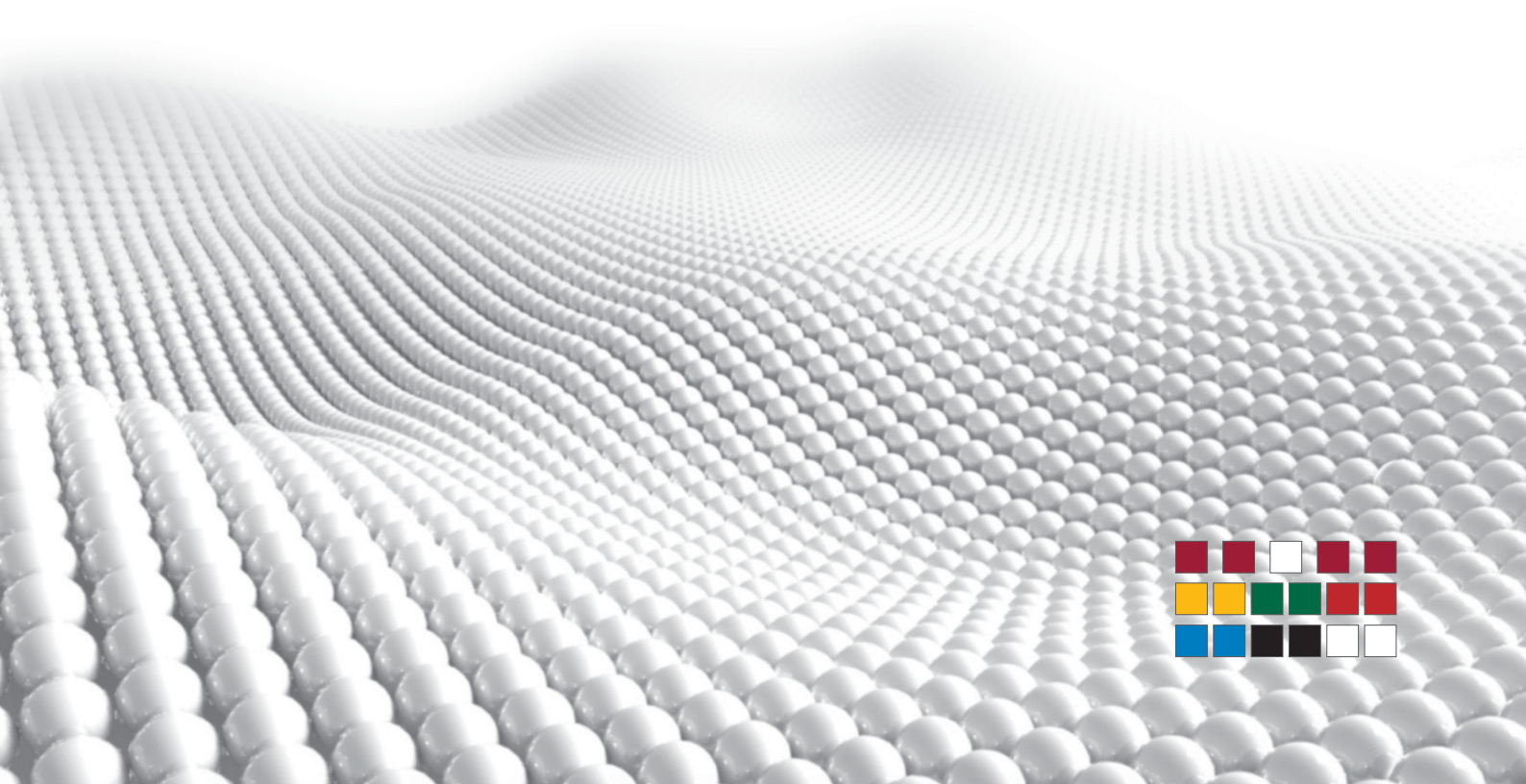
ISSN 2255-8950 (Online)
ISSN 2255-8942 (Print)

Volume 13(2025)

No. 4

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

VIDZEMES
AUGSTSKOLA

EDITORIAL BOARD

Words of Gratitude to Our Reviewers

On behalf of the editorial team, we would like to extend our deepest appreciation to all our reviewers. Your expertise, dedication, and thoughtful feedback are the backbone of our journal's quality. Each manuscript you evaluate benefits not only from your scholarly insight but also from the care you invest in guiding authors toward stronger, clearer, and more impactful work.

Peer review is often a demanding and time-consuming task, yet you approach it with professionalism and generosity. By sharing your knowledge, you help us uphold the highest standards of academic integrity and contribute to the advancement of research in our field.

We are truly grateful for your commitment and collaboration. Without your invaluable contributions, our journal could not thrive. Thank you for being an essential part of our community and for helping us build a platform where rigorous scholarship can flourish.

With sincere appreciation,

Juris Borzovs
Gintautas Dzemyda
Raimundas Matulevicius
Co-Editors-in-Chief

Reviewers in 2023-2025 (in random order)

Raita Rollande, Linas Petkevičius, Mebarek-Oudina Fateh, Ivan Sakhno, Aleksandrs Kolesovs, Tatjana Rubina, Pēteris Rivža, Dace Visnola, Mohamad Gharib, Baiba Holma, Artūrs Žogla, Inga Žilinskienė, Laila Niedrīte, Anita Jansone, Gintautas Grigas, Albina Auksoriutė, Darja Solodovņikova, Jolita Bernatavičienė, Liena Hačatrjana, Aušra Mackutė-Varoneckienė, Kęstutis Kubilius, Rinalds Vīksna, Mariia Bakhtina, Inguna Skadiņa, Antti Ainamo, Jānis Judvaitis, Ulrich Norbisrath, Kārlis Podnieks, Agris Šostaks, Māris Nartišs, Gražina Korvel, Evalds Ikaunieks, Mārcis Pinnis, Leo Seļāvo, Saulius Maskeliūnas, Tatjana Jevsikova, Sonata Vdovinskienė, Ivo Odītis, Māris Vītiņš, Stefano Bonnini, Irina Arhipova, Āris Dzērvāns, Andrejs Romanovs, Jānis Zuters, Oksana Ņikiforova, Maksims Ivanovs, Kuldar Taveter, Martynas Sabaliauskas, Vitalijs Bolsakovs, Andrejs Zujevs, Mindaugas Morkūnas, Aleksejs Zacepins, Vitalijs Komašilovs, Gatis Vītols, Svetlana Cipiševa, Asta Slotkienė, Jolanta Miliauskaitė, Māris Alberts, Lauris Cikovskis, Arnis Lektauers, Jose Juan Hernandez Cabrera, Olga Kurasova, Gintautas Dzemyda, Askars Salimbajevs, Jurģis Poriņš, Daiga Deksne, Julius Venskus, Jānis Sīlis, Tomas Krilavičius, Sarma Cakula, Gintautas Tamulevičius, Pijus Kasparaitis, Antra Kļavinska, Vjatšeslav Antipenko, Pēteris Vanags, Edmundas Trumpa, Artūrs Sproģis, Vladislav V. Fomin, Jurijs Merkurjevs, Alexander Suzdalenko, Ilya A. Galkin, Ingus Mitrofanovs, Kaspars Sudars, Vitalij V. Denisov, Juris Vīksna, Blerta Leka (Moçka), Linda Mihno, Radka Nacheva, Naveed Muhammad, Ernestas Filatovas, Rokas Gipiškis, Līga Zariņa, Viktor Medvedev, Artūrs Ņikuļins, Guntis Bārzdiņš, Guntis Vilnis Strazds, Kārlis Freivalds, Reelika Suviste, Pēteris Paikens, Artūrs Lavrenovs, Vladimir A. Oleshchuk, Ivan I. Piletski, Aleksandr Igumenov, Virginijus Marcinkevičius, Ivan S. Laktionov, Yash Gupta, Mārīte Kirikova, Boriss Mišņevs, Elīna Kalniņa, Dmitry Korzun, Rūta Levulienė, Imants Gorbāns, Oskars Java.

Table of Content

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, 740-757
https://doi.org/10.22364/bjmc.2025.13.4.01

Technology for Blackcurrant Plantations Control Using Drones

Janis BICEVSKIS[1], Reinis ODITIS[1], Ivo ODITIS[1], Zane BICEVSKA[2]

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

Janis.Bicevskis@lu.lv, ro17020@students.lu.lv,Ivo.Oditis@lu.lv, Zane.Bicevska@di.lv

ORCID 0000-0001-5298-9859, ORCID 0009-0003-5488-1608, ORCID 0000-0003-2354-3780, ORCID 0000-0002-5252-7336

Abstract. This article presents a technology-based solution for monitoring blackcurrant vegetation using drones and artificial intelligence. The proposed system, implemented in a blackcurrant farm in Latvia, includes a three-stage process: mapping, identification and segmentation, and classification. Drones capture aerial images of the plantation, which are processed using tools like WebODM and deep learning algorithms to create accurate field maps. Neural networks are employed for identification, instance segmentation and classification of blackcurrant leaves into categories such as healthy, nutrient-deficient, or diseased. The system incorporates several AI model families—YOLO and ResNet —selected based on performance, accuracy, and resource efficiency. The methodology enables high-throughput analysis of large horticultural areas, supporting growers in decision-making by providing precise, visual insights into plant health. The approach demonstrates the viability of integrating drone technology and AI for precision agriculture, particularly in the specialized context of blackcurrant farming. The proposed technology, with appropriate adjustments, can also be applied to the vegetation monitoring of other horticultural crops.

Keywords: Drone Technologies, Machine Learning, Plant Vegetation Monitoring

1. Introduction

Unmanned Aerial Vehicle (UAV), hereinafter referred to as a drone in this article, is defined according to the Dictionary of Military and Associated Terms (Drone, 2005) as: "A powered, aerial vehicle that does not carry a human operator, uses aerodynamic forces to provide vehicle lift, can fly autonomously or be piloted remotely, can be expendable or recoverable, and can carry a lethal or nonlethal payload". The term "drone," commonly used in mass media, was introduced even before World War II, as the first unmanned aerial vehicles were named after bees and wasps.

Today, drone manufacturing costs and production volumes have reached a level that allows their application across various fields, including precision agriculture. Scientific reviews available in the literature analyse the latest advancements in drone technology

used in precision farming. For example, Botta et al. (2022) compiled 184 publications using data from Google Scholar and SCOPUS, while Uzhinskiy (2023) reviewed 164 works focusing on the application of AI methods in agriculture.

The authors of these studies unanimously conclude that drones can be effectively used for crop vegetation monitoring, while agricultural operations should be performed using ground-based equipment. During a flight, a drone can capture images of designated field areas and transmit them for further analysis. Using machine learning methods, this enables the detection of vegetation conditions and issues that determine necessary agronomic actions. Thus, drones allow for rapid inspection of large agricultural areas and the collection of crucial data on crop health and required maintenance tasks.

This study is practically oriented, with the main objective being the development of a technology that integrates drones usage and artificial intelligence methods, described more in detail in (Oditis et al., 2025). The system is designed to alert farmers about plant diseases, pests, nutrient deficiencies, and other issues. The developed technology must be user-friendly and economically viable.

The following chapters provide a description of the technology designed to support blackcurrant cultivation using basic drones and imaging cameras. The proposed technology was tested on a blackcurrant farm in Latvia, confirming the validity of the chosen approach.

The structure of this study is as follows: Chapter 1 provides an overview of drone usage in precision agriculture worldwide. Chapter 2 focuses specifically on drone applications in horticulture. Chapter 3 presents the authors' proposed methodology for assessing blackcurrant plantations. Chapter 4 offers a visualization of blackcurrant plantation conditions. Chapter 5 discusses the obtained results and presents conclusions.

2. Drone application in horticulture

To feed the rapidly growing global population, agricultural enterprises must produce more food without increasing cultivated land areas. This can be achieved by applying advanced farming technologies. Some of these technologies are still in development, while others are already offered by commercial companies. Today, farms can utilize a range of advanced tools, such as satellite data, drones, autonomous platforms for agricultural operations, sensors, and robots, to obtain detailed information about crop and soil conditions and to perform specific agronomic tasks.

However, in many countries, including Latvia, the adoption of drone technology in agriculture is still in its early stages. Among various scientific and technological challenges being addressed to achieve sustainable development goals, the use of new technologies and methodologies in agriculture has attracted the interest of the engineering research community. The objective is to develop technologies suited for precision agriculture that enhance the long-term profitability and efficiency of agricultural production.

2.1. General overview of drone applications

According to (Botta et al, 2022) and (Uzhinskiy, 2023), data from the Food and Agriculture Organization (FAO) indicate that global food production must increase by 70% by 2050 to sustain the growing world population. However, in the European Union,

the number of people employed in agriculture has decreased by 35% over the past decade, and the expansion of agricultural land is largely unfeasible.

These factors have driven increased interest in advanced agricultural technologies, including sensors, robots, drones, digitalization, and artificial intelligence (AI). AI and machine learning are considered highly promising for detecting agricultural issues, monitoring crop health, forecasting yields and prices, mapping harvests, and optimizing pesticide and fertilizer use.

There are various research directions that discuss the use of modern technologies in agriculture: Internet of Things (IoT) technologies in agriculture (Xu et al., 2022), bibliometric analysis of drone use in farming (Rejeb et al., 2022), deep learning methods for controlled-environment agriculture (Ojo et al., 2022), robotic harvesting technologies (Mail et al., 2023), machine vision applications in agricultural robot navigation (Wang et al., 2022), AI in agriculture (Oliveira et al., 2023), Agriculture 4.0 (Dayioglu et al., 2021), (Abbasi et al., 2022).

Depending on the specific task, drones can offer similar capabilities to satellite image analysis but with higher precision and flexibility. They can perform tasks such as soil analysis (Huuskonen et al., 2018), (Zhou et al., 2023), (Bertalan et al., 2022 monitoring sowing density and crop development (Wilke et al., 2021), (Koh et al., 2019), weed and pest detection and classification (Ong et al., 2023), (Ong et al., 2023), (Tetila et al., 2020), (Mohidem et al., 2021), and yield prediction and maturity assessment (Kumar et al., 2023), (Zeng et al., 2021), (Shahi et al., 2023).

In rare cases, drones can also be used for harvesting, precision fertilization (Chen et al., 2022), (Song et al. 2023), (Su et. al. 2022), pesticide spraying (Anand et al., 2019), (Ivič et al., 2019), (Sinha, 2020) and even mechanical pest eradication. IoT and sensor technologies provide farmers with real-time data on soil parameters, temperature, atmospheric gases, weather conditions, and many other variables, often processed in cloud-based IT infrastructures for further analysis and forecasting (Dhanaraju et al., 2020), (Gagliardi et al., 2022), (Madushanki et al., 2019), (Bilotta et al., 2023).

2.2. Scope of drone applications

The use of artificial intelligence and cloud technology in drones has brought significant improvements to smart agriculture. These new technologies can capture high-resolution images, aerial maps, and thermal images, which can be utilized in various agricultural applications, including:

- Soil analysis: Drones can be used for soil sampling, analysing soil moisture levels, and assessing soil quality, helping farmers optimize fertilization and irrigation processes,
- Planting: Drones can be used for precise seed sowing and/or seedling planting, reducing labour and planting material costs,
- Crop spraying: Drones equipped with spraying systems can be used for the precise distribution of pesticides, herbicides, and fertilizers, minimizing environmental impact while saving time and financial resources,
- Irrigation management: Drones equipped with thermal sensors and infrared cameras can identify areas needing irrigation, helping to optimize water use and reduce waste,
- Yield mapping: Drones can generate yield maps, assisting farmers in optimizing crop management and increasing overall production,

- Livestock monitoring: Drones equipped with cameras can be used to monitor livestock health and behaviour, as well as track animal locations,
- Crop monitoring: Drones equipped with sensors and cameras can collect real-time data on crop health, growth, and yield, creating crop health maps,
- Field mapping: Drones can create high-resolution field maps, providing data on soil structure, topography, and plant populations, which can be used for informed decision-making regarding planting, fertilization, and other crop management practices,
- Pest and disease control: Drones can help to detect and map pest and disease spread in crops, helping farmers take timely action.

These drone applications have gained significant research attention over the past five years. Studies provide evidence of the potential of drones in agriculture. However, these results remain a future vision that is not yet accessible to practitioners. Implementing such technologies requires the involvement of highly qualified specialists and the establishment of modern infrastructure.

2.3. Drone usage for crop monitoring

Drones are increasingly being used for crop monitoring. The most monitored crops are:

- Cereals: Frequently monitored during growth stages for yield prediction and disease detection (Boursianis et al., 2022),
- Fruits and Vegetables: Crops such as grapes, citrus fruits, apples, tomatoes, and potatoes are monitored to detect pests, diseases, and assess yield,
- Oil Crops (soybeans, sunflowers): Vegetation monitoring, plant health assessment, and yield prediction,
- Specific crops (coffee, tea, cocoa, and tobacco) are primarily monitored for early detection of diseases or pest infestations, as well as yield optimization.

However, drone usage in crop monitoring faces several limitations, as outlined in studies by (Zou et al., 2021), (Shahi et al., 2023), and (di Gennaro et al., 2016). First, drones can cover only a limited area per flight, making large-scale farm monitoring challenging. Second, weather conditions, particularly wind and rain, can impact drone operability, limiting data collection in unfavourable conditions. Third, drone operation requires skilled personnel and specialized equipment, which can be costly and time-consuming to maintain. Additionally, regulations vary by country; for example, in Latvia, drones must remain within the certified operator's line of sight.

Other factors influencing drone efficiency include crop height, density, size, weather conditions, and sensor limitations. Despite these challenges, drone systems provide farmers with valuable insights and data to optimize crop management, improve productivity, and reduce pesticide and fertilizer usage.

2.4. Summary

The analysis of related works indicates that the application of drones in horticulture should begin with plant vegetation monitoring, while agricultural technological operations should remain reliant on ground-based traditional equipment. This is primarily due to the relatively low payload capacity of drones compared to conventional agricultural machinery.

3. Proposed solution of technology

This section presents the authors' proposed technology for monitoring blackcurrant vegetation using simple drones and artificial intelligence methods. The proposed solution is applied in a blackcurrant farming operation in Latvia.

3.1. Informal description of solution

The technology is offered to blackcurrant growers for monitoring plantations using drones, enabling the detection of healthy blackcurrant plants, fungal diseases, and nutrient deficiencies. The system consists of three main stages:

- Mapping: Prepares maps of blackcurrant plantations with the required precision (scale), links them to GPS (Global Positioning System) coordinates, records drone flight routes, and specifies operations/photography to be performed during flights.
- Identification and segmentation: Uses trained neural networks to extract blackcurrant leaf clusters from the mapped images. Instance segmentation then identifies individual blackcurrant leaves, which are passed to the classification stage for further analysis.
- Classification: Uses trained neural networks to recognize healthy leaves, leaves affected by fungal diseases, and leaves indicating nutrient deficiencies.

As a result, the study provides a blackcurrant plantation analysis tool that gathers information on plantation conditions and visualizes it for growers, aiding decision-making regarding necessary interventions.

3.2. Mapping of horticultural areas

The area captured in drone images is usually significantly smaller than the cultivated field area in horticulture. This is determined by the technical parameters of the drone's camera and the scale of the captured images. As a result, field maps must be "stitched" together from individual drone images, which, when combined, form a complete field representation.

Figure 1. Example with combination of two images from neighbour fields that can’t be stitched together (images taken from height of 45 meters)

Additionally, stitching large image segments, as illustrated in Figure 1, can sometimes be of poor quality. The figure shows the merging of two images taken by a drone from a height of 45 meters, where visible discrepancies occur. These inconsistencies are caused by the technical limitations of image capture—differences in altitude and angles between images taken from different positions.

3.3. Selection and storage of image capturing routes

Since the horticultural field is created from several smaller images captured by the drone, additional actions need to be taken before capturing the image:

- marking the drone's starting point and determining the GPS coordinates;
- selecting the drone's flight route and image capture points so that, for example, using the *WebODM* (Web (a)), the images can be "stitched" together into field maps;
- the images must be captured efficiently, without interrupting the drone's flight;
- routes must be encoded and saved in .csv format for later use;
- the drone camera settings need to be adjusted to capture images with sufficient precision.

The route can be created using the *Mission Planner* program. It is necessary to create a route for the drone to fly and capture images that can later be stitched together. The overall image route is calculated based on the drone's flight altitude and camera parameters (e.g., the angle the camera captures) to ensure adequate overlap between images. As a result, a route will be obtained for the drone to follow, an example of which can be seen in the image prepared by the authors in Figure 2, which illustrates the mapping issues—low trees and uncovered areas.

The automatic flight is provided by the *Litchi* program. It is available both on the computer's website to create the route and on mobile phones to fly the route.

Figure 2. Routes where a drone could fly into a tree (image links) or not the entire area covered (image right)

3.4. Drone image processing

WebODM is used to merge drone flight images by detecting overlapping images using GPS coordinates, stitching them together, and correcting perspective distortions. This creates a cohesive aerial view for further analysis.

The image processing stage focuses on detecting and isolating individual blackcurrant leaf instances through instance segmentation and subsequent identification methods. Instance segmentation algorithms are applied to drone images to accurately separate each leaf from the background and other plant structures. Each leaf instance is identified separately, which enables detailed classification in the next stage. Segmentation and identification are crucial for preparing high-quality data for precision agriculture.

Trained AI models are used for segmentation and identification, fine-tuned specifically for blackcurrant leaf detection. Fine-tuning enhances the models' ability to recognize unique leaf structures and subtle variations, improving classification accuracy. This customization ensures high-precision results in identifying leaves and detecting plant health conditions. This pipeline enables an efficient and automated leaf classification system, supporting precision agriculture applications.

3.5. Identification

For the identification step, a training and validation dataset was created, consisting of 102 images (Figure 3). Each image was annotated using the open-source web-based annotation tool *CVAT* (WEB (b)) The dataset annotations were exported in YOLO (*You Only Look Once*) and COCO formats to support various deep learning frameworks.

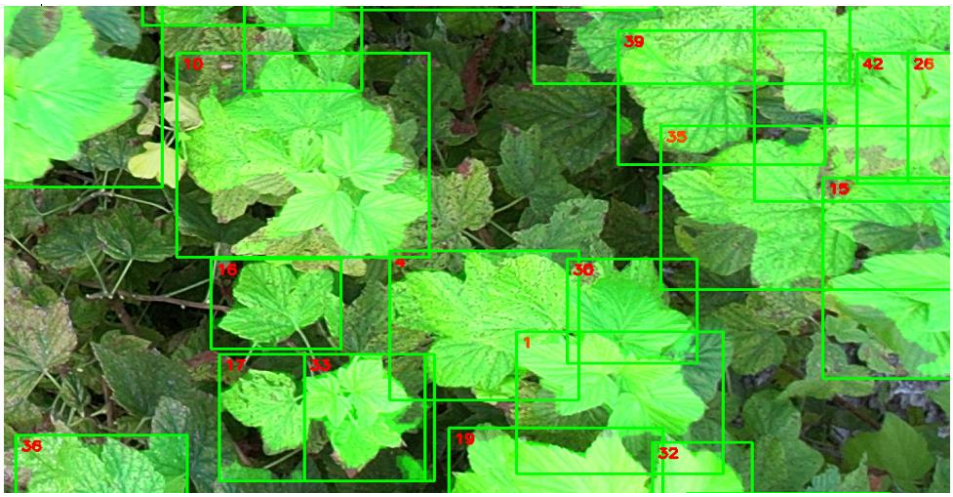


Figure 3. Leaf cluster marking in the field image

Several models were considered for analysing the images obtained in the identification step, focusing on accuracy, speed, and resource efficiency. The following models were selected for training: Faster R-CNN, ResNet50 FPN, Faster R-CNN X101-FPN, RetinaNet R101, YOLOv8x, YOLOv9e, and YOLOv10x. This model selection was based

on proven performance in object detection and instance recognition across various datasets.

