 23 October 2024 on ambient air quality and cleaner air for Europe (recast). European Parliament. URL : https://eur-lex.europa.eu/eli/dir/2024/2881/oj/eng.

Giancristofaro, R.A., Bonnini, S., Corain, L., Vidotto, D. (2015) Environmental odor perception: testing regional differences on heterogeneity with application to odor perceptions in the area of Este (Italy). *Environmetrics*, 26, 418–430. https://doi.org/10.1002/env.2341.

Honcharevskyi, K., Pikovets, V., Mushtaiev, O. (2013). Methodology for calculation of pollutant emissions into the atmosphere during mining (open method). *Ecological Sciences*, **2**(4), 68–93.

Konglok, S. A. Pochai, N. (2016). Numerical computations of three-dimensional air-quality model with variations on atmospheric stability classes and wind velocities using fractional step method. *IAENG International Journal of Applied Mathematics*, **46**(1), 112–120.

Kumar, A., Schafrik, S. (2020). Multiphase CFD modeling and laboratory testing of a Vortecone for mining and industrial dust scrubbing applications. *Process Safety and Environmental Protection*, 144, 330–336. https://doi.org/10.1016/j.psep.2020.07.046.

Lashko, Y., Chencheva, O., Laktionov, I., Rieznik, D., Halchenko, N. (2024). Mathematical and Computer Simulation of the Process of Movement of Respirable Dust Particles in the Working Area. *Baltic Journal of Modern Computing*, **12**(3), 270–285. https://doi.org/10.22364/bjmc.2024.12.3.04.

Liaskoni, M., Peter, H., Bartik, L., Perez, A., Karlický, J., Vlček, O. (2023). Modelling the European wind-blown dust emissions and their impact on particulate matter (PM) concentrations. *Atmospheric Chemistry and Physics*, 23, 3629–3654. https://doi.org/10.5194/acp-23-3629-2023.

Liu, W., Huang, X., Chen, H., Han, L. (2022). Analyzed and Simulated Prediction of Emission Characteristics of Construction Dust Particles under Multiple Pollution Sources. *Computational Intelligence and Neuroscience*, 7349001. https://doi.org/10.1155/2022/7349001.

Liu, Q., Nie, W., Hua, Y., Jia, L., Li, C., Ma, H., Wei, C., Liu, C., Zhou, W., Peng, H. (2019). A study on the dust control effect of the dust extraction system in TBM construction tunnels based on CFD computer simulation technology. *Advanced Powder Technology*, **30**(10), 2059–2075. https://doi.org/10.1016/j.apt.2019.06.019.

Naserzadeh, Z., Atabi, F., Moattar, F., Nejad, N. M. (2017). Effect of barriers on the status of atmospheric pollution by mathematical modeling. *Bioscience Biotechnology Research Communications*, **10**(1), 192–204. https://doi.org/10.21786/bbrc/10.1/29.

On Approval of the Procedure for Determining the Values of Background Concentrations of Pollutants in the Atmospheric Air, Order of the Ministry of Ecology and Natural Resources of Ukraine No. 286 (2021) (Ukraine). https://zakon.rada.gov.ua/laws/show/z0700-01#Text.

Oyjinda, P., Pochai, N. (2017). Numerical Simulation to Air Pollution Emission Control near an Industrial Zone. *Advances in Mathematical Physics*, 1–7. https://doi.org/10.1155/2017/5287132.

Tong, R., Cheng, M., Zhang, L., Liu, M., Xiaoyi, Y., Li, X., Yin, W. (2018). The construction dust-induced occupational health risk using Monte-Carlo simulation. *Journal of Cleaner Production*. 184. https://doi.org/10.1016/j.jclepro.2018.02.286.

Xiu, Z., Nie, W., Yan, J., Chen, D., Cai, P., Liu, Q., Du, T., Yang, B. (2020). Numerical simulation study on dust pollution characteristics and optimal dust control air flow rates during coal mine production. *Journal of Cleaner Production*, 248, 119197. https://doi.org/10.1016/j.jclepro.2019.119197.

Zhou, G., Liu, Y., Kong, Y., Hu, Y., Song, R., Tian, Y., Jia, X., Sun, B. (2022). Numerical analysis of dust pollution evolution law caused by ascensional/descensional ventilation in fully mechanized coal mining face based on DPM-DEM model. *Journal of Environmental Chemical Engineering*, **10**(3). https://doi.org/10.1016/j.jece.2022.107732.

Received January 28, 2025, revised April 27, 2025, accepted June 4, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 2, 453-485
https://doi.org/10.22364/bjmc.2025.13.2.09

Supply Chain Analytics: A Performance Evaluation of Machine Learning, Statistical, and Time-Series Models

Vimal DWIVEDI[1], Karuna KADIAN[2], Bryan GARDINER[1],
Rachana PANDEY[2], Anjali LATHWAL[2]

[1] Intelligent Systems Research Centre, Ulster University, Northern Ireland, UK
[2] Indira Gandhi Delhi Technical University for Women, India

v.dwivedi@ulster.ac.uk, karunakadian@igdtuw.ac.in,
b.gardiner@ulster.ac.uk, rachana033mtcse21@igdtuw.ac.in,
anjalilathwal@igdtuw.ac.in

ORCID 0000-0001-9177-8341, ORCID 0009-0004-8929-5999, ORCID 0000-0001-5642-6850,
ORCID 0000-0002-9404-2548, ORCID 0000-0001-9810-8893

Abstract. Supply chain analytics is pivotal in enhancing supply chain performance through data-driven decision-making. This study evaluates the effectiveness of various analytical models in improving supply chain performance using the DataCo Smart Supply Chain dataset. This research comprehensively evaluates machine learning, statistical, and time-series models for forecasting accuracy, demand prediction, late delivery risk, shipment duration, and route mapping optimisation. The study methodically compares and contrasts the outcomes of various modelling techniques, providing valuable insights into the most suitable approaches for optimising supply chain operations. The results indicate that decision tree and random forest models excel in supply chain forecasting and sales prediction. Similarly, Random Forest, XGBoost, and Gradient Boost models accurately predict late delivery risk, while Exponential Smoothing, SARIMA, and ARIMA models effectively predict shipment duration. To validate these findings, rigorous statistical testing, cross-validation, and alignment with industry standards were employed, ensuring the reliability and applicability of the results. This research contributes significantly to supply chain analytics, offering practitioners and researchers guidance on selecting appropriate methodologies for enhanced supply chain performance.

Keywords: Supply chain management, Machine Learning, Simulations

1 Introduction

The supply chain is a complex network consisting of various organisations and enterprises involved in the production and delivery of goods and services to consumers, as illustrated in Figure 1, adapted from (Anitha and Patil., 2018). Efficient management of the supply chain is essential for the success of any company, as it requires the identification of inefficiencies and the implementation of strategies to optimise the network (Samir, 2023). However, this task is complicated by the participation of multiple stakeholders and the need to manage a wide range of resources.

In recent years, advancements in big data and analytics have provided organisations with new tools to enhance their supply chain management practices (Raman et al., 2018), Supply chain analytics, which leverages data analytics and quantitative methodologies, play a crucial role in optimising supply chains by applying sophisticated algorithms and statistical models to large datasets (Surie and Reuter, 2014). This approach enables businesses to improve efficiency, enhance customer satisfaction, and reduce operational costs.

This study seeks to systematically evaluate the performance and accuracy of ma- chine learning, time series, and statistical models across four specific use cases in supply chain analytics utilising the “DataCo supply chain” dataset (Constante et al., 2021). These use cases include supply chain forecasting, late delivery risk, shipment duration prediction, and route mapping. The selection of these particular four use cases was guided by the practical strengths and constraints of the “DataCo Supply Chain” dataset we considered in our study. Dataset, though, is quite rich in information but has significant gaps in customer-level or product-level fields, namely “Customer Demographics”, “Product Description”, etc. Such limitations restrict the exploration to other areas, like customer behavior analysis or location-based trends. Our use cases mainly deal with the operational and feasibility significance based on the available dataset. Through a comprehensive comparison of these models, this research aims to identify the most effective models for each scenario, thereby supporting informed decision-making in supply chain management. To ensure a thorough evaluation, the assessment criteria are aligned with the ISO 25010 standard, incorporating essential quality attributes such as functionality, usability, efficiency, reliability, maintainability and portability as defined by the standard.

This paper presents an experimental research study complemented by survey elements. The experimental aspect focuses on evaluating and comparing various models within supply chain analytics, while the survey offers a comprehensive review of the current state of knowledge, setting the context for the proposed investigations. The integration of a survey within an experimental framework provides a holistic understanding of the field and introduces new empirical insights.

To achieve the best prediction models, data pre-processing, exploratory data analysis (EDA), feature selection, and algorithm selection was performed. The models were evaluated based on r2 score, root mean square error (RMSE), and mean square error (MSE). The findings from this research can assist businesses in selecting appropriate models for supply chain analytics.

The remainder of the paper is structured as follows: The rest of this Section introduces the study, outlines its objectives, and the significance of supply chain analytics. Section 2 reviews previous research in supply chain analytics and identifies gaps in the existing literature. Section 3 details the research methodology, including data pre-processing, EDA, feature selection, algorithm selection, prediction models, and evaluation metrics. Section 4 presents the findings for each use case, supported by tables and graphs. Section 6 summarise the findings, discuss critical implications for supply chain management, highlight limitations, and propose future research directions.

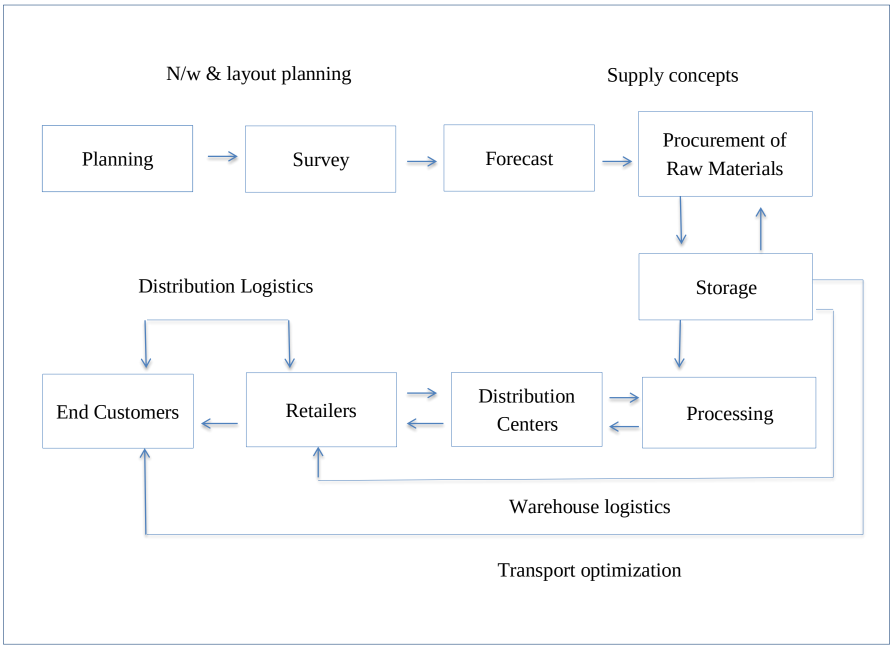


Fig. 1. Process of Supply Chain Management

1.1 Background and motivation for the study:

Advanced analytics in supply chain management has become increasingly important due to its ability to enhance both efficacy and efficiency (Zekhnini et al., 2020). Supply chain analytics leverages statistical and computational tools to analyze data, thereby improving decision making in areas such as inventory management, transportation, and procurement.

Machine learning, statistical, and time-series models are commonly employed in supply chain management to predict key performance indicators (KPIs) such as sales forecasting, delivery timeframes, and transportation costs (Hahn, 2019). However, there is a lack of comprehensive research evaluating the performance of these models across different supply chain use cases.

The study focuses on determining the accuracy and effectiveness of various models within the four specified supply chain analytics use cases. The goal is to identify models that not only perform well but also meet industry standards for accuracy and efficacy.

1.2 Research questions and objectives:

– This study aims to address the following research questions:

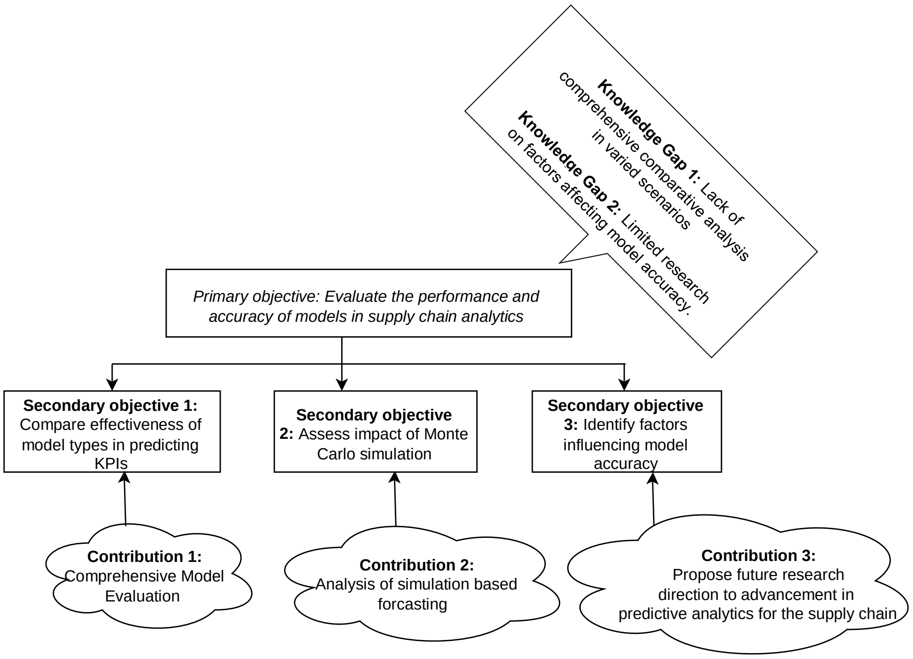

Knowledge Gap 1: Lack of comprehensive comparative analysis in varied scenarios
Knowledge Gap 2: Limited research on factors affecting model accuracy.
Primary objective: Evaluate the performance and accuracy of models in supply chain analytics
Secondary objective 1: Compare effectiveness of model types in predicting KPIs
Secondary objective 2: Assess impact of Monte Carlo simulation
Secondary objective 3: Identify factors influencing model accuracy
Contribution 1: Comprehensive Model Evaluation
Contribution 2: Analysis of simulation based forcasting
Contribution 3: Propose future research direction to advancement in predictive analytics for the supply chain

Fig. 2. Goal tree of research objectives and their corresponding contributions

1. Which type of model (machine learning, time series or statistical) performs best in each of the four use cases of supply chain analytics?
2. How does the performance of these models change when incorporating Monte Carlo simulation?
3. What are the key factors that most significantly impact the accuracy of these models in each use case?

– The primary objectives and sub-objectives of this research are listed below and visually depicted in Figure 2:

1. To assess the performance and accuracy of machine learning, statistical and timeseries models in predicting key performance indicators (KPIs) across various supply chain management scenarios.
2. To compare these models in four critical supply chain analytics use cases, ensuring alignment with the ISO25010 quality standards.
3. To analyse the impact of Monte Carlo simulation on model performance.
4. To identify critical factors influencing model accuracy in each use case, thereby

providing a comprehensive evaluation of model efficacy in supply chain analytics.

1.3 Research Methodology:

The research methodology employed in this study involves a systematic process that includes a literature survey, data collection, data pre-processing, exploratory data analysis (EDA), feature selection, algorithm selection, prediction model development and model evaluation using appropriate metrics (see Figure 3).

1.3.1 Literature Survey: A comprehensive literature survey was conducted, covering leading journals and conference proceedings in the field of supply chain management, analytics, artificial intelligence (AI), and machine learning (ML). This survey informed the research objectives, guided the framework for model selection and evaluation, and highlighted the role of Monte Carlo simulation in enhancing forecast accuracy. The literature survey was carried out to understand the current state of the art of the research in supply chain analytics and to locate gaps or areas that are not well explored. It also helped us to select the most relevant use cases and the respective methods, and ensured that our study not only offered something new but also built on the foundation of what already exists.

1.3.2 Data Collection: The primary dataset used in this study is the "DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS" which provides comprehensive information relevant to supply chain operations. The dataset was acquired and prepared for analysis, with careful attention to maintaining data quality and integrity.

1.3.3 Model Development: The model development process involved data preprocessing, exploratory data analysis (EDA), and feature selection to prepare the dataset for modeling.

1.3.4 Algorithm Selection: Drawing on insights from the literature review and the specific requirements of each use case, a comprehensive set of models and algorithms were selected for evaluation.

1.3.5 Prediction Model Development: Prediction models were developed for each use case using the selected models and algorithms. These models were trained on the pre-processed dataset, with techniques and parameters tailored to each model type to ensure accuracy and reliability.

1.3.6 Model Evaluation: The performance of prediction models was evaluated using standard assessment measures such as the accuracy, R2 score, RMSE, and MSE. These metrics were chosen based on the type of task and model. For regression problems such as forecasting and shipment duration prediction, we used metrics such as RMSE and MAE to measure the performance of the model. Late delivery risk and other classification-type tasks' performance are measured by the accuracy metric across balanced and imbalanced data. All these metrics are standard, interpretable, and are the scientifically validated way to assess the performance of classification and regression models. The evaluation process involved comparing the performance of various models within each use case to identify those with the highest accuracy and reliability.

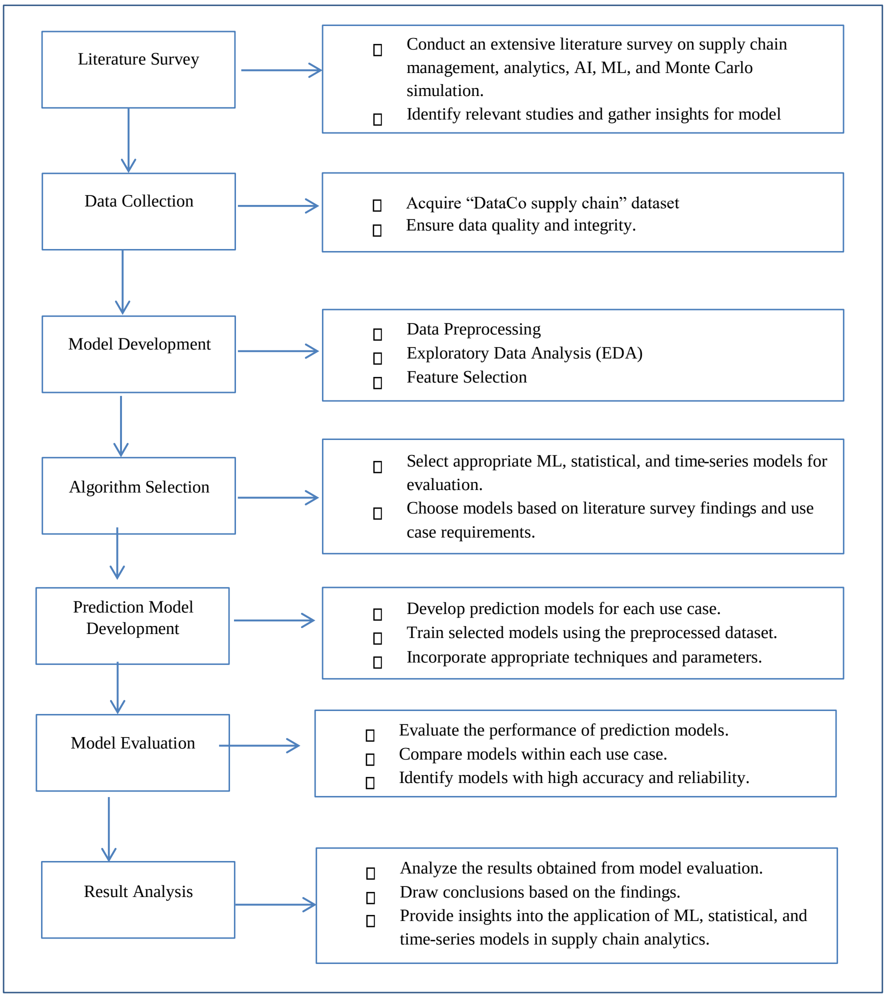

Literature Survey
□ Conduct an extensive literature survey on supply chain management, analytics, AI, ML, and Monte Carlo simulation.
□ Identify relevant studies and gather insights for model
Data Collection
□ Acquire “DataCo supply chain” dataset
□ Ensure data quality and integrity.
Model Development
□ Data Preprocessing
□ Exploratory Data Analysis (EDA)
□ Feature Selection
Algorithm Selection
□ Select appropriate ML, statistical, and time-series models for evaluation.
□ Choose models based on literature survey findings and use case requirements.
Prediction Model Development
□ Develop prediction models for each use case.
□ Train selected models using the preprocessed dataset.
□ Incorporate appropriate techniques and parameters.
Model Evaluation
□ Evaluate the performance of prediction models.
□ Compare models within each use case.
□ Identify models with high accuracy and reliability.
Result Analysis
□ Analyze the results obtained from model evaluation.
□ Draw conclusions based on the findings.
□ Provide insights into the application of ML, statistical, and time-series models in supply chain analytics.

Fig. 3. Steps used in research methodology

The way this study is done makes sure we look carefully at how well time-series, statistical, and machine-learning models work for supply chain analysis. It also discusses the research goals and gives valuable advice using these models.

1.4 Brief overview of the data and the four use cases selected:

In this study, we utilised a dataset of “DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS” (Constante et al., 2021), which encompasses various supply chain components such as sales, transportation, and inventory. Developed by DataCo Global, this dataset supports an intelligent supply chain solution designed for extensive data analysis, particularly in sectors such as clothing, sports equipment, and electronics. The dataset, named DataCoSupplyChainDataset.csv, enables users to analyse and process structured data related to Provisioning, Production, Sales, and Commercial Distribution activities using machine learning methods and R software. Additionally, the system can correlate structured and unstructured data from Clickstream logs, tokenized in a file called tokenized access logs.csv, enabling the identification of patterns and trends within the supply chain.

The DataCoSupplyChain.csv file provides a comprehensive explanation of each variable in the DataCoSupplyChainDataset.csv. This system is classified under Data Mining, Management of Supply Chain, Big Data Analytics and ML. The latest version, Version 5, was published on March 13, 2019. By leveraging DataCo's intelligent supply chain technology, businesses can gain a competitive advantage by optimising their supply chain processes and making data-driven decisions, ultimately fostering business grow.

Four use cases in supply chain analytics were selected from this dataset:

1. Supply Chain Forecasting Sales Prediction: In this use case, we predict sales for a specific period using historical sales data.
2. Late Delivery Risk: Here, we estimate the probability of a shipment being delayed, considering factors such as weather conditions, transportation mode and delivery distance.
3. Predicting Shipment Duration: This use case involves predicting the time required to deliver a shipment based on variables such as transportation mode, distance and weather conditions. Relevent studies on predicting shipment duration can be found in (Mariappan et al., 2022) and (Sahoo et al., 2021).
4. Route Mapping: Route mapping: This use case focuses on identifying the optimal shipment route based on transportation mode, delivery distance, and road conditions.

For each of these use cases, statistical, machine learning and time-series models were applied to forecast the relevant Key Performance Indicators (KPIs) (Estampe et al., 2013). Additionally, Monte Carlo simulations can be performed to evaluate the accuracy of these models under varying conditions. We also analyse the key factors influencing the accuracy of the models for each use case.

1.5 Selection of use cases for evaluation:

This study examines the effective application of machine learning, time-series and statistics models to enhance the understanding of supply chain dynamics. Four key areas were selected for evaluation: sales prediction, late delivery risk assessment, shipment duration forecasting, and route optimisation. These use cases were chosen due to their significant impact on such situations on supply chain management.

1.5.1 Supply Chain Forecasting - Sales Prediction: Accurate sales forecasting is critical in supply chain management as it directly influences business planning, production, inventory management, and logistics. Effective forecasting ensures the customer demands are met while optimising costs and resource allocation. Therefore, it is essential to evaluate the effectiveness of different models in predicting sales within the supply chain.

1.5.2 Late Delivery Risk: The risk of late deliveries is a major concern in supply chain management, as it can lead to customer dissatisfaction, increased costs, and scheduling disruptions By assessing the accuracy of models in predicting delivery delays, businesses can proactively address issues, ensuring timely deliveries and enhancing customer satisfaction.

1.5.3 Predicting Shipment Duration: Accurately predicting shipment duration is essential for effective supply chain management, scheduling, and resources optimisation. Evaluating the precision of models in forecasting delivery times allows businesses to improve customer service, optimise inventory management and streamline transportation processes.

1.5.4 Route Mapping: Optimising delivery routes is essential for essential for efficient logistics, cost reduction, and overall supply chain performance. By assessing the effectiveness of models in identifying optimal routes considering various factors, the study aims to determine their potential in improving transportation efficiency and enhancing the overall supply chain.

By focussing on these four scenarios, the study aims to provide a comprehensive assessment of model performance acrossvarious aspects of supply chain analytics. Each scenario addresses a distinct supply chain management challenge, highlighting the flexibility of the models. The outcomes of these evaluations will identify the strengths and potential limitations of each model, guiding the selection of the the most appropriate tools for specific supply chain tasks.

While this study focused on evaluating the effectiveness of machine learning, statistical, and time-series models in four specific deliver chain analytics use cases, there remains several other potential use cases that warrant further exploration. Future research should investigate those additional areas to provide a more comprehensive evaluation supply chain analytics. Potential use cases include demand planning, inventory management, production, transportation planning, risk management, and compliance.

Each of these domains present unique challenges and opportunities for utilising analytical models to enhance supply chain efficiency and effectiveness. By broadening the scope of analysis to include these additional use cases will enable academics and practitioners to gain a deeper understanding of the capabilities and limitations of various modelling approaches in addressing the complexities of supply chain management. This broader investigation will contribute to advancement and refinement of supply chain analytics practices, ultimately leading to improved decision-making and performance in supply chain operations.

1.6 Motivation for Model Selection:

Several key factors influenced the selection of models and algorithms in this research paper. Firstly, the study aimed to evaluate the performance of machine learning (ML), statistical and time-series models, given their proven effectiveness across various domains. By comparing these diverse modeling techniques within the context of supply chain management, the research sought to gain insights into their effectiveness and applicability.

Secondly, a diverse set of models was incldued to explore the strengths and weaknesses of each approach. Machine learning models such as Linear Regression, Lasso Regressor, Ridge Regressor, KNN Classifier, Gaussian Naive Bayes, SVM Classifier, Gradient Boost, Decision Tree, Random Forest, LSTM Model, and Gated Recurrent Unit are well-known for their ability to capture complex patterns and relationships in data. Statistical models, including the Chi-square test, OLS, T-test, Multiple Linear Regression, and Kruskal-Wallis test, provide rigorous methods for analysing correlations and drawing statistically significant inferences. Time-series models such as the Exponential Smoothing Model, SARIMA, and ARIMA are specifically tailored to accommodate temporal data and capture time-dependent trends.

Thirdly, the literature review revealed that machine learning models are commonly employed in supply chain analytics research. However, this study sought to explore the performance of alternative modelling approaches, such as statistical and time-series models, to provide a more comprehensive evaluation of their suitability for supply chain analytics.

Lastly, the selected use cases of supply chain forecasting, late delivery risk, predicting shipment duration, and route mapping were carefully chosen to cover a broad spectrum of supply chain management challenges. By evaluating multiple models across these diverse use cases, the study aimed to understand how different modelling strategies perform under distinct supply chain scenarios.

The objective was to conduct a comprehensive evaluation of machine learning, statistical, and time-series approaches in supply chain analytics, compare their performance, and assess their suitability for addressing real-world supply chain challenges. By incorporating a wide range of models, the research aimed to provide valuable insights into their strengths and limitations, assisting researchers and practitioners in making informed decisions regarding model selection in supply chain management.

2 Literature review

2.1 Literature Review Methodology

2.1.1 Study Identification To ensure a comprehensive understanding of existing research in supply chain analytics, a structured literature review approach was undertaken. The review focused on identifying studies related to the application of machine learning, statistical, and time-series models in supply chain forecasting, risk prediction, shipment duration estimation, and route optimization.

Search Keywords and Queries: The following search strings were used, combining key concepts with Boolean operators:

> (”supply chain analytics” OR ”supply chain forecasting” OR ”supply chain prediction”) AND (”machine learning” OR ”time series” OR ”statistical models” OR

”predictive analytics”) AND (”model evaluation” OR ”model comparison” OR ”performance assessment”)

Search Sources: The literature search was conducted across Scopus, Web of Science, IEEE Xplore, and Google Scholar. Table 1 mentions the number of articles across the respective sources considered as part of the review. These databases were selected for their extensive coverage of peer-reviewed journals and conference proceedings relevant to supply chain management and predictive analytics.

Table 1. Study sources and identified paper count.

Search Engine	Databases / Coverage	Article Count
Scopus*	Peer-reviewed journals, conference proceedings	104
Web of Science*	Peer-reviewed journals, conference proceedings	90
IEEE Xplore*	Conference papers, technical magazines	65
Google Scholar*+	Grey literature, technical reports, working papers	482
	Total	741
	Total (after duplicate removal)	452

*Includes journals, proceedings, and book chapters
+Includes technical reports, white papers, and working papers

2.2 Study Selection

To further refine the articles identified in the initial search, a two-phase screening process was applied: the inclusion phase and the exclusion phase. This process ensured that only relevant and high-quality articles aligned with the study’s objectives were retained.

A two-stage screening process was applied:

Table 2. Inclusion Criteria.

Criteria ID	Inclusion Criteria
IC1	The article applies machine learning, statistical, or time-series models for supply chain forecasting, risk prediction, shipment duration estimation, or route optimisation.
IC2	The study includes empirical validation of predictive models using real-world or benchmark datasets.
IC3	The article provides performance evaluation metrics (e.g., accuracy, RMSE, F1-score) for model comparison.
IC4	The article is published in peer-reviewed journals, conference proceedings, or technical reports.
IC5	The publication is within the date range of January 2013 to December 2024.