Faster R-CNN models (ResNet50 FPN, X101-FPN) are known for high precision, particularly in complex scenes. Feature Pyramid Network (FPN) enables the model to analyse objects at multiple scales effectively.

RetinaNet R101 was chosen for its effectiveness in handling imbalanced datasets. It uses focal loss, which improves the detection of less frequent classes by reducing the influence of dominant classes.

YOLO series models (YOLOv8x, YOLOv9e, YOLOv10x) have high-speed performance, making them ideal for real-time applications in resource-constrained environments while maintaining strong identification performance.

This diverse model combination allows for a comprehensive comparison of accuracy vs. performance trade-offs, ensuring the most optimal identification model is selected.

3.6. Instance segmentation

The segmentation step allows extracting blackcurrant leaves from the leaf clusters identified in the previous identification step. Instance segmentation is essential to achieving high accuracy by isolating each leaf as a separate object within the image. By precisely delineating each leaf instance, a higher level of granularity and detail is ensured in the subsequent classification process. This approach enables algorithms to analyse each leaf individually, thereby improving classification accuracy by accounting for subtle differences such as leaf shape, size, and potential disease symptoms. Instance segmentation plays a crucial role in ensuring high-quality data acquisition and precise result interpretation.

A dedicated instance segmentation dataset was created using images captured with a Nikon D3300 DSLR camera equipped with a 24.2 MP DX-format CMOS sensor. The images were taken in the same blackcurrant fields where automated drone missions were conducted. The dataset includes images of blackcurrant leaves from various angles to enhance diversity and improve model training by simulating different perspectives. A total of 87 annotated images were compiled, with annotations created using the same tool as the identification dataset—CVAT (Computer Vision Annotation Tool). The dataset contains a single object type: "blackcurrant leaf," with each object annotated using segmentation mask contour points.

Several popular segmentation models were considered, with an in-depth analysis of YOLO models (YOLOv5-seg, YOLOv7-seg, YOLOv8-seg, YOLOv9-seg), as referenced in (Oditis et al., 2025). Additionally, the SAM model family (SAM, SAM2) and Mask R-CNN were evaluated theoretically based on literature sources (Chegini, 2023).). Each model offers different approaches to instance segmentation, with unique advantages and limitations.

YOLO models, known for their speed and efficiency, making them ideal for real-time applications like video surveillance and robotics (Oditis et al., 2025). However, their accuracy in complex segmentation tasks may be lower compared to more sophisticated models.

SAM (Segment Anything Model) family (SAM, SAM2) offers general-purpose segmentation, capable of segmenting any object with minimal input (e.g., a point or bounding box). It does not require task-specific training data, making it highly versatile. However, it lacks real-time processing speed.

Mask R-CNN is well-known for its high accuracy, especially in detecting overlapping and complex objects. Heavy computational requirements make it less suitable for real-time applications. Mask R-CNN is best suited for precision tasks like medical image analysis or autonomous driving.

A summarized model comparison is provided in Table 1. Based on the analysis, the YOLO model family was selected for instance segmentation, and further training will be conducted to determine the most optimal model for the task.

Although Mask R-CNN demonstrates higher accuracy, the specific requirements of the task call for instance segmentation of a single object type, prioritizing the solution's speed. This decision is based on the fact that a single field image covers approximately 260 sectors to be analysed, which require instance segmentation. Additionally, when surveying one hectare of field, about 7,500 images are obtained, meaning that instance segmentation must be performed on approximately 1.9 million sectors per hectare. These data strongly suggest the use of the YOLO solution for further analysis.

Table 1. Comparison of Instance Segmentation Models

Model	**Strengths**	**Weaknesses**	**Suitable Applications**	**Real-time Performance**
YOLO (YOLOv8-seg, YOLOv9-seg)	Fast, efficient for real-time tasks	Lower accuracy in complex segmentation tasks	Video surveillance, robotics	Excellent for real-time tasks
SAM (SAM, SAM2)	Universal object segmentation with minimal input	Lacks real-time processing speed	General segmentation tasks	Not suitable for real-time tasks
Mask R-CNN	High accuracy, especially for segmenting overlapping and complex objects	Resource-intensive, slow, not suitable for real-time tasks	Medical image analysis, autonomous driving	Low real-time performance but excellent accuracy

3.7. Classification

This chapter describes the blackcurrant leaf classification process and the conceptually chosen solutions. The classification process involves multi-class data classification, utilizing pre-trained models adapted to the specific dataset. These models are trained on pre-processed datasets to ensure accuracy and efficiency in blackcurrant leaf analysis. The use of artificial intelligence tools provides a generalized solution adaptable to various data types and classification criteria.

By analysing drone-acquired field images and evaluating blackcurrant bushes, three leaf classes were identified: “Healthy Leaf”, “Leaf with Nutrient Deficiency”, “Leaf with Fungal Disease”. It was also observed that some leaf instances exhibit characteristics of multiple classes, making multi-class classification necessary. In this chapter, the only object type under consideration is the blackcurrant leaf. This classification structure is sufficient to test the effectiveness of the selected approach. If needed, the set of classes

can be expanded without altering the classification process, implementation, or planned solution. The class definitions were determined with input from a domain expert.

The training and validation dataset was collected from the same blackcurrant fields where automated drone missions were conducted. After image acquisition, data annotation was performed, where leaf instances were manually segmented and assigned to their respective classes. The dataset consists of:

- 118 images of leaves classified as "Healthy Leaf",
- 109 images of leaves classified as "Leaf with Nutrient Deficiency",
- 102 images of leaves classified as "Leaf with Fungal Disease",
- 57 images containing instances belonging to multiple classes.

(See Figure 4 for reference.)

Figure 4. Leaf examples: fungal disease, healthy leaf, leaf with nutrient deficiency

For the classification task, three families of artificial intelligence models popular for multi-class image classification were examined:

- ResNet, including ResNet50, ResNet101, and ResNet152;
- EfficientNet covering models from EfficientNet-B0 to EfficientNet-B7;
- VGGNet, including VGG16 and VGG19.

ResNet (Residual Network) architectures are designed to address issues in deep neural networks, such as the vanishing gradient problem. They use "skip" connections, allowing information to bypass certain layers, enabling the training of very deep networks without performance degradation, which is common in traditional deep networks .

EfficientNet is a highly efficient neural network architecture optimized for both accuracy and resource usage. It employs a compound scaling approach, simultaneously adjusting the network's width, depth, and resolution to enhance performance.

VGGNet is a classical deep neural network for image classification, known for its simple and structured architecture. It primarily relies on convolutional layers with 3x3 filters. However, this design results in high computational complexity and memory requirements, making it slower compared to more modern architectures (Simonyan, 2015).

For the classification process, the ResNet family was selected, with ResNet50, ResNet100, and ResNet512 models considered during training. This choice was based on ResNet’s high accuracy and performance, which surpass those of the VGGNet models. From the comparison in Table 2, it is evident that ResNet effectively mitigates the gradient vanishing problem, allowing the training of deeper networks without performance loss.

While EfficientNet is highly efficient in resource utilization, ResNet provides an optimal balance between accuracy and speed, which is crucial for blackcurrant leaf classification, where high reliability and processing speed are required. These advantages make ResNet the most suitable model family for successfully executing the given task.

Table 2. Comparison of Classification Model Families

Model family	**Strengths**	**Weaknesses**	**Application areas**	**Efficiency / Resource usage**
ResNet	Solves the gradient vanishing problem in deep networks using "skip" connections	Deeper architecture may be challenging to train with small data	Image classification, data analysis, computer vision	Efficient in training deep networks without performance loss
EfficientNet	Optimizes both accuracy and resource usage through proportional scaling	Higher complexity in the optimization process	Mobile applications, AI solutions	High efficiency, low resource requirements
VGGNet	Simple architecture, clear and intuitive	High computational complexity, slower compared to modern models	Image classification, computer vision, early research applications	Low efficiency and high memory requirements
Mask R-CNN	High accuracy, especially in segmenting overlapping and complex objects	Resource-intensive, slow, not suitable for real-time tasks	Medical image analysis, autonomous driving	Low real-time performance, but excellent accuracy

3.8. Summary

The proposed technology utilizes several artificial intelligence methods for a specific application – monitoring the vegetation of blackcurrants. The selected methods proved to be sufficiently effective for this particular application.

4. Fields history

For every farm, it is beneficial to maintain a field record journal that logs all activities within a specific agricultural area, including completed horticultural operations, the use of crop materials, and plant protection products. Several information systems already exist to support such functionalities.

This study, however, focuses on collecting field images, offering a different perspective on historical data—both visually and through insights derived from image analysis. The goal of our research is to develop a solution that allows for visual tracking of field changes over time while also providing timely detection of plant health issues identified through image analysis.

4.1. Field surveying process

To ensure field history tracking:

- identify the surveyed fields (this information is used to plan drone flight and photography routes),
- conduct field surveys using drones to capture images and link them to specific geolocations,
- analyse the images to identify plant health issues,
- visualize the extracted information on a map for better interpretation and decision-making.

This type of solution is designed for horticulturists. Their main interest is tracking long-term changes in fields, especially in crop cultivation involving perennial plants such as blackcurrants. In such cases, even historical images taken from the same vantage point can provide valuable insights into field conditions, moisture levels, pest infestations, disease development, and more.

Although the functionalities may seem simple, this type of solution comes with certain technical challenges. First, to obtain images suitable for whole-field analysis, several hundred photos per hectare must be captured, making the imaging process time-consuming and requiring multiple drone flights. Second, storing historical images can be space-intensive, considering that blackcurrant plantations typically cover between 2 to 20 hectares. Third, analysing such a vast number of images is time-consuming, as it involves identifying and classifying thousands of leaves.

To address these technical challenges, it is important to recognize that a complete photographic record of the entire field is not necessary to assess the condition of blackcurrant plantations. A similar approach is used in soil fertility assessment, where sampling is conducted systematically at predetermined intervals—for example, every 20 meters. In this case, evaluating one hectare would require only 36 images, significantly reducing storage and analysis demands. This approach would still provide sufficiently representative information about the field's condition and the spread of potential issues.

4.2. Field browser functions

The field browser provides the following key functions:

- Field survey planning – defining field boundaries, preferably using the territorial division applied by the Rural Support Service.

- Drone flight planning – setting flight routes and photography points, including altitude and camera parameter configuration.
- Automated drone flight execution – performing flights autonomously and saving captured images.
- Image analysis – conducting segmentation, identification, and classification using a pre-trained neural network.
- Blackcurrant plantation condition monitoring – allowing selection of a specific field for evaluation, enables to view images in different resolutions and timeframes (field images can be viewed on different dates, switching between them changes the displayed field area and resolution).

A sample result of the field browser operation is shown in Figure 5. A map of a 5-hectare blackcurrant plantation was created using a drone, with a portion of the area undergoing in-depth analysis—including segmentation, identification, and classification. This analysis provides an overview of the plantation's condition, revealing healthy leaves – 97.4%, fungal disease presence – 1.7%, nutrient deficiency detected – 1.0% of the surveyed area. This assessment of the plantation's condition provides valuable insights for the agronomist, enabling informed decisions on necessary crop management actions.

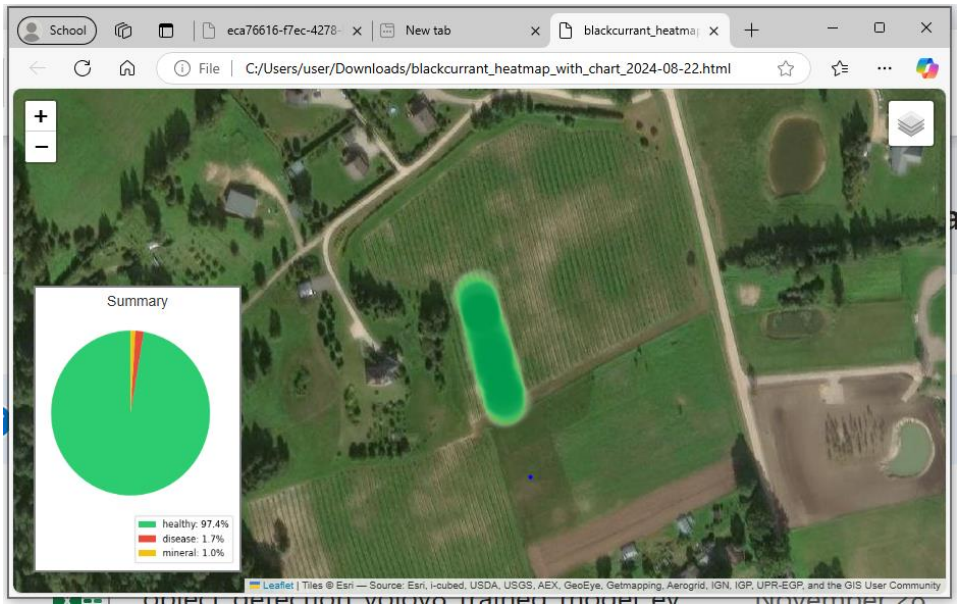


Figure 5. Blackcurrant Field Condition (Nutrient Deficiency: 1.0%, Fungal Disease: 1.7%, Healthy Leaves: 97.4%).

5. Discussion

The study results demonstrate new technologies for monitoring blackcurrant vegetation. Instead of traditional visual assessments by horticulturists, an automated system is proposed, offering several advantages: precise crop evaluation using AI methods,

applicability to large blackcurrant plantations, economic efficiency, and ease of implementation.

However, these achievements should be considered a first step toward precision agriculture in blackcurrant cultivation, requiring further development. The proposed approach relies on blackcurrant leaf analysis, which means it can only detect diseases affecting leaves, while issues affecting roots and stems—such as blackcurrant clearwing moth (*Synanthedon tipuliformis*) and blackcurrant bud mite (*Eriophyes ribis*)—remain undetected.

Similarly, yield prediction requires an alternative approach, possibly analysing entire blackcurrant bushes rather than just leaves. Pest infestations can also only be partially identified through leaf analysis.

Nevertheless, other blackcurrant cultivation challenges, such as crown rust, can be identified using similar methods by analysing different plant parts, segmenting them in images, and classifying them based on the specific problem being addressed.

5.1. Additional applications of the method

A diligent horticulturist monitors not only the spread of blackcurrant diseases but also frost damage, flowering progress, yield ripening time and volume predictions, and other vegetation-related events. Although these aspects were not the primary focus of this study, they could be addressed by modifying the proposed method—for example, by segmenting and classifying flower buds and berry clusters accordingly.

Beyond plant vegetation monitoring, the method can also be applied to optimizing agricultural operations. By identifying disease-affected field areas, maintenance tasks such as targeted spraying can be carried out only in infected regions. This would lead to significant savings in materials and labour resources.

A promising direction for further development is integrating the method into dynamic robotic management. By transferring real-time data from the blackcurrant plant analysis module to an agricultural operations execution robot, it would be possible to perform precise interventions only where necessary, further increasing efficiency and sustainability.

5.2. Limitations of the method’s application

When analysing the benefits of the proposed method, it is also important to highlight its limitations.

One key limitation is the lack of precision in determining nutrient deficiencies. While the method can detect a deficiency, it cannot specify which particular nutrient—potassium, phosphorus, or nitrogen—is lacking. Currently, this type of analysis is performed using soil and plant agrochemical testing, which involves manually collecting soil and leaf samples. This process requires significant labor resources.

A more advanced approach involves spectral analysis of plants, which can provide more precise nutrient deficiency diagnostics. However, this requires more complex imaging cameras and advanced processing methods, which are not yet widely available due to high costs and a lack of specialists.

Additionally, leaves are just one indicator of plant health, but they do not reveal all potential issues. For example, pest infestations, such as the blackcurrant clearwing moth, which primarily affects the stems rather than the leaves, cannot be detected using this method. Furthermore, yield estimation—a crucial aspect from an economic perspective—

is not covered by this approach. Addressing these gaps requires further research and development in this field.

6. Conclusions

Key conclusions of the conducted research and its application results:

1. Integration of Drones and AI is Effective for Precision Agriculture. The proposed system successfully integrates drone-based imaging with artificial intelligence methods to monitor blackcurrant plantations. This approach enables accurate, large-scale assessment of plant health conditions while reducing the need for manual inspection.
2. Modular Architecture Ensures Flexibility and Scalability. The system's modular structure—mapping, identification and segmentation, and classification—allows for flexibility in adapting the pipeline to different crops or environmental conditions. Each module can be fine-tuned or replaced independently to improve performance.
3. Mapping functions can be implemented using standard solutions available in commercial drone systems or through adapting open-source solutions for blackcurrant cultivation. Identification and segmentation solutions must be developed individually in collaboration with industry experts—in this project's case, blackcurrant growers. This includes training a neural network for leaf recognition and transmitting the identified leaf information for classification. The classification task involves training a neural network to recognize specific characteristics of blackcurrant leaves, which is the project's final goal and of interest to horticulturists.
4. YOLO Models Balance Speed and Accuracy for High-Volume Analysis. While models like Mask R-CNN offer higher segmentation accuracy, the YOLO model family was selected for its superior processing speed, making it more suitable for handling the large volume of images required in agricultural drone surveys.
5. ResNet Models Provide Robust Classification Capabilities. The ResNet family of models proved optimal for classifying blackcurrant leaves due to their ability to train deep networks efficiently. Their balance of accuracy and computational efficiency makes them suitable for real-world deployment in agricultural settings.
6. System Supports Informed Decision-Making for Growers. By providing detailed visualizations and health assessments of blackcurrant plantations, the system aids growers in making timely decisions regarding interventions such as fertilization or disease management.
7. Field-Validated Data Collection Enhances Reliability. The use of annotated datasets created from real blackcurrant fields ensures that the models are trained and validated with realistic, domain-specific data, improving the accuracy and practical applicability of the system.
8. Potential for Wider Application. Although developed specifically for blackcurrants, the solution demonstrates potential for adaptation to other types of crops and agricultural monitoring tasks, supporting broader applications in precision horticulture.

Following the identification of fungal diseases and nutrient deficiencies, potential future applications include yield prediction, pest detection, and identification of other

plant diseases. A key challenge remains improving the speed of image analysis, as computational demands may exceed the capabilities of simple and inexpensive hardware.

Acknowledgment

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

Abbasi, R., Martinez, P., Ahmad, R. (2022). The digitization of agricultural industry—A systematic literature review on agriculture 4.0. *Smart Agric. Technol.* 2022, **2**, 100042.

Anand, K., Goutam, R. (2019). An autonomous UAV for pesticides praying. *Int. J. Trend Sci. Res. Dev*., **3**, 986–990

Bertalan, L., Holb, I., Pataki, A., Negyesi, G., Szabo, G., Kupasne Szaloki, A., Szabo, S. (2022). UAV-based multispectral and thermal cameras to predict soil water content—A machine learning approach. *Comput. Electron. Agric.*, **2,** 107262.

Bilotta, G., Genovese, E., Citroni, R., Cotroneo, F., Meduri, G.M., Barrile, V. (2023). Integration of an Innovative Atmospheric Forecasting Simulator and Remote Sensing Data into a Geographical Information System in the Frame of Agriculture 4.0 Concept. *AgriEngineering*, **5**, 1280–1301

Botta, A., Cavallone, P., Baglieri, L., Colucci, G., Tagliavini, L., Quaglia, G. (2022) A Review of Robots, Perception, and Tasks in Precision Agriculture. *Appl. Mech.*, *3*(3), pp. 830-854; https://doi.org/10.3390/applmech3030049

Boursianis, A. D., Papadopoulou, M. S., Diamantoulakis, P., LiopaTsakalidi, A., Barouchas, P., Salahas, G., Karagiannidis, G., Wan, S., Goudos, S. K., (2022). Internet of Things (IoT) and Agricultural Unmanned Aerial Vehicles (UAVs) in smart farming: A comprehensive review. *Internet of Things* (Netherlands), **18**, 100187, 22. https://doi.org/10.1016/j.iot.2020.100187

Chegini, H. (2023). A Deep Comparison on two Deep Learning Models: SAM and MaskRCNN. Retrieved 2025.05.06 from https://medium.com/@h.chegini/a-deep-comparison-on-two-deep-learning-models-sam-and-maskrcnn-176eee1d1103

Chen, P., Ouyang, F., Zhang, Y., Lan, Y. (2023). Preliminary Evaluation of Spraying Quality of Multi-Unmanned Aerial Vehicle (UAV) Close Formation Spraying. *Agriculture 2023*, **12**, 1149

Dayioglu, M.A., Turker, U. (2021). Digital transformation for sustainable future agriculture 4.0: A review. *J. Agric. Sci*., **27**, 373–399.

Dhanaraju, M., Chenniappan, P., Ramalingam, K., Pazhanivelan, S., Kaliaperumal, R. (2022). Internet of Things (IoT)-Based Sustainable Agriculture. *Agriculture 2022*, **12**, 1745.

di Gennaro S. F., Battiston E., di Marco S., Facini O., Matese A., Nocentini M., Palliotti A., and Mugnai L., (2016). Unmanned Aerial Vehicle (UAV)-based remote sensing to monitor grapevine leaf stripe disease within a vineyard affected by esca complex, *Phytopathologia Mediterranea*, **55(**2), pp. 262–275, 2016

Drone (2025). The free Dictionary by Farlex. Retrieved 2025.05.06 from https://www.thefreedictionary.com/drone

Huuskonen, J., Oksanen, T. (2018). Soil sampling with drones and augmented reality in precision agriculture. *Comput. Electron. Agric.*, **154**, pp. 25–35.

Ivič, S., Andrejčuk, A., Družeta, S. (2019). Autonomous control for multi-agent non-uniform spraying. *Appl. Soft Comput.*, **80**, pp. 742–760.

Koh, J.C.O., Hayden, M., Daetwyler, H. (2019). Estimation of crop plant density at early mixed growth stages using UAV imagery. *Plant Methods*, **15**, 64.

Kumar, C., Mubvumba, P., Huang, Y., Dhillon, J., Reddy, K. (2023). Multi-Stage Corn Yield Prediction Using High-Resolution UAV Multispectral Data and Machine Learning Models. *Agronomy*, **13**, 1277

Madushanki, R., Halgamuge, M., Wirasagoda, S., Syed, A. (2019). Adoption of the Internet of Things (IoT) in Agriculture and Smart Farming towards Urban Greening: A Review. *Int. J. Adv. Comput. Sci. Appl.*, **10**, pp. 11–28.

Mail, M.F., Maja, J.M., Marshall, M., Cutulle, M., Miller, G., Barnes, E. (2023). Agricultural Harvesting Robot Concept Design and System Components: A Review. *AgriEngineering 2023*, **5**, pp. 777–800./ Retrieved 2025.05.06 from https://play.google.com/store/apps/details?id=com.aryuthere.visionplus

Mohidem, N.A., Che Ya, N.N., Juraimi, A.S., Fazlil Ilahi, W.F., Mohd Roslim, M.H., Sulaiman, N., Saberioon, M., Mohd Noor, N. (2021). How Can Unmanned Aerial Vehicles Be Used for Detecting Weeds in Agricultural Fields? *Agriculture 2021*, **11**, 1004.

Oditis R., Oditis I., Freivalds K., Bicevskis J. (2025). Blackcurrant leaf analysis using instance segmentation and multi-class classification. *Baltic J. Modern Computing*, **13**(2), 486–501

Oliveira, R.C.d., Silva, R.D.d.S.e. (2023). Artificial Intelligence in Agriculture: Benefits, Challenges, and Trends. *Appl. Sci. 2023, 13*, 7405.

Ojo, M.O., Zahid, A. (2022). Deep Learning in Controlled Environment Agriculture: A Review of Recent Advancements, Challenges and Prospects. *Sensors 2022*, **22**, 7965.

Ong, P., Teo, K.S., Sia, C.K. (2023). UAV-based weed detection in Chinese cabbage using deep learning. *Smart Agric. Technol. 2023*, **4**, 100181.

Rejeb, A., Abdollahi, A., Rejeb, K., Treiblmaier, H. (2022). Drones in agriculture: A review and bibliometric analysis. *Comput. Electron. Agric.* 2022, **198**, 107017.

Simonyan K. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. Retrieved 2025.05.06 from https://arxiv.org/abs/1409.1556

Sinha, J.P. (2020). Aerial robot for smart farming and enhancing farmers' net benefit. *Indian J. Agric. Sci.* 2020, **90**, pp. 258–267.

Shahi, B., Xu, C.Y., Neupane, A., Fleischfresser, D., O'Connor, D., Wright, G., Guo, W. (2023). Peanut yield prediction with UAV multispectral imagery using a cooperative machine learning approach. *Electron. Res. Arch*., **31**, pp. 3343–3361.

Song, C., Liu, L., Wang, G., Han, J., Zhang, T., Lan, Y. (2023). Particle Deposition Distribution of Multi-Rotor UAV-Based Fertilize Spreader under Different Height and Speed Parameters. *Drones*, **7**, 425.

Su, D., Yao, W., Yu, F., Liu, Y., Zheng, Z., Wang, Y., Xu, T., Chen, C. (2019). Single-Neuron PID UAV Variable Fertilizer Application Control System Based on a Weighted Coefficient Learning Correction. *Agriculture 2022*, **12**, 1019.

Tetila, E.C., Machado, B.B., Astolfi, G., de Souza Belete, N.A., Amorim, W.P., Roel, A.R., Pistori, H. (2020). Detection and classification of soybean pests using deep learning with UAV images. *Comput. Electron. Agric. 2020*, **179**, 105836.

Uzhinskiy, A. (2023) Advanced Technologies and Artificial Intelligence in Agriculture. *AppliedMath*, **3**(4), pp. 799-813; https://doi.org/10.3390/appliedmath3040043

Wang, T., Chen, B., Zhang, Z., Li, H., Zhang, M. (2022). Applications of machine vision in agricultural robot navigation: A review. *Comput. Electron. Agric. 2022*, **198**, 107085.

WEB (a) Drone Mapping Software. Retrieved 2025.05.06 from https://www.opendronemap.org/webodm/

WEB (b) Computer Vision Annotation Tool. Retrieved 2025.05.06 from https://venturebeat.com/ai/intel-open-sources-cvat-a-toolkit-for-data-labeling/

Wilke, N., Siegmann, B., Postma, J., Muller, O., Krieger, V., Pude, R., Rascher, U. (2021). Assessment of plant density for barley and wheat using UAV multispectral imagery for high-throughput field phenotyping. *Comput. Electron. Agric. 2021, 189*, 106380.

Xu, J., Gu, B., Tian, G. (2022). Review of agricultural IoT technology. *Artif. Intell. Agric*. 2022, **6**, pp. 10–22.

Zeng, L., Peng, G., Meng, R., Man, J., Li, W., Xu, B., Lu, Z., Sun, R. (2021). Wheat Yield Prediction Based on Unmanned Aerial Vehicles-Collected Red–Green–Blue Imagery. *Remote Sens.* 2021, **13**, 2937.

Zhou, J., Xu, Y., Gu, X., Chen, T., Sun, Q., Zhang, S., Pan, Y. (2023). High-Precision Mapping of Soil Organic Matter Based on UAV Imagery Using Machine Learning Algorithms. *Drones* 2023, **7**, 290.

Zou, K., Chen, X., Zhang, F., Zhou, H., Zhang, C. (2021). A field weed density evaluation method based on uav imaging and modified u-net, *Remote Sensing*, **13(**2), pp. 1–19. Available: https://doi.org/10.3390/rs13020310

.

Received February 19, 2025, revised September 1, 2025, accepted October 14, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, 758-777
https://doi.org/10.22364/bjmc.2025.13.4.02

Data-driven Simulation in Transportation Management through Cross-Sectoral Collaboration

Galina MERKURJEVA[1], Jurijs MERKURJEVS[1],
Andrejs ROMANOVS[1], Vitalijs BOLSAKOVS[1], Rolands FELDMANIS[2]

[1]Institute of Information Technology, Riga Technical University, Kipsalas Street 6A, Riga, Latvia
[2]Latvian Association of Agricultural Cooperatives, Republikas Square 2, Riga, Latvia

Galina.Merkurjeva@rtu.lv, Jurijs.Merkurjevs@rtu.lv,
Andrejs.Romanovs@rtu.lv, Vitalijs.Bolsakovs@rtu.lv,
Rolands.Feldmanis@llka.lv

ORCID 0000-0002-6710-5128, ORCID 0000-0001-7178-5640, ORCID 0000-0003-1645-2741,
ORCID 0000-0003-0540-5317, ORCID 0009-0001-1794-0563

Abstract: This article proposes an innovative integrated framework and research methodology for data-driven simulation in transportation management through cross-sectoral collaboration. The presented study is based on a real case of synergy in the transportation of products in two sectors - agriculture and forestry. The purpose of this study is to simulate the synergistic improvement of the planning and organization of transportation of agricultural and forest products by small businesses. The benefit for participants in both economic sectors is to ensure a more efficient use of their transport and labor resources by using the opportunities of intersectoral cooperation and emergent data driven-modeling and simulation technologies. The data-driven models such as symbolic regression are used to identify patterns in data and translate the underlying relationships into the modeling formalisms needed to build computer simulation models. The proposed study was tested in real-life conditions.

Keywords: Agriculture; Forestry; Cross-sector Collaboration, Data-driven Modeling; Web-based Simulation; Transportation Management.

1. Introduction

Modern digital platforms are radically transforming business models and changing inter-industry and inter-company relationships, opening up new areas of cooperation today (Veile et al., 2022). Simulation technologies provide an experimental approach to exploring the potential of cross-sectoral collaboration in transportation management and to developing and planning further transportation decisions (Castañer and Oliveira, 2020). Collaboration is a necessary prerequisite for synergy in the transportation of agricultural and forest products, which expands the collaboration process and takes the results beyond simple cooperation. Moreover, rapidly evolving data-driven computing technologies enable companies to realize the value of data and exploit the opportunities supported by data.

The objective of the presented study is to develop an integrated approach and methodology for data-driven modeling and simulation in transportation management

through cross-sectoral collaboration. It is designed to process and interpret relevant data from various business areas (e.g. operations, economics) and provide meaningful and cost-effective analysis of the most important decisions.

Various methods and tools, including data science and modeling technologies, are considered and employed to model and analyze this synergy. The experimental part of the study was carried out using the example of synergy in the transportation of products from two industries – agriculture and forestry. In this example, the company is engaged in managing the transportation of not only agricultural or forest products separately, but both of them. The benefits of improved planning and organization of transportation of agricultural and forest products for small enterprises and farms are that it ensures a more efficient use of their transport and labor resources (in particular, vehicle drivers) by using the opportunities of inter-sectoral interaction and advanced computing technologies.

The modeling concept and research methodology are based on the well-established simulation-based approach to applications in logistics and supply chain management. In particular, practical examples of various case studies using the simulation approach to solve complex logistics problems can be found in (Merkuryev et al, 2009). This paper advances that methodology by integrating data-driven modeling and web-based simulation technologies. It also applies to a new application area, namely the planning and organization of transport across two industrial sectors.

Data-driven modeling involves using historical and other data to create models that identify trends and patterns in the data, and represent data in a structural form (Habib and Ayankoso, 2021). With the advancement of computational intelligence and machine learning techniques, as well as the wide availability of accumulated data, the use of data-driven models has increased in different application areas (Belmont Guerrón and Hallo, 2022; Merkuryeva, 2024).

Data-driven simulation is an emerging trend in the development of computer simulation technologies. It is based on the integration of different data models, complemented by computational intelligence and machine learning methods, changing the simulation from a model-based paradigm to data-driven one (Mütsch et al., 2023). In this study, data-driven modeling provides a data basis for guiding and controlling computer simulations throughout its life cycle, starting from defining the model structure and parameters for different experimental scenarios.

In the field of transport logistics in agricultural and forestry industrial sectors, one of the strongest influencing factors is the seasonality of demand for transportation. The seasonality of demand for agricultural transportation naturally depends on the harvest time, whereas the seasonality of demand for forestry transportation is associated with greater accessibility of forest roads during cold periods. This seasonality of demand presents important economic challenges. During busy seasons, large amounts of specialized and expensive resources are required in the short term and there is usually a shortage of skilled labor. At the same time, during low seasons, these logistics companies have idle resources and are forced to pay high wages to unemployed skilled workers in order to retain them for the next season.

Currently, small agricultural and forestry companies in developing EU countries are typically focused only on transporting products from a specific sector (agriculture or forestry) and have a conservative attitude towards collaboration opportunities. These businesses usually do not consider the possibility of sharing skilled labor and transportation resources in order to improve their economic efficiency. At the same time, small businesses, as a rule, have a fairly low level of digitalization and are not sufficiently equipped with modern information and communication technologies. Thus there is room

for improvement through mutual collaboration of these businesses by using innovative data technology and data-driven modeling.

The development of innovative solutions to improve the planning and organization of transportation of agricultural and forest products by expanding cooperation between small enterprises was carried out within the framework of a research project, in which representatives of the industry from two sectors took part (Bolsakov et al., 2024). This project focused on small agricultural, forestry and transport companies. Consequently, the presented research methodology was tested in real-life conditions.

The following section provides a review of the literature on modeling and organization aspects of transport and logistics services in agriculture and forestry. Section 3 introduces the conceptual framework of the study and research methodology, including web-based data management, data-driven modeling, system dynamics modeling, stochastic discrete-event simulation, and a multi-user web environment. Section 4 describes the experimental part of the study and discusses the obtained experimental results. The conclusions of the study are given in Section 5.

2. Literature Review

An analysis of the organization of transport services and supplies in small enterprises in forestry and agriculture was conducted using the example of Latvia. Considering the seasonal demand for transportation in both sectors, it was found that the planning and organization of logistics in such enterprises could certainly be improved by sharing available vehicles and labor to transport produce as needed. Such a collaborative organization of transport logistics at SME level between different industrial sectors must not only be technically and economically justified, but also take into account all current relevant theoretical and technological developments. In addition, a comprehensive analysis of existing and new business processes in practice requires consideration of a large volume of detailed transport documentation and logistics data, followed by modeling possible scenarios for planning and organizing transportation.

To improve transport logistics in the forestry sector, both analytical and algorithmic models have been studied in (Alonso-Ayuso et al., 2020; Troncoso and Garrido, 2005; Sfeir et al., 2019; Alayet et al., 2018). The proposed models often do not take into account the factor of seasonal demand fluctuations when offering new solutions for the optimal use of existing transportation resources.

The literature suggests many factors directly and indirectly affecting forestry transportation companies as well as the environment. For example, the model proposed in (Mathur and Warner, 1997) takes into account direct transportation costs, as well as indirect costs such as investments in public infrastructure and financial assessment of the impact on the environment and ecology. The impact on the labor force and the labor market is assessed in (Boukherroub et al., 2013). Furthermore, a hybrid multi-criteria model for assessing the effectiveness of sustainable management of forest enterprises from economic, social and environmental points of view is proposed in (Deng et al., 2023).

Often, the proposed models (Boukherroub et al., 2013; Walsh et al., 2003; Bajgiran et al., 2016) do not focus only on the transportation process itself, but more specifically on the supply chain management of all products involved. This enables efficient planning and organization of the work processes of the participating enterprises for the benefit of all of them. In particular, models with cost uncertainty are described in (Walsh et al., 2003) and models with shared transport resources are presented in (Francois et al., 2017).

For modeling such a socio-techno-economic system, analytical models are often used (Buongiorno, 1996; Bajgiran et al., 2016; Francois et al., 2017; Oke et al., 2018). To solve various transportation problems, simulation models that imitate physical processes and allow computer experiments to be carried out with them are proposed (as, for example, in (Walsh et al., 2003; Boukherroub et al., 2013)).

The literature on logistics modeling in the agricultural sector mostly carries out process analysis. Priority is given to the transportation of agricultural products from farms to warehouses and further in the logistics chain to processing or transshipment points. Both system dynamics and discrete-event simulation models are used to analyze the transportation of agricultural products. Examples of several models based on system dynamics and their comparison are given in (Oliveira et al., 2022; Chen et al., 2022). Examples of discrete-event simulation models for the analysis of farm-scale grain transportation systems and the modeling of grain logistics from farms to ports are described in (Fioroni et al., 2015; Turner et al., 2019).

There are also studies in the literature, including quite recent ones, that use well-known discrete-event modeling formalism (Cavone et al., 2017), such as Petri nets, to analyze supply and production chains in the agricultural and forestry sectors. Hybrid Petri net models have been used to assess the logistic efficiency of forestry production chains (Cardoso et al., 2009) and to plan agricultural work processes under uncertainty (Ozgun and Kirci, 2015; Guan et al., 2008), and colored Petri nets were applied to model and analyze agricultural supplies to the European Union (Pavlenko et al., 2020).

In research on the organization of agricultural production, much attention is also paid to analytical models for ensuring the efficiency of logistics services. A large number of articles, for example (Xiao and Lang, 2009; Lamsal et al., 2016; Mehmann and Teuteberg, 2016; Mogale et al., 2019; Mardaneh et al., 2021; Trunina et al., 2021; Sgurev et al., 2022), present linear programming models for optimizing the transportation of agricultural products. Although there are quite effective methods for solving such optimization problems, these models contain many assumptions and simplifications that usually do not take into account the volatility and/or seasonality of demand. The use of complex optimization algorithms to organize efficient logistics is discussed in (Xiao and Lang, 2009; Lopez and Qassim, 2023), but these algorithms are always specific to certain transportation factors.

Although the linear programming models mentioned normally do not take the factor of seasonal demand into account, the fact that demand for agricultural products is highly variable, unstable and difficult to predict is one of the most important aspects of transport logistics in the agricultural sector mentioned in literature. There are studies covering both forecasting and modeling of demand for agricultural logistics services (e.g., (Li and Lu, 2015)) and transport models that take into account particularly short periods of high demand for agricultural crops (Sgurev et al., 2022).

Various key performance indicators (KPI) have been identified that need to be analyzed when organizing agricultural transportation. Cost efficiency (Diaz and Perez, 2000; Mehmann and Teuteberg, 2016; Nourbakhsh et al., 2016; Lomotko et al., 2019; Mardaneh et al., 2021) is the KPI most frequently selected, followed by energy consumption (Trunina et al., 2021) and supply chain efficiency (Nourbakhsh et al., 2016).

Collaborative data collection, analysis and processing for decision support in the agricultural and forestry industries are discussed implicitly or explicitly in several research papers (Zhu et al., 2009; Mehmann and Teuteberg, 2016; Gupta and Garima, 2017; Mardaneh et al., 2021; Oliveira et al., 2022). Although the methods proposed in these studies differ, the common point is that digitalization of agricultural and forestry transport companies is relevant and important in order to increase the efficiency of these companies

and to improve their logistics processes. One of the widely recommended methods is the accumulation of process data in order to develop accurate and reliable models for further optimization of business processes.

The possibility of using a resource sharing approach in transportation is also being considered. In particular, relevant qualitative and quantitative productivity estimates based on an extensive study of the forestry sector are given in (Palatova et al., 2023). The impact of such resource sharing on sustainable economic development in the agricultural sector is analyzed in (Lin et al., 2022). However, the possibility of sharing resources between agricultural and forestry companies to create a flexible and diversified transport organization in such a case has not yet been widely discussed. In fact, the intersectoral cooperation and synergies between agriculture and forestry (Noordwijk et al, 2018) are analyzed mainly from the perspective of environmental sustainability.

3. Research Methodology

3.1. Rationale

Logistics and transportation of agricultural and forest products together constitute one fifth of all freight on Latvian roads (National Statistical System of Latvia, 2024). The current situation, where both agricultural and forestry transportation companies may suffer from seasonal factors and do not support each other, is due to several factors. Different technical specifications for the trucks, trailers and semi-trailers used in the transportation processes are the main cause. Additionally, transporting timber products, such as unprocessed logs, may also require additional skills from drivers as they will have to load the vehicle themselves.

Latvian forestry companies use three-axle trucks with three-axle trailers. These are also equipped with an hydraulic loader for cutting or open storage. Agricultural transport companies on the other hand use tractors with semi-trailers to transport bulk cargo, such as grain, fertilizers or peat. They do not require self-loading, as they are usually loaded from an elevator or agricultural machinery.

The agricultural enterprises may be interested in transporting timber during the winter period. Some even tried to do this with minor modifications to semi-trailers, but these were isolated cases. To provide agricultural trucks with the ability to transport timber, specialized semi-trailers equipped with an hydraulic loader can be used. By renting or purchasing such a semi-trailer, agricultural enterprises can ensure the loading and transportation of timber logs using the existing capacities and drivers when they stand idle during the off-season.

Furthermore, it can be envisaged that small businesses and farmers will be equipped with digital technologies and tools enabling the proper collection and analysis of relevant data for effective planning and organization of transportation of agricultural and forest products, thus taking advantage of the opportunities for cooperation between the two sectors.

3.2. Conceptual research framework

Simulation is widely recognized in literature and in practice as an effective and flexible approach to design efficient logistics and supply chain management. It should be noted that this approach can only provide reliable experimental results and estimates if the initial modeling data are reliable and if the model itself is based on a clearly formalized system

structure and well defined processes. Preparing input data for small businesses in a target application can be challenging due to the lack of access to many of these data. The application of data mining using powerful machine learning methods to understand and interpret the available data and their integration with simulation technologies appears to be a quite reasonable solution for this problem in the context of this study.

The idea of integrating data science and modeling technologies is not new to either management or logistics. For example, integrated solutions based on cluster analysis and simulation-based optimization have been introduced for planning and scheduling product deliveries to a distribution center (Merkuryeva, 2012). The integration approach was subsequently supplemented by fitness landscape analysis (Merkuryeva and Bolshakov, 2015) and extended by applying computational intelligence methods to simulation-based planning and optimization in multi-echelon supply chains (Merkuryeva et al., 2011).

In our study the modeling concept and research methodology are enhanced by integrating data-driven modeling and web-based simulation. Extracting useful information from historical and operational data of a real system, identifying dependencies within these datasets, and transforming them into modeling formalisms (e.g., formulas) enables the definition of the structure and parameters of simulation models for subsequent experiments. In this approach, the simulation is dependent on and driven by the actual behavior of the system.

The proposed conceptual framework (Fig. 1) describes the principal components of the research methodology, as well as the interrelationships and data flows among them. It provides a structured foundation for the study, ensuring alignment between the proposed methodological approach and the application example presented in Section 4.

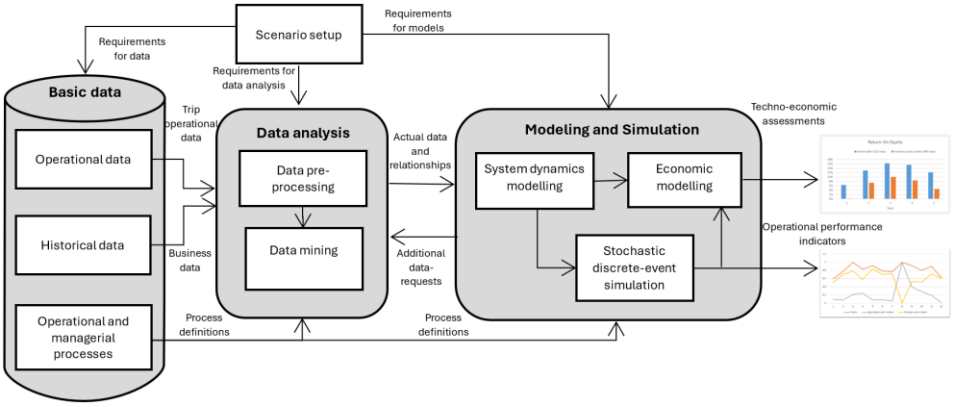


Figure 1. Conceptual framework of the study

An essential data preprocessing component in data analysis, especially in machine learning, involves the cleaning of raw data and the normalizing of attribute values to the same units of measurement. Then, a machine learning-based symbolic regression method (Kronberger et al., 2024) is used to find dependencies in historical and operational datasets and represent them in the form of analytical expressions. The obtained analytical models serve as the basis for the construction of system dynamics models.

In particular, the operational processes include transportation, loading, unloading and dispatch of vehicles, their maintenance and replacement. The system dynamics model is combined with an economic model to assess the economic viability of potential cooperation scenarios using technical and economic estimates. Additionally, the stochastic

discrete-event simulation model provides a virtual simulation of real processes of transportation of agricultural and forest products. This simulation in turn takes into account random factors and events that influence the implementation of the developed scenarios, as well as the assessment of corresponding transport solutions using operational efficiency KPIs. Feedback loops between the basic components make the proposed framework both flexible and interactive.

Scenario setting plays a key role in the study. Possible options for cooperative scenarios and transport solutions for the use and configuration of transport resources are based on previous experiences of the involved companies and take the benefits of the cooperation between the two industrial sectors into account. These scenarios may include both existing and potential workflows, as well as available or leased vehicles.

Validation of the proposed models and tools was carried out during the development process. The purpose is to ensure that at a certain stage of development the results of modeling and simulation meet the initially defined requirements (Zenina et al., 2020). For example, in the case of symbolic regression, 80% of the available data were used to train the model, and then 20% were used to test the resulting model. Finally, the results of the experimental scenarios were tested in real-life conditions, allowing them to be compared with practical ones.

Since no single tool to address all these components exists, various tools and methods are proposed and combined to implement the developed research framework in the target application.

3.3. Methods and tools

The proposed research methodology offers a combination of methods and tools for web data management, data mining, system dynamics modeling, stochastic discrete-event simulation and multi-user web environment.

Web-based data management. As mentioned above, the study focuses on small businesses and farms in the agricultural and forestry sectors. Most of them have one to several trucks and carry out local transportation in the agricultural sector. These small business often have a relatively low level of digitalization and lack modern information technology tools.

The required data are obtained using the digital web platform for agricultural logistics management (Graudvedis, 2024). This web platform allows collecting and storing business data of small businesses and operational data on completed trips. Moreover, it provides data management for overall control of transport logistics and improved cost efficiency.