Table 3. Exclusion Criteria.

Criteria ID	Exclusion Criteria
EC1	The article focuses solely on supply chain strategy, policy, or qualitative frameworks without predictive modelling.
EC2	The article lacks empirical results or does not report performance evaluation of predictive models.
EC3	The study focuses on unrelated domains (e.g., manufacturing processes, IoT hardware) without supply chain forecasting context.
EC4	Non-English language publications.
EC5	Duplicate publications or extended versions of already included papers.

- *Title and Abstract Screening:* Articles unrelated to predictive modelling or lacking empirical validation were excluded, reducing the pool to 97 articles.
- *Full-Text Review*: Studies were further assessed for relevance based on methodological alignment (use of ML/statistical/time-series models), application to supply chain forecasting, risk, duration, or routing, and availability of performance evaluation results.

Final Sample Size: 32 articles as specified in the Table 4 were selected as the final operational sample, which formed the basis of this study's literature review and comparative analysis.

Following this systematic search and selection process with specific inclusion and exclusion criteria outlined in Table 2 and 3, the literature was thematically analyzed to provide a comprehensive understanding of current methodologies, identified gaps, and opportunities for further research. The subsequent sections summarize key findings from the reviewed studies, beginning with general approaches to supply chain forecasting, followed by examining specific modeling techniques and simulation methods.

Table 4. Number of articles selected from each selection phase.

Search Stages	Identified Articles
Study Identification Phase	741 452 *(after duplicate removal)*
Study Selection Phase (Title & Abstract Screening)	97 Primary Studies
Full-Text Review Phase	32 Selected Studies (Operational Sample)

2.3 Previous research on supply chain management and forecasting

Supply chain management involves optimising and coordination of all processes in the production and delivery of goods and services. Forecasting plays a crucial role in predicting demand, determining production needs, managing inventory levels, and setting delivery schedules. Although various methods have been proposed to improve prediction accuracy in supply chain management, a comprehensive comparison across a broad spectrum of models (machine learning, statistical, and time-series) across different use cases remains underexplored.

Time-series models, such as ARIMA, have been extensively employed due to their ability to identify trends and seasonal patterns. For instance, (Fattah et al., 2018) utilised ARIMA to forecast demand in the food industry. However, such studies, while effective, often lack a comparative analysis across different model types in a unified framework, which is the focus of this study

Machine learning algorithms are continuing to gain traction in supply chain forecasting. Studies such as (Guanghui et al., 2012) and (Kilimci et al., 2019) investigate the use of algorithms such as Support Vector Regression and deep learning methods. Despite these advancements, there remain a significant gap in the literature regarding a comprehensive evaluation of these models against traditional statistical and time-series models across multiple use cases - a gap that this study aims to address.

Recent research, including the work of (Terrada et al., 2022), has introduced hybrid forecasting models that combine different methodologies. However, these studies often do not provide a detailed comparative analysis of the effectiveness of machine learning, statistical, and time-series models in various supply chain scenarios. Addressing this gap is crucial to answering the research questions posed in this study.

The impact of Monte Carlo simulation on model performance is another area that has been insufficiently explored in existing literature, including research by (Terrada et al., 2022) and others. This research seeks to fill this gap by investigating how Monte Carlo simulation influences the performance of various model types in supply chain analytics.

While previous studies has often focused on evaluating model performance under specific conditions, it has not thoroughly explored the key factors influencing model accuracy across different supply chain management use cases. This study aims to address this significant gap in the existing body of research by identifying and analysing these critical factors.

2.4 Summary of the study's machine learning models

According to (Terrada et al., 2022), machine learning has become a vital tool for evaluating and modelling complex data in various domains. In the literature, supervised learning models such as SVM, linear regression, random forests, decision trees, and ANN have been employed yo address issues including demand forecasting, transportation planning, and inventory control. (Dash et al., 2019) highlight the importance of a well-functioning supply chain and the benefits of using Artificial Intelligence (AI) to optimise inventory forecasting and customer demand prediction. The authors argue that AI can enhance asset utilisation, increase revenue, and reduce costs by providing highly accurate predictions, optimising R & D and manufacturing processes, improving promotional strategies, and enhancing the customer experience. For instance, (Wan, 2021) proposes a hybrid model that combines the XGBoost algorithm with Random Forest to detect product fraud. This model can assist businesses better align their strategies with market demands and boost revenue. The hybrid model outperforms traditional machine learning techniques in terms of the F1 score, achieving increases of 0.49, 0.49, and 27.9 points over Gaussian Naive Bayes, SVM, and Logistic Regression. The study tested the proposed model using DataCo's innovative supply chain dataset.

In contrast, unsupervised learning models, such as clustering and association rule mining, are utilised to analyse data patterns and identify hidden structures within datasets. Additionally, statistical models, including the Chi-square test, Ordinary Least Squares (OLS), T-test, Multiple Linear Regression and Kruskal-Wallis test, have been used in literature to model and analyse supply chain data (Morgenthaler, 2009).

Time series models such as SARIMA, ARIMA, and exponential smoothing are commonly employed to analyse and predict time-series data in the supply chain. ARIMA, a popular time-series model, has been widely used to forecast various supply-chain factors, including demand, inventory levels, and delivery schedules. SARIMA extends ARIMA's capabilities to account for seasonality in the data. Exponential smoothing, another time-series technique, has been applied to demand forecast in the supply chain. For instance, (Nguyen et al., 2021) propose two data-driven strategies to enhance decision-making in supply chain management. The first strategy utilises an LSTM network-based approach to forecast multivariate time-series data by integrating internal and external business sources of data to improve performance. The second strategy involves detecting sales anomalies using a combination of a one-class SVM algorithm with an LSTM Autoencoder network-based technique. These methods were evaluated on real-world data from the fashion retail industry and benchmarking datasets, demonstrating superior performance compared to previous research. Additionally, (Shih et al., 2019) discusses two data-driven approaches for estimating the supply of blood components at blood centres to decrease blood wastage and shortage. The study compared time-series and machine learning techniques using five years of historical blood supply data from the Taiwan Blood Services Foundation (TBFS). The findings suggest that decision support systems for executives and pathologists at blood centres, hospitals and blood donation facilities can benefit from time-series forecasting techniques, particularly seasonal ESM and ARIMA models, which were found to outperform machine learning algorithms in terms of accuracy.

In addition to the aforementioned predictive models, simulation-based techniques have also been employed to enhance model robustness and manage uncertainties in supply chain analytics. The following subsection focuses specifically on the role of Monte Carlo simulation in this context.

2.5 Previous research on Monte Carlo simulation and its impact on prediction accuracy:

Monte Carlo simulation is a widely used technique in statistics and modelling, which leverages random number generation to simulate and analyse the behaviour of complex systems or processes. in supply chain analytics, Monte Carlo simulation is frequently employed to assess uncertainties related to demand forecasts, inventory levels, and overall supply chain performance. For example, (Schmitt et al., 2009) discusses a project aimed at evaluating the risk of supply chain disruptions and developing mitigation strategies. The authors constructed a simulation model utilising Arena and Monte Carlo simulations to evaluate the disruption risk at various supply chain nodes and the impact of these disruptions on customer service. This approach enabled a straightforward risk assessment and outcome estimation for adverse events. The initiative made several insightful points, such as no specific measures for mitigation at the strategic level. The project also highlighted the absence of specific strategy-level mitigation measures and the importance of historical distribution data in the database. The model's results indicate the need for operational changes within the network to maintain system performance, even under stable conditions. By employing Monte Carlo simulation, researchers can evaluate the effects of different scenarios on supply chain performance.

Researchers have discovered that integrating Monte Carlo simulation as part of a supply chain forecasting system, in combination with advanced modelling techniques such as time-series analysis and machine learning, can significantly enhance prediction accuracy. For instance, (Mohamed et al., 2020) explores the challenge of calculating the gradient of an expectation of a function which is particularly relevant in applied statistics, machine learning, and computational finance. The study presents three main estimators for computing gradients: score-function, measure-valued gradient and pathwise. It also provides guidance on selecting between these estimators based on problem characteristics, such as the number of parameters and whether gradients are decentralised, exhibit high variance, or do not exist (e.g., in cases lacking foundational knowledge in differential calculus). The study further addresses variance reduction techniques and recommends testing the unbiasedness of the gradient estimator. The report concludes with case examples and research suggestions.

Additionally, (Foerster et al., 2018) introduces DICE as a general method for constructing order gradient estimators suitable for stochastic computation graphs. DICE aims to address limitations in current methods for estimating higher-order gradient estimates. The study proposes a practical implementation of DICE for deep learning frameworks, testing its validity and utility in a multi-agent reinforcement learning scenario. The findings demonstrate the accuracy of DICE estimators, and the authors suggest that DICE could facilitate further exploration and adoption of higher-order learning approaches in metalearning, reinforcement learning, and other stochastic computation graph applications. Similarly, (Bousqaoui et al, 2017) compares the performance of several machine learning algorithms in supply chain management using Monte Carlo simulation, while (Thete, 2022) present a stochastic model for demand forecasting in Python, employing the Time-Series SARIMA model alongside Monte Carlo simulation. Overall, previous research has demonstrated that Monte Carlo simulation can enhance the accuracy and resilience and robustness of supply chain predictions, particularly when combined with advanced modelling techniques like machine learning and time-series analysis.

3 Methodology

The following section provides a detailed overview of the proposed methodology and the processes involved. The flowchart illustrating this method is depicted in Figure 4.

3.1 Data source and preprocessing

3.1.1 Data Source.

The “DataCo supply chain” dataset, which was publicly available on Kaggle, was used in this study. The dataset contains transactional data from an online retail organisation that has dealt with electronic products for over one year. The dataset has 180,519 rows and 53 columns. The data is structured and includes numerous features of a customer’s transaction. The methodology we adopted for this study is designed to evaluate and compare various supply chain analytics models systematically. This approach ensures a complete understanding of model performance in different scenarios within the supply chain domain.

3.1.2. Data Preprocessing

To ensure the accuracy, integrity, and readiness of the dataset for model development, a systematic preprocessing approach was undertaken. This phase encompassed data cleaning, transformation, and feature selection—each essential for ensuring analytical reliability and facilitating optimal model performance.

The following key steps were implemented:

1. **Handling Missing Values:** The dataset contained missing values across several attributes. The Product Description column, which lacked any usable entries, was removed entirely. Similarly, the Order Zipcode field exhibited over 85% missing data and was excluded due to insufficient representativeness. In contrast, the Customer Lname column had a moderate level of missing values, which were imputed using the placeholder value "Unknown" to preserve data completeness without introducing bias.
2. **Data Type Conversion:** Several fields were stored in inappropriate data types. For instance, Order Date and Shipping Date, originally in object format, were converted to datetime format. This transformation enabled the extraction of temporal features and supported subsequent time-series analysis.
3. **Feature Selection:** Features were selected based on a combination of domain relevance, statistical correlation, and model interpretability. This process ensured that only the most informative and non-redundant attributes were retained for training. By

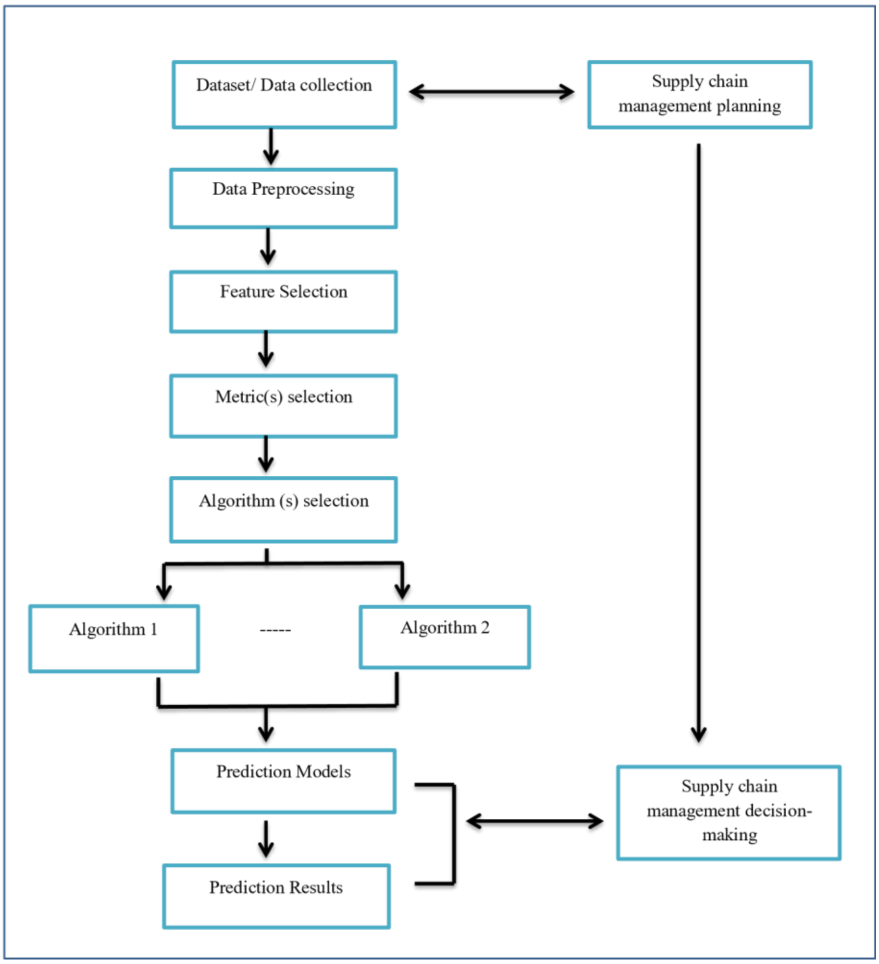


Fig. 4. Supply Chain Analytics - Performance Evaluation Flowchart

reducing dimensionality, the models were rendered more efficient, interpretable, and robust to overfitting.

4. **Data Cleaning:** The dataset was further refined by eliminating duplicate records, correcting inconsistencies, and addressing outliers. Duplicate rows were identified by checking for identical entries across all fields and were subsequently removed. Outliers were detected using z-score normalisation ($|z| > 3$) and visually confirmed via box plots. Depending on their contextual validity, they were either removed or capped. Additional data inconsistencies were resolved using domain knowledge to maintain data integrity.
5. **Data Integration:** Relevant information distributed across multiple tables was

consolidated into a unified dataset using unique identifiers such as Order ID and Order Item ID. This integration facilitated coherent analysis across the various supply chain scenarios addressed in the study

The resulting structured and cleaned dataset formed the analytical foundation for the four use cases examined in this research: *Supply Chain Forecasting – Sales Prediction, Late Delivery Risk, Predicting Shipment Duration, and Route Mapping.*

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an essential step in gaining insights into the dataset, helping to identify patterns, irregularities and relationships essential for the modelling process. This phase informs data preparation and model selection, ensuring that the analysis is grounded in a thorough understanding of the underlying data characteristics.

EDA is particularly important in supply chain analysis, as highlighted by (Colicchia et al., 2018) and (Dohale et al., 2021). It aids in understanding the data, detecting outliers, and perform descriptive statistics. EDA is necessary according to (Morgenthaler, 2009) and (Li Vigni et al., 2013) as it helps to understand the significant aspects of a dataset such as its distributions, patterns, outliers, correlations and missing values. EDA gives data insights and helps make informed decisions about data preparation, feature engineering, and modelling, resulting in improved results. This, in turn, enhances the overall modelling process and results.

In this section, the dataset from DataCo is explored in detail, focusing on its properties and characteristics. The dataset contains 180,519 records with 53 attributes, which includes a mix of categorical, numerical and date-time data. The dataset contains information on customers, orders, products and shipping details, and is analysed through four supply chain use cases.

The EDA process begins with an investigation of the summary statistics of the numerical columns, providing an overview of the central tendencies, variances and other key metrics. The distribution of these variables is then visualised using histograms, box plots or density plots, focusing on metrics such as shipping, sales per customer, and benefit per order. For categorical variables such as Delivery Status, delivery, Customer Segment, etc., bar plots are used to display the frequency of each category.

Relationships between variables are analysed using scatter plots, correlation matrices or heat maps. For instance, the relationship between Sales per customer and Order Item Total is explored to determine any correlation. Additionally, potential links between Late delivery risk and Delivery Status are investigated.

Outliers, missing values and inconsistencies are also identified during EDA. Outliers are detected using box or scatter plots, while missing values are highlighted through summary statistics or visualisation techniques such as heat maps.

Finally, the distribution of data across different categories is examined. For example, the distribution of shipping days across various customer segments or order regions is analysed to uncover patterns or differences. Similarly, the distribution of Late delivery risk across different market segments or shipping modes is examined to identify trends or patterns.

Overall, EDA provides a comprehensive understanding of the dataset, helping to identify patterns, trends, and guiding the selection of appropriate models for addressing the research questions.

3.3 Feature Selection and Engineering

The primary goal of feature selection and engineering in this study was to refine the dataset to enhance model accuracy, reduce computational complexity, and increase interpretability. This process involved identifying the most impactful features and transforming the data into formats more suitable for predictive modelling. According to (Akbar et al., 2021) and (Brintrup et al., 2019), effective feature selection and engineering are crucial in supply chain analytics, where data is often heterogeneous, multidimensional, and context-dependent.

This section outlines the techniques used for selecting and engineering features, as well as the rationale behind the inclusion and exclusion of specific attributes.

3.3.1 Feature Selection Techniques

1. **Correlation Analysis:** Pearson correlation coefficients were computed to assess linear relationships between independent features and target variables. Features with very low correlation ($|r| < 0.1$) were considered weak predictors and excluded from the final model pipeline. This helped reduce noise and avoid redundancy in the feature space, particularly in regression-based tasks such as sales forecasting and shipment duration.
2. **Recursive Feature Elimination (RFE):** RFE was applied to systematically remove the least impactful features by recursively training models and ranking features based on importance. This method was utilised for both classification and regression tasks. For each use case, the top 10 features were retained based on performance scores across cross-validation folds, ensuring generalisability and relevance.

   For transparency, the top 10 features selected by the RFE process for each use case are summarised below. These features consistently demonstrated strong predictive capacity across multiple cross-validation folds:

– **Sales Prediction:** Order Quantity, Product Price, Order Item Total, Shipping Mode, Customer Segment, Order Priority, Product Category, Profit Margin, Order Region, Shipping Duration.

– **Late Delivery Risk:** Shipping Mode, Distance, Weather Conditions, Delivery Segment, Product Category, Order Priority, Carrier Name, Shipment Date, Customer City, Delivery Status.

– **Shipment Duration:** Distance, Weather, Shipment Date, Shipping Mode, Carrier Name, Order Priority, Product Category, Customer Location, Delivery Segment, Day of the Week.

– **Route Mapping:** Distance, Road Conditions, Transport Mode, Order Region, Order

ID, Delivery Zip Code, Traffic Pattern, Carrier Name, Delivery Segment, Customer Latitude/Longitude

The feature ranking was determined individually for each model and averaged across repetitions to ensure consistency.

3. **Principal Component Analysis (PCA):** To manage high-dimensional data and improve computational efficiency, PCA was used to reduce the number of variables while retaining the variance in the dataset. Though interpretability was reduced, PCA was particularly useful in shipment duration and route mapping tasks, where interaction effects were more complex. This aligns with approaches in supply chain dimensionality reduction discussed in (How and Lam, 2018) and (Wang et al., 2020).
4. **Feature Importance (Tree-based Models**): Feature importance scores from Random Forests were used to rank variables based on their contribution to prediction accuracy. This was especially effective for categorical and nonlinear interactions in late delivery risk and route optimisation models. Features consistently ranked low were excluded from further processing to reduce overfitting and improve model robustness.

Features were excluded if they exhibited high multicollinearity, contained excessive missing data (e.g., *Product Description*), had low variance, or lacked contextual relevance as judged by domain knowledge. This curation step ensured that only meaningful predictors were retained, enabling cleaner, more interpretable models.

3.3.2 Feature Engineering Techniques

1. **Date Features:** Calendar-based features such as day of the week, day of the month, and month were extracted from Order Date and Shipping Date fields. These features were critical in capturing seasonal trends and lead time patterns.
2. **Categorical Features:** Categorical attributes like Market, Delivery Segment, and Customer City were encoded using one-hot encoding. This allowed models to effectively leverage the categorical structure of the dataset without introducing ordinal assumptions.
3. **Numerical Features:** Numerical variables were standardised using the StandardScaler to bring them onto a comparable scale. This step was essential to ensure balanced contribution to distance-based models and neural networks.
4. **Text Features:** The Product Description field was sparsely populated and thus removed from modelling, though initial experiments involved tokenisation, stopword removal, and stemming to evaluate its utility. Future studies could revisit this attribute with a more complete dataset.

3.4.3 Key Performance Indicators (KPIs) The KPIs selected for this study were chosen based on their significance in supply chain operations and alignment with business-critical outcomes. Each KPI corresponds to one of the four use cases and reflects a distinct operational objective:

– *Sales Forecasting (Demand Prediction):* Vital for inventory control, procurement planning, and reducing stockouts or overstocking.
– *Late Delivery Risk:* Linked directly to customer satisfaction and operational efficiency; enables proactive intervention in logistics.
– *Shipment Duration*: Important for scheduling, cost estimation, and service-level management.
– *Route Mapping:* Supports optimisation of transportation paths, contributing to cost reduction and timely delivery.

These KPIs are widely recognised in both academic and industry literature as fundamental to effective supply chain decision-making. Their inclusion in this study ensures that the modelling outputs are not only technically sound but also practically applicable in real-world supply chain contexts.

3.4 Selection and Implementation of ML, time series and statistical model:

Various models were slected based on their ability to address different supply chain analytics challenges. The right approach is used in the implementation process to guarantee that each model is set up and evaluated correctly for the particular use case. This systematic approach is critical for making a fair and useful comparison of model performances.

This study compares and contrasts several time-series, statistical, and machine learning algorithms for supply chain analytics. The selection and application of these statistical and machine learning models are detailed in this section. The experimental results for the four use cases - supply chain forecasting (sales prediction), late delivery risk, shipment duration prediction - are all discussed.

3.4.1 Selection of ML, Statistical and Time series Model: In this research, various machine learning, statistical and time-series models were evaluated to assess their performance in supply chain analytics. The machine learning models considered include Linear Regression, Lasso Regressor, Ridge Regressor, KNN Classifier, Gaussian Naive Bayes, SVM Classifier, Gradient Boost, Decision Tree, Random Forest, LSTM Model, and Gated Recurrent Unit. The statistical models considered include the chi-square test, OLS, T-test, Multiple Linear Regression and Kruskal-Wallis test. The Exponential Smoothing Model, SRIMA, and ARIMA are the time-series models considered.

The choice of a particular algorithm was guided by the specific type of use case and the datatype employed. For shipment duration and demand forecasting tasks, time-series models, e.g., ARIMA, SARIMA, and Exponential Smoothing, are suitable as they can capture temporal patterns. Classification tasks like late delivery risk are modelled using SVM, Decision Trees, and Random Forests, as these are quite robust with imbalanced and categorical data. Linear, Lasso, and Ridge regressors are suitable for predicting continuous outcomes. Deep neural networks, namely, LSTM and GRU, were used due to their ability to model sequential dependencies, particularly in time-series contexts. Various statistical tests were conducted to validate assumptions and relationships within the data for the enhanced robustness of our models. In summary, each selected method was chosen based on its compatibility with the problem complexity, data type, and interpretability needs.

3.4.2 Implementation of Machine Learning and Statistical Models The data preprocessing and exploratory data analysis (EDA) provided critical insights into the dataset, including the presence of missing values, data imbalance, outliers, and feature redundancies. These findings directly informed both the feature selection process and the subsequent model implementation strategy. Specifically, the correlation analysis, Recursive Feature Elimination (RFE), and Random Forest feature importance methods enabled the identification of high-impact attributes tailored to each predictive task. For each use case, models were chosen based on the type of target variable (classification or regression), the nature of the features (categorical, numerical, temporal), and the desired balance between interpretability and performance. The refined feature sets ensured alignment with domain knowledge and model compatibility. The selected models were then trained on the cleaned and feature-engineered dataset, and evaluated using appropriate metrics: Root Mean Square Error (RMSE), Mean Square Error (MSE), and R-squared (R2) for regression tasks, and accuracy for classification tasks.

– **Supply Chain Forecasting – Sales Prediction:**

This regression task aimed to forecast sales using structured transaction data. Key features selected via RFE included *Order Quantity, Product Price, Order Item Total, Shipping Mode, Customer Segment, Order Priority, Product Category, Profit Margin, Order Region, and Shipping Duration*. A variety of machine learning, statistical, and time-series models were applied, including Linear Regression, Lasso, Ridge, Decision Tree, Random Forest, and SARIMA. Among these, the Random Forest and Decision Tree models achieved the highest accuracies of 0.99984 and 0.9, respectively. SARIMA emerged as the most effective time-series model with an accuracy of 71.88%. Monte Carlo simulation further validated SARIMA under stochastic demand scenarios, maintaining consistent accuracy.

– **Late Delivery Risk:**

This classification task focused on predicting the probability of shipment delays due to variables such as transport issues or environmental disruptions. The most impactful features included *Shipping Mode, Distance, Weather Conditions, Delivery Segment, Product Category, Order Priority, Carrier Name, Shipment Date, Customer City, and Delivery Status.* Applied models included SVM, Random Forest, XGBoost, and Gaussian Naive Bayes, alongside statistical techniques like Chi-Square Test. Among them, Random Forest and XGBoost achieved top accuracies of 0.90443 and 0.90442, respectively. Time-series modelling with ARIMA yielded 98.28% accuracy, which remained robust (93.23%) under Monte Carlo simulations simulating delayed scenarios.

– **Predicting Shipment Duration:**

This regression-based use case estimated delivery times using features such as *Distance, Weather, Shipment Date, Shipping Mode, Carrier Name, Order Priority, Product Category, Customer Location, Delivery Segment, and Day of the Week.* Machine learning models including Random Forest and XGBoost were evaluated, with Random Forest achieving an accuracy of 0.81351. Statistical and time-series models such as Multiple

Linear Regression, Exponential Smoothing, and SARIMA were also assessed. The Exponential Smoothing model yielded the highest accuracy at 99.37%, a result that remained consistent during Monte Carlo simulations simulating transportation delays and weather impacts.

Route Mapping:
The final use case explored route optimisation using a classification approach informed by spatial and logistical data. The most relevant features included *Distance, Road Conditions, Transport Mode, Order Region, Delivery Zip Code, Traffic Pattern, Carrier Name, Delivery Segment, Customer Latitude, and Customer Longitude.* Gradient Boost, Random Forest, and Multilayer Perceptron models were employed, along with ARIMA for time-dependent routing decisions. The Random Forest model achieved the highest accuracy of 0.98205, while ARIMA reached 99.13% accuracy in timeseries forecasting. Monte Carlo simulations confirmed the robustness of ARIMA and its derivatives under variable routing conditions, such as fluctuating delivery windows and traffic disruptions.

Overall, this implementation strategy ensured that models were not only statistically valid but also operationally meaningful, enhancing their potential application in realworld supply chain management contexts.

3.5 Monte Carlo simulation methodology and implementation

Monte Carlo simulation are crucial in supply chain management for addressing uncertainty and testing model robustness across various scenarios. This method helps assess how well adapted and reliable our models represent variation among these obstacles. For a given variable or set of variables, to construct its probable distribution (if it has one), we use the computational procedure known as Monte Carlo simulation, which produces random samples. This technique involves generating random samples to simulate different scenarios and assess model performance under varying conditions. This section outlines the methodology and implementation of Monte Carlo simulations for the identified use cases.
– Steps in Monte Carlo simulation Methodology:

1. Data Preprocessing: Data cleaning is performed to remove missing value and irrelevant information, preparing the dataset for analysis.
2. Exploratory Data Analysis (EDA): EDA techniques are employed to identify patterns, relationships, and categories within the dataset, guiding the selection of models.
3. Feature Selection: Relevant features are selected based on their relationships and importance, helping to optimise model performance
4. Algorithm Selection: Appropriate machine learning, statistical, or time-series algorithms are selected based on the dataset’s characteristics and the problem at hand.
5. Model Development: After the algorithm is developed, models are trained on the training data and tested on a test dataset to evaluate their performance.
6. Monte Carlo Simulation: Once the model is built, the next step must be to perform a Monte Carlo simulation. Random samples are drawn from the probability distribution

of variables, and different scenarios are simulated to assess model performance.

7. Results Evaluation: The final step in the Monte Carlo simulation approach is to evaluate outcomes. The model’s performance is measured using accuracy-based metrics such as root mean square error (RMSE), R2 and average (mean) values.

- Monte Carlo simulation methodology is implemented for the identified use cases, as explained below:

1. Data Preprocessing: The values of sales from the above dataset were converted into datetime, resampled to monthly frequency and summed for analysis.
2. Exploratory Data Analysis: EDA was conducted to uncover trends and patterns in the data.
3. Feature Selection: Relevant features for model development were selected through feature selection.
4. Algorithm Selection: For model development, the algorithms selected were ARIMA, SARIMA and Exponential Smoothing.
5. Model Development: Models were trained and tested, with performance evaluated using RMSE, R2 and error of mean square to gauge all models ’performance in terms of accuracy.
6. Monte Carlo Simulation: A Monte Carlo simulation was performed using random samples drawn from the probability distribution of a given variable.

Varying Simulation Conditions: The Monte Carlo simulations were designed to reflect uncertainty in key input variables such as demand volume, shipping lead times, and weather-related disruptions. By repeatedly sampling from probability distributions defined for these variables, we generated a range of realistic operational scenarios. This allowed us to assess how model performance metrics (e.g., RMSE, R2) responded to these fluctuations, thereby testing the robustness of each model under variable conditions reflective of real-world supply chain environments.