An example of statistical data on the transportation of agricultural products obtained from the digital management platform is shown in Fig. 2. In particular, the data records in the table include loading and delivery data, the total length of the route, the types and volumes of transported products, the time of contact between the dispatcher and the driver, overtime work, as well as data on the driver and the vehicle, and fuel consumption.

Data-driven modeling. Symbolic regression is one of the most powerful machine learning methods that allows extracting key underlying relationships as analytical expressions directly from data without making assumptions about the model structure (Kronberger et al., 2024). Moreover, it is applicable to small datasets.

One of the widely discussed methods in the literature for constructing symbolic regression is genetic programming with the representation of evolution by tree-like data structures (Affenzeller et al., 2009). This method is applicable in various fields, since it allows to obtain mathematical relationships regardless of the data context. For example, it

was used to forecast river floods in (Merkuryeva et al., 2015) and to forecast demand for pharmaceutical products in (Merkuryeva et al., 2019). To construct symbolic regression models for the output variables in the study, the HeuristicLab software environment (Kronberger et al., 2012) was used.

Vehicle	Carrier	Reference ID	Cargo type	Units of measure	Volume	Total distance, KM	Loaded distance, KM	Empty distance, KM	Date	Filled fuel, L	Notes
RZ-79	Pan-Trans	22LV709762	Cutter chip	Bulk m³	99	250	125	125	06.12.2022	0	Spruce
RZ-79	Pan-Trans	22LV709790	Cutter chip	Bulk m³	99	250	125	125	07.12.2022	0	Spruce
RZ-79	Pan-Trans	22LV709853	Cutter chip	Bulk m³	99	250	125	125	11.12.2022	430	Pine
RZ-79	Pan-Trans	22LV709854	Cutter chip	Bulk m³	99	250	125	125	11.12.2022	0	Spruce
RZ-79	Pan-Trans	22LV709873	Cutter chip	Bulk m³	99	250	125	125	12.12.2022	0	Spruce
RZ-79	Pan-Trans	22LV709874	Cutter chip	Bulk m³	99	250	125	125	12.12.2022	0	Spruce
RZ-79	Pan-Trans	22LV709897	Cutter chip	Bulk m³	99	250	125	125	13.12.2022	410	Spruce
RZ-79	Pan-Trans	22LV709898	Cutter chip	Bulk m³	99	250	125	125	13.12.2022	0	Spruce
RZ-79	Pan-Trans	22LV7557	Cutter chip	Bulk m³	99	250	125	125	08.09.2022	0	Pine
KB7960	RANPO	22R1776	Grains	t	24.012	62	31	31	14.08.2022	0	Wheat
KB7960	RANPO	22R2002	Grains	t	25.68	62	31	31	15.08.2022	0	Wheat
KB7960	RANPO	22R2694	Grains	t	24.78	62	31	31	18.08.2022	0	Wheat
KB7960	RANPO	22R2980	Grains	t	16.54	62	31	31	19.08.2022	0	Wheat
KB7960	RANPO	22R3108	Grains	t	26	62	31	31	19.08.2022	0	Wheat
KB7960	RANPO	22R3240	Grains	t	25.2	62	31	31	20.08.2022	0	Wheat
KB7960	RANPO	22R3280	Grains	t	25.06	62	31	31	20.08.2022	0	Wheat
KB7960	RANPO	22R3305	Grains	t	25.53	62	31	31	20.08.2022	0	Wheat
KB7960	RANPO	22R3396	Grains	t	25.44	62	31	31	21.08.2022	0	Wheat
KB7960	RANPO	22R3430	Grains	t	25.02	62	31	31	21.08.2022	0	Wheat
KB7960	RANPO	22R3452	Grains	t	25.34	62	31	31	21.08.2022	0	Wheat
KB7960	RANPO	22R4105	Grains	t	25.82	62	31	31	24.08.2022	0	Wheat
KB7960	RANPO	22R4287	Grains	t	18.488	62	31	31	25.08.2022	0	Rapeseed

Figure 2. Examples of data records

An example of the analytical extract encoded in a tree-like data structure is shown in Fig. 3. The leaf nodes of a tree data structure represent specific input variables multi-plied by constant coefficients, and the internal nodes represent binary operations on their values or the results of other binary operations. The resulting symbolic regression model is presented by an analytical expression. For more information, see Section 4.

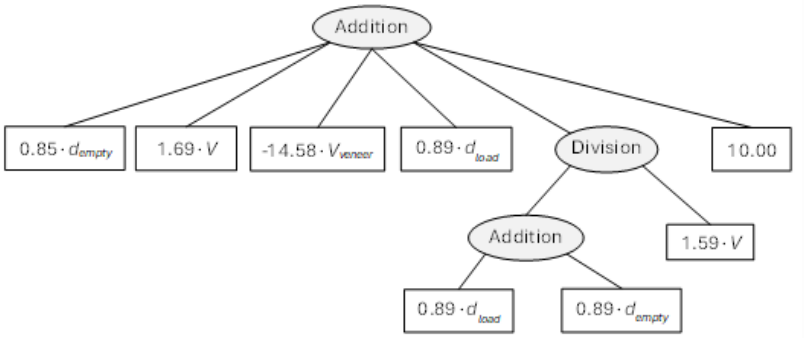


Figure 3. An example of the analytical extract encoded in a tree-like data structure

In fact, the use of the symbolic regression method has the advantage that the obtained dependences in the form of analytical expressions can be directly integrated into simulation models of the system.

The symbolic regression model is deterministic and reflects the past behavior of the system being analyzed by mathematical expressions. In the presented study, these mathematical expressions are updated as new data becomes available. The obtained dependences in the form of analytical expressions are then directly integrated into the system dynamics model for calculating nodes.

System dynamics modeling. The conceptual system dynamics model is constructed by identifying causal relationships between input variables and key performance indicators of the system. The model assumes that the company can use both its own trucks and trailers and rented vehicles.

The model contains more than 20 nodes that represent input, intermediate and output variables. The input variables define parameters of trips to be performed (e.g., numbers of owned and rented semi-trailers, average transportation distance, loading time, etc.), as well as parameters that affect costs and revenues, such as fuel costs and prices, taxes and wages. The intermediate variables represent the results of intermediate calculations of economic and operational performance indicators, e.g., time-dependent costs and total travel distance. The output variables are final calculations of economic and operational performance indicators such as fixed and variable costs, revenue, average vehicle utilization, and driver workload.

In Fig. 4, a cause-and-effect diagram created in AnyLogic shows how the various variables of the modeled system are interrelated.

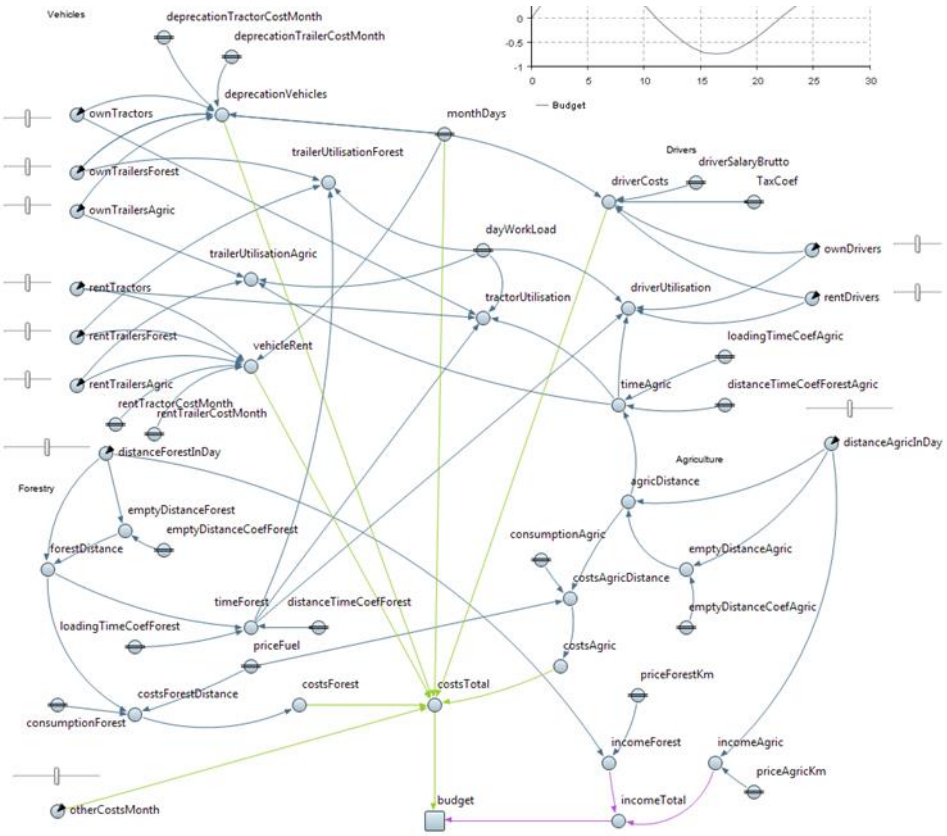


Figure 4. Causal loop diagram of the modeled system

The node connection logic and node calculations are based on assumptions derived from the results of symbolic regression analysis as well as from business process analysis. The dependencies and relationships between variables that can be derived from data-

driven models (such as the relationship between empty and loaded route distances) are substantiated through data analysis. Other relationships that could not be identified in this way are determined based on analysis of the cost and income structure of the analyzed enterprise.

The computational modeling is carried out separately for the transportation of timber and for the transportation of agricultural products. The developed system dynamics model is deterministic and does not take into account randomness in the behavior of the system and its processes. If the computational model does not have a value for a particular output or it is unclear how to calculate it, it can be obtained using a symbolic regression formula derived from the available data.

Stochastic Discrete-Event Simulation. A stochastic discrete-event simulation model is developed to generate, simulate and evaluate transportation decisions for potential vehicle trips according to a given scenario. The model takes into account the input and output data of the system dynamics model and is partly based on its calculations. Specifically, input data include the company's geographical location, employee wages, costs and depreciation of vehicles (tracks, trailers and semi-trailers), equipment maintenance and repair costs, fuel consumption and price.

In particular, the model ensures modeling of stochastic input variables and processing of random events. For example, the influence of seasonality on the distance of transportation of agricultural products is described by the probability distributions, which depend significantly on the month of the year. Finally, the model provides values of key performance indicators.

Moreover, the model provides industry representatives with graphs of economic efficiency and operational performance indicators for a given scenario, and also displays the transport processes themselves online. This provides a holistic view (Fig. 5) for a better understanding of potential organizational scenarios and corresponding transport solutions.

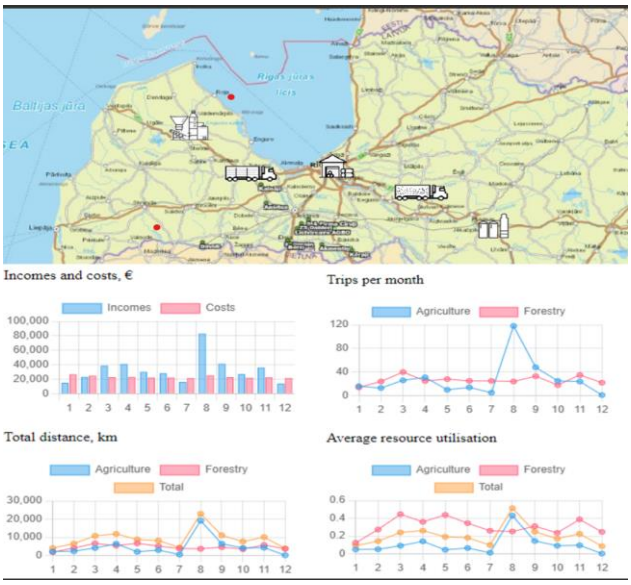


Figure 5. Screenshot from discrete-event model visualization

Multi-user Web-based Environment. To make modeling and simulation services widely available to all stakeholders (e.g. managers, planners, drivers, etc.), a multi-user web-based modeling approach is proposed. Web modeling allows these services to be used over the internet and provides multi-user access to models, simulation experiments, their results and visualizations.

To enable multiple users to conduct independently experiments online, it was proposed to split the simulation model application into front-end and back-end system components. The front-end component is designed to provide the user with web access to the model, data entry, visualizations and process diagrams. The back-end component is designed to work with the simulation model, store user data and modeling results, and process the corresponding system behavior. The multi-user approach in the back-end component is organized by storing the model configuration for each user in a separate session. The simulation model updates these sessions in a loop (see Fig. 6).

The software is developed using the general-purpose PHP 7.4 scripting language for Windows web development.

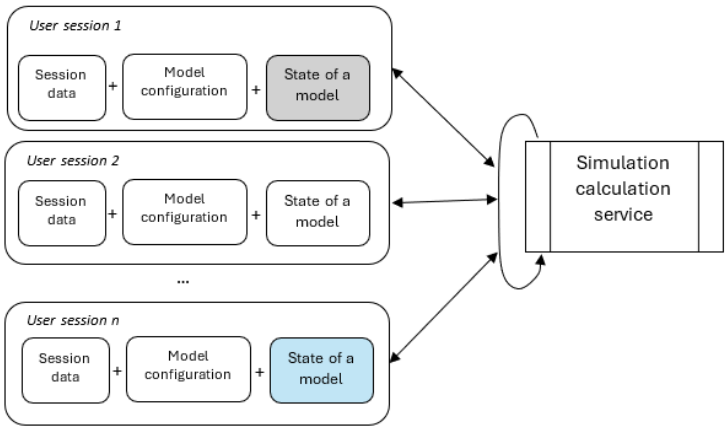


Figure 6. Multi-user web-based simulation

Economic Feasibility Assessment. A cooperative business model was developed to assess the economic feasibility of using different types of trailers. It is estimated that transport utilization and cost optimization strategies determine operating cost variations in the range of approximately 30%. The cost of purchasing equipment and driver comfort requirements affect the investment amount by approximately 15%.

The model allows appropriate adjustment of parameters to compare alternative solutions and determine the expected return on investment as well as the cost of carbon emissions. Moreover, it provides the calculation of the internal rate of return, the borrowing rate, as well as the calculation of the return on equity under certain conditions. Finally, the model uses the experience and knowledge of the industry experts who participated in the study, and the calculations are performed in Excel spreadsheets.

4. Case Study

This section presents an example of application of the proposed research methodology to model and simulate the benefits of cooperative scenarios in the transportation of products from two industrial sectors by a small agricultural company. The description of the experimental results obtained follows the proposed methodology step by step.

Preliminary data analysis. The initial experimental dataset contained data from over 61 forest companies with a wide range of used vehicles from 1 to 29 and 12 agricultural companies with 1 to 8 vehicles. These data sets include historical performance data, such as delivery date and location; travel distance for empty and loaded trips; type, weight and volume of transported products; vehicle and driver identification; supplementary time needed if any, etc.

The demand for transportation in the agricultural sector is largely dependent on seasonal factors. The probability distributions of the delivery distances for the agricultural products are shown in Fig. 7. The likelihood of short or nearly identical trips is higher during peak seasons, and the spread of distances is higher during off-peak seasons. This may be due to the sporadic, irregular nature of trips. The histogram of timber delivery distances for timber transportation in different months is presented in Fig. 8.

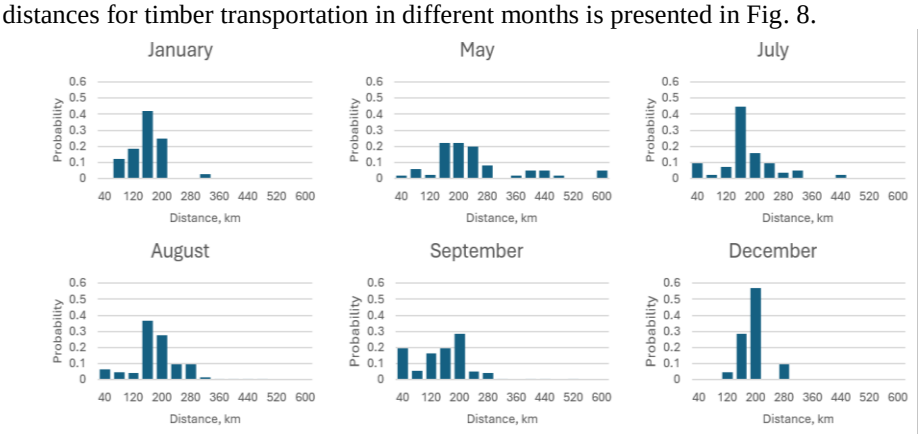


Figure 7. Histograms of delivery distances (in km) for agricultural products transportation in different months

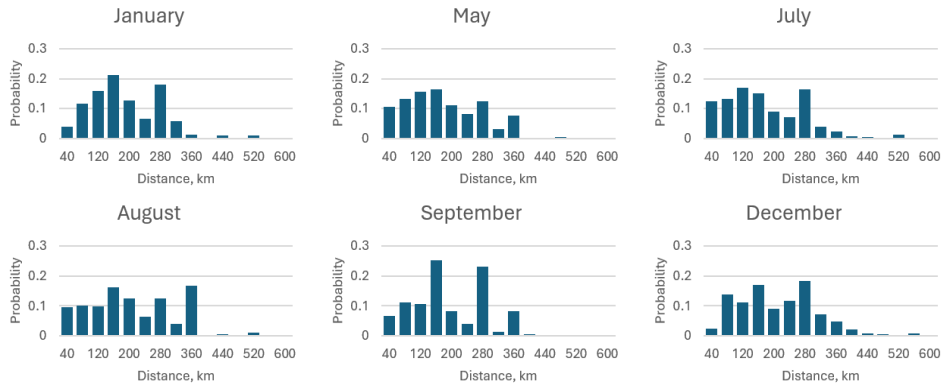


Figure 8. Histograms of delivery distances (in km) for forest products transportation in different months

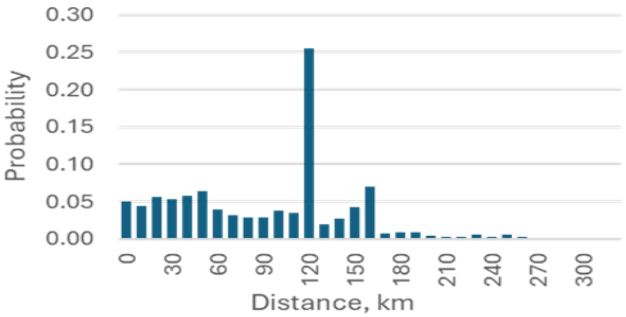


Figure 9. Histogram of length of loaded trips

In addition, a preliminary analysis of the data shows a rather complex nature of the empirical distributions of other random variables. For example, a fairly large portion of loaded trips have a length of about 120 km, and the possible distances of the remaining trips are evenly distributed over the range of possible values (see Fig. 9).

Among the different types of products analyzed in the life example by using traditional statistical techniques, the greatest dependence was found for grain.

Data-driven modeling. The seasonal factors influence the potential order volumes, distances and deliveries included in the model. At the stage of data-driven modeling using symbolic regression, various dependencies were obtained, such as the cost of the trip, the number of registered transport trips, the number of trips by product groups as well as the average volume of transported products in different months, etc. For example, the resulting symbolic regression model of the trip cost is presented by an analytical expression showing the dependence of this cost on the distances of travel with and without goods, as well as on the volume of transported goods, which turned out to be the most significant factors:

$$C_{trip} \approx 0.85\, d_{empty} + 1.69\, V - 14.58\, V_{veneer} + 0.89\, d_{load} + \\ + \frac{0.89\, d_{load} + 0.89\, d_{empty}}{1.59\, V} + 10, \qquad (1)$$

where C_{trip} is the cost of the trip; d_{empty} and d_{load} – the distance of travel without and with transported timber, respectively; V is the volume of transported timber; V_{veneer} is the total volume of veneer blocks in the transported cargo. The formula (1) was obtained directly as a result of the analytical extract, encoded in a tree-like structure presented in Fig.3.

Similarly, mathematical expressions were obtained for various factors and parameters required for further development of the simulation models. In particular, to justify the calculation of the cost of a specialized semi-trailer in simulation experiments, based on the analysis of the collected data, the following model to estimate the provisional unit cost per km was obtained:

$$C_{dist} \approx \frac{2.092 \cdot d_{empty} + \log\left(115.2 \cdot d_{load} + 933.2 \cdot d_{empty}\right)\left(0.149 \cdot V + 0.055 \cdot t_{load}\right)}{1.7993 \cdot d_{load}} - \\ - \frac{1}{0.0172\, d_{load}^2} + \frac{1}{0.0142\, d_{load}^3} + 0.942, \qquad (2)$$

where C_{dist} represent a unit cost in euros per km.

To assess the fitness of the obtained symbolic regression models, a cross-validation method was employed by splitting the data into a training (80%) and a testing (20%) subset. Accordingly, numerical accuracy metrics scores were estimated for each subset. E.g., for the data model expressed by equation (1) above, the following estimates were obtained: normalized root mean square error of 0.0229 and 0.0265, respectively; and Pearson coefficients (R2) of 0.973 and 0.977, respectively. Therefore, this model is considered reliable.

Setting up cooperative scenarios. The baseline scenario is one in which a small agricultural company is engaged in the transportation of agricultural products. It owns a certain number of trucks and semi-trailers, which are used for transportation purposes in certain months and employs a certain number of drivers. Key costs are estimated, including transportation, labor, and equipment maintenance.

This scenario is compared with new cooperative scenarios.

In the new scenario (called 'Scenario 1'), a company that primarily transports grain has one specialized semi-trailer that can be used to transport wood. This scenario was introduced to simulate the synergies of transporting products (i.e. grain and timber) from the two industrial sectors by a company and the impact on its performance.

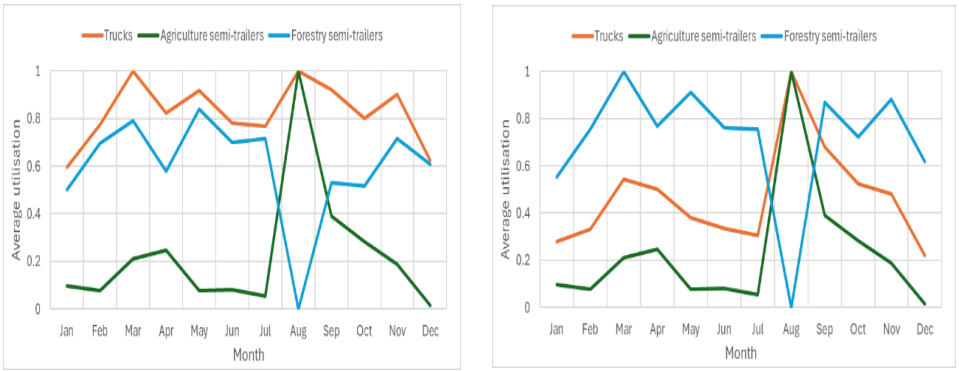
In the following scenarios (called 'Scenario 2'), this company is primarily engaged in the transportation of grain. But it can use or share a certain number of semi-trailers (more than 1), either owned or rented, for the transportation of timber. The priority is the transportation of grain, but if the drivers are free and the specialized semi-trailers are idle, they are used to transport timber. In this scenario, different configurations of the enterprise's transport vehicles and drivers can be analyzed, taking into account the benefits of diversifying the transported products from the two sectors.