4 Results

In this research study, the performance of multiple models and algorithms were evaluated across four distinct use cases in supply chain analytics using the DataCo supply chain dataset (Constante et al., 2021). The analysis aimed to identify the most efficient models for each use case, with a focus on accuracy, performance and alignment with the ISO25010 quality model. To ensure internal validity of our experiments, variables and conditions were meticulously controlled during the modelling process, minimising biases and errors. External validity has been addressed by selecting a diverse and representative dataset, helping to generalise the findings to broader supply chain contexts. Criterion validity was established through alignment with the ISO25010 quality model, ensuring that the assessment criteria are industry-relevant and standardised.

For each use case, a range of machine learning, statistical, and time-series models were tested, both with and without Monte Carlo simulation. The primary objective was to ascertain the most efficient models and algorithms for each use case, focusing on accuracy, performance, and alignment with the ISO25010 quality model for product quality evaluation. This comprehensive approach ensures a thorough assessment that extends beyond basic performance metrics.

4.1 Supply Chain Forecasting - Sales Prediction

The initial focus was on predicting sales within the supply chain. Several machine learning models, statistical models, and time-series models were explored. The highes performing accuracy results for each category are presented in Table 5. The robustness of these results is supported by rigorous statistical testing and cross-validation techniques, ensuring their reliability and validity.

Table 5. Accuracy of Models for Supply Chain Forecasting - Sales Prediction

Model Category	Best Accuracy
Machine Learning	Decision Tree (0.99989)
Statistical	T-test (1)
Time Series	SARIMA (0.6962)
Time Series + Monte Carlo	SARIMA (0.7188)

From Table 5, it is evident that the model Decision Tree is achieving the highest accuracy among all machine learning models, the T-test yielded the best results among the statistical models, and SARIMA outperformed other time series models. Moreover, incorporating Monte Carlo simulation further improved the accuracy of the SARIMA model.

4.2 Late Delivery Risk

In the second use case, the focus was on predicting the risk of late delivery. Machine learning, statistical, and time-series models were applied and their performance evaluated. Table 6 summarises the highest accuracy results for each category.

Table 6. Accuracy of Models for Late Delivery Risk Prediction

Model Category	Best Accuracy
Machine Learning	Random Forest (0.90443)
Statistical	T-test (1)
Time Series	ARIMA (0.9828)
Time Series + Monte Carlo	ARIMA (0.9324)

According to Table 6, a similar trend in model performance to that demonstrated in Table 6 is presented. The Random Forest model demonstrates the highest accuracy among the machine learning models, the T-test achieved the best results among the statistical models, and the ARIMA model outperforms other time-series models. After we added Monte Carlo simulation, it further improved the ARIMA model's accuracy.

4.3 Predicting Shipment Duration

The third use case centered on predicting shipment duration. To assess accuracy, various machine learning, statistical and time-series models were employed. A summary of the accuracy results is provided in Table 7.

Table 7. Accuracy of Models for Predicting Shipment Duration

Model Category	Best Accuracy
Machine Learning	Random Forest (0.81351)
Statistical	Chi-square test (0.74826)
Time Series	Exponential Smoothing (0.99372)
Time Series + Monte Carlo	Exponential Smoothing (0.99372)

From Table 7, it is observed that the model Random Forest is achieving the highest accuracy among the machine learning models, the Chi-square test yielded the best results among the statistical models, and the Exponential Smoothing model outperformed other time-series models. The inclusion of Monte Carlo simulation did not significantly impact the accuracy in this particular use case.

4.4 Route Mapping

The fourth use case involved route mapping. To evaluate accuracy, machine learning models, statistical models, and time-series models were applied. The summarised accuracy results for each category are represented in Table 8.

Table 8. Accuracy of Models for Route Mapping

Model Category	Best Accuracy
Machine Learning	Random Forest (0.98205)
Statistical	Multilayer Perceptron (0.92686)
Time Series	ARIMA (0.99135)
Time Series + Monte Carlo	ARIMA (0.99135)

According to Table 8, the Random Forest model achieved the highest accuracy among the machine learning models, the Multilayer Perceptron model demonstrated the best results among the statistical models, and the ARIMA model outperformed other time-series models. The inclusion of Monte Carlo simulation also did not significantly impact the accuracy in this use case.

To visually compare performance across the four case studies, Figure 5 displays a grouped bar graph illustrating the accuracy of different model categories. In conclusion, the research study employed a range of models and algorithms to address various use cases in supply chain analysis. The analysis identified the best-performing models and algorithms based on accuracy, with Decision Tree, Random Forest, ARIMA, and Exponential Smoothing outperforming others across different scenarios. These results highlight the effectiveness of machine learning, time-series and statistical models in enhancing supply chain analytics and supporting informed decision-making processes. Note: For the code discussed in this paper and detailed results, and comparison table see the Appendix section.

While the dataset provides a comprehensive view of supply chain operations, the authors recognise limitations in its representativeness. To enhance the generalisability of the findings, future studies are encouraged to replicate the methodology across diverse datasets. To further validate the results, supplementary material are available on GitHub, including detailed experimental setups and additional analyses, promoting transparency and reproducibility of this research.

In conclusion, the study presents a detailed comparative analysis of various models in supply chain analytics, identifying the best-performing models based on accuracy and robustness. These findings, validated through rigorous experimental designs and alignment with industry standards, offer significant insights for enhancing decision-making processes in supply chain management. However, the scope for further validation on diverse datasets remains, inviting additional research to confirm and expand upon these conclusions.

5 Discussion

5.1 Factors Influencing Model Accuracy

Across the four use cases, several key factors were observed to influence model accuracy:

- **Data Quality and Completeness:** Features with missing or inconsistent data (e.g., Product Description, Order Zipcode) negatively impacted accuracy if not adequately handled.
- **Feature Relevance and Selection:** Models using domain-informed feature selection (e.g., RFE, correlation filtering) performed significantly better, particularly in shipment prediction and late delivery classification.

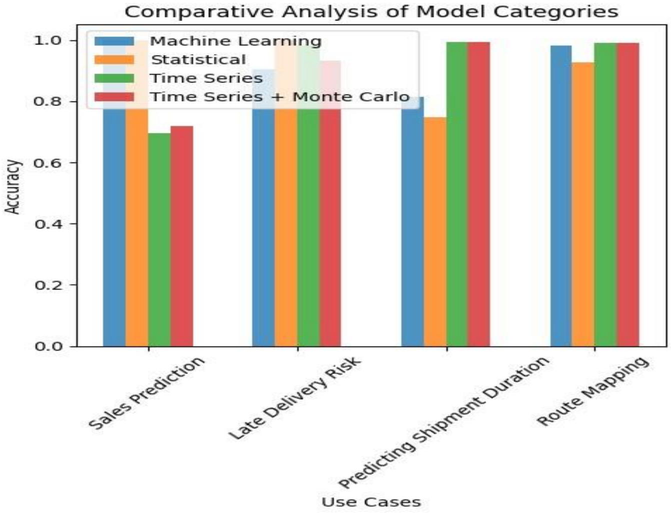


Fig. 5. Comparative Analysis of Model Categories for each Use Case

- **Model Complexity:** Ensemble models (e.g., Random Forest, XGBoost) consistently outperformed simpler linear models due to their ability to capture nonlinear relationships, especially under varied data distributions.
- **Simulation Conditions:** Model performance varied with changing assumptions in Monte Carlo simulations—models that were less sensitive to fluctuations in input distributions showed better reliability under operational uncertainty.

These factors highlight the importance of preprocessing, model selection, and simulation design in achieving robust and generalisable results in supply chain analytics.

5.2 Implications for Research and Practice

The results of this study provide meaningful contributions to both academic research and supply chain management practice.

Implications for Research: This study advances the scholarly understanding of predictive modelling in supply chain analytics by offering a systematic, multi-model evaluation across four critical use cases. Unlike prior research that often focuses on isolated techniques or single use-case analysis, this work offers a comparative framework that integrates machine learning, statistical, and time-series approaches within a common dataset. Furthermore, the inclusion of simulation-based testing (Monte Carlo simulations) under varying conditions enhances the methodological robustness, offering insights into model stability and generalisability—an area that remains underexplored in supply chain analytics literature.

In addition, the alignment of model evaluation with the ISO/IEC 25010 quality framework introduces a novel validation layer rarely employed in data analytics studies. This bridges the gap between performance-focused evaluation and system-level quality attributes such as reliability, usability, and efficiency, thereby setting a precedent for more structured model assessment in future analytics research.

Implications for Practice: From a practical standpoint, this research offers decisionmakers guidance on selecting suitable predictive models for specific supply chain objectives. For instance, the superior performance of Random Forest and XGBoost in tasks involving classification and regression suggests their applicability in operational scenarios such as delivery risk prediction and demand forecasting. Similarly, time-series models like SARIMA and Exponential Smoothing demonstrate robustness in handling historical trend-based forecasting.

The study also underscores the importance of feature selection and preprocessing in achieving high model performance. Supply chain professionals and data analysts can benefit from incorporating structured feature selection techniques and simulation-driven stress testing to improve model reliability in real-world applications

Furthermore, by highlighting the influence of data quality, task alignment, and model complexity on predictive accuracy, this research contributes to the development of more intelligent, data-driven supply chain management systems. It encourages a move toward adaptive analytics solutions that are not only accurate but also transparent and deployable within operational decision-support tools.

Overall, this study lays the groundwork for the practical implementation of AI and machine learning models in supply chain contexts, fostering a more evidence-based and resilient approach to supply chain optimisation.

5.3 Validation Framework

To ensure the robustness and applicability of our results, we adopted the ISO/IEC 25010 standard as a guiding framework for evaluating model quality. ISO/IEC 25010 is a widely recognised international standard that defines a quality model for software and system evaluation, encompassing eight quality characteristics: functionality, reliability, usability, efficiency, maintainability, portability, compatibility, and security.

In the context of supply chain analytics, where predictive models serve as decision support systems, this standard offers a structured and holistic approach to validate the performance and operational relevance of the developed models. For this study, we focused on a subset of ISO/IEC 25010 characteristics that are most applicable to data driven predictive models:

- **Functionality:** The ability of the model to fulfil its intended purpose—accurately predicting KPIs such as sales, shipment duration, or delivery risk.
- **Reliability:** Evaluated through statistical validation methods (e.g., cross-validation, RMSE, MSE), assessing whether the model consistently produces stable outputs under different data splits or simulation scenarios.
- **Efficiency:** Considered in terms of computational performance and resource usage during training and inference.
- **Usability:** Measured indirectly by the interpretability and ease of deployment of models, particularly those used in operational decision-making (e.g., tree-based models).

- **Maintainability and Portability**: Reflected in the modular pipeline design and compatibility with standard machine learning environments (e.g., Python, scikitlearn).

The inclusion of the ISO/IEC 25010 framework supports the generalisability and practical applicability of our findings, aligning the evaluation of analytical models with recognised software quality principles. This contributes to a more rigorous and standardised validation process, ensuring that the models not only perform well statistically but also meet broader expectations for operational deployment in supply chain contexts.

6 Conclusion

This research examined the effectiveness of various machine learning, time-series and statistical models in supply chain analytics, focusing on four use cases: Sales Prediction, Late Delivery Risk, Predicting Shipment Duration and Route Mapping. The analysis included comprehensive data preprocessing, exploratory data analysis, feature selection, selection of algorithm, and prediction model development. WThe models were evaluated using metrics such as r2 score, mean square error, and RMSE, all applied to the DataCo supply chain dataset.

In the first use case, Supply Chain Forecasting – Sales Prediction, the Gaussian Naive Bayes model exhibited the highest accuracy among all models, achieving an accuracy of 91.35%. Conversely, the Gradient Boost and LSTM models demonstrated the lowest accuracy, yielding negative scores. For the second use case, Late Delivery Risk, the Random Forest and XGBoost models performed exceptionally well, with accuracies of 90.44% and 90.43%, respectively. The Lasso Regressor and LSTM Models, however, showed the lowest accuracy with negative scores. In the third use case, Predicting Shipment Duration, the Exponential Smoothing Model demonstrated the highest accuracy of 99.372%, while the KNN Classifier and SVM Classifier recorded the lowest accuracies, both showing negative scores. Finally, for the fourth use case, Route Mapping, the Random Forest model achieved an accuracy of 39.74%, while the Lasso Regressor performed the least accurately with a negative score.

The study reveals that the performance of the models varied across different use cases. In some scenarios, machine learning models outperformed statistical and timeseries models, whereas in other cases, statistical and time-series models outperformed machine learning models. Notably, the Exponential Smoothing model consistently demonstrated high accuracy across all use cases, making it a reliable choice for supply chain analytics.

In conclusion, thi study underscores the importance of selecting the appropriate model for each use case in supply chain analytics. Understanding the strengths and weaknesses of different models can assist supply chain management professionals in making informed decisions and optimising their supply chain operations. Further research could explore the performance of additional models and incorporate real-time data to enhance the accuracy and applicability of supply chain analytics.

6.1 Limitations and direction of future research

While the research provides valuable insights into the performance of various models in supply chain analytics, it is essential to acknowledge certain limitations and suggest directions for future exploration.

6.1.1 Limitation

- *Dataset Limitations:* The dataset, though it includes over 100’000 records, may still be small for supply chain applications, typically needing larger, more diverse datasets. The dataset also lacks diversity in product types, customer profiles, and geographic coverage. We understand that these issues can impact the ability of the model to generalize and exhibit good performance in real-world cases, but the clean, structured, and accessible nature of the data on shipments, orders, and deliveries is good for validating the initial hypothesis and prototyping models. However, less diversity in the dataset, specifically in product, customer, and geographic information, might hamper the ability of the model to generalize in real-world environments. This can be addressed in future work by integrating a larger number of diverse datasets, spanning different customer segments, locations, and product types.
- Scope of Use Cases: Our research focused on four specific use cases in supply chain analytics. While we chose these because of their relevance and applicability, other use cases in supply chain management should have been covered in this study. This limitation might be affecting the generalizability of our findings to these unexplored areas.

6.1.2 Future Research Direction

To fix the weak points and improve the range, we provide the study:

- *Growing the Use Cases and Dataset:* This means that we can use the big datasets to cover all supply chains, which can be done in future studies. This will provide more correct results and improve the model performance. Also, studying use cases other than the four this study looked at would give a bigger view of how useful and correct the model is.
- *Combining Models and Approaches:* Employing a combination of various models and approaches could potentially boost model accuracy. Integrating different modeling techniques might offer more nuanced insights and improved predictive capabilities. The use of hybrid models—such as combining time-series techniques with machine learning—offers strong potential for improving prediction accuracy and capturing both temporal patterns and complex nonlinear relationships in certain use cases, like forecasting and shipment duration prediction.
- *Exploring Additional Performance Metrics:* Investigating several performance variables, such as recall, precision, and F1-score, could provide a more rounded assessment of model effectiveness. These metrics would offer additional dimensions to model evaluation, especially in scenarios where accuracy alone might not be the sole indicator of model performance.

- *Innovative Strategies in Supply Chain Optimization*: There is scope for exploring other strategies in optimizing supply chain activities. Future research could dive into the potential of emerging techniques like reinforcement learning and transfer learning. These advanced methods could offer novel solutions to complex supply chain challenges.

This research lays the groundwork for future studies in supply chain analytics. By acknowledging the limitations of our study and proposing these future research directions, we aim to encourage a more expansive and thorough exploration of this field. The continued investigation and improvement of models and methodologies will undoubtedly contribute to more robust and effective supply chain management strategies.

References

Anitha, P., Patil, M. M. (2018). A review on data analytics for supply chain management: a case study. *International Journal of Information Engineering and Electronic Business*, 14(5), 30.

Babaee Tirkolaee, E., Sadeghi, S., Mooseloo, F. M., Vandchali, H. R., Aeini, S. (2021). Application of machine learning in supply chain management: A comprehensive overview of the main areas, *Math. Probl. Eng*. 2021, 1–14.

Bousqaoui, H., Achchab, S., Tikito, K. (2017). Machine learning applications in supply chains: An emphasis on neural network applications, *2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)*, Rabat, Morocco, 2017, pp. 1-7.

Brintrup, A., Pak, J., Ratiney, D., Pearce, T., Wichmann, P., Woodall, P., McFarlane, D. (2019). Supply chain data analytics for predicting supplier disruptions: A case study in complex asset manufacturing, *Int. J. Prod. Res.* **58**, 3330–3341.

Colicchia, C., Creazza, A., Menachof, D. A. (2018). Managing cyber and information risks in supply chains: Insights from an exploratory analysis, Supply Chain Manag.: Int. J. **24**, Dec. 2018.

Constante, F., Silva, F., Pereira, A. (n.d.). Dataco smart supply chain for big data analysis, Data.Mendeley.com, 5.

Dash, R., McMurtrey, M., Rebman, C., Kar, U. K. (2019). Application of artificial intelligence in automation of supply chain management, *J. Strateg. Innov. Sustainab.* **14**, Jul. 2019.

Dohale, V., Ambilkar, P., Gunasekaran, A., Verma, P. (2021). Supply chain risk mitigation strategies during COVID-19: Exploratory cases of “make-to-order” handloom saree apparel industries, *Int. J. Phys. Distrib. Logist. Manag*., ahead-of-print, Apr. 2021.

Estampe, D., Lamouri, S., Paris, J.-L., Brahim-Djelloul, S. (2013). A framework for analysing supply chain performance evaluation models, *Int. J. Prod. Econ.* **142**, 247–258.

Fattah, J., Ezzine, L., Aman, Z., El Moussami, H., Lachhab, A. (2018). Forecasting of demand using ARIMA model, *Int. J. Eng. Bus. Manag*. **10**, ID 184797901880867.

Ghanadian, S. A., Ghanbartehrani, S. (2021). Evaluating supply chain network designs: An approach based on SNA metrics and random forest feature selection, *Univ. J. Oper. Manag.* **1**, 15–35.

Guanghui, W. (2012). Demand forecasting of supply chain based on support vector regression method, *Procedia Eng.* **29**, 280–284.

Hahn, G. J. (2019). Industry 4.0: A supply chain innovation perspective, *Int. J. Prod. Res*. **58**, 1425–1441.

How, B. S., Lam, H. L**.** (2018). Sustainability evaluation for biomass supply chain synthesis: Novel principal component analysis (PCA) aided optimisation approach, *J. Clean. Prod.* **189**, 941–961.

ICML (2018). DiCE: The infinitely differentiable Monte Carlo estimator, *Proc. Int. Conf. Mach. Learn.*, 2018.

Kilimci, Z. H., Akyuz, A. O., Uysal, M., Akyokus, S., Uysal, M. O., Bulbul, B. A., Ekmis, M. A. (2019). An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain, Complexity 2019, 1–15.

Li Vigni, M., Durante, C., Cocchi, M. (2013). Exploratory data analysis, in *Data Handling in Science and Technology,* pp. 55–126.

Mariappan, M. B., Devi, K., Venkataraman, Y., Lim, M. K., Theivendren, P. (2022). Using AI and ML to predict shipment times of therapeutics, diagnostics and vaccines in e-pharmacy supply chains during COVID-19 pandemic, *Int. J. Logist. Manag*., Jan. 2022.

Mohamed, S., Rosca, M., Figurnov, M., Mnih, A. (2020). Monte Carlo gradient estimation in machine learning, *J. Mach. Learn. Res*. **21**, 1–62.

Morgenthaler, S. (2009). Exploratory data analysis, *Wiley Interdiscip. Rev. Comput. Stat*. **1**, 33–44.

Nguyen, H. D., Tran, K. P., Thomassey, S., Hamad, M. (2021). Forecasting and anomaly detection approaches using LSTM and LSTM autoencoder techniques, *Int. J. Inf. Manag.* **57**, 102282.

Raman, S., Patwa, N., Niranjan, I., Ranjan, U., Moorthy, K., Mehta, A. (2018). Impact of big data on supply chain management, *Int. J. Logist. Res. Appl*. **21**, 579–596.

Saci, S. (2023). What is a supply chain digital twin?, Mar. 2023.

Sahoo, R., Pasayat, A. K., Bhowmick, B., Fernandes, K., Tiwari, M. K. (2021). A hybrid ensemble learning-based prediction model to minimise delay in air cargo transport, *Int. J. Prod. Res.*, 1–17.

Shih, H., Rajendran, S. (2019). Comparison of time series methods and machine learning algorithms for forecasting Taiwan’s blood supply, *J. Healthc. Eng*. 2019, 1–6.

Surie, C., Reuter, B. (2014). Supply chain analysis, in *Springer Texts in Business and Economics,* pp. 29–54.

Terrada, L., El Khaili, M., Ouajji, H. (2022). Demand forecasting model using deep learning methods for supply chain management 4.0, *Int. J. Adv. Comput. Sci. Appl.* **13**, 2022.

Thete, J. (2022). A stochastic model for demand forecasting in Python, Mar. 2022.

Wan, F. (2021). XGBoost based supply chain fraud detection model, Mar. 2021.

Wang, J., Swartz, C. L. E., Corbett, B., Huang, K. (2020). Supply chain monitoring using principal component analysis, *Comput. Chem. Eng*. **59**, 12487–12503.

Winter Simulation Conference (WSC) (2009). Quantifying supply chain disruption risk using Monte Carlo and discrete-event simulation, *Proc. Winter Simul. Conf.*, 2009.

Zekhnini, K., Cherrafi, A., Bouhaddou, I., Benghabrit, Y., Garza-Reyes, J. A. (2020). Supply chain management 4.0: A literature review and research framework, *Benchmarking: Int. J.,* **28**(2).

Appendices

All the source code discussed in this paper for the implementation of the proposed methodology in the supply chain management of DataCo Smart supply chain Dataset are available under the following Github repository:
https://github.com/Rachana-pandey11/Supply-Chain-Analytics

Received August 20, 2024, revised May 26, 2025, accepted May 26, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 2, pp. 486–501
https://doi.org/10.22364/bjmc.2025.13.2.10

Blackcurrant Leaf Analysis Using Instance Segmentation and Multi-label Classification

Reinis ODĪTIS, Ivo ODĪTIS, Kārlis FREIVALDS, Jānis BIČEVSKIS

University of Latvia, Raiņa bulvāris 19, Riga, Latvia

ro17020@students.lu.lv, ivo.oditis@lu.lv, karlis.freivalds@lu.lv, janis.bicevskis@lu.lv

ORCID: 0009-0003-5488-1608, ORCID: 0000-0003-2354-3780, ORCID: 0000-0003-2684-559X, ORCID: 0000-0001-5298-9859

Abstract. In agriculture, the health of plants can often be assessed by examining the appearance of their leaves. Traditionally, this evaluation is carried out by humans through visual observation. However, drone technology enables plant monitoring through aerial photography, removing the need for direct human presence. Furthermore, artificial intelligence offers the potential to replace human expertise in this process. In this study, the authors explore the use of machine learning methods to evaluate the condition of blackcurrants using visible light (RGB) images. The research reviews similar approaches where machine learning techniques have been applied to analyze plant leaves, aiming to identify various issues in a timely and efficient manner. Specifically, this study employs the YOLO model for leaf instance segmentation, followed by multi-label classification of the segmented leaf instances using the ResNet model. The study concludes that this method, while not perfect, provides sufficient accuracy to effectively identify field-level health issues and support targeted crop management strategies.

Keywords: Blackcurrant leaf analysis, Instance segmentation, Multi-label classification, Plant health monitoring, Drone-based agriculture

1 Introduction

In modern and precise agriculture, there is a continual search for new and cost-effective methods to diagnose plant conditions, enabling timely and accurate decisions regarding fertilization, spraying, and other interventions. One of the simplest methods for problem identification is visual plant observation. This can be done by walking through the fields and visually observing the plants; however, it is more convenient to perform this using automatically captured images. Field images can be obtained quickly and easily using many of the commercially available drones. These images provide farmers with sufficient visual information on plant condition without the need for physical field inspections. However, the process becomes more efficient when plant health information is obtained automatically.

Various comprehensive reviews (section 2) indicate that plant leaf images are widely utilized for assessing plant health through machine learning methods. While some studies explore the methodologies for image acquisition, the primary emphasis is placed on identifying potential issues and their nature using these techniques. These studies also encompass several shrub species, such as raspberries and quince. However, blackcurrants and the challenges (leaf identification in images and classification) associated with their cultivation remain unaddressed.

The publication examines the classification of blackcurrant leaves using current image analysis methods (section 3). To identify issues with blackcurrants, the process is divided into two stages: instance segmentation and classification. Images of blackcurrant bushes are segmented to isolate individual blackcurrant leaves using machine learning methods (YOLO model). The segmented images are then classified using the ResNet model to distinguish between healthy plants, diseased plants, and plants lacking sufficient nutrients.

The significance of this work is highlighted by the fact that the analysis of blackcurrant leaves and disease identification has not been addressed in previous studies. Additionally, there are no publicly available datasets of blackcurrant leaf images containing both healthy and damaged leaves with annotations. The authors could not ignore real-world challenges and worked with naturally obtained images of blackcurrant bushes, performing segmentation and classification of individual leaf instances. As a result, a solution prototype was developed that could be integrated into industrial applications.

2 Related works

Currently, the analysis of agricultural crops using photographs is being widely studied. Several comprehensive reviews of studies in this field of agriculture may be found (Debangshi (2021), Hafeez et al. (2023), Istiak et al. (2023), Anam et al. (2024)). While images from various multispectral cameras are increasingly being examined, much attention is still focused on what can be inferred about plants from visible light (RGB) images (Hafeez et al. (2023)). When evaluating plant leaves, various plant properties are assessed, such as plant health (Janani and Jebakumar (2023)), changes in plant biomass (Fei et al. (2023)), age (Bai et al. (2023)), and others.

The processing of plant leaf images typically involves two steps: isolating plant leaves from larger images and segmentation. In the following subsections, the authors review segmentation and multi-label classification methods used in other similar studies.

2.1 Instance segmentation

Instance segmentation is a crucial step in the classification process, as it facilitates the accurate identification of individual leaves within a single image. It is a computer vision technique combining object detection and semantic segmentation to identify and delineate individual object instances within an image precisely. Unlike semantic segmentation, which assigns a class label to each pixel without distinguishing separate instances, instance segmentation provides both class labels and unique boundaries for each instance of the same class. This method allows for more precise data collection for each leaf, ultimately leading to improved overall data quality for the entire image.

Following the review of studies (Gu et al. (2022)), two instance segmentation methods were examined: YOLO (You Only Look Once), particularly YOLOv8-seg, and R-CNN (Region-based Convolution Neural Networks). Both models are widely recognized for their effectiveness in segmentation tasks (Charisis and Argyropoulos (2024)). However, given the high volume of data involved in analyzing blackcurrant fields—where potentially thousands of leaves need to be segmented and classified—computation time becomes a critical factor. Given the findings from recent research (Sapkota et al. (2024)) indicating that YOLOv8 demonstrates superior inference speed and effectiveness in segmentation tasks compared to Mask R-CNN, and additional evidence (Khan et al. (2023)) from a study where YOLOv8 was tested on a maize disease detection dataset, demonstrating that it not only achieved high precision but also proved to be highly effective. This further supports that YOLOv8 is not only fast but also highly accurate, making it an excellent fit for the task.

2.2 Multi-label classification

Classification is a supervised machine learning task that involves predicting a categorical label for an input based on its features. In classification, a model is trained on labeled data, where each input is associated with a specific class. The goal is to learn patterns from the data that allow the model to assign correct labels to unseen examples.

As demonstrated in the works of other authors (Hosny et al. (2023), Elfatimi et al. (2024)), multi-label classification algorithms have proven to be highly effective for the identification of issues in plant leaves. In the context of multi-label classification, the task extends to scenarios where one object can be associated with more than one class. For instance, in the case of blackcurrant leaf health analysis, leaves can be associated with multiple diseases and may be nutrient deficient. Hosny et al. (2023) identifies several models applicable for multi-label classification of leaves, highlighting among them ResNet, VGG, and EfficientNet.

3 Proposed method

In this study, we propose a blackcurrant leaf analysis method using advanced deep learning techniques for instance segmentation and classification. The proposed approach utilizes instance segmentation to detect and isolate individual blackcurrant leaves accurately from complex backgrounds and overlapping foliage within an image, thereby ensuring precise delineation of leaf boundaries. Once segmentation is complete, each leaf instance undergoes further analysis via a multi-label classification model, which evaluates its health condition based on critical features. As discussed in section 2, YOLO models (YOLOv8n-seg, YOLOv9c-seg) are employed for instance segmentation, while ResNet architectures (ResNet-50, ResNet-101, ResNet-152) are utilized for multi-label classification (the model selection is justified in section 3.7). The proposed methodology encompasses the steps represented in figure 1.

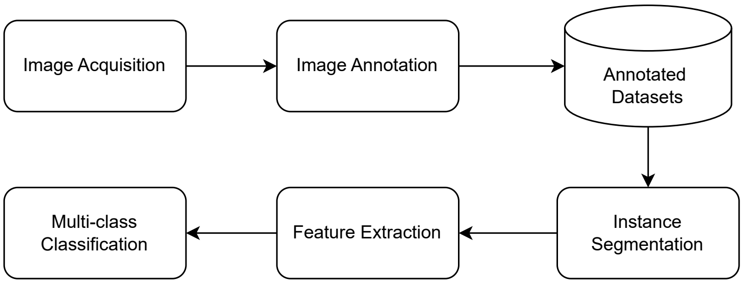


Fig. 1. Steps of leaf disease detection

3.1 Agricultural Context

Blackcurrants (*Ribes nigrum*) are valuable crops cultivated in temperate regions, thriving in well-drained soils with moderate moisture. However, they are highly susceptible to diseases such as powdery mildew, leaf spot, and rust, which manifest as discoloration, spotting, and necrosis, significantly reducing yield and quality. Early signs of these diseases are often identified through visual indicators on leaves, necessitating regular physical field monitoring. One of the most efficient ways to acquire such imagery is through drones equipped with automated missions, typically capturing images from a height of 3–5 meters to ensure sufficient detail for disease identification. The proposed method automates the identification of these diseases in images, associating leaf instances with disease classes. This information can be mapped onto aerial photos to provide a comprehensive view of field health. To be practically useful for farmers, the data should delineate the boundaries of affected areas, allowing targeted interventions. While instance-level metrics may not always achieve perfect accuracy, maintaining reliable field-level averages ensures the overall utility of the method for effective crop management.