It should be noted that small agricultural companies typically employ as many drivers as trucks. Having more drivers than vehicles can result in unnecessary driver downtime.

Simulation-based analysis of cooperative scenarios. The results of the stochastic simulation by month during the year for the baseline and cooperative scenarios for the agricultural company are presented in Fig. 10. In particular, the figure presents the results for the cases with 1 and 3 drivers (and trucks, respectively) and different configurations of agricultural and forestry semi-trailers, simulated over the year. Fig. 10a shows the results of the simulation experiments for the case of one driver, one truck, one agricultural trailer and one forestry trailer. Fig. 10b shows the results for three drivers, three trucks, three agricultural trailers and one forestry trailer. In Fig. 10a and 10b, the dark green line represents the utilization rate of agricultural trailers, whereas the light blue line represents the utilization rate of forestry trailers and the brown line the utilization rate of trucks (and drivers). Note that in the baseline scenario, where the company is only engaged in the transportation of agricultural products and does not have a specialized semi-trailer for transporting timber, the loading of trucks, as well as drivers, will correspond to the dark green lines. Finally, the results of the simulation experiments comparing different transportation scenarios over the year are summarized in Table 1.

In scenario 1, the average utilization rates of trucks and specialized semi-trailers for timber transportation increase significantly compared to the baseline scenario, reaching values of 0.47 (up from 0.23) and 0.72 (up from 0.00), respectively. The highest average utilization rate can be achieved in a configuration of six semi-trailers. It will provide the greatest diversification of deliveries. But it will also be the most expensive configuration in terms of both fixed and operating costs. Thus, assuming that the company can operate in accordance with the demand model obtained from the analyzed data, the provision of

timber transportation with just one semi-trailer will already significantly improve the use of the company's resources.



(a) (b)

Figure 10. Utilization rates of different vehicles with 1 (a), with 3 drivers (b), simulated over a year

Table 1. Results of simulation experiments for the baseline and cooperative scenarios

Number of semi-trailers		Trips per year		Total distance in thousands km per year		Average vehicle utilization rate		
Agriculture	Forestry	Agriculture	Forestry	Agriculture	Forestry	Trucks	Agriculture semi-trailers	Forestry semi-trailers
Baseline Scenario								
3	0	173	0	28.7	0	0.23	0.23	0.00
Scenario 1								
3	1	173	176	28.7	30.2	0.47	0.23	0.72
Scenario 2								
3	2	173	348	28.7	60	0.70	0.23	0.71
3	3	173	439	28.7	75.7	0.83	0.23	0.60
2	3	115	485	19.1	83.9	0.81	0.23	0.66

Furthermore, maximum resource utilization throughout the year can be achieved by having a diversified access to both agricultural and forestry transport (as described in different possibilities of scenario 2). This means that if a transport company is primarily engaged in the transportation of agricultural products, then by diversifying the types of goods transported with forest products, the existing and shared transportation resources (and the corresponding drivers) will be provided with additional work, even outside the harvest season. Indeed, the average vehicle transportation load for trucks and forestry semi-trucks increases in the simulation model for each of the studied numbers of semi-trucks used for transporting agricultural and forest products. However, assessing the cost-effectiveness of using additional vehicles requires a more complete examination of the

model's input data for individual item costs, as well as clarification of fixed and variable costs for different trailer categories.

5. Conclusions

Modern digital platforms and modern computing techniques are radically transforming the business models of small and medium-sized businesses and changing the relationships between them. One of the new opportunities that can be considered is intersectoral and intercompany cooperation.

At the same time, modern digital technologies provide opportunities for the accumulation of large volumes of data. These data are widely used to extract useful information and understand the cause-and-effect relationships in the behavior of a real system. The latter enables the implementation of data-driven approaches to support decision making, amongst others in transportation management.

This article proposes a new integrated research framework and methodology for data-driven modeling and simulation in transportation management through cross-sectoral collaboration. It is designed to improve the planning and organization of transportation of agricultural and forest products by small agricultural and forestry businesses. The presented study is based on a real life case of synergy in the transportation of products in two sectors - agriculture and forestry. The benefit for participants in both economic sectors is to ensure a more efficient use of their transport and labor resources (in particular, vehicle drivers) by using the opportunities of intersectoral cooperation and emergent data driven-modeling and simulation technologies. The synergy effect is demonstrated through simulation-based analysis of cooperative scenarios between involved participants.

The proposed methodology offers a combination of methods and tools for web data management, data mining using powerful machine learning methods, system dynamics modeling, stochastic discrete-event simulation and a multi-user web environment. Data-driven models such as symbolic regression are used to identify patterns in data and translate the underlying relationships into the modeling formalisms needed to build computer simulation models. The system dynamics model is coupled with an economic model to assess the economic viability of potential cooperative scenarios and transport solutions. The discrete-event simulation model provides virtual simulation of real processes of transportation of agricultural and forest products, taking into account random factors and events that influence the implementation of these scenarios, as well as their assessment using operational performance indicators.

In the article, the life example is given for a small agricultural company that owns its own fleet of vehicles and thus has its starting point in the industries that have a need for transportation and are creatively looking for cross-sectoral collaboration. But the advantages of such inter-sector synergy may also be explored by small businesses in the transportation industry or third-party logistics companies that provide transportation of a wide range of goods. The proposed work may also be useful for researchers working in the field of data-driven modeling and simulation. Future research will also explore the potential for wider use of artificial intelligence and blockchain technologies to improve transport management in agriculture and forestry for small businesses, such as collecting, storing and analyzing data on handling conditions during transportation (e.g. temperature control, storage conditions) and ensuring that transported products meet required standards.

Finally, data-driven simulation, which is based on the integration of various data models, supplemented by computational intelligence and machine learning methods, can

certainly be regarded as a new trend in the development of computer modeling technologies.

Acknowledgment

The experimental part of the research was conducted within the framework of the EIP-AGRI project 18-00-A01612-000022 "Innovative solutions in planning and organization of agricultural and forestry produce transportation" of the National RURAL Development Program (Latvia).

The authors would like to express their sincere gratitude to Mr. Wouter Faes, Visiting Lecturer at the University of Hasselt (Belgium) and Riga Technical University (Latvia), for his valuable advice in the area of cross-sectoral collaboration and management.

References

Affenzeller, M., Winkler, S., Wagner, S., Beham, A. (2009). *Genetic Algorithms and Genetic Programming: Modern Concepts and Applications*. Chapman & Hall/CRC.

Alayet, C., Lehoux, N., Lebel, L. (2018). Logistics approaches assessment to better coordinate a forest products supply chain, *Journal of Forest Economics*, **30**, 13-24. DOI: 10.1016/j.jfe.2017.11.001.

Alonso-Ayuso, A., Escudero, L. F., Guignard, M., Weintraub, A. (2020). On dealing with strategic and tactical decision levels in forestry planning under uncertainty, *Computers & Operations Research*, **115**, 104836. DOI: 10.1016/j.cor.2019.104836.

Bajgiran, O. S., Zanjani, M. K., Nourelfath, M. (2016). The value of integrated tactical planning optimization in the lumber supply chain, *International Journal of Production Economics*, **171**(1), 22-33. DOI:10.1016/j.ijpe.2015.10.021

Belmont Guerrón, P., Hallo, M. (2022). An Evaluation of Machine Learning Approaches to Integrate Historical Farm Data, *Baltic J. Modern Computing*, **10**(4), 623–644. DOI: 10.22364/bjmc.2022.10.4.03.

Bolsakovs, V., Romanovs, A., Merkurjevs, J., Feldmanis, R. Innovative Solutions in Planning and Management of Transportation of Forestry and Agriculture Products: Project Summary. *In Proc. of the 65th IEEE International Scientific Conference on Information Technology and Management Science*, 2024. DOI: 10.1109/ITMS64072.2024.10741930.

Boukherroubm T., Ruizm A., Guinet, A., Fondrevelle, J. (2013), An Integrated Approach for the Optimization of the Sustainable performance: a Wood Supply Chain, *IFAC Proceedings Volumes*, **46**(9), 186-191. DOI: 3182/20130619-3-RU-3018.00205

Buongiorno, J. (1996). Forest sector modeling: a synthesis of econometrics, mathematical programming, and system dynamics methods, *International Journal of Forecasting*, **12**(3), 329-343. DOI: 10.1016/0169-2070(96)00668-1

Cardoso, M. C., Gomes, R. R. M., Silva, E. A., Gomes de Souza, M. F. (2009). Evaluation of the wood hauling logistic performance in farm forest areas using Petri net, *Revista Árvore* **33**(6), 1159-1167. DOI: 10.1590/S0100-67622009000600018

Castañer, X., and Oliveira, N. (2020). Collaboration, Coordination, and Cooperation Among Organizations: Establishing the Distinctive Meanings of These Terms Through a Systematic Literature Review. *Journal of Management*, **46**(6), 965-1001. DOI: 10.1177/0149206320901565

Cavone, G., Dotoli, M., Seatzu, C. (2017). A Survey on Petri Net Models for Freight Logistics and Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems*, **19**(6), 1795-1813.

Chen, M., Zhichuan, W., Zhu, H., Xia, J., Yue, X. (2022). Research on System Dynamics of Agricultural Products Supply Chain Based on Data Simulation. In: *2022 5th International Conference on E-Business, Information Management and Computer Science (EBIMCS)*, 54-66.

Deng, D., Ye, C., Tong, K., Zhang, J. (2023). Evaluation of the Sustainable Forest Management Performance in Forestry Enterprises Based on a Hybrid Multi-Criteria Decision-Making Model: A Case Study in China. *Forests*, **14**, 2267. DOI: 10.3390/f14112267

Diaz, J. A., Perez, I. G. (2000). Simulation and Optimization of Sugar Cane Transportation in Harvest Season. In: *Proceedings of the 2000 Winter Simulation Conference*, 1114-1117

Fioroni, M. M., Franzese, L. A. G., Santana, I. R., Lelis, P. E. P., Silva, C. B., Telles, G. D., Quintans, J. A. S., Maeda, F. K., Varani, R. (2015). From Farm to Port: Simulation of the Grain Logistics in Brazil. In: *Proceedings of the 2015 Winter Simulation Conference*, 1936-1947.

Francois, J., Moad, K., Bourrières, J.-P., Lebel, L. (2017). A tactical planning model for collaborative timber transport, *IFAC-PapersOnLine*, **50**(1), 11713-11718. DOI: 10.1016/j.ifacol.2017.08.1695

Graudvedis (2024). A platform for agricultural transport. A tool for efficient logistics solutions. http://groudvedis.selflogistic.lv (accessed on 10 February 2025).

Guan, S., Nakamura, M., Shikanai, T., Okazaki, T. (2008). Hybrid Petri nets modeling for farm work flow. *Computers and Electronics in Agriculture*, **62**(2), 149-158. DOI: 10.1016/j.compag.2007.12.006

Gupta, S., Garima. (2017). Logistics Sprawl in Timber Markets and its Impact on Freight Distribution Patterns in Metropolitan City of Delhi, India. *Transportation Research Procedia*, **25**, 965-977. DOI: 10.1016/j.trpro.2017.05.471

Habib, M. K., Ayankoso, S. A. (2021). Data-Driven Modeling: Concept, Techniques, Challenges and a Case Study. *Proceedings of the 2021 IEEE International Conference on Mechatronics and Automation (ICMA)*. DOI: 10.1109/ICMA52036.2021.9512658

Kronberger, G., Burlacu, B., Kommenda, M., Winkler, S. M., Affenzeller, M. (2024). *Symbolic Regression*. New York, Chapman and Hall/CRC. DOI: 10.1201/9781315166407

Kronberger, G., Wagner, S., Kommenda, M., Beham, A., Scheibenpflug, A., Affenzeller, M. (2012). Knowledge Discovery Through Symbolic Regression with HeuristicLab. In *Proc. of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases*, Part II (ECML PKDD'12), 824-827.

Lamsal, K., Jones, P. C., Thomas, B. W. (2016). Harvest logistics in agricultural systems with multiple, independent producers and no on-farm storage. *Computers & Industrial Engineering*, **91**, 129–138. DOI: 10.1016/j.cie.2015.10.018

Li, Y., Lu, S. (2015). Research on Prediction of Regional Grain Logistics Demand Based on the Grey-regression Model. In: *IEEE International Conference on Grey Systems and Intelligent Services 2015*, 97-101.

Lin, K., Ishihara, H., Tsai, C., Hung, S., Mizoguchi, M. (2022). Shared Logistic Service for Resilient Agri-Food System: Study of E-Commerce for Local and B2B Markets in Japan. *Sustainability*, **14**(3), 1858, DOI: 10.3390/su14031858

Lomotko, D., Arsenko, D., Konovalova, O., Ischuka, O. (2019). Methods of infrastructure management for optimization of grain transport organization. *ICTE in Transportation and Logistics 2018*, *Procedia Computer Science* **149**, 500–507.

Lopez, J. J. U., Qassim, R. Y. (2023) A novel modelling approach for the redesign of supply chains: An application to soybean grain supply chains. *Research in Transportation Business & Management*, **51**, 101037. DOI: 10.1016/j.rtbm.2023.101037

Mardaneh, E., Loxton, R., Meka, S., Gamble, L. (2021). A decision support system for grain harvesting, storage, and distribution logistics. *Knowledge-Based Systems* **223**, 107037. DOI: 10.1016/j.knosys.2021.107037

Mathur, S. K., Warner, J. E. (1997). *Economics of Raw Timber Transportation: a Feasibility Study*, Technical Report. Texas Transportation Institute.

Mehmann, J., Teuteberg, F. (2016). The fourth-party logistics service provider approach to support sustainable development goals in transportation e a case study of the German agricultural bulk logistics sector. *Journal of Cleaner Production*, **126**, 382-393.

Merkuryev, Y., Merkuryeva, G., Piera, M. A., Guasch, A. (Eds.) (2009), *Simulation-based Case Studies in Logistics: Education and Applied Research*, SpringerVerlag, London.

Merkuryeva, G. (2012). Integrated Delivery Planning and Scheduling Built on Cluster Analysis and Simulation Optimisation. In *Proc. of the ECMS European Conference on Modelling and*

Simulation, European Council for Modeling and Simulation, 2012, 164-168. DOI: 10.7148/2012-0164-0168.

Merkuryeva, G. (2024). Emerging Technologies for Data-Driven Pharmaceutical Supply Chain Management. *Proceedings of the 36th European Modeling & Simulation Symposium, 2024*, 033, https://doi.org/10.46354/i3m.2024.emss.033

Merkuryeva, G., Bolshakov, V. (2015). Simulation-based fitness landscape analysis and optimisation of complex systems. *Technological and Economic Development of Economy*, **21**(6), 899–916, DOI:10.3846/20294913.2015.1107654.

Merkuryeva, G., Merkuryev, Y., Sokolov, B. V., Potrjasaev, S., Zelentsov, V. A., Lektauers, A. (2015). Advanced river flood monitoring, modelling and forecasting, *Journal of Computational Science*, **10**, 75-85.

Merkuryeva, G., Merkuryev, Y., Vanmaele, H. (2011). Simulation-Based planning and optimization in multi-echelon supply chains. *Simulation*, **87** (8), 698-713.

Merkuryeva, G., Valberga, A., Smirnov, A. (2019). A demand forecasting in pharmaceutical supply chains: A case study. *Procedia Computer Science*, **149**, 3-10.

Mogale, D. G., Cheikhrouhou, N., Tiwari, M. K. (2019). Modelling of sustainable food grain supply chain distribution system: a bi-objective approach. *International Journal of Production Research*, **58**(18), 5521-5544. DOI: 10.1080/00207543.2019.1669840

Mütsch, F., Gremmelmaier, H., Becker, N., Bogdoll, D., Zofka, M. R., Zöllner, J. M. (2023). From Model-Based to Data-Driven Simulation: Challenges and Trends in Autonomous. Computer Vision and Pattern Recognition, *CVPR 2023 VCAD workshop*, 2023, Vancouver, Canada, https://arxiv.org/abs/2305.13960

National Statistical System of Latvia (2024). Freight traffic by road by group of goods 2008 - 2023. Available online: https://data.stat.gov.lv/pxweb/en/OSP_PUB/START__NOZ__TRK__TRKA/TRK150/ (accessed on 12th June 2024)

Nourbakhsh, S. M., Bai, Y., Maia, G. D. N., Ouyang, Y., Rodriguez, L. (2016). Grain supply chain network design and logistics planning for reducing post-harvest loss. *Biosystems Engineering* **151**, 105-115.

Oke, M. O., Adebayo, O. J., Adenipekun, A. E. (2018). Determining the Best Transportation System for a Lumber Company Using the Transportation Algorithms, *Academic Journal of Statistics and Mathematics*, **4**(10), 1-11.

Oliveira, A. L. R., Marsola, K. B., Milanez, A., Fatoretto, S. L. R. (2022). Performance evaluation of agricultural commodity logistics from a sustainability perspective. In: *Case Studies on Transport Policy* **10**, 674–685.

Ozgun, A., Kirci, M. (2015). Petri net models for agricultural management tasks. *Proceedings of 2015 Fourth International Conference on Agro-Geoinformatics*, 235-241. DOI: 10.1109/Agro-Geoinformatics.2015.7248129

Palatova, P., Rinn, R., Machon, M., Palus, H., Purwestri, R. C., Jarsky, V. (2023). Sharing economy in the forestry sector: Opportunities and barriers. *Forest Policy and Economics* **154**, 103000. DOI: 10.1016/j.forpol.2023.103000

Pavlenko, O., Shramenko, N., Muzylyov, D. (2020). Logistics Optimization of Agricultural Products Supply to the European Union Based on Modeling by Petri Nets. In: *New Technologies, Development and Application III. NT 2020. Lecture Notes in Networks and Systems* **128**, 596–604. DOI: 10.1007/978-3-030-46817-0_69

Sfeir, T. A., Pécora, J. E., Ruiz, A., LeBel, L. (2019). Integrating natural wood drying and seasonal trucks' workload restrictions into forestry transportation planning, *Omega* **98**, 102135. DOI: 10.1016/j.omega.2019.102135

Sgurev, V., Doukovska, L., Drangajov, S. (2022). Intelligent Logistics at Harvest Time in Grain Production. *International Conference Automatics and Informatics (ICAI*) (Varna, Bulgaria, 2022), 135-139. DOI: 10.1109/ICAI55857.2022.9960136

Troncoso, J. J., Garrido, R. A. (2005). Forestry production and logistics planning: an analysis using mixed-integer programming, *Forest Policy and Economics* **7**(4), 625-633. DOI: 10.1016/j.forpol.2003.12.002.

Trunina, I., Moroz, M., Zahorianskyi, V., Zahorianskaya, O., Moroz, O. (2021) Management of the Logistics Component of the Grain Harvesting Process with Consideration of the Choice of Automobile Transport Technology Based on the Energetic Criterion. In: *2021 IEEE International Conference on Modern Electrical and Energy Systems (MEES)* (Kremenchuk, Ukraine, 2021), 1-5. DOI: 10.1109/MEES52427.2021.9598768

Turner, A. P., Sama, M. P., McNeill, L. S. G., Dvorak, J. S., Mark, T., Montross, M. D. (2019). A discrete event simulation model for analysis of farm scale grain transportation systems. *Computers and Electronics in Agriculture*, **167**, 105040. DOI: 10.1016/j.compag.2019.105040

Van Noordwijk, M., Duguma, L. A., Dewi, S., Leimona, B., Catacutan, D. C., Lusiana, B., Öborn, I., Hairiah, K., Minang, P. A. (2018). SDG synergy between agriculture and forestry in the food, energy, water and income nexus: reinventing agroforestry?, *Current Opinion in Environmental Sustainability*, **34**, 33-42. DOI: 10.1016/j.cosust.2018.09.003

Veile, J. W., Schmidt, M.-Ch., Voigt, K.-I. (2022). Toward a new era of cooperation: How industrial digital platforms transform business models in Industry 4.0. *Journal of Business Research*, **143**, 387-405, https://doi.org/10.1016/j.jbusres.2021.11.062

Walsh, K. D., Sawhney, A., Bashford, H. H. (2003). Simulation of the residential lumber supply chain. *2003 Winter Simulation Conference* (New Orleans, LA, USA, 2003), **2,** 1548-1551. DOI: 10.1109/WSC.2003.1261601

Xiao, L., Lang, B. (2009). A Hybrid Intelligent Algorithm for Grain Logistics Vehicle Routing Problem. *2009 International Conference on Environmental Science and Information Application Technology*, (Wuhan, China, 2009), 556-559. DOI: 10.1109/ESIAT.2009.312

Zenina, N., Merkuryev, Y., Romanovs, A. (2020). The general principles of the transportation simulation model development and validation. *WSEAS Transactions on Systems and Control*, **15**, 81-92. DOI: 10.37394/23203.2020.15.10

Zhu, Y., Li, X., Zhen, T. (2009). Analysis and Design of grain logistics Distribution and Optimization Research Based on Web GIS. In: *Third International Symposium on Intelligent Information Technology Application Workshops*, 7-9.

Received February 27, 2025, revised August 14, 2025, accepted October 12, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, 778-805
https://doi.org/10.22364/bjmc.2025.13.4.03

Effectiveness of Image Protection Software Against Image Generation Tool Training

Valerija JANUSEVA, Solvita ZARINA

Faculty of Science and Technology, University of Latvia, Riga, Latvia

valerijajanuseva@gmail.com, solvita.zarina@lu.lv

ORCID 0009-0008-4753-767X, ORCID 0000-0001-8884-2971

Abstract. Unauthorised use of artworks in training image generation models poses a growing challenge for online copyright protection. Contemporary artists' work is used without their consent both by artificial intelligence (AI) companies and Internet users, creating an urgent need for effective protective measures. The article examines and compares software solutions designed to safeguard artworks from such unauthorised use, examining both technical effectiveness and the output's visual quality. We evaluate different image protection tools and their change intensity levels by applying them to illustrations created by one of the authors and then training generative models on both protected and unprotected versions. The resulting images are evaluated by a voluntary survey of 71 respondents including artists and artwork viewers (non-artists). The discussion and conclusions assess image protection software based on research findings, provide recommendations for artists, outline future research opportunities, and demonstrate that participation increased respondents' awareness of the importance of protecting artworks.

Keywords: Image Protection Software, Protection of Artworks, Adversarial Perturbations, Testing, AI Model Training, Image Generation

1. Introduction

Image generation using artificial intelligence (AI) models, which has become widely accessible with the release of tools like Midjourney and Stable Diffusion in 2022, is now commonly used for entertainment, personal, and professional purposes. However, significant ethical and legal issues have been discovered – the image generation models were trained on millions of visual data samples obtained from the Internet without the consent of their creators, thus negatively impacting artists' careers and infringing on their copyright (Heikkilä, 2022).

In response to unauthorised use of artworks, several image protection tools have been developed that introduce subtle visual changes into images, affecting the ability of AI models to mimic artistic styles. Three promising tools are Glaze, Nightshade, and Mist. Although previous research shows their potential and technical feasibility, it focuses on each tool's individual performance and uses a separate experiment design. Furthermore, not all experiments include digital illustrations and art, which artists often publish online, and which are particularly vulnerable to unauthorised use for AI training.

This motivates us to address these gaps by conducting comparative technical experiments, analysing the visual evaluation of the results, and investigating the impact of the software on a less studied medium of artwork creation. Thus, the aim of this article is to investigate and compare the effectiveness of such image protection software tools in limiting the ability of AI to replicate an artist's individual style. We hypothesise that training generative models on protected images negatively impacts their ability to mimic artists' style effectively and that the introduced changes do not affect the perceived visual quality of the artwork. Therefore, artists can be advised to use these tools to protect their works. Thus, the objectives of the research are to conduct an experiment by training small image generation models with both unprotected and protected image samples, and to compare and evaluate the effectiveness of protection software with the participation of artists and viewers of artworks in the study.

To test the hypothesis, we proceed as follows: (a) we review the current situation and provide an insight into the relevant literature on image generation and its principles, as well as on image protection software, (b) we conduct an experiment in which we train eight small-scale image generation models, one with unprotected and seven with protected datasets, each using different protection software and change intensity settings, (c) we evaluate the effectiveness of protection software through a user survey among 71 respondents including artists and viewers of artworks (non-artists) and compare the survey results with computational measurements of style similarity. Finally, based on the results of the experiment and the evaluations obtained, we compare the three tools and provide artists with recommendations on how to protect images, as well as discuss future research directions and acknowledge the survey's impact on raising artists' awareness on image protection.

2. Insight into image generation

2.1. Review of the current situation

In the past three years, the rapid growth of generative AI tools has taken the technology world by storm. It is now possible to type in a prompt with just a few words or phrases and, within seconds, generate a high-quality image that may be almost indistinguishable from authentic human-made photographs or works of art. "The obvious source of these systems' popularity is that they offer something entirely new: being able to generate an image just by describing it, without having to go to the trouble of learning a skill – such as illustration, painting or photography – to actually make it." (McCormack et al., 2023).

AI-generated content is now regularly featured on social media platforms, and it has even won prizes in art and photography competitions (Roose, 2022; Parshall, 2023). Large companies such as Coca-Cola have used generative AI for creating video adverts (Coca-Cola, 2024). The popular online platform for films and TV series, Netflix, plans to introduce AI-generated advertising in the middle of streaming from 2026 (Harding, 2025).

In March 2025, the company OpenAI released the ChatGPT model GPT-4o with capabilities to generate improved quality images in a "wide range of styles" (OpenAI, 2025). Internet users quickly realised that the model could convincingly replicate the style of animated films from the renowned Japanese animation studio, Studio Ghibli. This led

to a flood of images generated in the artistic style of the films on Instagram, Facebook, and X (formerly Twitter). The release of this model pushed the average number of weekly active ChatGPT users to over 150 million for the first time (Sriram, 2025). In Latvia, a Japanese cuisine restaurant Shōyu Ramen generated a Ghibli-like animated promotional reel that was published on its Instagram account (Shōyu, 2025).

Many ethical concerns surround these practices. The AI company Midjourney has published a list of 16,000 artists whose artworks were collected to train their image generation tool. The list includes not only historical artists but also contemporary illustrators, some of whom have worked for corporations like Nintendo and Hasbro (Ho, 2023). The company OpenAI itself admits that it is unable to develop its products without copyright infringement – otherwise, there would not be enough data to train its models sufficiently (Milmo, 2024).

The aforementioned Studio Ghibli has not yet publicly commented on the imitation of the artistic style of its films using models such as GPT-4o. The studio is known for its traditional and hand-drawn animation techniques. In 2016, Hayao Miyazaki, the studio's founder, criticised the proposal to develop and use a machine that draws like a human in filmmaking, saying: "We humans are losing faith in ourselves." (MPNFW, 2016). Given the studio's principles, it is difficult to imagine that they would willingly permit their works to be used in AI model training, but it cannot be ruled out in the absence of an official statement. In the meantime, Japanese politicians have begun to discuss the legal ramifications of this situation. While it would be the studio's own responsibility to initiate legal proceedings, the politicians have commented: "If AI-generated content is determined to be similar to or reliant on preexisting copyrighted works, then there is a possibility that it could constitute copyright infringement" (Mullicane, 2025).

Professional artists invest significant time in mastering various forms of art and developing distinctive artistic styles. Many are unhappy that their work is being used for AI training without their consent or financial compensation. Legal actions have been initiated against several companies, such as Stability AI, and Midjourney. Although the court initially dismissed the artists' claims, it recognised the possibility of illegal use of copyrighted material in 2024 after a re-filing with amended arguments (Porterfield, 2023; Cho, 2024).

The legal status of generative AI currently remains unresolved. Since image generation models do not store the actual dataset but convert it into model weights during training, it may be challenging to prove copyright infringements. Furthermore, not all major AI companies disclose the datasets they use for model training, and some, such as Stability AI, argue that their actions constitute pastiche rather than copyright infringement (Wyn Davies, 2024). Clear legal guidelines, therefore, still need to be established.

In 2024, the European Union (EU) introduced the AI Regulation 2024/1689, which will come into force in 2026. Article 105 of the Regulation acknowledges that the development of generative AI poses problems for artists, authors, and other creatives. Although generative AI tools are not classified as high-risk, they do belong to a category that is subject to transparency requirements. In 2026, generative AI companies will have to disclose that content is generated by AI and publish sufficiently detailed descriptions of training datasets containing copyrighted material (European Parliament, 2024). Although this does not completely solve the fundamental problem, it is a step towards protecting creators.

In the EU, works of art are automatically protected by copyright upon creation, and the copyright lasts for up to 70 years after the death of the author (WIPO, 2003). It is not mandatory to go through a formal application process, but optional registration is possible if desired (Your Europe, 2025). It should be noted that artificial images generated by AI tools are not currently protected by copyright. In the United States, courts have ruled in several cases that the author of an artwork must be a human in order to be protected by copyright (Brittain, 2023).

2.2. Principles of image generation

This section outlines how image generation models work, in order to provide an understanding of how exactly they interact with and learn from visual data.

AI models are neural networks consisting of multiple layers of artificial neural nodes that mimic the behaviour of biological neurons. These neural networks require large training datasets to learn to perform a variety of tasks as accurately as possible, including image generation. There are different methods for training neural networks.

Generative adversarial networks (GANs), developed in 2014, were among the first deep learning architectures capable of generating new images. GANs consist of two dynamically updated neural networks – a generator and a discriminator. The generator is trained on a dataset of images to create new artificial images with the aim of increasing the probability that the discriminator will make a mistake. The discriminator is a classifier that predicts whether the provided image is a real or an artificial one. The two networks compete during training until the generator is able to produce such convincing images that the discriminator struggles with classification. At this point, the GAN can be used for image generation tasks (Goodfellow et al., 2014).

A variational autoencoder (VAE) is another type of machine learning architecture that consists of two components – an encoder and a decoder. The encoder takes input data from a dataset and tries to understand its features. The data is compressed into the latent space, which is a low-dimensional space where only meaningful information about the input data is retained. Instead of outputting a single point of data, VAE outputs the standard distribution in the latent space, which shows how much the values can vary. The decoder selects a single value from the distribution and attempts to reconstruct the original input by generating new data. Incorporating variation into the process provides more diverse generation possibilities and ensures that the model does not simply memorise the original dataset (Bergmann and Stryker, 2024).

Diffusion models are currently the most commonly used architecture for image generation. To learn to generate new images, diffusion models gradually add Gaussian noise to the input image until it contains only noise that bears no resemblance to the original. The models then learn the structure of the input images by reversing the diffusion process - removing the noise in a structured way to gradually reveal more image features, e.g. eyes, lips, etc., in a photo of a human. After training, the model can generate images from randomly selected noise that are not direct copies of the images in the original dataset but are very similar in structure (Sohl-Dickstein, 2015). Diffusion models can be combined with large language models (LLMs) to create a guided diffusion model, e.g. text-to-image models such as Stable Diffusion and Midjourney.

Diffusion models offer higher quality and more stable results than GAN models, but they are computationally intensive and therefore much slower. Stable Diffusion models use a modified and improved approach - using the principles of VAEs, the diffusion process is implemented in latent space rather than in the pixel space of the image. Such latent diffusion models require fewer computational resources while providing higher output quality (Rombach et al., 2022).

2.3. Adapting image generation models for specific purposes

Although large generative models are trained on datasets with millions of images, fine-tuning them for specific tasks often requires far fewer samples. For example, models like Stable Diffusion can be adapted for a specific purpose using a method called Low-Rank Adaptation (LoRA) with as little as 20 image samples (Holostrawberry, 2025). LoRA significantly reduces the required computational cost and time, making fine-tuning of models accessible to users with consumer-grade GPUs (Martineau, 2024).

Originally developed for adapting LLMs to specific tasks such as analysing legal documents, LoRA is now also widely used in image generation. While previous techniques required retraining all model weights to fine-tune the model, this method focuses only on a subset consisting of the model's attention layers while freezing the rest. The targeted layers are responsible for ensuring that the generated images match the text prompts (Hu et al., 2022).

Numerous platforms on the Internet, such as CivitAI, host user-made LoRA models with a wide variety of model customisation targets, such as the representation of popular fictional characters and celebrities, different poses, facial expressions, objects, clothing and backgrounds. There are also many models that aim to imitate the artistic style of contemporary artists, often without their consent.

On CivitAI, for example, a user has published a LoRA model that imitates illustrations by American comic artist Evan Stanley from IDW Publishing. The developer of this LoRA model has commented that commissioned work from an artist is too expensive, so they have trained this model to give others a free alternative. The model was used to generate over 62 thousand images and has been downloaded almost a thousand times. The model description does not indicate whether permission to use Evan Stanley's work in training was obtained from her, but the nature of the model suggests that this was probably not the case (AcanthAI, 2024).

2.4. Image generation practices and artist involvement on social media platforms

With the rise of generative AI, several social media platforms have begun training their own AI models on their users' data, including images. Many have updated their terms of service to allow such use. Often it is enabled by default and opting out is difficult or even impossible. Instagram, for example, with over two billion registered user accounts (as of February 2025), trains Meta AI services with its users' data. Opting out requires finding and filling out a hidden form in the app settings, in which you have to provide a reason for opting out. It only applies to future posts and is not available in many regions, including the United States (Jiménez, 2024).

Social media platform X (formerly Twitter) introduced similar rules in November 2024 to train its AI tool Grok (Pauley, 2024). So did Pinterest, a platform that was often used by artists as a source of inspiration, but which they are now abandoning due to the flood of AI-generated images (Dupré, 2025). Pinterest trains its image generator, Pinterest Canvas, with user data by default, regardless of the date the image was published (Pinterest, 2025).

In contrast, some platforms have adopted a more ethical stance. Adobe's artist platform Behance, for example, states that no user data is used to train its AI tool Adobe Firefly. They explain that it is only trained with public domain samples, as well as Adobe Stock data, with compensation paid to the authors (Adobe, a).

Nowadays, artists often rely on social media to promote their artwork and find employment opportunities. This is especially for emerging artists. For example, Latvian illustrator Paula Bobrova was discovered through her Instagram profile and was subsequently hired for the animated film "Flow", which later won the Golden Globe and Oscar awards (Dumbere, 2024; NFC, 2025a; NFC, 2025b). Bobrova created sketches of animal characters and the film's logo. This example illustrates why many artists, especially digital illustrators, cannot afford not to publish their work online, even if it sounds like it is the only viable way to avoid their artwork being used in AI training without permission.

Overall, given the uncertainties in the legislation, the misleading policies of social media platforms, and the ease of fine-tuning image generation models, we conclude that publicly available works by any artist are at risk of being incorporated into AI model training datasets. Therefore, artists need to protect their images and thus secure their copyright. The following chapter looks at the solutions currently available to protect images for this purpose.

3. Image protection software overview

This chapter deals with software solutions for the protection of artworks on the Internet. Three protection software – Glaze, Nightshade and Mist – are first described and then compared. The functionality and technical requirements of the individual software are given so that their practical suitability for artists without technical knowledge and/or necessary computer resources can be assessed. We also show how the images processed with the software look with different protection intensity settings.

3.1. Overview and comparison of Glaze, Nightshade and Mist

Glaze is an image protection software developed in 2023 by the Department of Computer Science at the University of Chicago to combat the unauthorised use of artwork when training and fine-tuning image generation tools to mimic artistic styles. The tool provides protection by masking the artist's work with a sufficiently different artistic style, chosen from a set of public domain images. This is achieved by using a pre-trained style transfer model and adding the resulting generated image to the original in the form of barely perceptible adversarial perturbations (UChicago, a).

If an image generation model is trained on multiple "glazed" images, it begins to associate the artist with the incorrect artistic style, and the resulting AI-generated images fail to successfully imitate the artist's original works. The Glaze team's study suggests that sufficient protection can be achieved if only 25% of an artist's online portfolio is "glazed" (Shan et al., 2023a).

Figure 1 illustrates the available perturbation intensities of Glaze. The difference from the original image is obtained by superimposing the processed image on the original and performing a contrast correction operation to improve visibility. The darker the pixels, the fewer differences there are with the original, while the lighter and brighter pixels indicate stronger perturbations, which become visibly discernible at higher settings. The following images for Figure 2 and Figure 3 were created in a similar way.

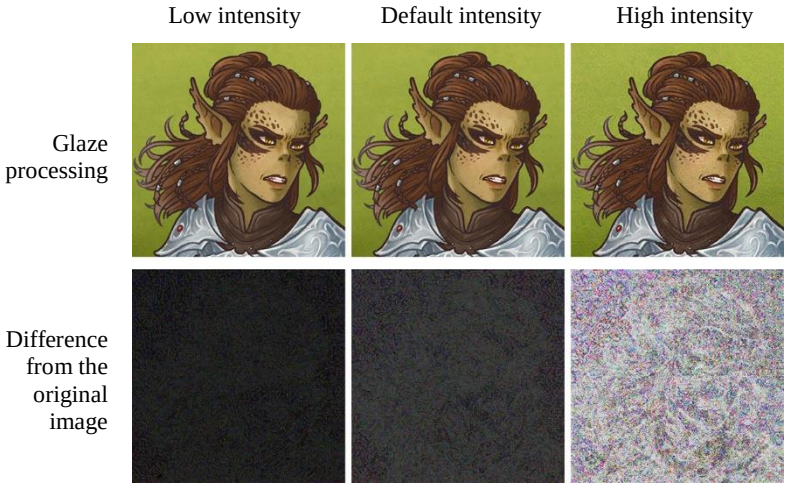


Figure 1. Sample processing with different Glaze perturbation intensities.

However, Glaze does not guarantee complete security for artworks. There have already been attempts to test the resilience of imperceptible perturbations against removal. Already in 2023, a group of professors and graduate students from Pennsylvania State University, Stony Brook University and the University of Illinois, Urbana Champaign in the USA proposed a "purification" method called IMPRESS, attempting to evaluate several contemporary protection methods, including Glaze (Cao et al., 2023). While this method demonstrated partial success on historical artistic styles, the developers of Glaze argue that it fails on contemporary works – its primary protection target (Shan et al., 2023b). Another study suggested several types of image manipulation that are effective in removing Glaze's changes (Hönig et al., 2024). Subsequent release of Glaze version 2.1 improved the robustness of the software against the "purification" methods (UChicago, 2024).

Nightshade, also developed by the University of Chicago's Department of Computer Science in 2023, is designed as an offensive tool that turns images into "poisoned" samples that disrupt and degrade the performance of image generation models by exploiting

"concept sparsity" – the relatively limited representation of specific concepts (e.g. "cat", "forest", "impressionism") in the training datasets of large-scale models (UChicago, b).

Nightshade improves on primitive types of data poisoning attacks by subtly manipulating the features of the image at the pixel level while preserving the description that matches the visual content. For example, an image showing a dog is correctly described as a "photo of a dog" but contains small perturbations that change the representation of its features closer to the appearance of a cat in the eyes of the AI model. A successful Nightshade attack, which changes the model's understanding of a concept to an incorrect one, is achievable even with 50 "poisoned" samples and it has an additional effect on related concepts. For instance, "poisoning" the concept of "dog" affects the model's perception of "puppy", "husky" or "wolf", teaching it to generate creatures that look closer to cats (Shan et al., 2024).

Figure 2 illustrates Nightshade version 1.0.2 processing with different intensity levels. As can be seen, the low and default intensity images are visually similar to each other, but the high intensity image results in noticeable artefacts.

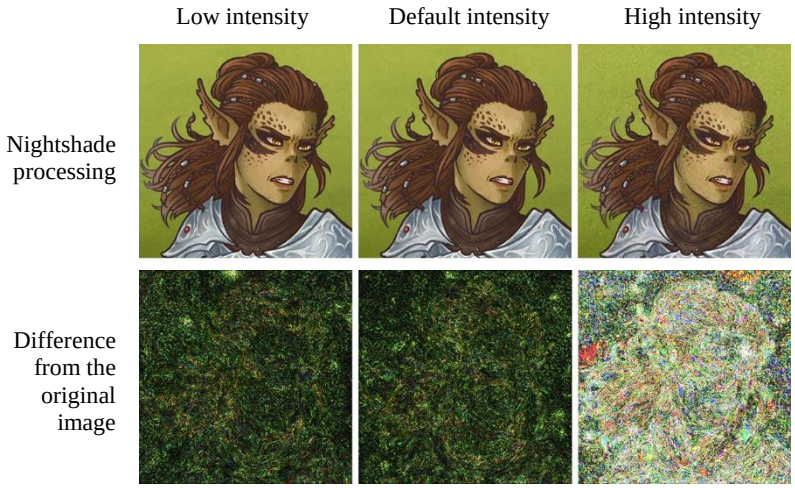


Figure 2. Nightshade "poisoned" samples with different levels of intensity.

Mist, developed in 2023 by the Psyker Group, is a targeted attack aimed at model classifiers, in contrast to Glaze and Nightshade, which aim to mislead models without degrading output quality. It works by embedding a specially selected chaotic pattern into images that causes the AI models to generate lower quality images, rendering the output unusable.

Mist uses two specially designed patterns that have high black and white contrast, frequent repetition and similarity to Moiré patterns. One incorporates the Mist logo, while the other – the NeurIPS logo. An image generation model trained on Mist-processed images begins to reproduce the chaotic patterns, thereby degrading the overall visual quality of all generated images (Zheng et al., 2023).

Figure 3 demonstrates images processed with different Mist version 2.0 intensity settings, ranging from 0 to 32. The default intensity is 12, while 1 was chosen as low

intensity and 32 as high. Compared to Glaze and Nightshade, Mist perturbations are significantly more visible. This is especially true for the highest intensity, where the used NeurIPS logo is easily perceptible.

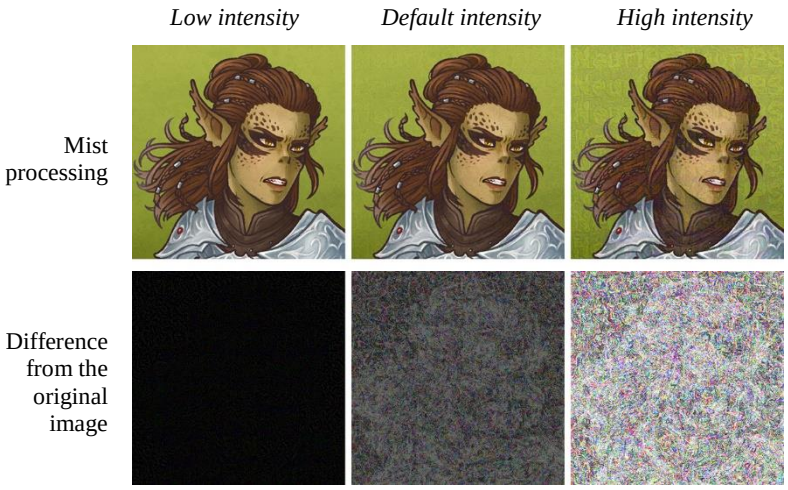


Figure 3. Mist processing samples with different levels of intensity.

Table 1 provides an overview of the three image protection tools and outlines the advantages and limitations of each using five criteria: Approach (image protection technique), Availability (required operating systems and/or existence of an online version), Requirements (minimum video random-access memory), Advantages, and Limitations. Furthermore, it should also be noted that all three tools share a drawback – they are hardware-intensive, which may be unsuitable for artists whose personal computers lack the necessary GPUs, or for those who do not use a computer to create art (for example, some use only iPads or traditional materials).

Table 1. Comparison of artwork protection tools Glaze, Nightshade and Mist

Criteria	*Glaze*	*Nightshade*	*Mist*
Approach	Facilitating incorrect style learning by masking the image with public domain artists' art styles	"Poisoning" models by exploiting concept sparsity	Introducing chaotic patterns into images to degrade the output's quality
Availability	• Windows/macOS • Web version WebGlaze	• Windows/macOS	• Windows/Linux • Google Colab notebook
Requirements	5 GB VRAM	5 GB VRAM	6 GB VRAM
Advantages	• Least visible perturbations • Easy to use, accessible (in terms of platform) • Receives the most updates	• Easy to use • Effective with relatively few poisoned samples • Affects related concepts	• Accessible (in terms of platform) • Approach theoretically provides equal protection for all art styles
Limitations	• Difficult to get a WebGlaze invite • Effectiveness depends on the user's art style	• Lack of an online version • Possibly unsuitable for illustrations depicting more than one clear subject	• Cannot run CPU mode on devices with non-NVIDIA GPUs • Errors and lack of user-friendly customisation in the Colab version • Most visible perturbations

3.2. Other types of protection software

Apart from these adversarial perturbation application tools, there are several other approaches to protect images from unauthorised use in the training of image generation models.

ArtShield applies an invisible watermark to protect images from being automatically scraped for training datasets. It mimics the watermarks used by AI image generation models to prevent AI-generated images from entering training datasets. The watermark is embedded by converting the image from RGB to YUV channels and applying discrete wavelet transforms (Xie, 2023). However, ArtShield does not protect artwork from users who can manually download any image from the Internet to train AI models with it.

Nepenthes and Iocaine are anti-scraping methods that function as "digital tarpits". They trap web crawlers that do not respect "robots.txt" anti-crawl directives in an endless maze where they are fed incomprehensible data created by a Markov babbler that produces text mimicking the structure of English sentences but lacking any meaning (Belanger, 2025). Although these methods were developed to protect websites from text scraping, they could potentially be repurposed by artists using personal portfolio websites.

Cara and ArtGram are social media platforms that were created for publishing artworks while disallowing the posting of AI-generated images. Cara is integrated with Glaze, which registered users can use instantly to protect their published works, while ArtGram claims to protect users' works with unique identifying signatures. In addition, Cara offers the possibility of posting job opportunities for artists, and ArtGram offers an online store with materials for artists and other creatives.

Finally, Have I Been Trained (https://haveibeentrained.com/) is a website where users can check whether their artworks appear in publicly available datasets. Using this website, a user can request to exclude their works from future model training, but, of course, there is no way to retroactively remove them from already trained models.

It is clear from this overview of adversarial perturbation applications Glaze, Nightshade and Mist that their use may in many cases be difficult for artists who (a) are not specialists in the field of computer science and (b) do not have access to sufficiently powerful computer equipment. On the other hand, other types of protection software can only partially protect images. Currently, there is a lack of practical comparative experimentation with the three tools, which we aim to provide further in this article. It is also important to determine which of these methods artists find the most effective and visually acceptable.