3.2 Image acquisition

All images used in the datasets for training and validation, as described in subsection 3.4, were captured using a Nikon D3300 DSLR camera with a 24.2 MP DX-format CMOS sensor. The images were acquired from a local blackcurrant field, taken from various angles to ensure diversity in the dataset and enhance the robustness of the model training by simulating different perspectives of the leaf instances.

3.3 Image annotation

The acquired images were annotated using tailored methods to support the different stages of analysis. For instance segmentation, each image was manually annotated by outlining the boundaries of individual blackcurrant leaves using the Computer Vision Annotation Tool (CVAT). These annotations, provided in the form of point coordinates, were exported in the YOLOv8 segmentation 1.0 format. This process enables the model to accurately detect and isolate leaves from the background and overlapping objects, which is critical for precise instance segmentation.

For multi-label classification, each leaf instance was analyzed manually, and a label was assigned to describe its association with one or more of the predefined health status classes. The classification scheme, which includes the classes Healthy Leaves (HL), Nutrient Deficient Leaves (NDL), and Mycosphaerella ribis-Affected Leaves (MRL), was established in collaboration with a domain expert following an evaluation of local blackcurrant fields. The class labels were compiled in a structured file, which includes the image names and their corresponding class associations. This process ensures that both segmentation and classification tasks are fully supported by the annotations.

3.4 Datasets overview

This section provides an overview of the datasets employed in the study, which were designed to support different stages of the proposed method. Two distinct datasets were prepared: one for multi-label classification and the other for instance segmentation.

The multi-label classification dataset consists of 287 images, each containing a single blackcurrant leaf instance placed against a black background. These images are distributed across three subsets: 229 images for training, 49 for validation, and 49 for testing. The health status of each leaf instance was labeled based on the classification scheme developed in consultation with a domain expert, as outlined in the section 3.3. The dataset includes a single annotation file that lists image names and their associations with the three identified classes: Healthy Leaves (HL), Nutrient Deficient Leaves (NDL), and Mycosphaerella ribis-Affected Leaves (MRL), visually represented in figures 2, 3, and 4. A leaf instance may exhibit traits associated with multiple classes, thus requiring the model to handle overlapping features through a multi-label classification approach.

A single annotation file accompanies every subset, containing detailed information about the class associations for each image. The distribution of samples across classes in the multi-label classification dataset is presented in table 1. Because each instance can belong to multiple classes, the total number of labels in the table is greater than the total number of images.

The instance segmentation dataset consists of 87 images, each featuring blackcurrant leaves as the sole object type. It is divided into two subsets: a training set with 71 images and a validation set with 16 images. Compared to the classification dataset, the instance segmentation dataset is significantly smaller. This is because annotating this dataset is a complex and time-consuming process, requiring the precise outlining of each leaf instance. Each image was manually annotated to provide segmentation information in the form of points delineating each leaf. To analyze the potential impact and

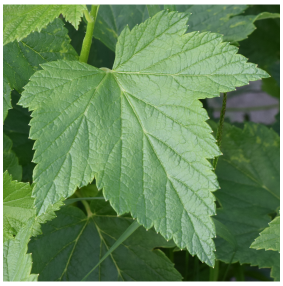


Fig. 2. HL class example

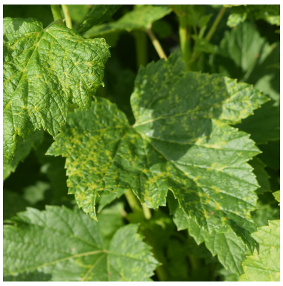


Fig. 3. NDL class example

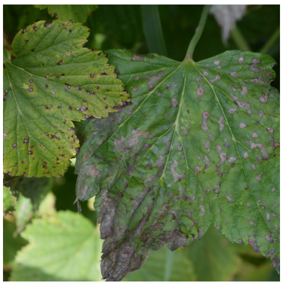


Fig. 4. MRL class example

Table 1. Distribution of class occurrences across training, validation, and test sets

Class	**Training (229 images)**	**Validation (49 images)**	**Test (49 images)**
Healthy Leaves (HL)	92	21	19
Nutrient Deficient Leaves (NDL)	85	19	16
Mycosphaerella ribis-Affected Leaves (MRL)	80	14	18
Total Labels	257	54	53

trends related to the dataset, the model was trained on different portions of the training set (1/4, 1/3, 1/2, 3/4, 1/1), with detailed results available in subsection 4.1. These detailed annotations enable the model to learn how to accurately identify and isolate individual leaves from complex backgrounds, ensuring reliable segmentation for further feature extraction and analysis.

3.5 Instance segmentation

For the instance segmentation task, authors selected YOLOv8n-seg and YOLOv9c-seg models, whose training was implemented using the Ultralytics package (LLC (2023)).

The ultralytics package provides an optimized pipeline for YOLO models, including configured setups for essential hyperparameters such as:

1. **Learning Rate (lr)**: Automatically initialized and dynamically adjusted in response to observed gradient behaviors.
2. **Optimizer**: The default Adam optimizer, selected to enhance gradient-based learning, was applied as configured within the Ultralytics framework for YOLO models.
3. **Batch Size**: Automatically determined according to available GPU memory, resulting in a batch size of 8 for this study.

For each model, the training was conducted in three separate configurations, each using images of a different size: 256x256 pixels for the first configuration, 512x512 pixels for the second, and 1024x1024 pixels for the third. This approach allowed the authors to assess how the models performed with varying resolutions and to understand the impact of image size on instance segmentation accuracy.

To determine the optimal number of epochs, the authors analyzed the model loss metrics by training the models with an image size of 256x256 over 100 epochs. The results are summarized in table 2.

Table 2. Initial and Final Loss Values: Box, Segmentation, Classification

Model	Loss metric	Initial value	Final value
YOLOv8n-seg	Box loss	2.3889	0.88616
YOLOv8n-seg	Segmentation loss	5.1496	1.1622
YOLOv8n-seg	Classification loss	3.4797	0.50577
YOLOv9c-seg	Box loss	2.0545	0.71893
YOLOv9c-seg	Segmentation loss	4.017	1.1462
YOLOv9c-seg	Classification loss	2.1793	0.50037

All metrics demonstrated substantial reductions within the initial 60–70 epochs, after which improvements plateaued, indicating diminishing returns. Consequently, the authors set the training limit at 70 epochs. The training loss metrics can be seen in figure 5.

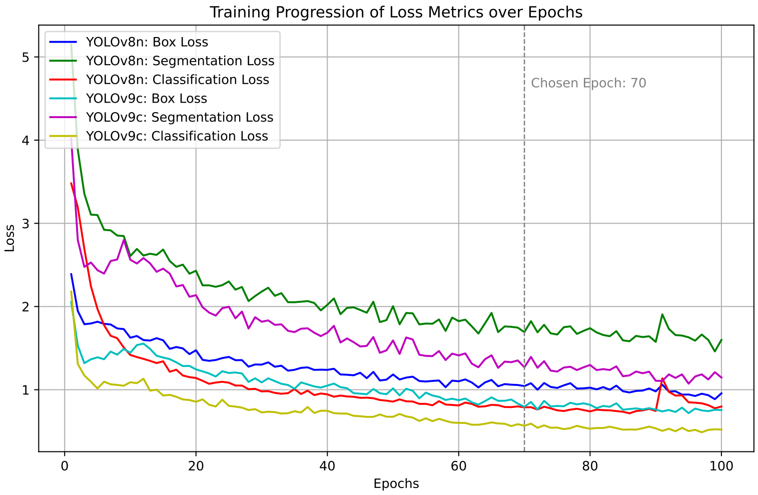


Fig. 5. Training progression of Box Loss, Segmentation Loss, and Classification Loss over 100 epochs

More detailed results, including additional insights and analysis of model performance, are covered in section 4.1.

3.6 Instance isolation

In the proposed method, instance isolation involves isolating the segmented leaf instances from the background to prepare them for classification. Following instance segmentation, the segmentation mask of each detected leaf is extracted and resized to match the original image's dimensions. Using bitwise operations, the segmentation mask is overlaid onto the original image, effectively isolating each leaf by removing unwanted background pixels. The bounding box of the mask is computed to define the region of interest, and the leaf is cropped accordingly. To standardize the input for the classification step, each cropped leaf instance is placed on a black background, ensuring a consistent visual format. These cropped instances, now devoid of any irrelevant features or noise, are then passed to the multi-label classification model to assess their health status.

3.7 Multi-label classification model selection

Based on the related researches (section 2.2), the authors selected three classification models for further investigation: ResNet, VGG and EfficientNet. The authors first conducted a comparative analysis of ResNet and VGG models. Following this initial evaluation, the results were further compared with EfficientNet, a more recent model designed to enhance the scalability of architectures like ResNet and similar convolutional neural networks. The comparison of solutions in this study is based on the analysis of existing research and performance reports, without conducting direct training of the models on the proposed dataset.

Based on the results from training on ImageNet, ResNet is a better choice than VGGNet for this particular task. Specifically, the research compared ResNet-152 and VGG-16, where ResNet-152 achieved a top-1 accuracy of 0.870 and a top-5 accuracy of 0.963, while VGG-16 showed a lower top-1 accuracy of 0.715 and a top-5 accuracy of 0.901 (Wani et al. (2020)).

In a comparison between ResNet and EfficientNet, the authors analyzed the study (Sinha and Patil (2024)) where a comparative analysis of CNN, EfficientNet, and ResNet was conducted for grape disease prediction. Both EfficientNet and ResNet demonstrated strong performance, with ResNet slightly outperforming EfficientNet. According to the study, ResNet achieved the highest accuracy of 98%, while EfficientNet closely followed with an accuracy of 97%. Both models were fine-tuned using transfer learning on a dataset containing high-resolution images of grape leaves affected by diseases such as black rot, leaf blight, and grapevine measles. Although EfficientNet is known for its efficiency in model scaling, the residual learning mechanism in ResNet provided a marginal advantage in this specific task, resulting in better overall classification performance. This study was chosen due to the nature of the task and its similarities to blackcurrant leaf analysis, making the findings highly relevant.

After comparing the models, ResNet has been selected for this study. The training will be conducted on ResNet-50, ResNet-101, and ResNet-152 to determine which architecture yields the best results for blackcurrant leaf analysis.

3.8 Multi-label classification model training

Classification is used to identify and categorize diseases in blackcurrant leaf instances based on distinguishing features. For this task, the authors trained and evaluated ResNet-50, ResNet-101, and ResNet-152.

The training process utilized PyTorch to implement a multi-label classification model based on ResNet architectures (ResNet-50, ResNet-101, and ResNet-152). The models were initialized with their default pre-trained weights provided by PyTorch to leverage feature representations learned from large-scale datasets. The final fully connected layer of each ResNet model was replaced to output predictions for three classes, making them suitable for multi-label classification. Input images were resized to 256x256 pixels and transformed into tensors for processing. The model was trained using the BCEWithLogitsLoss loss function , which combines sigmoid activation with binary cross-entropy, ensuring efficiency and numerical stability for multi-label tasks (Ansel et al. (2024)). The Adam optimizer was selected for its ability to adapt learning rates and incorporate momentum (Kingma and Ba (2017)). To assess performance, evaluation metrics such as accuracy, precision, recall, and specificity were computed.

To identify the optimal hyperparameters for model training, a grid search was conducted separately for each ResNet architecture—ResNet-50, ResNet-101, and ResNet-152—over three key hyperparameters: the number of epochs, learning rate, and batch size. The following ranges were tested: epochs in [20, 30, 50, 70, 100], learning rates in [0.01, 0.001, 0.0005, 0.0001], and batch sizes in [4, 8, 16, 32, 64]. Performance was evaluated using precision, accuracy, recall, specificity, and validation loss.

After completing the grid search, the best-performing hyperparameter configurations for each model were identified as follows: (1) ResNet-50: 70 epochs, a learning rate of 0.0001, and a batch size of 16; (2) ResNet-101: 70 epochs, a learning rate of 0.0001, and a batch size of 8; and (3) ResNet-152: 70 epochs, a learning rate of 0.0001, and a batch size of 16. These optimal configurations were then used to train and evaluate each model. Training was conducted on a dataset comprising 229 images, with validation performed on a separate set of 49 images. The final model evaluation was conducted using a test set of 49 images to assess overall performance. A detailed overview and result analysis can be found in section 4.2.

4 Results and discussion

4.1 Instance segmentation results

The instance segmentation results presented in table 3 highlight the performance of YOLOv8n-seg and YOLOv9c-seg across varying input image sizes.

The evaluation metrics include mean Average Precision (mAP) at IoU thresholds of 0.50 (mAP50) and 0.50–0.95 (mAP50–95), as well as precision for bounding box (B)

Table 3. Instance Segmentation results

Model	Image size	mAP50(B)	mAP50-95(B)	Pre(B)	mAP50(M)	mAP50-95(M)	Pre(M)
YOLOv8n-seg	256x 256	0.59164	0.38272	0.67455	0.58450	0.34303	0.74249
YOLOv8n-seg	512x 512	0.65656	0.49653	0.73628	0.65931	0.47840	0.76106
YOLOv8n-seg	1024x 1024	0.67211	0.52423	0.74736	0.66956	0.51023	0.75789
YOLOv9c-seg	256x 256	0.64149	0.46194	0.74662	0.65124	0.47340	0.72102
YOLOv9c-seg	512x 512	0.68866	0.54670	0.87470	0.69849	0.51147	0.79023
YOLOv9c-seg	**1024x 1024**	**0.72130**	**0.68643**	**0.88182**	**0.69930**	**0.53121**	**0.81230**

and mask (M) predictions. Higher mAP values indicate better performance of the model in accurately detecting and localizing objects. The formulas for these performance metrics are provided in table 4.

Table 4. Performance parameters

Indicator	Formula
mAP50	$\frac{1}{N}\sum_{i=1}^{N}\mathrm{AP}_i(IoU = 0.50)$
mAP50-95	$\frac{1}{10}\sum_{IoU=0.50}^{0.95}\frac{1}{N}\sum_{i=1}^{N}\mathrm{AP}_i(IoU)$
Accuracy (Acc)	$\left(\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FN+FP}}\times 100\right)\%$
Sensitivity (Sen)	$\left(\frac{\mathrm{TP}}{\mathrm{TP+FN}}\times 100\right)\%$
Specificity (Spe)	$\left(\frac{\mathrm{TN}}{\mathrm{FP+TN}}\times 100\right)\%$
Precision (Pre)	$\left(\frac{\mathrm{TP}}{\mathrm{TP+FP}}\times 100\right)\%$

Across all tested image sizes, YOLOv9c-seg consistently outperformed YOLOv8n-seg in both bounding box and mask segmentation metrics. At the highest resolution (1024 × 1024), YOLOv9c-seg achieved a bounding box mAP50 of 0.72130 and a mask mAP50 of 0.69930, compared to YOLOv8n-seg's corresponding scores of 0.67211 and 0.66956. Additionally, YOLOv9c-seg demonstrated better precision metrics, exceeding YOLOv8n-seg's box precision by 16.2% and mask precision by 5.4%.

Increasing the image resolution resulted in improved performance for both models, with mAP50 and mAP50–95 metrics rising consistently. For example, YOLOv9c-seg's bounding box mAP50–95 increased from 0.46194 at 256 × 256 to 0.68643 at 1024 × 1024, while YOLOv8n-seg showed a comparable increase from 0.38272 to 0.52423. This trend suggests that higher-resolution inputs provide richer feature details, enhancing segmentation performance.

In figures 6 and 7, the YOLOv9c-seg model with an image resolution of 1024 × 1024 was used to segment leaf instances. The results in the provided images align well with those presented in table 3, demonstrating strong detection and segmentation of leaves. The high precision (88.18% for bounding boxes and 81.23% for masks) ensures that most detected leaves are correctly classified, which is crucial for agricultural applications such as plant health monitoring.

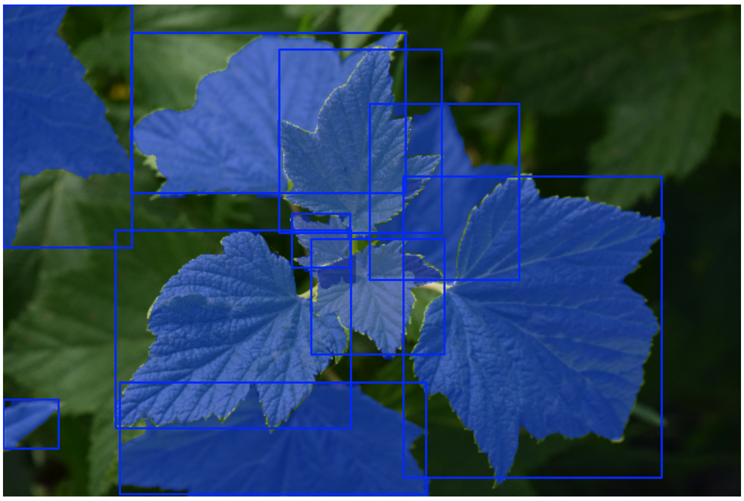


Fig. 6. Instance segmentation results using YOLOv9c-seg (1024 × 1024 resolution) on healthy leaves

However, the model encounters difficulties in recognizing damaged or diseased leaves, as seen in Figure 7, where some leaves with spots or holes show inaccuracies in both the bounding box and the segmentation mask. This limitation suggests that, in real-world scenarios, some unhealthy leaves might be missed, potentially delaying disease detection in crops. Additionally, overlapping detections in dense foliage indicate that the model may have difficulty distinguishing individual leaves in clustered environments, which could impact tasks such as automated pruning recommendations.

While the overall accuracy is promising, further improvements in fine-grained segmentation would enhance the model's ability to support precision agriculture by reliably identifying both healthy and unhealthy leaves.

Overall, the results demonstrate that YOLOv9c-seg is better suited for the given instance segmentation task, particularly at higher resolutions. Since YOLOv9c-seg achieved

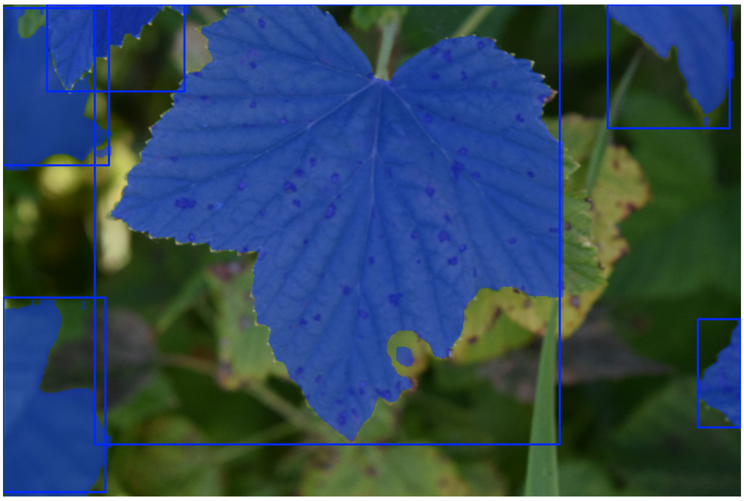


Fig. 7. Instance segmentation results using YOLOv9c-seg (1024 × 1024 resolution) on leaves with visible damage

the best performance during training, the model was further trained using different portions of the training set (1/4, 1/3, 1/2, 3/4, 1/1) to analyze its robustness and adaptability. The training results are summarized in table 5 and visualized in figure 8.

Table 5. YOLOv9c-seg instance Segmentation results based on dataset fraction

Dataset fraction	mAP50(B)	mAP50-95(B)	Pre(B)	mAP50(M)	mAP50-95(M)	Pre(M)
1/4	0.49543	0.26224	0.72417	0.48971	0.23260	0.71650
1/3	0.51329	0.28320	0.74446	0.48492	0.25204	0.70009
1/2	0.53370	0.30844	0.77216	0.51555	0.26884	0.75563
3/4	0.59152	0.37815	0.78589	0.57172	0.33265	0.79359
1/1	0.72130	0.68643	0.88182	0.69930	0.53121	0.81230

The results indicate a clear trend of performance improvement as the dataset fraction increases. With only 1/4 of the training data, YOLOv9c-seg achieved a bounding box mAP50 of 0.49543 and a mask mAP50 of 0.48971, which gradually improved with larger dataset portions. Notably, at the full dataset, the model reached its highest performance, with a bounding box mAP50 of 0.72130 and a mask mAP50 of 0.69930. This demonstrates that increasing the dataset size leads to higher segmentation accu-

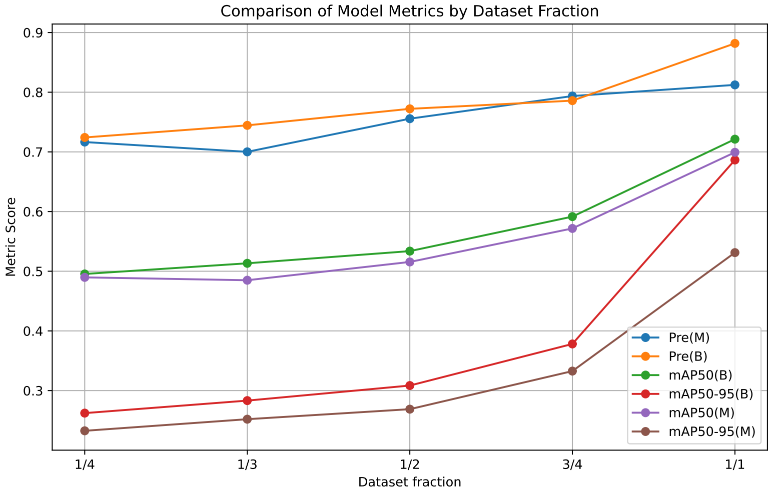


Fig. 8. Comparison of YOLOv9c-seg instance segmentation metrics by dataset fraction

racy, as no decline in performance was observed at any stage. Based on this trend, it is reasonable to assume that further expanding the dataset would yield even better results. However, since this study presents a conceptual rather than an industrial approach, and given that instance segmentation annotation is a time-consuming process, the dataset was not further extended.

4.2 Multi-label classification results

The performance of the multi-label classification models is summarized in table 6. The results were evaluated using accuracy, sensitivity, specificity, and precision, with all performance metrics defined in table 4. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) were used to compute these metrics, where TP and TN indicate correct predictions, while FP and FN represent misclassifications.

Table 6. Multi-label classification results

Model	Acc (%)	Sen (%)	Spe (%)	Pre (%)	Epochs	Batch size	Lr
ResNet-50	0.896	0.913	0.982	0.913	70	16	0.0001
ResNet-101	0.931	0.948	0.981	0.965	70	8	0.0001
ResNet-152	0.940	0.967	0.988	0.982	70	16	0.0001

ResNet-152 achieved the highest performance among all models, with accuracy, sensitivity, and specificity reaching 94.0%, 96.7%, and 98.8%, respectively. However, ResNet-101 also demonstrated strong results, attaining an accuracy of 93.1% and precision of 96.5%, indicating its effectiveness in distinguishing between the three classes. ResNet-50, while exhibiting slightly lower accuracy (89.6%), maintained competitive sensitivity and precision values. All models were trained for 70 epochs, with variations in batch size and learning rate having a limited impact on overall performance.

The results suggest that increasing model complexity does not yield substantial improvements, as even the smaller ResNet-50 model performed well. The similarity in metrics across all three architectures indicates that the dataset's characteristics, such as size and class distribution, may have a greater influence on performance than the choice of ResNet depth. This suggests that ResNet-50 provides a suitable trade-off between accuracy and computational efficiency, making it a practical choice for this classification task.

5 Conclusion

This study demonstrated the potential of a combined instance segmentation and multi-label classification approach for analyzing blackcurrant leaf health using RGB images. While the achieved accuracy was not the highest, it was sufficient to identify underlying issues in blackcurrant fields. YOLOv9c-seg excelled in instance segmentation, particularly at higher resolutions, enabling precise detection of individual leaves, and ResNet-50 provided a reliable balance between classification performance and computational efficiency. Instance segmentation training with different fractions of the dataset showed a tendency for the metrics to improve, with no decline even when using the full dataset, indicating that a larger dataset could further enhance overall performance. Importantly, even though the method may not yield perfect results for every individual instance, the aggregated data is effective for identifying patterns of health issues and marking their spatial distribution across the field. This capability supports targeted interventions and resource-efficient crop management. Future work should explore larger datasets and enhanced models to further improve accuracy and adaptability.

6 Acknowledgements

This work has been conducted within the research project ”Competence Centre of Information and Communication Technologies” of The Recovery and Resilience Facility, contract No. 5.1.1.2.i.0/1/22/A/CFLA/008 signed between IT Competence Centre and Central Finance and Contracting Agency, Research No. 1.8 “Innovative Use of Drones in Horticulture.”

References

Anam, I., Arafat, N., Hafiz, M. S., Jim, J. R., Kabir, M. M., Mridha, M. (2024). A systematic review of uav and ai integration for targeted disease detection, weed management, and pest control in precision agriculture, *Smart Agricultural Technology* **9**, 100647. https://doi.org/10.1016/j.atech.2024.100647

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M. Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., Chintala, S. (2024). Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation, *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*, ACM. https://doi.org/10.1145/3620665.3640366

Bai, Y., Shi, L., Zha, Y., Liu, S., Nie, C., Xu, H., Yang, H., Shao, M., Yu, X., Cheng, M., Liu, Y., Lin, T., Cui, N., Wu, W., Jin, X. (2023). Estimating leaf age of maize seedlings using uav-based rgb and multispectral images, *Computers and Electronics in Agriculture* **215**, 108349. https://doi.org/10.1016/j.compag.2023.108349

Charisis, C., Argyropoulos, D. (2024). Deep learning-based instance segmentation architectures in agriculture: A review of the scopes and challenges, *Smart Agricultural Technology* **8**, 100448.
https://doi.org/10.1016/j.atech.2024.100448

Debangshi, U. (2021). Drones-applications in agriculture, *Chronicle of Bioresource Management* **5**(Sep, 3), 115–120.

Elfatimi, E., Eryiğit, R., Elfatimi, L. (2024). Deep multi-scale convolutional neural networks for automated classification of multi-class leaf diseases in tomatoes, *Neural Computing and Applications* **36**(2), 803–822.
https://doi.org/10.1007/s00521-023-09062-2

Fei, S., Xiao, S., Li, Q., Shu, M., Zhai, W., Xiao, Y., Chen, Z., Yu, H., Ma, Y. (2023). Enhancing leaf area index and biomass estimation in maize with feature augmentation from unmanned aerial vehicle-based nadir and cross-circling oblique photography, *Computers and Electronics in Agriculture* **215**, 108462.
https://doi.org/10.1016/j.compag.2023.108462

Gu, W., Bai, S., Kong, L. (2022). A review on 2d instance segmentation based on deep neural networks, *Image and Vision Computing* **120**, 104401.
https://doi.org/10.1016/j.imavis.2022.104401

Hafeez, A., Husain, M. A., Singh, S., Chauhan, A., Khan, M. T., Kumar, N., Chauhan, A., Soni, S. (2023). Implementation of drone technology for farm monitoring & pesticide spraying: A review, *Information Processing in Agriculture* **10**(2), 192–203.
https://doi.org/10.1016/j.inpa.2022.02.002

Hosny, K. M., El-Hady, W. M., Samy, F. M., Vrochidou, E., Papakostas, G. A. (2023). Multi-class classification of plant leaf diseases using feature fusion of deep convolutional neural network and local binary pattern, *IEEE Access* **11**, 62307–62317.
https://doi.org/10.1109/CVPR.2016.90

Istiak, M. A., Syeed, M. M., Hossain, M. S., Uddin, M. F., Hasan, M., Khan, R. H., Azad, N. S. (2023). Adoption of unmanned aerial vehicle (uav) imagery in agricultural management: A systematic literature review, *Ecological Informatics* **78**, 102305.
https://doi.org/10.1016/j.ecoinf.2023.102305

Janani, M., Jebakumar, R. (2023). Detection and classification of groundnut leaf nutrient level extraction in rgb images, *Advances in Engineering Software* **175**, 103320.
https://doi.org/10.1016/j.advengsoft.2022.103320

Khan, F., Zafar, N., Tahir, M. N., Aqib, M., Waheed, H., Haroon, Z. (2023). A mobile-based system for maize plant leaf disease detection and classification using deep learning, *Frontiers in Plant Science* **14**.
https://doi.org/10.3389/fpls.2023.1079366

Kingma, D. P., Ba, J. (2017). Adam: A method for stochastic optimization.
https://doi.org/10.48550/arXiv.1412.6980

LLC, U. (2023). Ultralytics yolo: Real-time object detection and segmentation. Accessed: 2024-11-26.
https://ultralytics.com

Sapkota, R., Ahmed, D., Karkee, M. (2024). Comparing yolov8 and mask r-cnn for instance segmentation in complex orchard environments, *Artificial Intelligence in Agriculture* **13**, 84–99.
http://dx.doi.org/10.1016/j.aiia.2024.07.001

Sinha, S. V., Patil, B. M. (2024). Comparative analysis of cnn, efficientnet and resnet for grape disease prediction: A deep learning approach, *International Journal of Intelligent Systems and Applications in Engineering* **12**(3), 600–609.
https://ijisae.org/index.php/IJISAE/article/view/5291

Wani, M. A., Bhat, F. A., Afzal, S., Khan, A. I. (2020). *Basics of Supervised Deep Learning*, Springer Singapore, Singapore, pp. 13–29.
https://doi.org/10.1007/978-981-13-6794-6_2

Received November 28, 2024 , revised April 30, 2025, accepted June 13, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 2, 502-527
https://doi.org/10.22364/bjmc.2025.13.2.11

Cyber Threat Risks in Higher Education Institutions: an Example of the Estonian Academy of Security Sciences

Kate-Riin KONT

Estonian Academy of Security Sciences, Internal Secyrity Institute, Kase 61, 12012 Tallinn, Estonia

kate-riin.kont@sisekaitse.ee

ORCID 0000-0002-9184-2363

Abstract. Purpose of this study is to identify the most common characteristics that make users vulnerable, either individually or in groups, and to determine whether there is a relationship between user behaviour and victimisation of a cyber-attack. This research should help characterise people who are more likely to become victims of various phishing and social attacks. For this purpose, students, employees and lecturers of the Estonian Academy of Security Sciences were investigated. A five-scale questionnaire was used as the methodology of the study, which takes into account the following behaviour scales: risky behaviour, conservative behaviour, risk exposure behaviour and risk perception behaviour. Survey scales already used in previous studies were applied to the students, academics and administrative staff of the Estonian Academy of Security Sciences (hereinafter Academy). These scales and questions are quite well suited for identifying cyber risks that are threatening the patrons of higher education institutions. The results of the study show there are significant differences within the samples and according to Internet usage habits and positions in the Academy.