In order to assess how exactly artists and artwork viewers (non-artists) evaluate the protection of images, the authors trained several small AI models and then developed a survey. This is discussed in the next chapter.

4. Research Design

Here we describe the preparation of our custom dataset, the fine-tuning of the Stable Diffusion model and the process of sample generation with the trained LoRA models. We conclude this chapter by describing how the survey was designed and conducted. Finally, we describe the computational metrics used to compare with survey results.

4.1. Dataset preparation

To ensure the authenticity of the results, it was decided to create an independent dataset instead of using publicly available collections. This was further motivated by findings of

the Glaze team that using public domain artworks might not be as effective in testing, as these works could already be present in the training datasets of large-scale models. Additionally, this choice provided an opportunity to evaluate the protection performance, particularly for digital illustrations, which are more difficult to protect than, for example, online reproductions of paintings executed initially on canvas in oil or other material techniques.

As established in Section 2.3, LoRA models adjust only a small selection of model weights and require around 20 or more training samples to achieve style imitation results (Holostrawberry, 2025). Therefore, for our dataset, we chose 20 digital drawings created by one of the authors of the article within the last four years using the digital art software Paint Tool Sai and the graphic tablets Wacom Intuos Pro Medium and Huion Kamvas Pro 16. All drawings depict stylised portraits of various characters on a coloured background rendered in a consistent digital technique. These particular drawings were never published online, which ensures their absence from any online training datasets. The drawings were cropped and saved as .png files with a resolution 512 x 512 pixels.

A description was created for each drawing and stored in separate .txt files. These descriptions were initially generated using the open source image description model Large Language and Vision Assistant (LLaVA) and manually corrected to improve accuracy (Hugging Face, 2023). The descriptions included distinguishing features of the depicted figures such as gender, age, hairstyle, hair and eye colour, clothing, accessories, facial expression, pose and visible additional objects. The files of the drawings and descriptions were numbered uniformly in pairs (e.g. 1.png and 1.txt), and the prepared dataset was uploaded to Google Drive for further processing. Figure 4 presents selected sample images from the prepared dataset.

Female, silver hair, bob, grey eyes, white shirt, white bow tie, smirking, beauty mark, blue earrings	Male, short hair, brown hair, serious, brown eyes, white shirt, red vest, blue tie, gold bracelet, raised fist	Female, green hair, fluffy hair, green eyes, black and white dress, striped gloves, surprised, slightly open mouth, hand in hair, small robot on shoulder	Male, medium hair, brown hair, brown beard, brown eyes, smiling with teeth, purple jacket, silver earring

Figure 4. Dataset samples. The activation word “valstyle” and the keyword “character” identified by Nightshade were also included in each prompt. Author of the images – Valerija Januseva.

Next, the dataset was processed with the selected protection software. In total, eight datasets were created for the experiment:

- Original images without protection,
- Glaze protection (default and high intensity),
- Nightshade protection,
- Nightshade and Glaze combination,
- Mist protection (low, default and maximum intensity).

The processing with Glaze and Nightshade was done using the respective desktop software. Due to the lack of powerful computer resources, running the software in CPU mode took quite a long time – processing an image with Glaze took two to three hours, and with Nightshade – one hour. The Mist desktop software had inexplicable runtime failures that could not be resolved. Therefore, the online Google Colab notebook was used instead. It took about two minutes and 34 seconds to process one image, and the entire dataset was completed in 51 minutes and 21 seconds. Using Google Colab required resolving some errors by modifying the requirements.txt file.

4.2. Training image generation models

For privacy and security reasons, model training was conducted using entirely open source resources and a private workspace. The widely used text-to-image model Stable Diffusion 1.5 was used as the base model. The Python code for fine-tuning was deployed in the format of a Jupyter Notebook on Google Colab, utilising the platform's free NVIDIA Tesla T4 GPU resources with 16 GB of VRAM. To access Stable Diffusion, a free Hugging Face account was created. The authentication token obtained was set as a secret parameter (HF_TOKEN) in the Google Colab environment. For fine-tuning, the LoRA method was chosen due to its computational efficiency and suitability for the style imitation task. Initially, an attempt was made to use LoRA training scripts made by GitHub user *kohya_ss* (Kohya, 2022). As it caused some runtime failures, a more Colab-friendly modification by GitHub user *hollowstrawberry* was implemented instead (Hollowstrawberry, 2023).

The hyperparameters were adjusted to optimise the quality of image generation. Table 2 summarises the final hyperparameter configuration used across all LoRA models. The selection of values was based on a combination of recommendations for LoRA training (Hollowstrawberry, 2023) and the authors' own experimentation. For example, a learning rate that is too low ($1\times10^{-4)}$ results in the generation of overly realistic faces that do not match the original works' stylisation, while a rate that is too high (1×10^{-3}) produces broken, chaotic output. It is also important to balance other hyperparameters accordingly. Therefore, we set the U-Net learning rate to 5×10^{-4}, while the text encoder learning rate was kept lower at 1×10^{-4} to balance visual and textual learning. We set repeats to 20, batch size to 2, and trained over 10 epochs, resulting in a total of 2,000 steps. On average, the training of one LoRA model required approximately 32 minutes.

Table 2. Hyperparameter configuration for model training

Hyperparameter	*Explanation*	*Value*
U-Net learning rate	Controls how much the model weights are updated with each step to learn visual elements and structure.	5×10^{-4}
Text encoder learning rate	Controls how much the model weights are updated with each step to learn textual descriptions. It is recommended to use a lower value than the U-Net learning rate.	1×10^{-4}
Repeats	Dataset image repetitions during training.	20
Batch size	The amount of training data in each training round.	2
Epochs	The number of iterations during training of the entire dataset fed to the model.	10
Steps	The total number of iterations the model processes each batch of data to update the weights. Total steps = epochs * (dataset size * repeats / batch size).	2,000

4.3. Sample generation

A separate LoRA model was fine-tuned with each of the dataset variants mentioned in Chapter 4.1. Multiple samples were generated by loading each model's weights as a .safetensors file into the base Stable Diffusion 1.5 model using the AutoPipelineForText2Image pipeline from the Hugging Face Diffusers library.

The images were created in the same Google Colab environment as before. Loading the base model took about a minute, while loading the LoRA weights into the base model took about two seconds. To try another LoRA model in the same session, the base model was refreshed, which took another 20 seconds each time.

Five prompts similar to the description format used in the training dataset were created for image generation. They contained the activation word "valstyle", as well as a list of character features chosen in such a way that their combination was distinct from the original dataset samples. For each model, ten images were generated per prompt, and one of them was randomly selected for use in the survey.

Figure 5 shows samples generated with the trained LoRA models. Each row corresponds to a text prompt used in generating the corresponding image. The columns

represent the LoRA models that were trained on the specified dataset and used for generating the respective images.

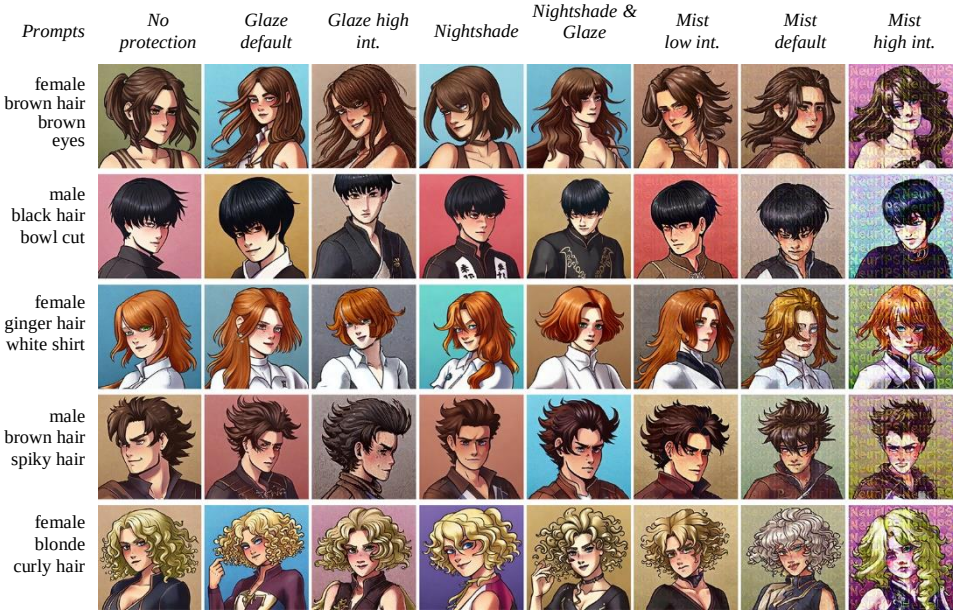


Figure 5. Sample images generated with the fine-tuned LoRA models. Each prompt also contained the activation word "valstyle" and the keyword "character".

4.4. Survey design

An anonymous user survey was conducted to evaluate the generated samples in terms of (a) similarity to original artworks, (b) overall image quality, (c) practical applicability and visual acceptability, and (d) agreement or disagreement with the use of image protection tools. To ensure that the survey questions met the above criteria for image analysis, a pilot test was conducted with a small group of selected participants from target groups of artists and non-artists. Based on the results of this pilot test, the wording of some questions was clarified, thereby improving the quality of the survey. The survey consisted of 41 questions and was divided into four parts.

1. The first section collected demographic data and also respondents' background, including involvement with creating art, to analyse whether there were differences between artists' and artwork viewers' (referred to as non-artists in the survey and in the description of the survey results) evaluations of artworks. It also included questions about respondents' experience with image generation and their views on the use of generative AI.

2. The second section was available only to artists and offered six questions asking about (a) the forms of art they work with, (b) the use of social media for publishing art and (c) awareness of protection software.

3. In the third section, respondents were shown both authentic drawings and style-mimicking AI-generated images. Respondents were asked to evaluate how close they think the generated samples were to the originals. They were told to consider (a) the stylisation of facial features, (b) the choice of colours, (c) the textures and (d) the overall image quality. The evaluation was based on a Likert scale.

4. In the fourth section, respondents were asked to rate the image quality as well as the practical applicability and visual acceptability of the illustrations after processing with protection software.

A total of 71 people took part in the survey, most of whom (99%) were between 18 and 36 years old. The participants included 45 artists as well as 26 artwork viewers (non-artists). The study was distributed among University of Latvia students and (professional and hobbyist) artists. These focus groups were selected to elicit the opinions of artists and non-artists. A deviation in group distribution occurred because some students were also hobbyist artists. The authors of the paper accepted this shift, believing it reflects the contemporary situation in which, thanks to the democratisation of digital art creation tools, representatives of other professions are actively working as hobbyist artists. Moreover, it coincided with the empirical observations of one of the authors, who is herself engaged in digital illustration as a hobby.

60% of participating artists worked both traditionally (drawing, painting, etc.) and digitally (digital illustration, graphic design, 3D modelling, etc.). Six artists worked only with traditional and twelve only with digital art creation techniques. One artist additionally specified their work with traditional printmaking techniques – lithography, letterpress and serigraphy.

Finally, to complement the survey with computational metrics, we have extracted features from images by using the VGG-19 convolutional network, as well as computed image embeddings using the open source openCLIP model (Simonyan and Zisserman, 2014; MLfoundations, a). For both, we have calculated the average cosine similarity between sets of images and expressed it in percentages. We used the original 20 artwork dataset as the reference to which compare each of the generated datasets, both with and without protection.

The next chapter analyses the results of the study and highlights the strengths and areas for future improvement of screen protection tools from the respondents' perspective.

5. Results

The results of the survey are presented in six figures and three tables. They show the respondents' attitudes towards image generation as well as their assessment of various aspects of the quality of protection offered by the generated images.

Figure 6 shows the frequency of use by respondent groups when asked how often they use AI image generation tools. Most respondents had a negative attitude towards image generation with AI tools. 21.1% were neutral, and only two respondents had a positive opinion. Despite the observed negative attitude, more than half of the respondents have

used image generation tools at least occasionally. 68.4% of respondents who have tried using AI tools at least a couple of times have used them strictly for entertainment, while 13.2% have utilised them for professional purposes. Several artists indicated that they were forced to use AI image generation tools for assignments at university or school. Some other artists used these tools to gain inspiration, as well as for quick visualisation of ideas and concepts required by their workplace. One non-artist had considered using image generation for text visualisation for children, but in the end decided not to use such tools. In addition, when respondents were asked about their habits when using AI, it was also found that 74.6% of them had observed cases where someone had tried to use artificial intelligence to imitate the artistic style of a contemporary artist they knew. Two artists stated that someone had specifically tried to imitate their artistic style.

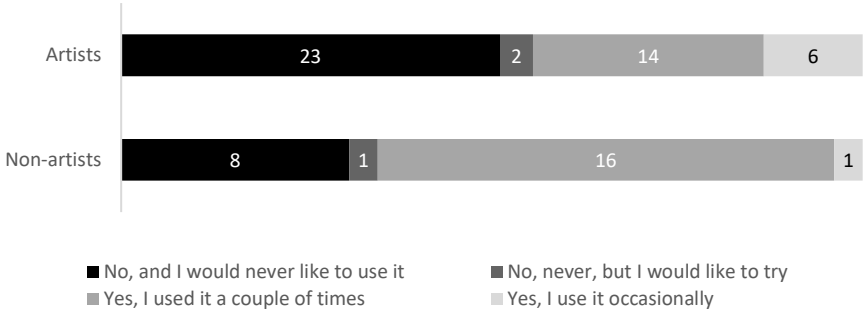


Figure 6. Frequency of use of image generation tools in the artist and non-artist groups.

80% of the surveyed artists publish images of their work online. As can be seen in Figure 7, the most popular social networking platforms among them are Instagram, X (formerly Twitter), and Tumblr. One artist uses a personal portfolio website.

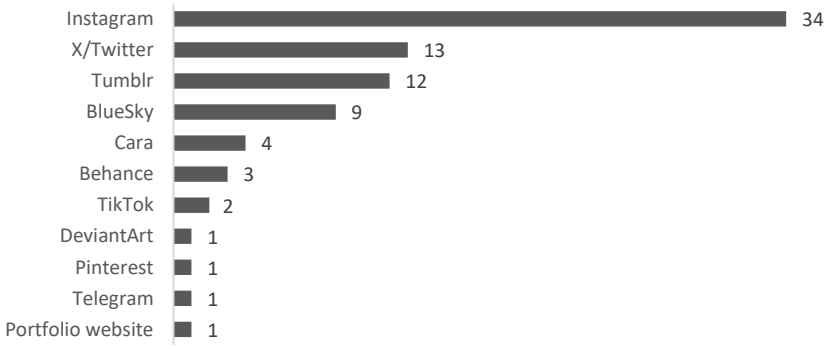


Figure 7. The choice of artists for publishing images on social media.

66.67% of artists do not protect their work when publishing it online. Those who do usually only add a watermark or signature. Only four artists use Glaze, two use Nightshade, and none use Mist. Four artists use the Cara platform, which integrates Glaze. The artist who created a personal portfolio website stated that their copyright is described there.

Figure 8 shows the distribution of average similarity ratings, divided into four individual criteria – facial feature stylisation, colour choices, textures and overall quality of the image. The lower the percentage rating, the worse the results of the style imitation. As can be seen, the models do not mimic the texture and image quality of the original illustrations as well as the colours and facial feature stylisation. Particularly poor texture and quality are observed for samples with high Glaze intensity and all Mist samples.

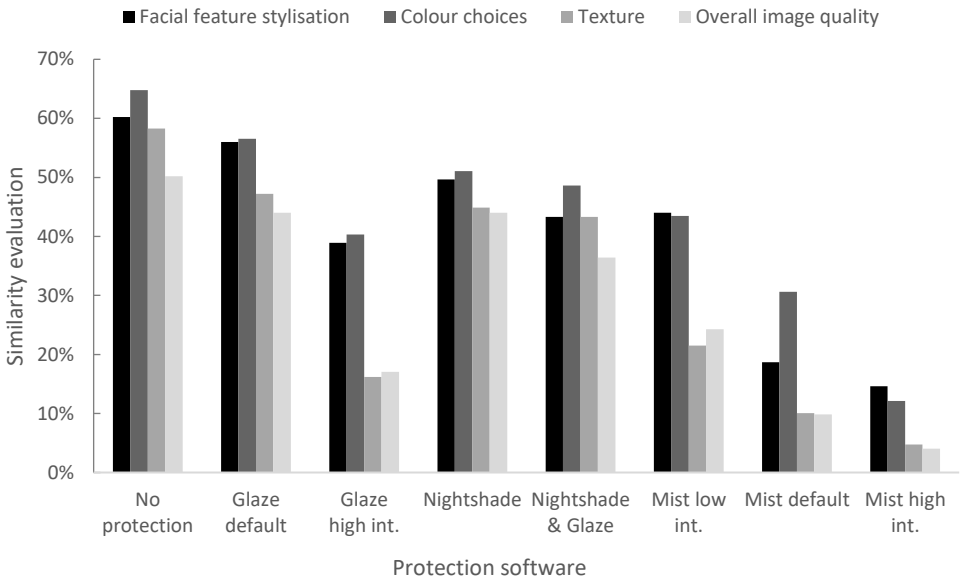


Figure 8. Evaluations of similarity to original illustrations.

Figure 9 shows the combined average similarity ratings of all respondents, divided into artist and non-artist groups. The higher the percentage, the greater the similarity to the original illustrations. The closest to the originals were the samples that were generated without any protection. However, their ratings were lower than expected – 56.18% (artists) and 62.13% (non-artists). As can be seen, the ratings for all Mist intensities and high Glaze intensity are lower compared to other software. The ratings for default Glaze, Nightshade and the combination of Nightshade and Glaze are closer to the generated images without protection, so the effectiveness of these settings is not as strong.

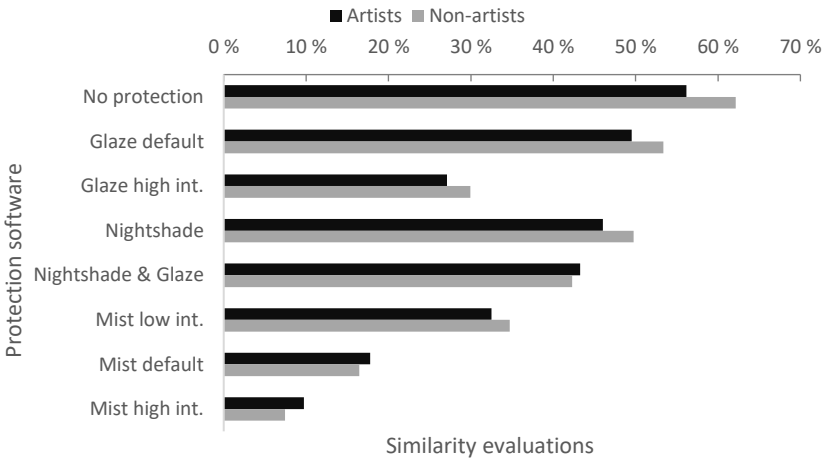


Figure 9. Assessments of the similarity of the generated images to the original illustrations.

Figure 10 shows the rating of the quality and practical applicability of protected images by group. The higher the percentage, the more respondents agreed to apply the respective perturbations to the images to achieve the specified level of protection. The low and default intensity Mist samples have the highest ratings – 80% and 82.22% of artists were satisfied with the visual intensity of Mist perturbations and the offered level of protection. Ratings of Glaze, Nightshade and their combination are generally lower, with the exception of high intensity Glaze protection. In almost all cases, artists were more willing to use higher intensity perturbations than non-artists.

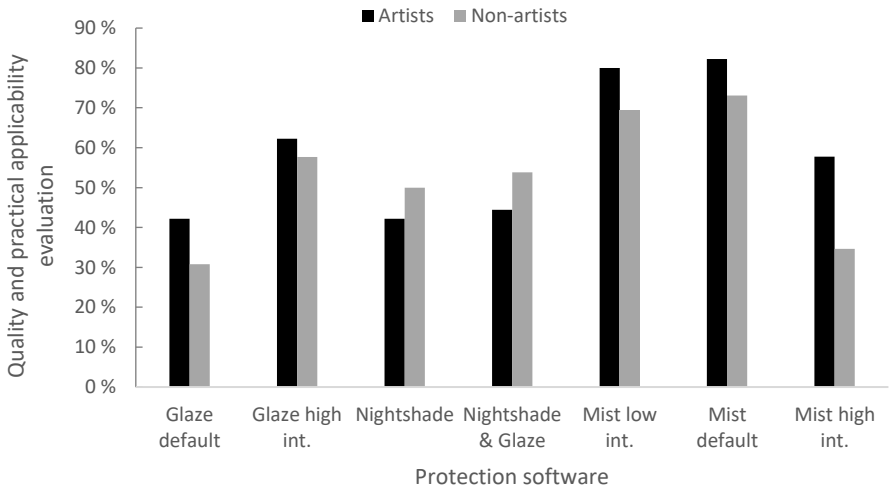


Figure 10. Assessment of the quality and practical applicability of protected images.

Figure 11 reveals the respondents' explanations as to why they would not agree to apply perturbations with protection software, divided into artist and non-artist groups. High intensity Mist perturbations are the only case where both artists and non-artists would reject the application because of artefacts of processing being very visible. Meanwhile, Glaze, Nightshade, and their combination do not provide sufficient protection, especially according to artists.

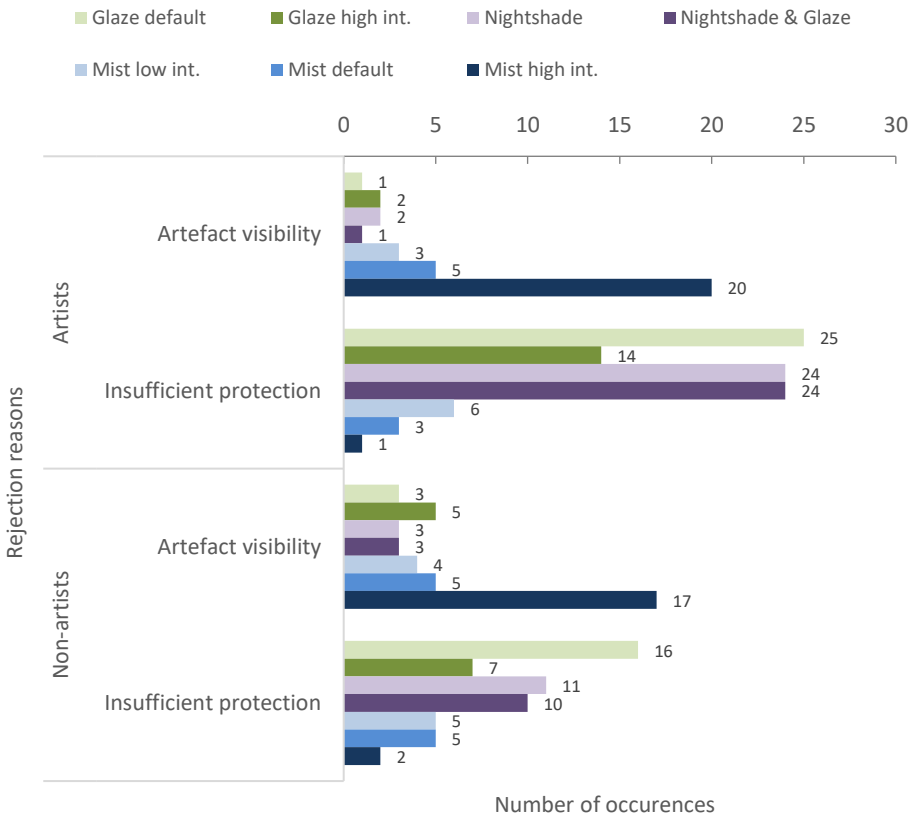


Figure 11. Reasons for rejecting the use of image protection software from the perspective of artists and non-artists.