Keywords: Cyber security, Cyber threats, Higher Education Institution, User behaviour, Risky behaviour, Conservative behaviour, Exposure behaviour, Risk perception behaviour

1. Introduction

The use of digital technologies in the education sector has increased worldwide in recent decades. Educational technologies have become an integral part of teaching and learning processes in the form of computer devices for content delivery, online learning applications, cloud storage, learning management systems, and computer-based assessment and training systems. Especially after the COVID-19 pandemic, many educational institutions have no choice but to use distance learning with the help of digital technologies to ensure the continuation of teaching. However, the increasing use of technology in education brings with it a number of challenges, including technical and human behavioural issues of cyber security, and insufficient cyber training for faculty, staff and students. In recent years, several cyber incidents against higher education institutions have exposed the problems that cyber threats can bring to educational institutions. These incidents have resulted in large-scale personal data leaks of students, staff and alumni, and ransomware attacks, which also cause serious financial losses, in which the target is blocked from accessing their data and a ransom is required to regain access or prevent data leakage. (Noran, 2021).

Cyber security means that digital data and the processing of that data are fully protected. Data creation, transmission, storage, presentation and all other data handling processes are protected. In other words, cyber security is a state where all kinds of threats that could affect digital data and the use of computers, smart devices, memory sticks and e-services do not materialise. In cyber and information security, there is a question about security. Safety is a subjective experience, there is no unequivocal description of safety. Everyone also experiences security in their own way: the same security threat can cause a much stronger feeling of insecurity in one person than in another. Security is made up of several factors: emotion, learned patterns, reality, and our ability to withstand disturbances and crisis situations. Security is an emotional state that varies depending on the situation a person is experiencing. Reality is an important factor in security, the way things around us are said to be “cold facts”. Learned values and patterns guide people as they value and prioritise safety. Resilience to disruption and crisis situations determine how people respond to disruption. With a good tolerance, one can deal with disturbances without panic. Understanding reality is a priority so that our sense of security is not based on false assumptions. Only with a proper understanding of reality can resilience be strengthened and models that develop safety be created. Our emotions are constantly appealed to in the media, by creating emotions one can influence people well. It does not matter whether our perceptions and fears are based on imaginary or real threats, either way, perceptions and fears drive our behaviour and greatly affect our well-being. To increase safety, development must take place in safety areas (Limnéll et al., 2014).

Although the hardware and software solutions used to ensure cyber security are constantly updated, it is still not possible to prevent information systems from becoming compromised, and the reason for this is precisely the behaviour caused by people’s ignorance. Although individuals may have knowledge about cyber security, this knowledge is not always reflected in appropriate behaviour. So, everyday cyber security is not a problem that can be solved by technological solutions alone. People’s behaviour in the field of cyber security must be evaluated as the weakest link.

This article is based slightly on a conference presentation given at the CYBER 2023: The Eighth International Conference on Cyber-Technologies and Cyber-Systems conference (Kont, 2023). The study examines the behaviour of students, lecturers (researchers) and employees of the Academy regarding hybrid threats and possibilities to prevent risks related to cyber security. This study is part of a larger research conducted within the framework of the cooperation programme on hybrid threats (HYBRIDC). The results of the study can be used to develop strategies and training to reduce errors related to the human factor in the cyber security of higher education institutions. Based on the two main studies, Ögˇütçü et al. (2016) and Benavides-Astudillo et al. (2022) of the Risky Behaviour Scale (RBS), the Conservative Behaviour Scale (CBS), the Exposure to Offence Scale (EOS) and the Risk Perception Scale (RPS), this study definitely aimed to obtain answers to the following questions:

- Is there a significant difference between the surveyed groups (females and males) concerning their average score according to different behavioural scales (RBS, CBS, EOS, RPS)?
- Is there a significant difference between the surveyed groups (students, academics and administrative staff) concerning their average score according to different behavioural scales (RBS, CBS, EOS, RPS)?
- Does the duration of time spent on the Internet affect the average of the scales (RBS, CBS, EOS, RPS)?
- Does the cyber security training attendance or non-attendance affect the average of the scales (RBS, CBS, EOS, RPS)?

The results were analysed using SPSS descriptive statistics analysis and ANOVA analysis with post-hoc tests to answer research questions, and the results are presented as tables. The results of the study can be used by either organisations or educational institutions to develop personalised and proactive training programmes or to prepare preventive strategies. This paper is structured as follows. In the literature review section, a brief overview of how a threat is conceptualised in cyber security is given, and an analysis of why security breaches in higher education institutions have become very frequent. In the research methodology section, the author briefly introduces the study design and sample characteristics. The results section answers the research questions. Finally, conclusions are given.

2. Literature review

2.1. Why are higher education institutions the targets of cyber-attacks?

Cyber security in a higher education institution is completely different than in the private sector because it is an open institution. There are many access points, there is a lot of personal information about employees and students. Information security training, awareness raising and cyber behaviour monitoring are not always top priorities for educational institutions. The contribution of lecturers, researchers and employees who engage in research and teaching work or provide administrative support to these activities are often considered to be the central figures of a higher education institution. IT employees deal with security to the extent that they have the human and time resources for it.

Higher education and academic institutions have become beneficial targets for cyber-attackers. In recent years, security breaches in higher education institutions have become very frequent. For example, the University of Maryland has repeatedly been the victim of cyber-attacks – in 2014, over 309,000 personal data records containing social security numbers, dates of birth and university ID numbers were breached. In 2015, the personal information of 288,000 students, faculty and staff at the University of Maryland was breached, and a month later it was breached again (Svitek and Anderson, 2014; Roman, 2018). The University of Maryland was also included in the list of recent actions by Russian hackers (Yerby and Floyd, 2018).

Attempts have also been made in Estonia to gain access to university emails with phishing letters. It was malware hidden in fake emails, which would have given access to the contents of the email account when activated. For example, in 2020, such an attack was made on the University of Tartu. In this case, too, it was most likely a campaign of Iranian state origin, known as Silent Librarian and Mabna Institute. With its expert action, the university was able to both detect the attack and prevent more damage (Einmann, 2020). On 4 September 2020, Tartu Health Care College was hit by a cyber-attack, which paralysed the use of the institution's service servers for several days. As a result of the break-in, backup copies, clones created for installing workstations, software installation files, the school's image bank accumulated over time and the original material of many educational videos were destroyed after resetting the data repositories (Karu, 2020).

Cyber-attacks on universities show that such an attack can be not only detrimental to relations between countries, but even life-threatening. Düsseldorf University Hospital failed to admit a woman brought by ambulance on 19 September 2020 after a cyber-attack froze the hospital's information system – the prosecutor started a murder investigation. It was the first time a human death was directly linked to a cyber-attack. However, it was not certain if the university hospital was the actual target of the attack or if it was collateral damage in an attack on the university. The ransom demands were aimed at Heinrich Heine University, not the hospital directly connected to it. The police contacted the attackers and informed them that the target of the attack was the hospital, not the university, and that the patient's life was in danger. After that, the attack was stopped and the authorities were given the encryption key, but it was too late (Busvine and Kaeckenhoff, 2020).

The world has been in a new security situation since 2022 when Russia started a war of aggression in Ukraine. Seppänen (2022) emphasises that "Typical information security threats for higher education institutions are phishing attacks and frauds, the aim of which is to obtain user IDs, gain access to higher education systems, or obtain, for example, financial benefits. The resulting data can be used, for example, to deliver scams and phishing messages, as emails sent on behalf of the university are generally considered trustworthy. Every month, hundreds of thousands of phishing messages are sent to individual universities, most of which do not pass technical security measures and are therefore not visible to students." However, in the home office, doing business with one's own devices can transfer information security threats from the private place to the workplace. Data security must be considered as a whole – it does not end when you leave the workplace or close the work laptop, but also in your free time (Seppänen, 2022).

Academic institutions are undertaking more cyber security research than before, while the higher education sector itself often leaves the issue of cyber security to IT technicians. The education sector has proven to be an interesting and, unfortunately, easy target for cyber-attacks. Prevention, collaboration, access rights management and training are examples of safeguards. Higher education often experiences the scenarios described above. Organisations that do not adequately protect and educate themselves may inadvertently expose the personal data of students and staff, as well as research data that is valuable to cybercriminals operating both domestically and internationally. Incidents like these are not stopping, so it is critical for higher education institutions to consider whether their cyber defence management is strong enough.

2.2. What is a cyber threat?

Threat orientation is emphasised in the cyber environment. This means that there is no security without threats and risks. The threat consists of stability and credibility. In a cyber environment, not all threats can be anticipated or taken into account. Threats like this have come to be called black swans (Limnéll et al., 2014). A cyber threat is an event in cyberspace that can potentially cause a loss of assets and undesirable consequences as a result (Bederna and Szadeczky, 2020). According to Shad (2019), it is the "action that may result in unauthorised access to, exfiltration of, manipulation of, or impairment to the integrity, confidentiality, or availability of an information system or information that is stored on, processed by, or transiting an information system" (Shad, 2019). A cyber threat is the possibility of a malicious act, the purpose of which is to damage or disrupt an information network system (Oxford Dictionary, 2019). Threats can be viewed at different levels (international, national, companies and individuals) and their significance varies by level (Iiskola, 2019).

From the state's point of view, the most serious cyber threats are directed at critical infrastructure. The background of the threat may be terrorism, crime, a state actor or a state-sponsored actor. States may use, for example, criminal organisations to achieve their own goals, in which case the actions and the state's involvement are also debatable (Iiskola, 2019). The threats can be classified as enabled by the cyber environment as part of hybrid influence. Hybrid influence is the non-military means of a state actor to influence another state's political and economic decision-making. For example, political opinion may be influenced via social media. In terms of the functioning of society, significant cyber threats are threats to communication services and networks, because the functioning of electricity networks and payment traffic, for example, depend on the functioning of communication services. The most common disruptions in communication services are caused by disruptions in electricity supply. The most serious cyber threat is attacks on energy production and health services due to the potential loss of human lives. Other serious threats are the automation and control systems of power plants or nuclear power plants, food transport and logistics systems, healthcare information systems, traffic control systems, banking and payment systems, and information systems enabling communication services as well as interference with satellite positioning and domain name services. Attacks on critical infrastructure may be attractive because of their effectiveness. Most of the critical infrastructure is held by private companies and organisations. Thus, national-level cyber security cannot be completely separated from corporate and organisational cyber security (Kansallinen riskiarvio, 2024).

Threats to the organisation consist of internal and external threats. External threats can be thought of as a deliberate and purposeful attack by someone outside the organisation. Internal threats can be divided into intentional and unintentional threats. Unintentional threats may be caused, for example, by ignorance, distraction or carelessness. They can also be conscious or unconscious (Limnéll et al., 2014). Intentional internal threats consist of insider crimes. Insider crime refers to a crime committed by a person who is able to utilise information and skills obtained from the organisation that others would not necessarily have had access to. Cyber-attacks by insiders accounted for about 14% of all attacks in 2013. In half of these attacks, the former employee used their old credentials or backdoors that had not been closed (Widup, 2013). In 2018, an insider was involved in 28% of attacks (Widup, 2018). It may be more difficult to protect against an attack by an insider. However, the threat is real, and its importance seems to be growing.

Opportunities also come with risks (Juvonen et al., 2014). According to Hubbard (2020), “Risk means a state of uncertainty where some of the possibilities involve a loss, catastrophe, or other undesirable outcome or event” (Hubbard, 2020). On the other hand, risk can be understood as the courage to make a decision and the freedom to choose between different options. Risks are perceived as scary when they cannot be controlled (Kuusela, 2005). Cyber risk is the risk of financial or reputational loss due to the non-functioning of information technology. Risks are included in all activities and cannot be completely eliminated, unlike threats. Risk is a condition of existence and the possibility of a negative event in the future. However, one can learn to live with risks and manage them. The aim of risk management is to minimise the probabilities and effects of risks. It is a method that aims to identify and assess risks and to choose, develop and implement alternatives. Risk management can also be seen as giving information about the organisation’s reliability and responsibility. As a result of successful risk management, resources can be effectively allocated to risk reduction (Limnéll et al., 2014). Technology also increases the possibilities of risk management. However, as addictions increase, the severity and impact of the risks may increase (Kuusela, 2005).

Cyber security is a balancing act between opportunities and threats. Threats are often emphasised in the phenomenon, when opportunities should be emphasised. There is no perfect cyber security, just as there is no perfect physical security either. The best way to prepare for threats is to put the basics in order (Limnéll et al., 2014).

2.3. Factors influencing cyber security and risk behaviour

As educational environments increasingly rely on digital platforms, understanding the factors that influence cyber security practices has become paramount. There are undoubtedly many factors that influence risk behaviour, of which this study focuses on status and gender differences, time spent online, and cyber training.

The assessment of cyber security awareness among students has revealed concerning trends regarding their engagement with protective measures. For instance, Saeed's study indicates that while students demonstrate a reasonable level of awareness regarding cyber security threats, their actual protective behaviors often fall short (Saeed, 2023). This gap between awareness and action highlights the need for targeted interventions that not only inform but also motivate individuals to adopt safer online practices. The implications of these findings extend beyond individual behavior to encompass broader institutional responsibilities. As highlighted by Triplett, educational institutions must proactively address cyber security challenges by implementing strategies that enhance awareness and encourage students to consider careers in cyber security (Triplett, 2023). Similarly, Concepcion's assessment of cyber security awareness among academic employees highlights the necessity of promoting cyber hygiene to secure sensitive information within educational institutions (Concepcion and Palaoag, 2024).

Cyber security awareness and training is a non-technical cyber risk prevention measure organisations use to strengthen the resilience of the socio-technical system at the human factor level (Pollini et al., 2022). Organisational cyber security programmes have traditionally focused on technical controls to protect infrastructure and equipment, while cybercriminals have focused on exploiting human vulnerabilities. Most internal users put the organisation at risk through negligence, error, or lack of knowledge. These non-malicious actions cause the majority of cyber incidents (Zimmermann and Renaud, 2019). Although employees are mostly informed about cyber risks and measures to mitigate them, it is common for them to take risky actions either out of haste, carelessness or fatigue. Awareness of the cyber risk landscape and the organisation's security policy, along with training on the actions and behaviours needed to mitigate risks, is essential to reduce human error and intentional misuse of information systems, including insider threats to organisations (Hadlington, 2018). Therefore, to mitigate cyber risks effectively, cyber security training has focused on raising awareness and educating end users about cyber risks (Jalali et al., 2019). Qashqari et al. (2020) argue that even with a strong security policy, people are considered the weakest link in information security, therefore, the study of human behaviour from a cyber security perspective is an important topic for organisations. Recent studies have identified that demographics are important factors that influence a person's attitude and behaviour towards cyber security. For organisations to develop effective cyber security training programmes, it is important to understand the security behaviours, similarities, and differences in the behaviour of different target groups.

Anwar et al. (2017) explain that gender is one of the most fundamental social groups that influences an individual's perceptions, attitudes, and outcomes, regardless of whether they are male or female. According to Ifinedo (2012), men have lower rates of security policy compliance compared to women, and the author recommends that those responsible for security and policy pay attention to gender differences in security policy compliance in organisations. The author also calls for the implementation of targeted security awareness programmes, training, and monitoring to eliminate differences in security behaviours between men and women. Verkijika (2019) suggests that women often have lower safety characteristics and related experiences compared to men.

The relationship between time spent online and risk behaviour has been poorly studied. Jeske and Van Schaik (2017) conducted a study on students' awareness of various online threats. Participants were presented with definitions of threats and asked how familiar they were with these threats. The responses showed that time spent online and length of online experience predicted awareness of online threats, which in turn predicted the use of computer security. Average time spent online positively predicted familiarity with threats, as did the duration of Internet use over several years. Familiarity was a significant predictor of positive cyber behaviour. Mediation analysis of the results showed a significant indirect effect, with time spent online and duration of Internet use fully mediating the relationship between awareness of threats and online behaviour. The study provided further evidence that time spent online and length of online experience (although not daily or weekly frequency of use) were significant predictors of threat awareness and online behaviour. These variables were also significant indirect predictors of computer security use, which was fully mediated by familiarity. Although the effects were generally quite small, the practical conclusion can be that computer security behaviour depends on familiarity, which is not achieved without a significant investment of time. This means that the time spent becoming aware of threats and learning about online opportunities is a time of learning, but at the same time a time of increased vulnerability until a certain level of familiarity with threats is achieved, which in turn triggers security behaviour. Duman (2022) studied the impact of students' daily Internet use on cyber security behaviour. The results showed that students' cyber security behaviour differs depending on the time spent online. Namely, students with lower daily Internet use (less than 1 hour or 1–2 hours per day) had significantly less risky cyber security behaviour compared to students who used the Internet for 3–4 hours or 5 hours per day or more. Based on this result, it can be concluded that students with lower daily Internet use have higher cyber security awareness. However, a similar study by Yiğiti and Seferoğlu (2019) did not find a significant relationship between students' cyber security behaviour and time spent online.

In conclusion, cyber security awareness and training are important measures in organisations to prevent cyber risks, as cybercriminals can often exploit people's vulnerabilities. Recent research underscores the multifaceted nature of cyber security risky behavior factors among students and staff in higher education institutions. The interplay of awareness, individual behaviors, institutional strategies, and social dynamics plays a crucial role in shaping cyber security practices. Based on the reviewed literature, it can be argued that gender differences play an important role in security behaviour – studies have found that men often have lower rates of compliance with security policies than women. The relationship between time spent online and cyber security risky behaviour has also been studied. Studies have shown that longer experience and time using the Internet are associated with greater awareness of risks and online behaviour. However, there are also studies that did not find a significant relationship between Internet use and more secure cyber behaviour.

3. Research methodology, the analysis techniques used and sample characterisation

The Academy was selected for this study primarily because the author is affiliated with the university. As Academy educates future professionals in internal security and operates under the direct supervision of the Estonian Ministry of the Interior, it is expected that both students and staff have received comprehensive training in cyber security. Therefore, it is both relevant and appropriate to assess the cyber security practices of Academy members (students and employees) in their online activities, exploring whether their behavior tends to be cautious or risky. This is the first study on information and cyber security behavior conducted at a higher education institution in Estonia, and its findings may provide insights into the online practices of individuals at other universities across the country.

This research uses a four-scale model to measure behaviour and awareness of cyber security. The Risky Behaviour Scale (RBS) measures the degree of risk of users of information systems related to behaviour, the Conservative Behaviour Scale (CBS) measures how careful users are when using information systems, the Exposure to Offence Scale (EOS) measures users' exposure to cyber security incidents due to their behaviour and the Risk Perception Scale (RPS) measures the level of danger or risk associated with information technology model and survey method to collect data (Benavides-Astudillo et al., 2022; Ceran, 2021; Ögˇütçü et al., 2016). Using the given model of scales, the challenges of the questionnaire survey were finding out the respondents' attitude towards the survey, successful and comprehensible wording of the answer options, clarity of the scale, and time required to answer the survey (Hirsijärvi, 2010). A major challenge of this study was the possible low response rate due to the novelty of the survey topic. Placing the questions in the context of Estonian higher education created challenges, as the topic is relatively unaddressed in higher education in the Baltic countries, so there was a high risk of misunderstandings. A questionnaire technique was used to achieve the objectives of the study and the survey was conducted online to obtain a large sample of both staff and students as efficiently and ethically as possible. The questionnaire consisted of 56 questions covering various aspects of cyber security, including the RBS (20 questions); the CBS (10 questions); the EOS (7 questions) and the RPS (17 questions). Additionally, 6 demographic questions were asked, and the questionnaire ended with a so-called open question, where the respondents were asked to express an opinion about the discussed topic or the questionnaire or simply leave a comment. The survey questions were selected based on instruments developed by other cyber security researchers (mainly based on Benavides-Astudillo et al. (2022), and IT-experts-suggested questions from the Academy. The digital team of the Academy added several suggestions to change the wording and order of the questions to make them more suitable for the context of the field of internal security. At the beginning of the questionnaire, the importance of cyber security in higher education was introduced and it was explained why higher education institutions have become attractive and important targets for cybercriminals. The significance and novelty of the questionnaire in the context of both Estonia and the Baltic states were also explained.

Answers were given according to a 5-point Likert type. The proposed scales were formulated depending on the questions asked. Total respondent scores were calculated by assigning 5 points for “Always”, 4 points for “Often”, 3 points for “Sometimes”, 2 points for “Rarely”, and 1 point for “Never” for the RBS and CBS questions. A higher score indicates that the respondent is very risk-tolerant. In RPS, “Very dangerous” is 5 points, “Dangerous” is 4 points, “Slightly dangerous” is 3 points, “Not dangerous” is 2 points and “I don’t know” is 1 point. As the scores increase, it is understandable that the respondent considers related technologies more dangerous (Benavides-Astudillo et al., 2022; Ceran, 2021; Ögˇütçü et al., 2016). In EOS, it is said that as the scores increase, the respondent is exposed to crime (negative experience) at a higher level. This was the bottleneck of the EOS scale of the questionnaire, which could not be adapted to the Academy questionnaire with the same points as in the other studies discussed. Since people who are very aware of cyber security and work at the Academy, the majority of them had never or rarely encountered the dangerous cyber situations mentioned in the questionnaire. Therefore, the author decided to invert this scale – 1 point for “Always”, 2 points for “Often”, 3 points for “Sometimes”, 4 points for “Rarely”, and 5 points for “Never” (see desciptive statistics Kont, 2024, pp. 94, 96, 97, 98).

As pointed out in the introduction section, several analysis techniques and tests were conducted to answer the research questions. The analysis of variance (ANOVA) is the most important of these tests and involves an additional post-hoc Tukey test. The meaning of ANOVA can be explained in several ways. First, as a comparison task, i.e. the study of how uniform the averages of groups are under a certain classification. Second, ANOVA as a prediction task, i.e. the study of how well the average variability of the characteristic under study can be described statistically through group membership in a certain classification. This means modelling the relationship between traits and the task of forecasting group means within the model. Variance analysis also makes it possible to deal with classifications based on several characteristics at the same time and in their mutual interaction, i.e. interaction. The average differences between women and men, the average differences depending on the job position, as well as whether the differences depending on the number of hours spent on the Internet per day or the completed cyber training are the same on average for women and men, etc. can be studied. If there is one argument characteristic, then it is a one-way ANOVA model (Tooding, 2014). The following items are the basis for making decisions:

1. If the significance value is >0.05, there is no significant difference.

2. If the significance value is <0.05, there is a significant difference.

Statistical significance indicates the probability that the result or effect that was discovered was obtained purely by chance. Statistical significance is expressed by a p-value that falls between 0 and 1 (because there is a probability). A small p-value (0.01) indicates that the probability of chance is small (in this case only 1%). Therefore, for a p-value of 0.01, one can be 99% confident that the result is not random. Usually, $p = 0.05$ is considered the limit of statistical significance of the results. However, $p > 0.05$ does not necessarily mean that the observed result is also substantively significant. Here is where formulas cannot be blindly trusted. The assessment of the importance of the obtained results from the point of view of the study must be given to the author of the study. Consequently, results with a p-value slightly higher than 0.05 should not be automatically discarded (McLeod, 2023).

If the significance value is smaller than 0.05, it means there is a significant difference, and a further test called a post-hoc test should be used to find out the difference (Candiwan et al., 2022, 233). It is used after performing a one-way ANOVA. ANOVA can say whether there are significant differences between the groups being studied, but it does not say which groups differ. Tukey's honestly significant difference test (Tukey HSD) is used to test the significance of differences in sample means. Tukey's HSD tests all pairwise differences while controlling for the probability of making one or more errors (Lane, 2012). A paired samples t-test is a hypothesis test for determining whether the population means of two dependent groups are the same. The test begins by selecting a sample of paired observations from the two groups. Thus, each observation in each group is paired (matched) with another observation from the other group. After that, the difference between each of these paired observations will be calculated and a one-sample t-test on these difference scores conducted (Stone, 2012)..

When there are only two characteristics to compare and ANOVA shows that there are significant differences between the scales, instead of the Tukey test, the Kruskal-Wallis test could be chosen, which shows more clearly where significant differences occur (Sonavala, 2024).

Invitations to participate were sent to the email addresses of 1,000 undergraduate students, 69 master's students, 439 faculty members and 271 staff members. A total of 277 employees and students from the Academy answered the questionnaire through LimeSurvey.

4. Results

4.1. General results

Table 1 describes how many people from each group were investigated. There were more women than men among the respondents. Among the age groups, most respondents were from the 41–50 age group (27%), followed by the 19–25 age group (30%) and 31–40 age group (19%). As much as 60% of respondents have completed cyber security training. Most people spend 1–5 hours a day on the Internet (52%), but there were also those who spent 11 or more hours a day on the Internet (3%). Outside of school, mobile Internet (48%) and private Wi-Fi networks (46%) are mainly used to access the Internet.

Table 1. Results of the respondents profile section (Kont, 2024, p. 93)

Characteristic	***Category***	***Number of respondents***	***Percentage***
Gender	Male	120	43%
	Female	157	57%
Age range	19–25	68	30%
	26–30	27	9%
	31–40	58	19%
	41–50	81	27%
	51–60	32	11%
	61-70	9	3%
	70+	2	1%
Position in the SKA	Vocational student	33	12%
	Under-graduate	98	35%
	Graduate	14	5%
	Lecturer	71	26%
	Administrative staff	42	15%
	Others	19	7%
Completed cyber security training	Yes	165	60%
	No	112	40%
Time range of Internet use	1-5 hours/day	145	52%
	6-10 hours/day	123	44%
	11 or more hours/day	9	3%
How do you access the Internet from outside your workplace?	Using Mobile Internet	133	48%
	Using public Wi-Fi network(Cafes, Shopping malls)	1	1%
	Using private Wi-Fi network (Home)	15	46%
	Using remote connection of my organization	128	5%

The question about their level of IT skills was not directly asked, because it is clear that basic computer literacy is required when studying, teaching, and working at a higher education institution. In case of problems, an employee of the IT department is always there to help. The author of the study was more interested in the level of their cyber skills and whether or not they had completed the relevant training.

Table 2 shows the descriptive statistics obtained in the survey for all four defined categories – the Risky Behaviour Scale (RBS), Conservative Behaviour Scale (CBS), Exposure to Offence Scale (EOS) and the Risk Perception Scale (RPS). A score of 1 is considered the lowest and a score of 5 the highest value for each question.

Table 2. Descriptive statistics according to scales

Scale	No of questions	Average	Mean	Std. deviation	Std. error	Min	Max	Range
RBS	20	2,612	52,250	6,882	0,413	35	72	37
CBS	10	4,051	40,513	5,082	0,305	24	50	26
EOS	7	1,389	9,722	2,037	0,122	7	22	15
RPS	17	3,498	59,462	8,448	0,507	17	85	68

4.2. Research Question 1: Is there a significant difference between the surveyed groups (females and males) concerning their average score according to different behavioural scales (RBS, CBS, EOS, RPS)?

To answer the first research question, it is necessary to get an initial overview of the sample results and check whether the ANOVA assumption is met. For this purpose, the following are included: 1) descriptives – descriptive statistics about the results of the sample, 2) homogeneity of variance test – a test to check the equality of variances of the general set. Table 3 and Table 4 characterise the descriptive statistics and homogeneity of variances results according to the respondent's gender. According to Table 3, it can be seen that in the sample the average level (mean) of the comparable groups (men and women) is very similar according to the scales, while the dispersion of knowledge (standard deviation) is somewhat different. Can the resulting difference be generalised to the general population?

Table 3. Descriptive statistics of scales according to gender

	Gender	N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
RBS	Female	157	52,00	6,21	,50	35,00	67,00
	Male	120	52,58	7,69	,70	35,00	72,00
	Total	277	52,25	6,88	,41	35,00	72,00
CBS	Female	157	40,20	4,93	,39	24,00	50,00
	Male	120	40,93	5,27	,48	25,00	50,00
	Total	277	40,51	5,08	,31	24,00	50,00
EOS	Female	157	9,53	1,92	,15	7,00	16,00
	Male	120	9,98	2,16	,20	7,00	22,00
	Total	277	9,72	2,04	,12	7,00	22,00
RPS	Female	157	59,50	8,87	,71	24,00	85,00
	Male	120	59,42	7,90	,72	17,00	77,00
	Total	277	59,46	8,45	,51	17,00	85,00

Note: N (sample size), Mean (sample mean value), Std.Deviation (sample standard deviation)

Table 4. ANOVA tables according to gender

		Sum of Squares	df	Mean Square	F	Sig.
RBS	Between Groups	22,49	1	22,49	,47	,492
	Within Groups	13051,32	275	47,46		
	Total	13073,81	276			
CBS	Between Groups	36,00	1	36,00	1,40	,238
	Within Groups	7093,20	275	25,79		
	Total	7129,21	276			
EOS	Between Groups	13,55	1	13,55	3,29	,071
	Within Groups	1132,05	275	4,12		
	Total	1145,60	276			
RPS	Between Groups	,44	1	,44	,01	,938
	Within Groups	19698,42	275	71,63		
	Total	19698,85	276			

Sig. = significance probability p; Levene Statistic = a statistic that expresses the magnitude of the difference

According to Table 4, since p > α for all four scales, it is proven that p > α and H₀ remains valid. Therefore, it can be said with confidence that there are no significant differences between men and women across the scales.

4.3. Research Question 2: Is there a significant difference between the surveyed groups (students, academics and administrative staff) concerning their average score according to different behavioural scales (RBS, CBS, EOS, RPS)?

To answer the second research question and perform a comparative analysis with other similar studies, the data of the respondents' position was grouped into three groups (students, academics, administrative staff) and four scales instead of the six surveyed groups (i.e. vocational students, undergraduate students, graduate students, lecturers, administrative staff and others).

Table 5. Descriptive statistics of scales according to the position

	Position	N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
RBS	Student	146	53,60	6,95	,57	35,00	70,00
	Academic	89	49,81	6,75	,72	35,00	72,00
	Administrative	42	52,71	5,53	,85	40,00	64,00
	Total	277	52,25	6,88	,41	35,00	72,00
CBS	Student	146	38,81	5,22	,43	24,00	50,00
	Academic	89	42,83	4,27	,45	30,00	50,00
	Administrative	42	41,52	3,91	,60	31,00	47,00
	Total	277	40,51	5,08	,31	24,00	50,00
EOS	Student	146	9,77	2,14	,18	7,00	22,00
	Academic	89	9,47	1,88	,20	7,00	16,00
	Administrative	42	10,07	1,98	,31	7,00	16,00
	Total	277	9,72	2,04	,12	7,00	22,00
RPS	Student	146	59,50	8,67	,72	24,00	85,00
	Academic	89	58,90	9,00	,95	17,00	81,00
	Administrative	42	60,52	6,26	,97	46,00	75,00
	Total	277	59,46	8,45	,51	17,00	85,00

Table 5 presents the descriptive statistics of the various member groups of the Academy. The sample shows that the average level (mean) of the comparable groups (students, academics, administrative staff) is very different, especially in the RBS and CBS scales. Can the resulting difference be generalised to the general population?

Table 6. ANOVA tables according to the position

		Sum of Squares	df	Mean Square	F	Sig.
RBS	Between Groups	806,53	2	403,26	9,01	,000
	Within Groups	12267,28	274	44,77		
	Total	13073,81	276			
CBS	Between Groups	945,63	2	472,81	20,95	,000
	Within Groups	6183,58	274	22,57		
	Total	7129,21	276			
EOS	Between Groups	11,09	2	5,54	1,34	,264
	Within Groups	1134,51	274	4,14		
	Total	1145,60	276			
RPS	Between Groups	75,79	2	37,89	,53	,590
	Within Groups	19623,07	274	71,62		
	Total	19698,85	276			

As shown in Table 6, there are at least two groups representing populations with different levels (RBS and CBS). Post-hoc tests must then be used to see which groups' mean values are significantly different. The Tukey test compares group means pairwise. First, the RBS average means of the two groups are compared. In Table 6, the most important behaviour scales are RBS and CBS, the other two scales are given only for the sake of clarity.

1. RBS (students and academics)
Ho: μStudents = μAcademic
H_1: μStudents < μAcademic
α = 0.05
Conclusion: According to the probability of significance (0.000, e.g. 0%), H_1 has been proved. The mean scores of these groups considering the RBS are statistically significantly different.

2. RBS (students and administrative staff)
Ho: μStudents = μAdministrative
H_1: μStudents < μAdministrative
α = 0.05
Conclusion: According to the probability of significance (0.729, e.g. 72.9%), Ho remains true. The mean scores of these groups considering the RBS generally do not differ.

3. RBS (academics and administrative staff).
Ho: μAcademic = μAdministrative
H1: μAcademic < μAdministrative
α = 0.05
Conclusion: According to the probability of significance (0.090, e.g. 9%), Ho remains true. The mean scores of these groups considering the RBS of these groups generally do not differ.

Table 7. The Tukey test according to position

Position			Mean Difference (I - J)	Std. Error	Sig.
Tukey HSD (RBS)	Student	Academic	3,79	,90	,000
		Administr.	,89	1,17	,729
	Academic	Student	-3,79	,90	,000
		Administr.	-2,91	1,25	,055
	Administrative	Student	-,89	1,17	,729
		Academic	2,91	1,25	,055
Tukey HSD (CBS)	Student	Academic	-4,02	,64	,000
		Administr.	-2,72	,83	,004
	Academic	Student	4,02	,64	,000
		Administr.	1,31	,89	,307
	Administrative	Student	2,72	,83	,004
		Academic	-1,31	,89	,307
Tukey HSD (EOS)	Student	Academic	,30	,27	,513
		Administr.	-,30	,36	,682
	Academic	Student	-,30	,27	,513
		Administr.	-,60	,38	,259
	Administrative	Student	,30	,36	,682
		Academic	,60	,38	,259
Tukey HSD (RPS)	Student	Academic	,60	1,14	,858
		Administr.	-1,02	1,48	,769
	Academic	Student	-,60	1,14	,858
		Administr.	-1,62	1,58	,561
	Administrative	Student	1,02	1,48	,769
		Academic	1,62	1,58	,561

Similar conclusions can be drawn for other behaviour scales. The mean scores of students and academics as well as students and administrative staff groups considering the CBS are statistically significantly different but between academics and administrative staff, the CBS of these groups generally does not differ. Thus, it can be concluded that the cause of the differences in the RBS and the CBS is the student group. These results coincide with the Öğütçü et al. (2016) study, which also revealed that it was the students who created a significant difference between the surveyed groups, while the Benavides-Astudillo et al. (2022) study determined that the academic group had significant differences with the administrative staff and student groups.

4.4. Does the duration of time spent on the Internet affect the average of the scales (RBS, CBS, EOS, RPS)?

To answer the third research question, respondents were grouped into three ranges according to how much time per day they use the Internet (i.e. 1 to 5 hours/day, 6 to 10 hours/day, and 11 or more hours/day). Table 8 presents descriptive statistics for various Internet users. It can be seen that there are no significant differences between the averages of the CBS and RPS scales, but for the EOS and RBS scales, the group that uses the Internet 11 or more hours/day clearly stands out in terms of the average indicator. The same difference can be noticed in the case of this group in the dispersion (standard deviation). Can the resulting difference be generalised to the general population?

Table 8. Descriptive statistics according to the Internet use time per day

Internet use per day		N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
RBS	1-5 hours/day	145	51,09	6,68	,55	35,00	68,00
	6-10 hours/day		53,07	6,60	,60	38,00	69,00
	11+ hours/day	9	59,67	8,35	2,78	44,00	72,00
	Total	277	52,25	6,47	,41	35,00	72,00
CBS	1-5 hours/day	145	40,19	5,25	,44	24,00	50,00
	6-10 hours/day	123	40,88	4,92	,44	25,00	49,00
	11+ hours/day	9	40,67	4,74	1,58	35,00	46,00
	Total	277	40,51	5,08	,31	24,00	50,00
EOS	1-5 hours/day	145	9,77	2,09	,17	7,00	22,00
	6-10 hours/day	123	9,56	1,91	,17	7,00	16,00
	11+ hours/day	9	11,11	2,47	,82	7,00	14,00
	Total	277	9,72	2,04	,12	7,00	22,00
RPS	1-5 hours/day	145	59,08	8,81	,73	17,00	85,00
	6-10 hours/day	123	60,00	8,23	,74	24,00	79,00
	11+ hours/day	9	58,22	5,02	1,67	50,00	68,00
	Total	277	59,46	8,45	,51	17,00	85,00

Table 9 shows the ANOVA analysis, which shows that there is a significant difference between the other scales and the RBS scale with a value of p = 0.000. Therefore, there is a significant difference between the Internet usage times of the participants. This confirms that at least two groups have a statistically significant difference in their level of RPS.

Table 9. ANOVA tables according to the Internet use time per day

		Sum of Squares	df	Mean Square	F	Sig.
RBS	Between Groups	773,64	2	386,82	8,62	,000
	Within Groups	12300,18	274	44,89		
	Total	13073,81	276			
CBS	Between Groups	31,44	2	15,72	,61	,546
	Within Groups	7097,76	274	25,90		
	Total	7129,21	276			
EOS	Between Groups	20,92	2	10,46	2,55	,080
	Within Groups	1124,67	274	4,10		
	Total	1145,60	276			
RPS	Between Groups	70,29	2	35,14	,49	,613
	Within Groups	19628,56	274	71,64		
	Total	19698,85	276			

Note: SS = Sum of Squares; DF = Degree of Freedom; MSe = Mean of squares; F = Error; SIG = Significant at .05 level of significance

To find out in which group this difference occurs, the Tukey test was applied between groups Internet use time per day, to see which groups' mean values in RBS are significantly different if the group means are compared pairwise (see Table 13). First, the RBS average means of the two groups are compared.

1. RBS (1–5 hours/day and 6–10 hours/day).
Ho: μ1–5 hours/day = μ6–10 hours/day
H_1: μ1–5 hours/day < μ6–10 hours/day
α = 0.05
Conclusion: According to the probability of significance (0.043, e.g. 4,3%), H_1 has been proved. The mean scores of these groups considering the RBS are statistically significantly different.

2. RBS (1–5 hours/day and 11 and more hours/day)
Ho: μ1–5 hours/day = μ11 and more hours/day
H_1: μ1–5 hours/day < μ11 and more hours/day
α = 0.05
Conclusion: According to the probability of significance (0.001, e.g. 0.1%), H_1 has been proved. The mean scores of these groups considering the RBS are statistically significantly different.

3. RBS (6–10 hours/day and 11 and more hours/day)
Ho: μ6–10 hours/day = μ11 and more hours/day
H_1: μ6–10 hours/day < μ11 and more hours/day
α = 0.05

Conclusion: According to the probability of significance (0.013, e.g. 1.3%), H_1 has been proved. The mean scores of these groups considering the RBS are statistically significantly different.

Table. 10. Results obtained in the Tukey test between the groups (1–5 hours/day, 6–10 hours/day and 11 or more hours/day) and the scale (RBS)

Position			Mean Difference (I - J)	Std. Error	Sig.
Tukey HSD (RBS)	1-5 hours/day	6 -10 hours/ day	-1,98	,82	,043
		11+ hours/ day	-8,58	2,30	,001
	6 -10 hours/ day	1-5 hours/day	1,98	,82	,043
		11+ hours/ day	-6,59	2,31	,013
	11+ hours/ day	1-5 hours/day	8,58	2,30	,001
		6 -10 hours/day	6,59	2,31	,013

Thus, it can be concluded that the cause of the differences in the RBS are the groups who spend 11 or more hours/day and 6–10 hours/day on the Internet. They are the weakest link. The results obtained by applying the post-hoc Tukey test on the RBS scale show that the respondents of the groups who spend 6–10 hours/day and 11 or more hours/day on the Internet are more tolerant of risky situations than the groups who spend less time on the Internet (see Table 9). In risky situations, they are more at risk than groups that spend less time on the Internet. These results coincide with Öğütçü et al. (2016) and Benavides-Astudillo et al. (2022) findings. In both studies, it was found that the significant difference occurs precisely on the RBS scale.

4.5. Does the cyber security training attendance or non-attendance affect the average of the scales (RBS, CBS, EOS, RPS)?

Finally, to answer the fourth research question, a test of whether participation in security training affects the mean scale scores was conducted. Respondents were surveyed in two groups, divided into those who had participated in cyber security training and those who had not.

Table 11 presents the descriptive statistics for those who answered that they have completed the training and for those who answered that they have not participated in the training. The smallest difference is the average level (mean) for the EOS scale, but there is a difference in the dispersion between those who have passed and those who have not, in every scale (standard deviation).

Table 11. Descriptive statistics according to the cyber security training

Passed Cyber Security Training		**N**	**Mean**	**Std. Deviation**	**Std. Error**	**Minimum**	**Maximum**
RBS	Yes	165	51,19	6,34	,49	35,00	72,00
	No	112	53,81	7,37	,70	35,00	70,00
Total		277	52,25	6,88	,41	35,00	72,00
CBS	Yes	165	42,00	4,71	,37	25,00	50,00
	No	112	38,32	4,83	,46	24,00	49,00
Total		277	40,51	5,08	,31	24,00	50,00
EOS	Yes	165	9,65	1,76	,14	7,00	16,00
	No	112	9,83	2,40	,23	7,00	22,00
Total		277	9,72	2,04	,12	7,00	22,00
RPS	Yes	165	61,58	7,32	,57	34,00	85,00
	No	112	56,34	9,04	,85	17,00	79,00
Total		277	59,46	8,45	,51	17,00	85,00

The fourth research question was also tested at a significance level of α = 0.05. Table 11 shows the ANOVA analysis, which indicates that there is a significant difference between the EOS with a significance value of $p > 0.05$ and other scales with a value of $p = 0.000$. Therefore, there is a significant difference between the participants who have completed cyber security training and those who have not in their level of RBS, CBS and RPS.

Table 12. ANOVA tables according to the passed cyber security training

		Sum of Squares	df	Mean Square	F	Sig.
RBS	Between Groups	459,57	1	459,57	10,02	,002
	Within Groups	12614,24	275	45,87		
	Total	13073,81	276			
CBS	Between Groups	902,78	1	902,78	39,87	,000
	Within Groups	6226,43	275	22,64		
	Total	7129,21	276			
EOS	Between Groups	2,21	1	2,21	,53	,467
	Within Groups	1143,39	275	4,16		
	Total	1145,60	276			
RPS	Between Groups	1833,60	1	1833,60	28,22	,000
	Within Groups	17865,25	275	64,96		
	Total	19698,85	276			

Since there are only two comparable characteristics, it was decided to use the MPAR (nonparametric) Kruskal-Wallis test instead of Tukey's test. It shows more clearly in which scales the largest and smallest mean differences occur between respondents who had undergone cyber security training and those who had not. According to Table 13, the mean rank of the respondents who passed cyber security training and who have not passed this training, has large differences in the CBS, 163.09 for respondents who have completed the training and 103.50 for respondents who have not. The other significant difference is in the RPS, where respondents who passed cyber security training scored 158.83 as a mean rank, and non-passed respondents gained 109.78 as a mean rank. However, these two scales are united by the fact that the average of respondents who have undergone cyber training is significantly higher than the average of those who have not completed the training. On the other hand, the results are opposite in the case of the RBS scale – those who passed the cyber security training gained 126.37 as a mean rank while those who have not completed the training gained 157.60 as a mean rank. Exposure Offence is not at all affected by completed cyber security training.

Table 13. Ranks according to the completion of cyber security training

		N	Mean Rank
RBS	Yes	165	126,37
	No	112	157,60
	Total	277	
CBS	Yes	165	163,09
	No	112	103,50
	Total	277	
EOS	Yes	165	139,08
	No	112	138,89
	Total	277	
RPS	Yes	165	158,83
	No	112	109,78
	Total	277	

According to the probability of ANOVA significance (Sig. = 0.467, e.g. 46.7%), Ho remains true for EOS. The mean rank of those who have completed cyber security training considering the Exposure Offence of these groups generally do not differ. This can be explained by the fact that the people working and studying at the Academy are certainly more careful about cyber offences than average – future specialists are prepared here for rescue, finance, justice, and police and border guard.

Finally, a selection of the respondents' own opinions about the survey and the topics covered in the questionnaire is presented:

- *In principle, everything can be dangerous, but some environments need to be used. It would be safest to live offline.*

- *My internet usage outside of work: mobile data and home wifi. As far as I know, different communication environments have different levels of security. I never share sensitive information on FB Messenger, but I have done so on Signal. Online shopping and entering data - if I do it in the safest places I know, I don't consider it a problem, but I would never go shopping in a less-known Estonian store or in a foreign online environment.*
- *If you understand where to press and what to share, there are no problems. The more you participate in Facebook sharing games, the more problems you have.*
- *Everything depends on the nature of the activity, the information used, previous awareness, etc.*
- *A common peasant mind must be maintained in the Internet environment as well as in a normal environment.*
- *Several of the aforementioned activities can be dangerous, but it is necessary to consider the justification and check the existence of security solutions (e.g. in the case of Internet banking), whether there is secure authentication and the correct website, before opening emails with advertising content, the authenticity of the sender and to be sure that there is any interest in such emails against letters, etc. On the other hand, the use of public Wi-Fi should be avoided in any case and rather use mobile data communication, which is now quite affordable, than looking for it from various service providers in Estonia. Checking the identity card number by security personnel when entering the building – again, the possibility of benefit and harm should be assessed – e.g. if the fire escape is secured by a contract security company, then it may be absolutely necessary, but it should be avoided in the case of arbitrary fire escapes. In the case of file sharing and chat programs and AI, it is simply necessary to avoid entering sensitive texts, in which case the benefits of sharing information outweigh the possible dangers.*
- *Even paid movies/music/software, etc. may contain malware. In addition, opening the email itself should not be dangerous, opening and saving or viewing a file/link, etc. attachment contained in the email is.*
- *In my opinion, all actions on the Internet are already very dangerous – and at the same time, you cannot stop using the Internet. Everything already goes through the Internet or services, etc. always need a permanent connection.*

5. Conclusions

In this study, students, lecturers (researchers) and employees of the Academy were investigated in terms of hybrid threats and cyber security-related risk prevention options such as risky behaviour, conservative behaviour, exposure to offence and risk perception. The present study is part of a larger study conducted within the framework of the hybrid threat cooperation programme (HYBRIDC).

Four research questions were raised and all of them were answered during the analysis of the results. Previous research highlights the role of gender in shaping cybersecurity attitudes and behaviours shows that males tend to have better awareness of online safety. In current study it can be seen that there is no significant difference in the cyber behaviour scales of women and men in the RBS, CBS score and RPS score types, although it must be emphasized that the mean of men (mean severe) is slightly higher than that of women. However, there is a significant difference in the EOS scale. According to the results, the more the respondents perceive threats, the more defensive their behaviour becomes. Therefore, based on the results, it can be said that men's exposure to danger has been higher and they are accordingly more careful.

Research shows that while students often have awareness of cyber threats, this is not always reflected in safe behavior. The current study suggests that there are no significant differences in EOS and RPS score types among faculty, administrative staff and students. However, it can be seen that the proportion of students using risky information technologies is higher than in other groups. For example, students' exposure to threat scale scores are higher than other groups. Conservative behaviour is also not as well developed for students as it is for academics and administrative staff. All this shows that students are more vulnerable to risks.

To explore the relationship between daily internet usage and cybersecurity behaviours, respondents were grouped into three categories based on usage time: 1–5 hours, 6–10 hours, and 11 or more hours per day. The EOS and RBS scales showed noticeable differences, particularly among those using the internet for 11 or more hours daily, who scored higher and showed greater variability. ANOVA analysis confirmed a statistically significant difference in the RBS scale across groups, indicating that internet usage time influences risky online behaviour. The reseults showed that the most significant differences occur between those using the internet 6–10 hours and 11+ hours per day - they are more tolerant of risky behaviour compared to those with lower usage. We can cinclude that as the use of technology increases during the day, people are exposed to more risks, but at the same time, their perception of danger and conservative behaviour increase.

Trained and knowledgeable employees reduce the likelihood of accidental and unintentional actions that could lead to violations of cyber security policies, and play a key role in minimizing information security risks and safeguarding the organization's critical assets and sensitive personal data. Only 60% of respondents have completed cyber security training, which is obviously too few considering that the Academy is the most important school in Estonia in the field of internal security. This fact in itself shows that more training needs to be done. While there is no significant difference between the EOS scores of the group that received security training and the group that did not receive such training, the CBS score and RPS score of the first group are significantly higher than the score of the second group. This result clearly shows that such training increases people's awareness.

The open answers of the respondents expressed a range of cautious and pragmatic views about internet use and digital security. Many believe that while everything online can potentially be dangerous, risk can be managed by being informed and selective about platforms and activities. They stressed the importance of common sense, secure environments, and avoiding untrusted sources—particularly when shopping or using public Wi-Fi. Several noted that secure authentication and awareness of sender authenticity are key when handling emails. Others highlighted the risks of sharing sensitive data via chat apps or participating in social media games. Despite widespread concerns, respondents acknowledged that avoiding the internet altogether is unrealistic given its necessity in modern life.

Future studies could build upon this recent study by expanding the sample size and replicating the research across diverse organizational and educational settings. This would enhance the generalizability of the findings and help validate the applicability of the research model in different contexts. Moreover, incorporating new respondent groups would allow for a broader understanding of cybersecurity behaviours across various demographics. Such replication efforts could yield valuable insights for the development of targeted information security training programmes and policies, enabling organizations to implement more effective, context-specific cybersecurity measures.

References

Anwar, M., He, W., Ash, I., Yuan, X., Li, L., Xu, L. (2017). Gender difference and employees' cybersecurity behaviors. *Computers in Human Behavior*, 69, 437–443. DOI: 10.1016/j.chb.2016.12.040.

Bederna, Z., Szadeczky, T. (2020). Cyber espionage through Botnets. *Security Journal*, 33, 43–62, DOI: 10.1057/s41284-019-00194-6.

Benavides-Astudillo, E., Silva-Ordoñez, L., Rocohano-Rámos, R., Fuertes, W., Fernández-Peña, F., Sanchez-Gordon, S., Bastidas-Chalan, R. (2022). Analysis of Vulnerabilities Associated with Social Engineering Attacks Based on User Behavior, in: *International Conference on Applied Technologies, Springer International Publishing*, Cham, pp. 351-364, DOI: 10.1007/978-3-031-03884-6 26.

Busvine, D., Kaeckenhoff, T. (2020). Prosecutors open homicide case after hacker attack on German hospital, *Reuters*, September 18, https://www.reuters.com/article/us-germany-cyber-idUSKBN26926X

Candiwan, C., Azmi, M., Alamsyah, A. (2022). Analysis of Behavioral and Information Security Awareness among Users of Zoom Application in COVID-19 Era. *International Journal of Safety and Security Engineering*, 12(2), 229-237. DOI: 10.18280/ijsse.120212.

Ceran, O., Karataş, S. (2021). Individual differences on conservative and risky behaviors about information security. *Bilişim Teknolojileri Dergisi*, 14(2), 161-170, https://dergipark.org.tr/en/download/article-file/990584

Concepcion, J. D., Palaoag, T. D. (2024). An Assessment of Cybersecurity Awareness among Academic Employees at Quirino State University: Promoting Cyber Hygiene. *Journal of Electrical Systems*, 20(7s), 769-775. 390. DOI: 10.52783/jes.3445.

Duman, F. K. (2022). Determining Cyber Security-Related Behaviors of Internet Users: Example of the Faculty of Sport Sciences Students. *European Journal of Education*, 5(1), 112-128. DOI: 10.26417/723gru15.

Einmann, A. (2020). Iranian intelligence attempted to access University of Tartu email accounts (Estonian). *Postimees*, April 14, https://www.postimees.ee/6949265/iraani-luure-uritas-ligipaasu-tartu-ulikooli-e-posti-kontodele.

Hadlington, L. (2018). The “human factor” in cybersecurity: Exploring the accidental insider. In Psychological and behavioral examinations in cyber security (pp. 46-63). IGI Global. DOI: 10.4018/978-1-5225-4053-3.ch003.

Hirsjärvi, S., Remes, P., Sajavaara, P. (2020). *Research and Write* (Estonian). Medicina, Tartu, 2010.

Hubbard, D. W. (2020). *The failure of risk management: Why it's broken and how to fix it,* John Wiley & Sons.

Ifinedo, P. (2014). Information systems security policy compliance: An empirical study of the effects of socialisation, influence, and cognition. *Information and Management*, 51(1), 69–79. DOI: 10.1016/j.im.2013.10.001.

Iiskola, J. (2019). *Measuring Cybersecurity* (Finnish), Haaga-Helio Amattikorkeakoulu OY. https://urn.fi/URN:NBN:fi:amk-201905159951

Jalali, M.S., Bruckes, M., Westmattelmann, D., Schewe, G. (2020). Why Employees (Still) Click on Phishing Links: Investigation in Hospitals. *Journal of Medical Internet Research*, 22(1):e16775. DOI: 10.2196/16775.

Jeske, D., van Schaik, P. (2017). Familiarity with Internet threats: Beyond awareness. *Computers & Security*, 66, 129-41. DOI: 10.1016/j.cose.2017.01.010.

Juvonen, M., Koskensyrjä, M., Kuhanen, L., Ojala, Penttinen, A., Porvari, P., Talala, T. (2014). *Corporate Risk Management* (Finnish). Finanssi- ja vakuutuskustannus Oy.

Kansallinen riskiarvio, Sisäinen turvallisuus (Finnish). (2024). *Sisäministeriön julkaisuja*, 4, 2023. http://urn.fi/URN:ISBN:978-952-324-602-7.

Karu, L. (2020). The cyber attack that hit the health care college disrupted the servers (Estonian). *Tartu Postimees*, September 15, https://tartu.postimees.ee/7062330/tervishoiukorgkooli-tabanud-kuberrunnak-loi-serverid-sassi.

Kont, K.-R. (2024). Cybersecurity behaviours of the employees and students at the Estonian Academy of Security Sciences. *Organizational Cybersecurity Journal: Practice, Process and People*, 4(2), 85-104. https://doi.org/10.1108/OCJ-02-2024-0001.

Kont, K.-R. (2023). Presentation at the Eighth International Conference on Cyber-Technologies and Cyber-Systems.
https://www.iaria.org/conferences2023/filesCYBER23/CYBER_80056.pdf

Kuusela, H., Ollikainen, R. (2005). Risks and Risk Management Thinking, (Finnish) in: Risks and Risk Management, Tampere University Press, Tampere. https://trepo.tuni.fi/bitstream/handle/10024/65418/riskit_ja_riskienhallinta_2005.pdf?sequence=1

Lane, D. M. (2012). *Tukey's Honestly Significant Difference (HSD)*, in: Neil J. Salkind, Encyclopedia of Research Design. DOI: 10.4135/9781412961288.

Limnéll, J., Majewski, K., Salminen, M. (2014). *Cybersecurity*, Docendo.

McLeod, S. (2023). P-Value And Statistical Significance: What It Is & Why It Matters. *Simply Psychology*. https://www.simplypsychology.org/p-value.html.

Mian, T. S., Alatawi, E. M. (2023). Exploring Factors to Improve Intentions to Adopt Cybersecurity: A Study of Saudi Banking Sector, Humanities & Natural Sciences Journal 4(9), 101–114. DOI: 10.53796/hnsj498.

Noran, S. F. (2021). Securing higher education against cyberthreats: from an institutional risk to a national policy challenge. *Journal of Cyber Policy*, 6(2). 137-154, DOI: 10.1080/23738871.2021.1973526

Ög˘ütçü, G., Testik, Ö. M., Chouseinoglou, O. (2016). Analysis of personal information security behavior and awareness. *Computers & Security*, 56, 83–93.
DOI: 10.1016/j.cose.2015.10.002.

Oxford Dictionary (2019). https://en.oxforddictionaries.com/definition/cyberthreat

Pollini, A., Callari, T. C., Tedeschi, A., Ruscio, D., Save, L., Chiarugi, F., Guerri, D. (2022). Leveraging human factors in cybersecurity: an integrated methodological approach. *Cognition, Technology & Work*, 24(2), 371-390.

Qashqari, A., Munshi, A., Alturkstani, H., Ghwati, H., Alhebshi, D. (2020). *The Human Factors and Cybersecurity Policy* [Ebook]. Hämtad från
http://paper.ijcsns.org/07_book/202004/20200401.pdf

Roman, J. (2015). Universities: prime breach targets.
https://www.databreachtoday.asia/universities-prime-breach-targets-a-7865

Seppänen, T. (2022). Changing security environment and information security in higher education institutions (Finnish) . *Current information security review*, 12. https://blogs.helsinki.fi/thinkopen/tietoturvakatsaus-2022-12/

Shad, M. R. (2019). Cyber threat landscape and readiness challenge of Pakistan. *Strategic Studies*, 39 (1), 1–19, DOI: 10.53532/ss.039.01.00115.

Sonawala, J. (2024). Exploring Statistical Analysis with the Kruskal-Wallis Test. https://www.linkedin.com/pulse/exploring-statistical-analysis-kruskal-wallis-test-jvalin-sonawala-i4occ/.

Stone, E. R. (2012). *t Test, Paired Samples*. In: Neil J. Salkind, Encyclopedia of Research Design, 2012, DOI: 10.4135/9781412961288.

Svitek, P., Anderson, N. (2014). University of Maryland computer security breach exposes 300,000 records, *The Washington Post*, February 19, https://www.washingtonpost.com/local/college-park-shady-grove-campuses-affected-by-university-of-maryland-security-breach/2014/02/19/ce438108-99bd-11e3-80ac63a8ba7f7942_story.html.

Tooding, L-M. (2014). *Social analysis methods and methodology training database* (Estonian). *Analysis of Variance*, 2014, https://samm.ut.ee/dispersioonanalyys.

Triplett, W. J. (2023). Addressing cybersecurity challenges in education. International *Journal of STEM Education for Sustainability*, 3(1), 47-67. DOI: 10.52889/ijses.v3i1.132.

Verkijika, S. (2019). “If you know what to do, will you take action to avoid mobile phishing attacks”: Self-efficacy, anticipated regret, and gender. *Computers In Human Behavior*, 101, 286-296. DOI: 10.1016/j.chb.2019.07.034.

Widup, S., Maxwell, K., Baker, W., Porter, C., Jacobs, J., Thompson, K., Spitler, M., Hylender, D., Brannon, S., Gilbert, K. (2013). 2013 verizon data breach investigations report, Technical report, Verizon. DOI: 10.13140/RG.2.1.4729.8647.

Widup, S., Spitler, M., Hylender, D., Bassett, G. (2018). 2018 verizon data breach investigations report, Technical report, Verizon. https://www.researchgate.net/publication/324455350_2018_Verizon_Data_Breach_Investigayions_Report.

Yerby, J., Floyd, K. (2018). Faculty and staff information security awareness and behaviors. *Journal of The Colloquium for Information Systems Security Education*, 6(1). 23-23, https://cisse.info/journal/index.php/cisse/article/view/90/CISSE_v06_i01_p05.pdf

Yiğit, M. F., Seferoğlu, S. S. (2019). Investigating students’ cyber security behavior in relation to big five personality traits and other various variables. *Mersin University Journal of the Faculty of Education*, 15(1), 186–215. DOI: 10.17860/mersinefd.437610.

Zimmermann, V., Renaud, K. (2021). The Nudge Puzzle: Matching Nudge Interventions to Cybersecurity Decisions, *ACM Transactions on Computer-Human Interaction* 28(1), 1–45. DOI: 10.1145/3429888.

Received February 22, 2025, revised May 3, 2025, accepted June 15, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 2, pp. 528–550
https://doi.org/10.22364/bjmc.2025.13.2.12

LLMs and XAI for Breast Cancer Transparency: A Review

Sobia DASTGEER, Povilas TREIGYS

Vilnius University, Institute of Data Science and Digital Technologies, Akademijos str. 4,
Vilnius, LT-08412, Lithuania

Sobia.dastgeer@mif.stud.vu.lt, Povilas.treigys@mif.vu.lt

ORCID 0009-0005-8016-7197, ORCID 0000-0002-6608-5508

Abstract. The rising global mortality rate of women due to breast cancer highlights the urgent need for advancements in its diagnosis and early detection. Early identification of breast cancer significantly improves patient prognosis and survival outcomes. Artificial intelligence (AI), particularly Deep Learning (DL) and Large Language Models (LLMs), shows transformative potential in enhancing the diagnostic and prognostic capabilities in breast cancer detection. However, their clinical adoption remains challenged due to their ”black-box” nature. Intelligent systems in healthcare, understanding the reasoning behind AI decisions is as critical as ensuring their performance, accuracy as well as patient safety and trust. Explainable AI (XAI) addresses these challenge by making AI reasoning transparent, allowing clinicians to interpret, validate, and trust model outputs. This paper reviews the application of XAI methods like SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), and Gradient-weighted Class Activation Mapping (Grad-CAM) in improving the transparency of DL models for breast cancer detection. This paper explores advanced XAI strategies that balance accuracy with interpretability, including attention-based mechanisms and LLM-driven explanations. In particular we discuss LLMs embedded within XAI systems, act as translational interfaces, decoding complex model outputs into clinician-friendly explanations. By adapting technical explanations to the end user's context and needs, LLMs enhance the accessibility and interpretability of complex model explanations. Collectively, these approaches help to bridge the gap between AI behavior and human understanding, ultimately improving transparency, trust and decision support especially in healthcare domain.

Keywords: Artificial Intelligence, Deep Learning, Breast Cancer, Healthcare, Explainable AI, Large Language Models

1 Introduction

Opaque decision-making systems have increased dramatically in the last fifteen years. Machine learning (ML) and deep learning (DL) models are used in a wide range of methods in this rapidly developing field. The majority of these models are referred to as ”Black-Box” due to their inherent complexity and lack of explanations of the decision-making process (Sabol et al., 2019). These ”black-box” systems use advanced machine-learning algorithms to evaluate and forecast individual data, frequently containing private or sensitive data. One of the main obstacles to their adoption in mission-critical application domains, including banking, e-commerce, healthcare, public services, and safety, is their interpretability (Malhi and Främling, 2023). The European Union’s General Data Protection Regulation (GDPR), effective in 2018, places strict limitations on automated decision-making systems that significantly affect users while mandating a right to explanation for affected individuals (Goodman and Flaxman, 2017). This regulation highlights the urgency for industries to adopt nondiscriminatory machine learning practices and the critical role of computer scientists in developing interpretable algorithmic frameworks that align with compliance and ethical standards.

The high failure rates of digital innovation adoption in the healthcare industry are noticeable (Guidotti et al., 2018). Artificial intelligence (AI) systems can analyze medical images, such as ”Computed Tomography” (CT), ”Magnetic Resonance Imaging” (MRI), ”Ultrasound”, ”X-ray”, and ”Infrared Scans” to identify specific anatomical structures and identify anomalies (Raghavan, Balasubramanian and Veezhinathan, 2024). As a result, the widespread use of AI has caused people to questions: ”How comfortable are we blindly trusting these AI-generated detection and results and When anything goes wrong, who is going to be responsible?”. Notably, the highly effective predictions of AI models come from Deep Neural Networks (DNNs), built from incredibly complicated non-linear statistical models with countless parameters. However, the complexity of DNNs which consist of numerous non-linear layers and millions of parameters often compromises the transparency and interpretability of these models.

The most notable example of AI application in healthcare is cancer prediction (Bray et al., 2018). According to World Health Organization(WHO) Breast Cancer (BC) is the most prevalent disease worldwide, with over 2.3 million new cases annually. The most significant risk factor for breast cancer is being a woman. Women are affected by breast cancer in about 99% of cases, whereas men are affected in 0.5–1% of cases. Women with breast cancer who live in high-income nations have a 60% higher chance of surviving than those who live in low- and middle-income countries (WHO, 2024). Furthermore, 70% of breast cancer deaths occur in resource-limited environments because of challenges in early diagnosis and treatment. A study by (McKinney et al., 2020) shows that using AI may significantly enhance breast cancer diagnosis statistically, but it doesn’t thoroughly examine how these developments fit into routine clinical procedures. The study ignores common problems, such as describing how the system works, making sure it is easy to use, and comprehending how it fits into collaborative practices that allow for a smooth transition into standard clinical work.

Another vital factor regarding interpretability is knowing why a system, service, or method needs to be interpretable. In some situations, explanations may not be required if no critical outcomes depend on the prediction’s outcome. For instance, if the objective is to determine whether an image contains a tomato, and this information has no significant consequences, in this situation, an interpretable model may not be required, and a black-box approach might be sufficient. Consequently, explaining and interpreting the model's outcome and functionality are essential to enhance the applicability of these systems across diverse clinical applications. Explainable AI (XAI) aims to provide researchers with a wide range of tools to understand the opaque nature of black-box AI systems, with a focus on transparency and the interpretability of AI models utilized to make decisions (Croce et al., 2024).

This article focuses on the current state of research, contributions made in this area of XAI and LLMs, and an investigation into what is still to be discovered. Our ultimate goal is to give a comprehensive taxonomy in the field of XAI, which helps those who are new in this field that they can use it as a guide to advance future research while also motivating professionals and experts from other fields to embrace the advantages of AI in their respective fields, free from assumptions about its interpretability. This study, explored the potential of integrating LLMs into XAI pipelines to enhance the interpretability of breast cancer prediction models. Initially, an image-based machine learning model is trained using a breast cancer dataset to predict diagnostic outcomes. Post-hoc explanation techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM) or SHapley Additive exPlanations (SHAP), are then employed to generate visual or textual insights into the model's decision making process. However, these explanations often lack clarity and are not easily understandable by non expert users. To address this gap, emerging research has investigated the use of LLMs to generate user friendly, human centered narratives that describe the reason behind the model's predictions. This integration helps to bridge the gap between complex model behavior and end user interpretability by making explanations more accessible, trustworthy, and actionable. Figure 1 illustrates the yearly publication trend from 1999 to 2024 in the healthcare domain using interpretable, explainable, and transparent AI approaches. The core contribution of this study are following:

- An analysis of the role of XAI with a focus on healthcare domain.
- Emerging trends and tools in XAI and LLMs for enhancing interpretability in AI.
- An exploration of how LLMs enhance the explanation by translating them into more understandable format.

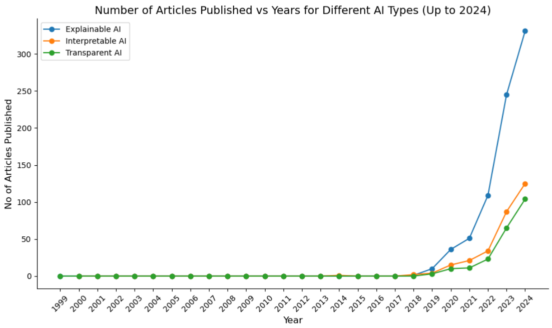


Figure 1: Yearly publication for interpretable, explainable and transparent AI in healthcare(Data derived from SCOPUS)

2 Survey Strategy

To ensure a comprehensive review, we structured our survey strategy into three phases: identifying relevant studies, implementing inclusion and exclusion criteria, and extracting the most suitable articles for detailed analysis.

Step 1: Identifying studies To systematically compile peer-reviewed research on the use of XAI in breast cancer diagnosis, we employed an automated search strategy using specific keywords. Our search targeted reputable academic databases known for publishing high-impact healthcare AI research, including PubMed, IEEE Xplore, Scopus, and Google Scholar.

Step 2: Inclusion and exclusion criteria

- Articles focusing on the use of XAI and LLMs methods and tools for breast cancer diagnosis across multiple imaging or data modalities (e.g., Mammography, MRI, Histopathology).
- Studies employing black-box models (e.g., non-interpretable AI) for breast cancer diagnosis that lack explicit explanation methods or recent approaches integrate LLMs to enhance interpretability by generating human understandable explanations from these opaque systems.
- This survey excluded those articles published prior to 2019, to prioritize recent advancements in XAI and breast cancer research.
- Articles are excluded other than English language, due to potential inconsistencies in translation and accessibility.

Step 3: Extracting suitable articles To ensure the quality and relevance of our review, we applied predefined inclusion and exclusion criteria. Selected studies had to be original research articles published in the aforementioned peer-reviewed journals and must have employed at least one explainable artificial intelligence (XAI) methodology or Large Language Model (LLM) in the context of breast cancer. We conducted a thorough screening of titles and abstracts, excluding studies that did not meet the inclusion criteria. Specifically, we excluded studies that focused on XAI or LLMs without addressing breast cancer, studies on breast cancer without XAI components, preprints pending peer review, duplicate entries, and non-research materials such as books, dissertations, and technical notes. After this screening process, 224 studies were excluded, resulting in a final selection of 54 articles that met all inclusion criteria and were included in our comprehensive analysis.

Table 1 presents the search strings used for article selection, covering publications from January 2020 to December 2024.

3 Fundamental Concepts and Background

Traditionally, radiologists analyze mammograms to identify and diagnose malignancies. This is often done in consultation with other medical professionals for a final decision, but in rural areas and developing countries, access to qualified experts is limited. The complex structure of breast tissue and the peculiarities of breast tumors further

Table 1: Review articles published from 2020 to 2024 were selected based on keywords, with the number of papers retrieved from different databases according to predefined inclusion and exclusion criteria.

Database	Keywords	Paper count
Scopus	(TITLE-ABS-KEY (”Explainable Artificial Intelligence” OR ”Explainable AI” OR ”XAI” OR ”Large Language Model” OR ”LLMs”)) AND TITLE-ABS-KEY (”Breast Cancer”) AND PUBYEAR > 2019 AND PUBYEAR < 2025	192
IEEE Xplore	(”Abstract”:”Explainable Artificial Intelligence” OR ”Abstract”:”Explainable AI” OR ”Abstract”:”XAI” OR ”Abstract”:”Large Language Models” OR ”Abstract”:”LLMs”) AND (”Abstract”:”Breast Cancer”)	2
Google Scholar	(”Explainable Artificial Intelligence” OR ”Explainable AI” OR ”XAI” OR ”Large Language Models” OR ”LLMs”) AND (”Breast Cancer”)	30
PubMed	(”Explainable Artificial Intelligence”[Title/Abstract] OR ”Explainable AI”[Title/Abstract] OR ”XAI”[Title/Abstract] OR ”Large Language Models”[Title/Abstract] OR ”LLMs”[Title/Abstract]) AND (”Breast Cancer”[Title/Abstract]) AND (2020[Date - Publication] : 2024[Date - Publication])	54

complicate the manual analysis process. In contrast to human inspection, AI-based automated image analysis expedites the screening process by saving time and effort by effectively collecting valuable and relevant information from vast amounts of images (Schaffter et al., 2020). To automate the identification of breast cancer, researchers have used a variety of imaging modalities, including CT, MRI, Ultrasound, Thermography, Mammography, and Histopathological imaging (Thakur et al., 2024).

3.1 Datasets & Breast Cancer Screening Approaches

To detect breast cancer, several imaging techniques have been developed. The different methods depend on many factors, like the cancer’s size, location inside the body and aggressiveness. Among the most widely recognized methods for diagnosing and determining breast cancer in its early stages are Mammography, Thermography, MRI, Positron Emission Tomography (PET), CT, Ultrasound and Histopathology (Karthiga et al., 2024). Table 2 provides an overview of publicly available breast cancer datasets, including their imaging modalities and corresponding access links. This section examines existing diagnostic techniques, reviews key studies and validating their effectiveness, and outlines evidence based guidelines for clinical application.

3.1.1 Manual Physical Breast Cancer Checkup A healthcare professional or the patient can perform a Breast Physical Examination (BPE), also known as a Clinical Breast Examination (CBE), to detect abnormalities in breast tissue, such as lumps, asymmetry, or skin changes (Mohamed et al., 2021). To assess texture, mobility, and potential masses, the examiner applies varying pressure levels to palpate the breasts and

Table 2: Public Datasets for Breast Cancer Imaging

Ref	Modality	Dataset name	Dataset availability link
(Spanhol et al., 2015)	Histopathology	BreakHis	https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/
(Mangasarian et al., 1995)	cytology	Breast Cancer Wisconsin(Diagnostic)	https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
(Araújo et al., 2017)	Histopathology	Breast Histology Dataset	https://rdm.inesctec.pt/dataset/nis-2017-003
(Suckling, 1994)	Mammograpghy	Mini-MIAS	http://peipa.essex.ac.uk/info/mias.html
(Rose et al., 2006)	Mammograpghy	DDSM	http://www.eng.usf.edu/cvprg/mammography/database.html
(Moreira et al., 2012)	Mammograpghy	INBreast	https://biokeanos.com/source/INBreast
(Ramos-Pollán et al., 2012)	Mammograpghy	BCDR	https://bcdr.eu/information/about
(Halling-Brown et al., 2020)	Mammograpghy	RSNA	https://www.england.nhs.uk/statistics/statistical-work-areas/diagnostics-waiting-times-and-activity/imaging-and-radiodiagnostics-annual-data/
(Saha et al., 2021)	Radiology	Duke-Breast-Cancer-MRI	https://www.cancerimagingarchive.net/collection/duke-breast-cancer-mri/
(Rodrigues, 2017)	Radiology	Breast ultrasound image	https://data.mendeley.com/datasets/wmy84gzngw/1
(Institute, 2025)	Radiology	The Cancer Genome Atlas Program (TCGA)	https://portal.gdc.cancer.gov/

surrounding tissues, including the axillary lymph nodes. This non-invasive technique is crucial for the early detection of breast cancer, particularly for individuals who are not yet eligible for routine mammography or those living in resource-limited settings. BPE are cost-effective and easily accessible screening methods that do not require specialized equipment, making them particularly valuable for initial clinical assessments (Sultania et al., 2017). Despite being sensitive compared to advanced imaging modalities such as Mammography or MRI, BPE serves as a critical complementary component of screening programs, helping identify suspicious abnormalities requiring advanced diagnostic assessment.

3.1.2 Mammograpghy Mammography is a key imaging technique used for the early detection of breast cancer, which uses low-dose X-rays to visualize internal breast tissue. It effectively identifies microcalcifications, lumps or structural distortions that may indicate malignancy frequently before symptoms show up (Welch et al., 2016). The breast is compressed between two plates to get high resolution images typically in craniocaudal and mediolateral oblique views. Although the test can identify abnormal regions, it cannot determine that they are cancer. Although challenges such as false-positive results, false negatives (notably in dense breast tissue) and patient discomfort exist, innovations like digital mammography and 3D tomosynthesis have enhanced diagnostic precision and minimized recall rates. Its sensitivity is still limited, especially in high risk patients which increases the likelihood of false positives and raises questions about the careless use of population based screening. Additionally, age and breast density affect mammography accuracy with younger people or those with dense tissue showing lower sensitivity (Geisel et al., 2018).

3.1.3 Magnetic Resonance Imaging (MRI) Breast MRI is an advanced, non-invasive diagnostic tool that utilizes powerful magnetic fields and radio waves to generate highly detailed, cross-sectional images of breast tissue. Compared to Mammography or Ultrasound, breast MRI excels in soft tissue contrast and it is beneficial for high risk patients such as those with BC gene mutations or dense breasts, and for evaluating complex cases where other imaging results are inconclusive (Kuhl, 2024). It is particularly used to assess tumor extent, monitor chemotherapy response and screen for cancer recurrence. The procedure often involves a gadolinium based contrast agent which enhances visualization of abnormal blood flow patterns associated with malignancies. While breast MRI boasts high sensitivity in detecting cancers it has lower specificity sometimes leading to false positives and unnecessary biopsies (Gao and Heller, 2020). Additionally, it is more time-consuming, costly and requires careful consideration for patients with certain implants, renal impairment or claustrophobia.

3.2 Challenges in breast cancer recognition using AI

We review several articles on breast cancer detection using AI and address the issues identified in these studies:

- **Limited Public Datasets**: The lack of publicly available datasets limits the progress of breast cancer diagnostic research and poses an obstacle to model development.
- **Unbalanced and insufficient data**: Unbalanced datasets and small sample sizes can negatively impact model performance. This makes it challenging to get reliable results.
- **Data loss in data preprocessing**: Techniques such as data scaling solve the problem of small data sizes. This often leads to data loss. This may affect the quality of the input data.
- **Bias in AI Algorithms**: Sometimes AI algorithms can produce biased results. This challenges the development of models that can be generalized to diverse communities.

4 Impact of Explainability on AI Systems

Machine learning and Deep learning models are often criticized as 'black boxes' due to their inherently opaque and complex structures (Dastgeer and Treigys, 2024). The opaque nature of their decision-making processes makes it challenging for researchers to justify their outputs in human-understandable terms. This lack of transparency has gained significant interest in XAI. A concept significance arises from its social need. Given the increasing focus on explainability in AI algorithms, we identify a few crucial areas where XAI might result in bringing about transformative change.

4.1 Explainable AI (XAI)

The concept of XAI has deep roots in computational history. Early research on this topic can be traced back to literature published over four decades ago. Early examples include rule based expert systems that explained their outcomes based on applied rules (Swartout, 1985). The term XAI refers to the characteristics that explain how the AI model makes its predictions (Shi et al., 2022). According to (Sadeghi et al., 2024), XAI focuses on creating an interface that makes AI decision-making accessible and helps people to understand. Interpretability focuses on creating human understandable rules that explain how a system makes its decisions. In the healthcare industry, where decisions can have critical consequences, it is essential to understand how AI algorithms generate their recommendations. Healthcare workers may find it difficult to assess and trust AI systems outputs if they lack explainability potentially leading to hesitation in adopting these technologies.

4.2 Need of Explainable AI (XAI)

AI models, often called ”black boxes” frequently produce unjustifiable, unexplainable and unaccountable outcomes. In recent years, the XAI field has received more attention. These days, it is crucial for AI systems not only provide precise diagnoses but also offer supplementary information that clarifies or supports the complex classifier decisions. A study (Moxey et al., 2010) highlights that physicians typically do not prefer black boxes in medical systems because they would rather know how the system generates this decision. According to (Lamy et al., 2019) the primary goal of XAI is to develop intelligent systems that can clearly and understandably communicate their choices, predictions, and behaviors to users. This approach aims to develop models that produce correct results and explain the reasoning behind them. This makes it easier for users to trust and communicate with AI systems, particularly in crucial fields like healthcare, finance, and law. XAI focuses on enhancing the transparency, accountability, and fairness of AI systems which help users to understand the model’s behaviour better and make defensible judgements based on its recommendations (Hassija et al., 2024). It draws attention to essential methods like LIME, SHAP, Grad-CAM, and other vital factors that advance explainability and interpretability. Figure 2 illustrates the overview of the problem of black-box AI in medical diagnosis and the proposed solution using XAI and LLMs to provide interpretable and user friendly explanations for clinical decision making.

4.3 Explainable Artificial Intelligence in Medical Diagnostics

The use of XAI to explain medical diagnostic conclusions has recently come into the spotlight of the scientific community. Therefore, it can be understood that the healthcare

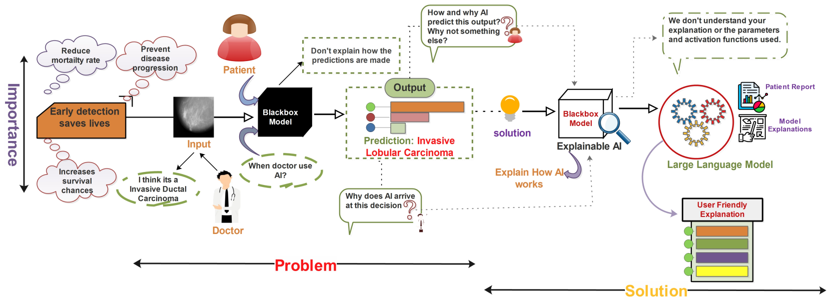

Importance
Reduce mortailty rate
Prevent disease progression
Early detection saves lives
Increases survival chances
Patient
Input
I think its a Invasive Ductal Carcinoma
Doctor
Blackbox Model
Don't explain how the predictions are made
When doctor use AI?
How and why AI predict this output? Why not something else?
Output
Prediction: Invasive Lobular Carcinoma
Why does AI arrive at this decision
solution
Blackbox Model
Explainable AI
Explain How AI works
We don't understand your explanation or the parameters and activation functions used.
Patient Report
Model Explanations
Large Language Model
User Friendly Explanation
Problem
Solution

Figure 2: Explainable AI Framework for Medical Diagnosis

sector is unique in its need where user acceptability of AI algorithms depends on both explainability and accuracy (Zhang et al., 2022). Medical professionals must ensure the models are appropriately trained and the parameters they rely on align with their expertise (Aziz et al., 2024). For example, suppose an ML model's post-hoc analysis findings indicate that fatigue is an indication of breast cancer. In that case a medical expert may instantly suggest that the ML model is unreliable. In applications such as sentiment analysis, spam detection, or recommendation systems the lack of user participation may not pose significant issues. This is because experts can independently analyze the outputs of XAI methods to debug models and identify gaps in the training data. However, in the medical domain the situation is different. Even if XAI methods provide plausible explanations, only clinicians can properly analyze the outputs and understand the causes of failure cases in ML models especially in critical areas like breast cancer diagnosis.

4.4 Four Principles of XAI

The increasing use of AI systems in high stakes fields such as healthcare, finance, and legal decision-making requires these principles since opaque ”black-box” models compromise accountability, safety, and trust (Angelov et al., 2021). According to (Phillips et al., 2020) an AI system must meet these four essential guidelines to be classified as an XAI:

- **Explanation**: AI systems must provide a transparent justification for their outcomes providing relevant contextual evidence or specific reasons.
- **Meaningful**: To ensure clarity, explanations should match the user's expertise level and be conveyed in an understandable, clear, and appropriate format.
- **Accuracy**: The system's explanation must clearly and accurately represent its internal processes and decision pathways, ensuring that they are neither oversimplification or misrepresentation.
- **Knowledge limits**: This principle asserts that AI systems must recognize limitations outside of their intended design where their responses may not be reliable.

The explanation promotes user confidence in AI results by allowing users to examine judgments such as medical diagnoses. For example, a medical professionals need explanations that differ from those of patients or regulators to achieve a meaningful explanation. Accurate explanations can help avoid misunderstandings that bias a model's reasoning essential for evaluating regulatory compliance and identifying algorithmic bias. Lastly, knowledge limits reduce the risk of damage by preventing overconfident or outside of scope predictions such as an AI that detects uncommon diseases it was never trained on.

5 Insights into Explainable AI (XAI)

Explainability in AI models has been attained using a variety of methods and strategies. Table 3 summarizes key studies that have applied XAI techniques in breast cancer diagnosis. The table lists the datasets used, the specific explanation techniques employed (such as Grad-CAM, SHAP, or LIME), machine learning models applied and the limitations identified in each study. This comparison provides insight into the current landscape of XAI applications in medical imaging helping to identify which methods are frequently used as well as their associated challenges.

5.1 Explainability Methods

The method for explaining AI behavior depends on the type of machine learning algorithm. Some algorithms produce inherently transparent models (e.g., Decision Trees, Bayesian Classifiers, Random Forest), while others like deep learning algorithms create complex black-box models that require specialized techniques to interpret their decisions for users to understand (Hall and Gill, 2019). The explainability method is categorized into two categories: How explanations are generated and When explanation are provided? Another key criterion for classifying XAI techniques is the scope of explanations which can be categorized into local explanations focusing on individual predictions and global explanations providing a broader understanding of overall model behavior.

5.1.1 Model-specific vs Model-agnostic Based on how explanations are generated, explainability methods in machine learning are categorized into two types: Model-specific and Model-agnostic. Model-specific approaches are designed to analyze particular types of models by examining their internal structure and parameters to generate insights (Ai and Narayanan. R, 2021). For instance, in Random Forest Models, feature importance is calculated using techniques directly tied to the model's structure. One such technique is the Gini importance metric which evaluates how much each feature reduces prediction uncertainty or impurity. Alternatively, permutation importance assesses a feature's impact by randomly shuffling its values and measuring the resulting decline in model performance. These techniques help identify which features contribute the most to the model's predictions.

Table 3: Comparison of Explainability methods used in different breast cancer studies

Ref no	Dataset	Method	Explanation Technique	Limitation
(Dhiman et al., 2024)	OCT images	TOPSIS & CSA	SHAP	The model's performance may vary if a different dataset is used
(Maheswari et al., 2024)	Fine Needle Aspirate (FNA) images	KNN, SVM, RF Naive Bayes	LIME & SHAP	When applied to another, unexplored situations with different features, the model's performance might not be as accurate or effective.
(Dihmani et al., 2024)	Infrared Image (DMR-IR)	Hybrid Particle Swarm Optimization & Hybrid Spider Monkey Optimization	SHAP	The study focused solely at one imaging modality; adding MRI or mammography might make feature extraction and interpretability more difficult.
(Briola et al., 2024)	Wisconsin Breast Cancer(tabular dataset)	XGBoost	SHAP	The effectiveness of federated learning depends on the consistency and quality of the data from all sources differences affect performance.
(Raghavan, B and v, 2024)	Infrared breast images	DenseNet201, VGG19 & EfficientNetB7	Attention guided grad cam	Using explanation maps resulted in a 42.5% decrease in performance, indicating a compromise between accuracy and interpretability.
(Kaushik et al., 2023)	Infrared breast imagery	DenseNet201	Grad-CAM	Clear and well-structured data is necessary for denoising autoencoders and classifiers to produce precise predictions.
(Rajpal et al., 2023)	DNA methylation data	MethylMarker Framework (Deep Neural Network-based)	SHAP	A single-omic approach to DNA methylation could miss information from combining data from several omics.
(Khater et al., 2023)	WBCD (tabular dataset)	KNN	LIME	Biases in the dataset might affect the model's decision.
(Paudel et al., 2023)	Categorical data	Support Vector Machine,Random Forest,MultiLayer Perceptron	LIME & SHAP	Although it is impressive to get high F1 scores (above 0.98), the study overlooks the possibility of overfitting and generality across other patient groups or datasets.
(Farrag et al., 2023)	INBreast (Mammogram dataset)	DeepLabv3	Grad-CAM	It is in doubt how well this study performs well because it does not compare with other cutting-edge segmentation models (such as U-Net, U-Net++, and Swin UNETR).

However, attention mechanisms in transformer models illustrate model-specific interpretability by revealing how the model processes and prioritizes input elements during prediction. For instance, in Natural Language Processing (NLP) tasks, self-attention layers generate heatmaps that visualize which words or phrases the model focuses on when making predictions. Similarly, in computer vision, attention maps identify salient regions in an image that contribute most to the model's decision. By directly exposing the model's internal focus attention based interpretations help to bridge the gap between the complexity of transformer layers and human understanding of their decision making process. In contrast, model agnostic techniques work across various model types and treat the model as a "black box," avoiding reliance on internal parameters. Methods such as LIME and SHAP provide interpretations by analyzing input output relationships rather than requiring deep expertise in the model's architecture (Wikle et al., 2023). This flexibility makes them particularly useful for explaining complex models like DNNs in a way that is more accessible to non experts.

5.1.2 Post-Hoc vs Transparent Explainability Based on when explanations are provided explainablity methods in machine learning are categorized into two methods:Post-hoc and Transparent. Post-hoc explainability methods systematically examine the internal logic and behavior of trained machine learning models after they generate predictions. These methods may also use surrogate modeling to deconstruct the mechanistic rationale behind the model's input output relationships (Rai, 2020). Transparent models also referred to as ante-hoc methods, intrinsically interpretable or glass-box models prioritize explainability by embedding interpretability directly into a model's architecture or training process. These approaches create inherently understandable systems ensuring transparency from the outset (Retzlaff et al., 2024). Rule Based Systems, such as Bayesian Rule Lists (Letham et al., 2015) are intrinsically interpretable models that classify data using a series of logical "if-then" conditions. These models operate by repeatedly identifying simple, human-readable rules that partition the data into subsets based on feature thresholds or categorical conditions. The interpretability arises from their clear, sequential logic which closely align with human decision making processes. Unlike black-box models, rule based systems provide outcomes that can be directly traced through the applied rules, allowing users to validate each step of the reasoning. Their transparency makes them particularly useful in domains like healthcare where stakeholders need clear justifications for predictions (Rudin, 2019).

5.1.3 Global vs Local Explanation Local interpretability methods aim to explain a model behavior for specific instances or regions within the input space, rather than providing a comprehensive understanding of model's decision making process (Hakkoum et al., 2024). These techniques generate instance specific explanations by approximating the model's behavior locally (e.g., near the data point of interest). While such explanations reveal how the model responds to particular inputs they do not necessarily provide insights into the model's broader decision patterns patterns or generalize to its overall functionality. Global explanation methods offer a holistic understanding of a machine learning model's behavior across the entire dataset rather than explaining an individual predictions (local explanations). These techniques reveal overarching patterns, feature importance and decision logic making them essential for auditing models, ensuring compliance, and building trust (Radensky et al., 2022). Figure 3 illustrates the classification of AI models into transparent (glass-box) models which are intrinsically interpretable (e.g., decision trees, linear models), and black-box models (e.g., deep neural networks, ensemble methods), which rely on post-hoc explanation techniques. The figure further classifies explainability methods by scope (global vs. local) and technique (model-agnostic vs. model-specific), such as feature importance scores, surrogate models or saliency maps. Transparent models enable direct inspection of their logic, while black-box systems require external methods to approximate or extract their decision-making patterns.

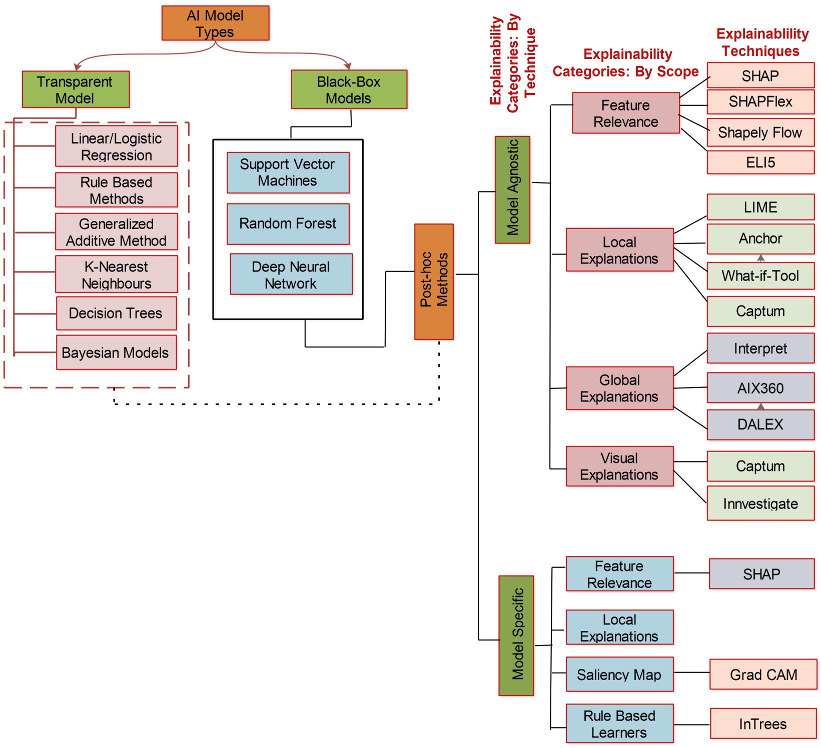


Figure 3: Taxonomy of AI Models and Explainability Approaches

5.2 Hurdles in Enabling XAI

There are several key challenges in achieving explainability for ML models. As noted by (Adadi and Berrada, 2018) all Black-box systems don't have to justify every decision they make because doing this could have a number of negative effects, including worse system performance and higher development costs. In the absence of comprehensive development framework for XAI, local interpretation methods has become a common practice to explain the cases being investigated. Complex machine learning model analysis requires a foundation in advanced statistics and mathematics concepts. The regions that are crucial to the model's predictions are identified by XAI techniques but the underlying characteristics that give these regions their significance are not explained. Healthcare systems have not yet been able to meet the functional and design requirements for the effective use of machine learning models in the medical field. End users actively participate in the creation of ML models in order to align technology with practical requirements and ensure that the end result meets user expectations. Although XAI methods provide meaningful insights, only healthcare care providers can appropriately assess the results and understand the reasons behind failures of the ML model, particularly in crucial domains such as breast cancer detection (Chaddad et al., 2023). Therefore, ML experts often have to rely on clinicians for debugging and improving their models.

6 The Urgency of Large Language Models in Medical Diagnostics

The opacity of black box models poses a critical challenge in high-stakes healthcare applications, such as breast cancer diagnostics, where clinicians and patients need clear, actionable insights. Traditional Explainable AI (XAI) method like SHAP and Gradients typically rely on visualizations (e.g., heatmaps, saliency maps) or numerical outputs to interpret model outcomes. While these approaches are valuable for ML experts, they may fail to communicate with end users such as radiologists, oncologists, or patients who lack specialized technical training. This gap undermines trust and limits the clinical adoption of AI tools despite their diagnostic potential. LLMs offer a transformative solution by converting complex model behaviors into contextual, natural language explanations (Williams et al., 2024). For instance, in breast cancer diagnostics, an LLM could elucidate why a deep learning model classified a mammogram as ”high risk” by summarizing key features (e.g., microcalcifications, tumor morphology in easily understandable terms. This aligns with broader trends in XAI research where frameworks like LLaVA-Med (Li et al., 2023), a multimodal LLM tailored for biomedical applications. It is trained to process medical images (e.g., Radiology scans, Histopathology slides) alongside textual data (e.g., clinical notes, lab reports) and generate natural language responses. Adapting similar approaches to healthcare could empower clinicians to validate AI driven insights and more effectively communicate the reasoning behind diagnoses to patients.

6.1 Enhancing AI Model Transparency via Large Language Models

In recent years, LLMs have shown great promise in enhancing XAI by translating complex machine learning outputs into coherent and accessible human language. The human centered approach further guides the refinement of these explanations by considering user's comprehension levels, contextual needs, and interaction preferences (Zhou et al., 2024). By involving end users in the development process through methods such as interviews and scenario based evaluations. The explanations generated by the LLMs are not only technically accurate but also socially relevant and easy to understand. This approach ultimately enhances the transparency of AI models in healthcare and supports more informed and confident decision making by medical professionals and patients alike.

6.2 Enhancing Breast Cancer Diagnosis Interpretability with Large Language Models

In recent developments, the integration of Deep Learning based image analysis with LLMs has emerged as a promising approach to enhance interpretability in AI-driven breast cancer diagnosis. While CNNs and vision transformers (ViTs) achieve state of the art performance in classifying Mammographic images (e.g., malignant vs. benign), post hoc explanation methods such as Grad-CAM and LIME typically generate saliency maps and feature importance scores. However, these technical outputs are often not easily interpretable by medical professionals or patients. To address this issue, emerging research has investigated the use of LLMs to generate human centered natural language explanations that better convey the model's reasoning in an accessible manner. By translating complex outputs into coherent narratives, LLMs help bridge the gap between model behavior and user understanding, improving transparency, trust and decision support particularly for medical professionals such as radiologists and oncologists. LLMs (e.g., GPT-4, LLaMA-3) can be incorporated as translation layers that convert structured XAI outputs including prediction confidence, salient image regions (e.g., spiculated masses, microcalcifications), and metadata (e.g., lesion size and location)into natural language explanations (Egli, 2023).

7 Tools for Fairness and Explainability in Interpretable AI

Machine learning practitioners frequently require tools to examine and evaluate their models (Rahman et al., 2023). Essential steps to enhance performance include assessing a model's effectiveness and investigating how input modifications impact its outcomes.

7.1 Tools for ensuring Fairness and reducing Bias

IBM AI Fairness 360 (AIF360): The open source toolkit IBM AI Fairness 360 (AIF360) (Varshney, 2018) offers a collection of measurements, algorithms and bias mitigation strategies to identify discrimination and resolve bias in machine learning models. It has tools for using preprocessing and postprocessing strategies to improve fairness, as well as the ability to measure bias in datasets and models (Blow et al., 2024).
The What-If Tool (WIT): WIT from Google is an open source TensorBoard web application that allows users to evaluate the performance and fairness of machine learning models (Wexler et al., 2019). The tool requires only a sample dataset and trained models.

Fairlearn : This toolkit is designed to help practitioners to evaluate and improve fairness of AI systems (Bird et al., 2020). Its accompanying Python library, supports fairness in AI by allowing practitioners to evaluate model outputs across different populations and includes specific algorithms designed to mitigate bias and fairness issues.

7.2 Agnostic Explainability Tools

SHapley Additive exPlanations (SHAP): SHAP is an Explainable Artificial Intelligence (XAI) method based on game-theoretic principles (Lundberg and Lee, 2017). It interprets machine learning models by considering individual features as team members working together to achieve a common goal, where the model's outcome represents the collective payoff. By calculating each feature's unique contribution to the result, SHAP provides both local and global explanations, offering insights into feature importance across the dataset as well as overall model behavior.

Local Interpretable Model-agnostic Explanations (Lime): LIME (Ribeiro et al., 2016) helps to enhance the interpretability of a machine learning models and make its individual outcome more understandable. It provides local explanations by approximating the model's behavior for a specific single instance, helps it to for understand how a particular outcome was made.

Anchors: It is an open-source toolkit that generates high-precision, rule-based explanations for individual model predictions (Ribeiro et al., 2018). It identifies minimal conditions (”anchors”) under which the prediction remains unchanged, thereby enhancing transparency and trust in AI systems through locally faithful and interpretable rules.

7.3 Explainability Methods for Deep Neural Networks

Captum: Captum is an open-source PyTorch library for model interpretability, offering a unified framework to implement and evaluate feature attribution methods. It supports gradient-based and perturbation-based algorithms to explain predictions across diverse models including complex architectures like graph neural networks in both classification and non-classification tasks (Stanchi et al., 2023).

Gradient-weighted Class Activation Mapping (Grad-CAM): Grad-CAM is a visual attribution method for CNNs that generates coarse heatmaps highlighting image regions critical to a model's class prediction. It computes gradients from a target class back to the final convolutional layer, weighting activation maps to reveal influential spatial features (Selvaraju et al., 2017).

Integrated Gradients: Integrated Gradients is an attribution method for DNNs that distributes a model's prediction to input features (Sundararajan et al., 2017). It satisfies two core principles: Sensitivity (non-zero attribution for output-changing features) and Implementation Invariance (identical attributions for functionally equivalent models). The approach requires no model modifications and leverages standard gradient computations.

7.4 Large Language Model Explainability Tools

ExBERT: ExBERT is an interactive visualization tool for exploring attention mechanisms in transformer models (e.g., BERT). It visualizes word-to-word attention across layers and heads, enabling granular analysis of how LLMs process linguistic relationships (Gajbhiye et al., 2021).

BertViz: BertViz is an open-source visualization tool enables interactive exploration of transformer attention mechanisms, visualizing token to token patterns and head dynamics in self or cross attention (Vig, 2019). It advances explainability by revealing how input tokens influence predictions and exposing biases or spurious correlations. Researchers leverage it to debug behaviors, improve model design, and enhance trust in AI outputs.

ExplainaBoard: ExplainaBoard is an open-source toolkit for interpretable evaluation of NLP models (Yuan et al., 2021). ExplainaBoard converts standard NLP evaluation into an interpretable, diagnostic, and comparative analysis, empowering researchers to decode model behavior beyond metrics.

Figure 4 tracks the rising GitHub star counts of machine learning explainability repositories over time. The x-axis represents the period from 2015 to 2025, while the y-axis quantifies repository popularity through accumulated stars. The repositories include well known tools include SHAP (slundberg/shap), LIME (marcotcr/lime), Captum (pytorch/captum), and other libraries dedicated to AI interpretability. SHAP exhibits the most pronounced growth trajectory, surpassing LIME and Captum in adoption. The trend suggests a rising interest in post-hoc explainability techniques, with certain repositories gaining significant traction over time.

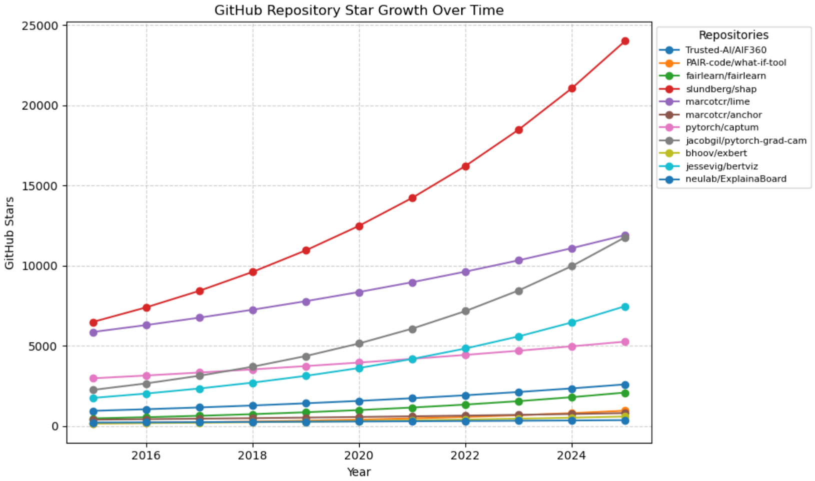


Figure 4: GitHub Repository Star Growth Over Time for Explainability Techniques

8 Conclusion

In this paper, we review the existing literature and provide a comprehensive analysis of XAI, with a focus on its applications in healthcare and cancer diagnostics, while also highlighting the emerging role of LLMs enhancing AI interpretability and user-centered explanations. This study explored the inherent interpretability challenges of DL models, clarifying why they are often described as 'black boxes. We discussed the limitations and challenges associated with current XAI methods, particularly in providing clear and meaningful explanations to end users. By leveraging the tools discussed in this study, practitioners can build interpretable models that promote the responsible and widespread adoption of AI in sensitive and high-impact domains such as healthcare. We highlight the critical need for trust between humans and AI, particularly in medical contexts, where even small errors in model predictions can have severe consequences. Futhermore, we explored the transformative potential of integrating LLMs into XAI systems, particularly in the context of AI-driven breast cancer diagnosis. Recent developments in the filed of interpretable machine learning, particularly in local interpretation methods, provide insights into the decision-making process of complex models by explaining individual predictions. It is crucial to explore approaches that make these explanations more accessible and comprehensible to a wider range of stakeholders, ensuring the effective translation of AI insights into actionable and understandable information. In future work, we will address the need to integrate visual tools with textual explanations, enabling end users to better understand the critical regions of an image by visualizing them using techniques such as saliency maps or heatmaps, ultimately enhancing transparency and trust in AI systems.

References

Adadi, A., Berrada, M. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (xai), *IEEE Access* **6**, 52138–52160.
http://doi.org/10.1109/ACCESS.2018.2870052

Ai, Q., Narayanan. R, L. 2021. Model-agnostic vs. model-intrinsic interpretability for explainable product search, *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 5–15.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., Atkinson, P. M. 2021. Explainable artificial intelligence: an analytical review, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **11**(5), e1424.

Araújo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Polónia, A., Campilho, A. 2017. Classification of breast cancer histology images using convolutional neural networks, *PloS one* **12**(6), e0177544.

Aziz, N. A., Manzoor, A., Mazhar Qureshi, M. D., Qureshi, M. A., Rashwan, W. 2024. Unveiling explainable ai in healthcare: Current trends, challenges, and future directions, *medRxiv* pp. 2024–08.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in ai, *Microsoft, Tech. Rep. MSR-TR-2020-32* .

Blow, C. H., Qian, L., Gibson, C., Obiomon, P., Dong, X. 2024. Comprehensive validation on reweighting samples for bias mitigation via aif360, *Applied Sciences* **14**(9), 3826.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., Jemal, A. 2018. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries, *CA: a cancer journal for clinicians* **68**(6), 394–424.

Briola, E., Nikolaidis, C. C., Perifanis, V., Pavlidis, N., Efraimidis, P. 2024. A federated explainable ai model for breast cancer classification, p. 194 – 201.
`http://doi.org/10.1145/3655693.3660255`

Chaddad, A., Peng, J., Xu, J., Bouridane, A. 2023. Survey of explainable ai techniques in healthcare, *Sensors* **23**(2), 634.

Croce, D., Smirnov, A., Tiburzi, L., Travaglini, S., Costa, R., Calabrese, A., Basili, R., Levialdi Ghiron, N., Melino, G. 2024. Ai-driven transcriptomic encoders: From explainable models to accurate, sample-independent cancer diagnostics, *Expert Systems with Applications* **258**.

Dastgeer, S., Treigys, P. 2024. Transforming black-box models into explainable ai for breast cancer recognition, *DAMSS: 15th conference on data analysis methods for software systems, Druskininkai, Lithuania, November 28-30, 2024.*, Vilniaus universiteto leidykla, pp. 19–20.

Dhiman, B., Kamboj, S., Srivastava, V. 2024. Explainable ai based efficient ensemble model for breast cancer classification using optical coherence tomography, *Biomedical Signal Processing and Control* **91**.
`http://doi.org/10.1016/j.bspc.2024.106007`

Dihmani, H., Bousselham, A., Bouattane, O. 2024. A new computer-aided diagnosis system for breast cancer detection from thermograms using metaheuristic algorithms and explainable ai, *Algorithms* **17**(10).
`http://doi.org/10.3390/a17100462`

Egli, A. 2023. Chatgpt, gpt-4, and other large language models: the next revolution for clinical microbiology?, *Clinical Infectious Diseases* **77**(9), 1322–1328.

Farrag, A., Gad, G., Fadlullah, Z. M., Fouda, M. M. 2023. Mammogram tumor segmentation with preserved local resolution: An explainable ai system, p. 314 – 319.
`http://doi.org/10.1109/GLOBECOM54140.2023.10436915`

Gajbhiye, A., Moubayed, N. A., Bradley, S. 2021. Exbert: An external knowledge enhanced bert for natural language inference, *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30*, Springer, pp. 460–472.

Gao, Y., Heller, S. L. 2020. Abbreviated and ultrafast breast mri in clinical practice, *Radiographics* **40**(6), 1507–1527.

Geisel, J., Raghu, M., Hooley, R. 2018. The role of ultrasound in breast cancer screening: the case for and against ultrasound, *Seminars in Ultrasound, CT and MRI*, Vol. 39, Elsevier, pp. 25–34.

Goodman, B., Flaxman, S. 2017. European union regulations on algorithmic decision-making and a “right to explanation”, *AI magazine* **38**(3), 50–57.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D. 2018. A survey of methods for explaining black box models, *ACM computing surveys (CSUR)* **51**(5), 1–42.

Hakkoum, H., Idri, A., Abnane, I. 2024. Global and local interpretability techniques of supervised machine learning black box models for numerical medical data, *Engineering Applications of Artificial Intelligence* **131**, 107829.

Hall, P., Gill, N. 2019. *An introduction to machine learning interpretability*, O’Reilly Media, Incorporated.

Halling-Brown, M. D., Warren, L. M., Ward, D., Lewis, E., Mackenzie, A., Wallis, M. G., Wilkinson, L. S., Given-Wilson, R. M., McAvinchey, R., Young, K. C. 2020. Optimam mammography image database: a large-scale resource of mammography images and clinical data, *Radiology: Artificial Intelligence* **3**(1), e200103.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A. 2024. Interpreting black-box models: a review on explainable artificial intelligence, *Cognitive Computation* **16**(1), 45–74.
`https://doi.org/10.1007/s12559-023-10179-8`

Institute, N. C. 2025. The cancer genome atlas (tcga). Accessed: 2025-01-26.
`https://www.cancer.gov/ccg/research/genome-sequencing/tcga`

Karthiga, R., Narasimhan, K., Amirtharajan, R. et al. 2024. Review of ai & xai-based breast cancer diagnosis methods using various imaging modalities, *Multimedia Tools and Applications* pp. 1–52.

Kaushik, R., Sivaselvan, B., Kamakoti, V. 2023. Integrating explainable ai with infrared imaging and deep learning for breast cancer detection, *OCIT 2023 - 21st International Conference on Information Technology, Proceedings*, p. 82 – 87.
`http://doi.org/10.1109/OCIT59427.2023.10431160`

Khater, T., Hussain, A., Mahmoud, S., Yasen, S. 2023. Explainable ai for breast cancer detection: A lime-driven approach, *Proceedings - International Conference on Developments in eSystems Engineering, DeSE*, p. 540 – 545.
`http://doi.org/10.1109/DeSE60595.2023.10469341`

Kuhl, C. K. 2024. Abbreviated breast mri: state of the art, *Radiology* **310**(3), e221822.

Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., Séroussi, B. 2019. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach, *Artificial Intelligence in Medicine* **94**, 42–53.
`https://doi.org/10.1016/j.artmed.2019.01.001`

Letham, B., Rudin, C., McCormick, T. H., Madigan, D. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, *Advances in Neural Information Processing Systems* **36**, 28541–28564.

Lundberg, S., Lee, S. 2017. Advances in neural information processing systems. 2017, *A unified approach to interpreting model predictions* pp. 4765–4774.

Maheswari, B. U., Aaditi, A., Avvaru, A., Tandon, A., De Prado, R. P. 2024. Interpretable machine learning model for breast cancer prediction using lime and shap.
`http://doi.org/10.1109/I2CT61223.2024.10543965`

Malhi, A., Främling, K. 2023. An evaluation of contextual importance and utility for outcome explanation of black-box predictions for medical datasets, *Communications in Computer and Information Science* **1901 CCIS**, 544 – 557.
`http://doi.org/10.1007/978-3-031-44064-9_29`

Mangasarian, O. L., Street, W. N., Wolberg, W. H. 1995. Breast cancer diagnosis and prognosis via linear programming, *Operations research* **43**(4), 570–577.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A. et al. 2020. International evaluation of an ai system for breast cancer screening, *Nature* **577**(7788), 89–94.

Mohamed, S. K., Sakr, N. A., Hikal, N. A. 2021. A review of breast cancer classification and detection techniques, *International Journal of Advanced Science Computing and Engineering* **3**(3), 128–139.

Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., Cardoso, J. S. 2012. Inbreast: toward a full-field digital mammographic database, *Academic radiology* **19**(2), 236–248.

Moxey, A., Robertson, J., Newby, D., Hains, I., Williamson, M., Pearson, S.-A. 2010. Computerized clinical decision support for prescribing: provision does not guarantee uptake, *Journal of the American Medical Informatics Association* **17**(1), 25–33.

Paudel, P., Saud, R., Karna, S. K., Bhandari, M. 2023. Determining the major contributing features to predict breast cancer imposing ml algorithms with lime and shap.
`http://doi.org/10.1109/ICECET58911.2023.10389217`

Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., Przybocki, M. A. 2020. Four principles of explainable artificial, *Technical report*, NIST Interagency/Internal Report (NISTIR). Gaithersburg, MD: National . . . .

Radensky, M., Downey, D., Lo, K., Popovic, Z., Weld, D. S. 2022. Exploring the role of local and global explanations in recommender systems, *Chi conference on human factors in computing systems extended abstracts*, pp. 1–7.

Raghavan, K., B, S., v, K. 2024. Attention guided grad-cam : an improved explainable artificial intelligence model for infrared breast cancer detection, *Multimedia Tools and Applications* **83**(19), 57551 – 57578.
`http://doi.org/10.1007/s11042-023-17776-7`

Raghavan, K., Balasubramanian, S., Veezhinathan, K. 2024. Explainable artificial intelligence for medical imaging: Review and experiments with infrared breast images, *Computational Intelligence* **40**(3).
`http://doi.org/10.1111/coin.12660`

Rahman, M. S., Khomh, F., Hamidi, A., Cheng, J., Antoniol, G., Washizaki, H. 2023. Machine learning application development: practitioners' insights, *Software Quality Journal* **31**(4), 1065–1119.

Rai, A. 2020. Explainable ai: From black box to glass box, *Journal of the Academy of Marketing Science* **48**, 137–141.

Rajpal, S., Rajpal, A., Saggar, A., Vaid, A. K., Kumar, V., Agarwal, M., Kumar, N. 2023. Xaimethylmarker: Explainable ai approach for biomarker discovery for breast cancer subtype classification using methylation data, *Expert Systems with Applications* **225**.
`http://doi.org/10.1016/j.eswa.2023.120130`

Ramos-Pollán, R., Guevara-López, M. Á., Oliveira, E. 2012. A software framework for building biomedical machine learning classifiers through grid computing resources, *Journal of medical systems* **36**, 2245–2257.

Retzlaff, C. O., Das, S., Wayllace, C., Mousavi, P., Afshari, M., Yang, T., Saranti, A., Angerschmid, A., Taylor, M. E., Holzinger, A. 2024. Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities, *Journal of Artificial Intelligence Research* **79**, 359–415.

Ribeiro, M. T., Singh, S., Guestrin, C. 2016. " why should i trust you?" explaining the predictions of any classifier, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Ribeiro, M. T., Singh, S., Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations, *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

Rodrigues, P. S. 2017. Breast ultrasound image, *Mendeley Data* **1**(10).

Rose, C., Turi, D., Williams, A., Wolstencroft, K., Taylor, C. 2006. Web services for the ddsm and digital mammography research, *Digital Mammography: 8th International Workshop, IWDM 2006, Manchester, UK, June 18-21, 2006. Proceedings 8*, Springer, pp. 376–383.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature machine intelligence* **1**(5), 206–215.

Sabol, P., Sincak, P., Ogawa, K., Hartono, P. 2019. Explainable classifier supporting decision-making for breast cancer diagnosis from histopathological images, Vol. 2019-July.
`http://doi.org/10.1109/IJCNN.2019.8852070`

Sadeghi, Z., Alizadehsani, R., CIFCI, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S. et al. 2024. A review of explainable artificial intelligence in healthcare, *Computers and Electrical Engineering* **118**, 109370.

Saha, A., Harowicz, M. R., Grimm, L. J., Weng, J., Cain, E., Kim, C., Ghate, S., Walsh, R., Mazurowski, M. A. 2021. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set], *The Cancer Imaging Archive* .

Schaffter, T., Buist, D. S., Lee, C. I., Nikulin, Y., Ribli, D., Guan, Y., Lotter, W., Jie, Z., Du, H., Wang, S. et al. 2020. Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms, *JAMA network open* **3**(3), e200265–e200265.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Shi, H., Yang, D., Tang, K., Hu, C., Li, L., Zhang, L., Gong, T., Cui, Y. 2022. Explainable machine learning model for predicting the occurrence of postoperative malnutrition in children with congenital heart disease, *Clinical Nutrition* **41**(1), 202–210.
`https://doi.org/10.1016/j.clnu.2021.11.006`

Spanhol, F. A., Oliveira, L. S., Petitjean, C., Heutte, L. 2015. A dataset for breast cancer histopathological image classification, *Ieee transactions on biomedical engineering* **63**(7), 1455–1462.

Stanchi, O., Ronchetti, F., Quiroga, F. 2023. The implementation of the rise algorithm for the captum framework, *Conference on Cloud Computing, Big Data & Emerging Topics*, Springer, pp. 91–104.

Suckling, J. 1994. The mammographic images analysis society digital mammogram database, *Exerpta Medica. International Congress Series, 1994*, Vol. 1069, pp. 375–378.

Sultania, M., Kataria, K., Srivastava, A., Misra, M. C., Parshad, R., Dhar, A., Hari, S., Thulkar, S. 2017. Validation of different techniques in physical examination of breast, *Indian Journal of Surgery* **79**, 219–225.

Sundararajan, M., Taly, A., Yan, Q. 2017. Axiomatic attribution for deep networks, *International conference on machine learning*, PMLR, pp. 3319–3328.

Swartout, W. R. 1985. Explaining and justifying expert consulting programs, *Computer-assisted medical decision making*, Springer, pp. 254–271.

Thakur, N., Kumar, P., Kumar, A. 2024. A systematic review of machine and deep learning techniques for the identification and classification of breast cancer through medical image modalities, *Multimedia Tools and Applications* **83**(12), 35849–35942.

Varshney, K. 2018. Introducing ai fairness 360, *IBM Research blog* .

Vig, J. 2019. Bertviz: A tool for visualizing multihead self-attention in the bert model, *ICLR workshop: Debugging machine learning models*, Vol. 3.

Welch, H. G., Prorok, P. C., O'Malley, A. J., Kramer, B. S. 2016. Breast-cancer tumor size, overdiagnosis, and mammography screening effectiveness, *New England Journal of Medicine* **375**(15), 1438–1447.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., Wilson, J. 2019. The what-if tool: Interactive probing of machine learning models, *IEEE transactions on visualization and computer graphics* **26**(1), 56–65.

WHO 2024. No-one should face breast cancer alone, *https://www.who.int/news-room/fact-sheets/detail/breast-cancer* .

Wikle, C. K., Datta, A., Hari, B. V., Boone, E. L., Sahoo, I., Kavila, I., Castruccio, S., Simmons, S. J., Burr, W. S., Chang, W. 2023. An illustration of model agnostic explainability methods applied to environmental data, *Environmetrics* **34**(1), e2772.

Williams, C. Y., Zack, T., Miao, B. Y., Sushil, M., Wang, M., Kornblith, A. E., Butte, A. J. 2024. Use of a large language model to assess clinical acuity of adults in the emergency department, *JAMA Network Open* **7**(5), e248895–e248895.

Yuan, W., Neubig, G., Liu, P. 2021. Bartscore: Evaluating generated text as text generation, *Advances in neural information processing systems* **34**, 27263–27277.

Zhang, Y., Weng, Y., Lund, J. 2022. Applications of explainable artificial intelligence in diagnosis and surgery, *Diagnostics* **12**(2), 237.

Zhou, H., Hu, C., Yuan, Y., Cui, Y., Jin, Y., Chen, C., Wu, H., Yuan, D., Jiang, L., Wu, D. et al. 2024. Large language model (llm) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities, *IEEE Communications Surveys & Tutorials* .

Received June 17, 2025 , accepted June 27, 2025