Table 3 summarises the survey results of the evaluation of the images. The higher the percentage value, the more successful the imitation of the original style. Since the generated images without protection only received a similarity score of 58.36%, the samples whose evaluations were at least 20% below this value were considered successful protection.

Table 3. Evaluation results of the similarity of the generated images with the originals. Lower percentages here indicate more effective protection by the tools – such results are marked in green

Protection	*Average evaluation*	*Facial features*	*Colour choices*	*Texture*	*Image quality*
None	58.36%	60.21%	64.79%	58.27%	50.18%
Glaze default	50.92%	55.99%	56.51%	47.18%	44.01%
Glaze high int.	28.13%	38.91%	40.32%	16.20%	17.08%
Nightshade	47.40%	49.65%	51.06%	44.89%	44.01%
Nightshade & Glaze	42.91%	43.31%	48.59%	43.31%	36.44%
Mist low int.	33.32%	44.01%	43.49%	21.48%	24.30%
Mist default	17.30%	18.66%	30.63%	10.04%	9.86%
Mist high int.	8.89%	14.61%	12.15%	4.75%	4.05%

Table 4 reveals the assessment of image quality and visual acceptability ratings of the protected images. Ratings above 50% are considered acceptable, i.e. most respondents would agree to the application of the appropriate protection to the artwork.

Table 4. Results of the evaluation of protected images. Ratings above 50% are marked in green

Protection/criteria	*Agree to use*	*Disagree to use*	*Most common disagreement reason*
Glaze default	38.03%	61.97%	Insufficient protection
Glaze high int.	60.56%	39.44%	Insufficient protection
Nightshade	45.07%	54.93%	Insufficient protection
Nightshade & Glaze	47.89%	52.11%	Insufficient protection
Mist low int.	76.06%	23.94%	Insufficient protection
Mist default	78,87%	21,13%	Artefact visibility
Mist high int.	49,30%	50,70%	Artefact visibility

Table 5 shows both the survey results and the results obtained by computing style similarity. The higher the evaluation, the closer that dataset is to the original artworks.

Table 5. Overview of the evaluation of unprotected and protected images

Protection	*Survey evaluation*	*VGG-based evaluation*	*CLIP-based evaluation*	*Agree to use*
None	58.36%	97.27%	95.77%	-
Glaze default	50.92%	96.17%	95.43%	38.03%
Glaze high int.	28.13%	94.51%	93.87%	60.56%
Nightshade	47.40%	97.72%	96.04%	45.07%
Nightshade & Glaze	42.91%	95.86%	95.56%	47.89%
Mist low int.	33.32%	95.16%	92.80%	76.06%
Mist default	17.30%	88.76%	88.87%	78.87%
Mist high int.	8.89%	73.64%	78.01%	49.30%

Both the VGG-based and CLIP-based similarity measures showed high stylistic similarity across all protection software and settings (ranging from approximately 73% to 97%). These scores are noticeably different from the subjective survey evaluations. This discrepancy suggests that the neural models capture some visual or structural features that do not fully correspond to human perceptions of artistic style or visual image quality. Therefore, the objective similarity scores should be interpreted as indicators of representational closeness in feature space, rather than how they appear to human viewers. Mist received lower similarity scores across all metrics, both model-based and survey evaluations, while Glaze and Nightshade are of higher similarity to the original illustrations.

Other studies (Shan et al., 2023a) indicate that Glaze has higher effectiveness for styles that are closer to traditional paintings, but is currently limited for simpler illustration

styles. It corresponds with the low ratings of Glaze in our evaluations – at least with the default intensity settings. The combination of Nightshade and Glaze leads to stronger protection against style imitation, but is still considered insufficient by almost half of the survey respondents.

The Nightshade protection results did not demonstrate concept "poisoning", most likely because the experiment conducted in this study was different from the experiments conducted by the Nightshade developers. However, the perturbations still affected the model training and slightly impaired the model's ability to imitate the style of the illustrations.

Mist received the highest scores for style imitation protection, image quality and practical applicability. Default intensity was preferred. 78.87% of respondents would agree to use it, and the overall similarity of the generated images to the original illustrations was rated at 17.30%. While the similarity rating for high intensity Mist is even lower (8.89%), only 49.30% of respondents would want to use images with such highly visible perturbations. Although Glaze provides a protection method that could better prevent "purification" attempts, it does not offer the same level of protection for all illustration styles. In contrast, the targeted perturbations offered by Mist protect all image types equally, which may be more appealing to artists.

Research results show that the hypothesis set at the beginning has been partially confirmed. Of the tools analysed, Mist affects the models' ability to generate images the most – it significantly reduces the quality of the generated outputs. On the other hand, the results of Glaze and Nightshade, as well as their combination, are too close to those of the unprotected artworks. Therefore, it is difficult to consider them completely successful, apart from using the high intensity settings of Glaze. However, despite the shortcomings, the majority of respondents acknowledge that when publishing works of art online, they need to be protected.

6. Discussion

This article contributes to the emerging field of digital artwork protection by providing a practical comparison of three image protection tools – Glaze, Nightshade, and Mist – which aim to protect artists from unauthorised imitation of their artworks by image generation models.

Although the developers of each software have conducted their own experiments, these used separate experiment designs and training data, making it difficult to draw precise comparisons between these three tools. Unlike their studies, our experiments were designed to compare the tools using a unified methodology. We evaluated both technical effectiveness (similarity between original and generated images) and visual usability (acceptability of the intensity of perturbations) of these tools. Furthermore, we tested these tools on digital illustrations and art, a type of artwork creation medium that has not been well researched.

Mist proved to be the most effective tool, reducing similarity to original artworks to 17.30% at default intensity while maintaining high acceptability (78.87%). In contrast, Glaze and Nightshade showed similarity scores above 42%, close to unprotected images (58.36%), and were rated as insufficient by most respondents (61,97% and 54,93%

respectively). The VGG-based and CLIP-based evaluations, although much higher in similarity, still correlate with the survey evaluations. These results suggest that protection methods like Glaze and Nightshade may not adequately protect digital illustrations, highlighting the benefits of Mist's targeted perturbations.

These findings indicate that image protection is not only theoretically possible but also supported by the artist community. Additionally, the trade-off between protection strength and visual acceptability was demonstrated: while high intensity Mist achieved the lowest similarity score (8.89%), fewer than half of respondents (49.30%) considered the outputs usable due to the visibility of artefacts. This highlights the importance of developing tools that balance technical effectiveness with visual quality.

We acknowledge that our study has limitations, primarily because our experiments were conducted on illustrations by a single artist. However, we have defined a methodology that could be useful for other researchers. Future work could continue our experiments with different artistic styles from several artists, which could be particularly useful for further evaluating the effectiveness of Glaze's style transfer.

Furthermore, lower results for models trained with unprotected samples indicate that the LoRA models used in the experiment could be improved to produce more accurate style imitation results relative to the original illustrations. The effectiveness of perturbation "purification" methods could also be further investigated.

It should be noted that the field of AI is still rapidly evolving, and the fight against unauthorised AI training could be a never-ending arms race. However, this research provides quantitative evidence of the advantages and limitations of existing image protection tools, and it establishes Mist as a potential candidate for further practical implementation. Our study also demonstrates that artists are now both aware of such technologies and are willing to use them. By combining technical experimentation with user-centred evaluation, this research strengthens the theoretical and practical foundation for protecting artworks in the era of generative AI.

7. Conclusions

This article investigated the effectiveness of image protection tools, such as Glaze, Nightshade, and Mist, against unauthorised imitation of artworks using AI image generation models. Using a combination of empirical research, technical experiments and a survey involving respondents including both artists and viewers of artworks (non-artists), this study offered insights into the current state of digital artwork protection. The results highlighted the need for protection tools, especially considering how easy it is to customise models with only 20-30 samples of artworks without the author's consent.

Of the tools tested, Mist demonstrated the most consistent performance, successfully deteriorating the quality of the model's output images even at low and default intensity settings, resulting in as low as 33.32% and 17.30% similarity to originals, respectively. Furthermore, Mist received lower similarity scores across all metrics, both model-based and survey assessments, whereas Glaze and Nightshade were more similar to the original illustrations. However, as the technical experiments have shown, the current version of the Mist software might be challenging to use for artists without technical knowledge. The tool needs to be improved to make it more user-friendly for everyone, regardless of

technical skills. However, 78.87% of the artists who participated in the survey were willing to use Mist's default intensity settings for image protection. This suggests that image protection is practically possible, and artists support it.

Overall, the findings confirm that practical image protection is feasible and supported by artists, though improvements in usability and balance of technical effectiveness and visual quality remain necessary. The study provides a theoretical and practical knowledge base for those interested in protecting their artwork and for further research.

Acknowledgements

We would like to express our sincere gratitude to Professor Juris Borzovs for suggesting the publication of this article and for his helpful advice on improving its contents. We are also grateful to the reviewers for their feedback and recommendations.

References

AcanthAI (2024). *Sonic IDW Style (Evan Stanley)*, available at https://civitai.com/models/596723/sonic-idw-style-evan-stanley.

Adobe (a). *Our approach to generative AI with Adobe Firefly*, available at https://www.adobe.com/ai/overview/firefly/gen-ai-approach.html#.

Belanger, A. (2025). *AI haters build tarpits to trap and trick AI scrapers that ignore robots.txt*, available at https://arstechnica.com/tech-policy/2025/01/ai-haters-build-tarpits-to-trap-and-trick-ai-scrapers-that-ignore-robots-txt/.

Bergmann, D., Stryker, C. (2024). *What is a variational autoencoder?*, available at https://www.ibm.com/think/topics/variational-autoencoder.

Brittain, B. (2023). *AI-generated art cannot receive copyrights, US court says*, available at https://www.reuters.com/legal/ai-generated-art-cannot-receive-copyrights-us-court-says-2023-08-21/.

Cao, B., Li, C., Wang, T., Jia, J., Li, B., Chen, J. (2023). *IMPRESS: Evaluating the Resilience of Imperceptible Perturbations Against Unauthorized Data Usage in Diffusion-Based Generative AI*, Advances in Neural Information Processing Systems, **36**, pp. 10657–10677.

Cho, W. (2024). *Artists Score Major Win in Copyright Case Against AI Art Generators*, available at https://www.hollywoodreporter.com/business/business-news/artists-score-major-win-copyright-case-against-ai-art-generators-1235973601/.

Coca-Cola (2024). *The Holiday Magic is coming*, available at https://www.youtube.com/watch?v=4RSTupbfGog.

Dumbere, L. (2024). *In the same boat* (in Latvian), available at https://ir.lv/2024/09/18/viena-laiva/.

Dupré, M. H. (2025). *Pinterest Changes User Terms So It Can Train AI on User Data and Photos, Regardless of When They Were Posted*, available at https://futurism.com/pinterest-data-photos-train-ai.

European Parliament (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council*, available at https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). *Generative Adversarial Nets*, Advances in Neural Information Processing Systems, **27**.

Harding, S. (2025). *Netflix will show generative AI ads midway through streams in 2026*, available at https://arstechnica.com/gadgets/2025/05/netflix-will-show-generative-ai-ads-midway-through-streams-in-2026/.

Heikkilä, M. (2022). *This artist is dominating AI-generated art. And he's not happy about it*, available at https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/.

Ho, K. K. (2024). *Database of 16,000 Artists Used to Train Midjourney AI, Including 6-Year-Old Child, Garners Criticism*, available at https://www.artnews.com/art-news/news/midjourney-ai-artists-database-1234691955/.

Hollowstrawberry (2023). *Kohya Colabs*, available at https://github.com/hollowstrawberry/kohya-colab.

Holostrawberry (2025). *Make your own Loras, easy and free, using Colab*, available at https://arcenciel.io/articles/1

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W. (2022). *LoRA: Low-Rank Adaptation of Large Language Models*, ICLR, **1**.

Hugging Face (2023). *LLaVa*, available at https://huggingface.co/docs/transformers/en/model_doc/llava

Hönig, R., Rando, J., Carlini, N., Tramèr, F. (2024). *Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI*, available at https://arxiv.org/pdf/2406.12027.

Jiménez, J. (2024). *Worried About Meta Using Your Instagram to Train Its A.I.? Here's What to Know*, available at https://www.nytimes.com/article/meta-ai-scraping-policy.html.

Kohya, S. (2022). *sd-scripts*, available at https://github.com/kohya-ss/sd-scripts.

Martineau, K. (2024). *Serving customized AI models at scale with LoRA*, available at https://research.ibm.com/blog/LoRAs-explained.

McCormack, J., Cruz Gambardella, C., Rajcic, N., Krol, S.J., Llano, M.T., Yang, M. (2023). *Is Writing Prompts Really Making Art?*, International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar), pp. 196–211.

Milmo, D. (2024). *'Impossible' to create AI tools like ChatGPT without copyrighted material, OpenAI says*, available at https://www.theguardian.com/technology/2024/jan/08/ai-tools-chatgpt-copyrighted-material-openai/.

MLfoundations. (a) *OpenCLIP*, available at https://github.com/mlfoundations/open_clip.

MPNFW (2016). Manhattan Project for a Nuclear-Free World. *Hayao Miyazaki's thoughts on an artificial intelligence*, available at https://www.youtube.com/watch?v=ngZ0K3lWKRc.

Mullicane, E. D. (2025). *"A Violation of the Law": After ChatGPT's AI Attack on Studio Ghibli, Lawmakers Are Looking to Take Legal Action*, available at https://screenrant.com/studio-ghibli-ai-artwork-chatgpt-japan-lawmakers-illegal/.

NFC (2025a). National Film Centre. *"Flow" becomes the first Latvian film to win the "Golden Globes" award*, available at https://www.nkc.gov.lv/en/article/flow-becomes-first-latvian-film-win-golden-globes-award.

NFC (2025b). National Film Centre. *"Flow" brings Latvia first-ever "Oscar"*, available at https://www.nkc.gov.lv/en/article/flow-brings-latvia-first-ever-oscar.

OpenAI (2025). *Introducing 4o Image Generation*, available at https://openai.com/index/introducing-4o-image-generation/.

Parshall, A. (2023). *How This AI Image Won a Major Photography Competition*, available at https://www.scientificamerican.com/article/how-my-ai-image-won-a-major-photography-competition/.

Pauley, C. (2024). *Elon Musk's X Can Now Use Your Data to Train Its AI*, available at https://web.archive.org/web/20250328021044/https://9meters.com/entertainment/social-media/elon-musks-x-can-now-use-your-data-to-train-its-ai.

Pinterest (2025). *Privacy Policy*, available at https://policy.pinterest.com/en/privacy-policy.

Porterfield, C. (2023). *Judge dismisses most of artists' copyright lawsuit against AI image generators*, available at https://www.theartnewspaper.com/2023/10/31/california-judge-dismisses-most-of-artists-ai-copyright-lawsuit/.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 10684–10695.

Roose, K. (2022). *An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy*, available at https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html/.

Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanochka, R., Zhao, Y. B (2023a). *Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models*, USENIX Security Symposium, **32**, pp. 2187–2204.

Shan, S., Wu, S., Zheng, H., Zhao, B. Y. (2023b). *A Response to Glaze Purification via IMPRESS*, available at https://arxiv.org/pdf/2312.07731.

Shan, S., Ding, W., Passananti, J., Wu, S., Zheng, H., Zhao, B. Y. (2024). *Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models*, IEEE Symposium on Security and Privacy, pp. 807–825.

Shōyu (2025). *Shōyu Annual Birthday Party*, available at https://www.instagram.com/shoyu.riga/reel/DIbzdgvNdSB/.

Simonyan, K., Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*, available at https://arxiv.org/pdf/1409.1556.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., Ganguli, S. (2015). *Deep unsupervised learning using nonequilibrium thermodynamics*, International Conference on Machine Learning, **32**, pp. 2256–2265.

Sriram, A. (2025). *Ghibli effect: ChatGPT usage hits record after rollout of viral feature*, available at https://www.reuters.com/technology/artificial-intelligence/ghibli-effect-chatgpt-usage-hits-record-after-rollout-viral-feature-2025-04-01/.

UChicago (2024). University of Chicago. *A Note on the new attack paper: "Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI"*, available at https://glaze.cs.uchicago.edu/update21.html.

UChicago (a). University of Chicago. *What is Glaze?*, available at https://glaze.cs.uchicago.edu/what-is-glaze.html.

UChicago (b). University of Chicago. *What is Nightshade?*, available at https://nightshade.cs.uchicago.edu/whatis.html.

WIPO (1979). *Berne Convention for the Protection of Literary and Artistic Works*, available at https://www.wipo.int/wipolex/en/text/283698.

Wyn Davies, C., Dennis, G. (2024). *Getty Images v Stability AI: the implications for UK copyright law and licensing*, available at https://www.pinsentmasons.com/out-law/analysis/getty-images-v-stability-ai-implications-copyright-law-licensing.

Xie, A. Z. (2023). *Stable Diffusion invisible watermarker - the math and our algorithm improvements*, available at https://artshield.io/blog/post/stable-diffusion-invisible-watermarker-the-math-and-our-algorithm-improvements.

Your Europe (2025). *Copyright*, available at https://europa.eu/youreurope/business/running-business/intellectual-property/copyright/index_en.htm#inline-nav-2.

Zheng, B., Liang, C., Wu, X. (2023). *Targeted Attack Improves Protection against Unauthorized Diffusion Customization*, available at https://arxiv.org/pdf/2310.04687.

Received August 20, 2025, revised October 15, 2025, accepted October 20, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, 806-833
https://doi.org/10.22364/bjmc.2025.13.4.04

Trends and Developments in Fuzzy Logic for Medical Diagnosis: A Bibliometric Analysis

Asefeh TAJODIN[1], Mehmet ÜNVER[2]

[1] Shiraz University, Shiraz, Iran
[2] Ankara University, Faculty of Science, Department of Mathematics, Ankara, Türkiye

s40030153@hafez.shirazu.ac.ir, munver@ankara.edu.tr

ORCID 0000-0003-4383-4139, ORCID 0000-0002-0857-1006

Abstract. The complexity and uncertainty inherent in medical diagnosis pose significant challenges for accurate and timely decision-making. To address these challenges, advanced technologies such as clinical decision support systems have gained prominence. Integrating fuzzy logic (FL) into these systems offers a practical solution for managing uncertainty, as it provides a more flexible framework than traditional binary (on–off, zero–one) approaches. In particular, FL has contributed to improved diagnostic accuracy, the handling of uncertainty and vagueness in patient-reported symptoms, and the support of decision-making when data is incomplete or ambiguous. This study presents a bibliometric analysis of the leading research trends in this field, offering insights into the evolution and impact of FL in medical diagnosis. By facilitating the representation of nuanced degrees of clinical parameters such as pain intensity or fever severity, FL enables the modeling of complex disease patterns, the development of predictive analytical models, and the construction of intelligent healthcare systems. This capability enhances diagnostic precision, supports early detection, and extends the applicability of decision support tools to complex and uncertain clinical scenarios. At the time of conducting this bibliometric analysis, searches in Scopus databases revealed no prior bibliometric studies focused on the application of FL in medical diagnosis.

Keywords: Clinical decision support systems, uncertainty modeling, artificial intelligence in healthcare, bibliometric mapping, research trends

1. Introduction

A primary challenge confronting both developed and developing nations pertains to the provision of adequate medical care to the indigent population (Awotunde et al., 2014). A significant number of hospitals grapple with a shortage of skilled medical professionals, prompting nations to allocate substantial resources to address this issue (Kanchanachitra et al., 2011). Despite these endeavors, numerous nations encounter difficulties in fulfilling the demand for quality healthcare services. In some regions, accessing affordable healthcare services can be time-consuming, yet certain illnesses require immediate attention due to their nature. Delayed treatment can result in the spread of infectious diseases and increased risks of mortality (Matinfar and Golpaygani, 2022).

Hence, it is crucial to consider reducing costs and providing prompt treatment within the healthcare system. To address these challenges, modern medical diagnostics increasingly rely on computer-related technologies, which are continually advancing (Charanbir et al., 2023).

The employment of FL is driven by the necessity to account for uncertainty and vagueness that are inherent in conventional medical diagnostic practices. Uncertainty constitutes a foundational component in disease diagnosis, wherein numerous variables contribute to the complexity of the diagnostic process (Ahmadi et al., 2018). This uncertainty is frequently associated with imprecise observations and subjective experiences reported by patients. For instance, patients may respond to questions regarding symptoms of weakness with responses that are ambiguous, such as "somewhat weak" (Ohayon, 1999). The pervasive uncertainty and lack of clarity in medical data pose challenges for traditional diagnostic methods, complicating the process of accurately identifying illnesses without error (Meyer et al, 2021). In this scenario, the integration of FL into decision support systems (DSSs) is regarded as the most effective method for addressing the complex nature of human illness. FL offers robust reasoning techniques capable of managing uncertainties and imprecision (Tang and Ahmad, 2024). Utilizing scientific principles that address imprecision, FL can provide a framework for managing the inconsistencies in medical data. The development of fuzzy models is informed by the expertise, observations, and experiences of medical professionals, thereby serving as the foundational element for medical diagnostic systems (Phuong and Kreinovich, 2001). The integration of FL technology facilitates the extraction of definitive conclusions from vague, imprecise, and ambiguous medical information, thereby enhancing the reliability and precision of diagnostic decisions.

In the domain of medicine, particularly in oriental medicine, numerous medical concepts are characterized by ambiguity. The vague nature of these concepts and their interconnections necessitates the application of "FL" (Pandey, 2016). Zadeh (1965) introduced a theory outlining the formalization of "fuzzy" (non-binary) properties. In fuzzy set theory, X is represented as a set of possible values ranging from 0 to 1, while a fuzzy property is described by a function $\mu \rightarrow [0.1]$. The value $\mu(x)$ denotes the extent to which x possesses the property (e.g., the degree of pain experienced by x). Fuzzy set theory differs from traditional binary logic in that it recognizes the gray area between truth and falsehood. This representation of therapeutic conditions and indications is less inflexible than a basic on-off or zero-one arrangement would imply (Phuong and Kreinovich, 2001). In fuzzy set theory, linguistic terms, symbolic and numerical values are mapped to each fuzzy variable and fuzzy set, respectively.

In practice, the membership degree serves as an input to the rule-based inference mechanism of diagnostic systems such as CADIAG-2. For example, if a patient's temperature is 38.8∘C, the membership degree for the fuzzy set "High Fever" is $\mu HF = 0.6$ indicating a 60% presence of this symptom. Each symptom (e.g., high fever, rash, fatigue) is represented by a corresponding fuzzy value in [0,1]. In CADIAG-2, these values are combined through fuzzy rules, such as

IF High Fever AND Rash THEN Measles (high likelihood),

where the logical AND is often implemented using the minimum operator

$$\mu_{\text{Measles}} = \min(0.7,0.9) = 0.7.$$

If symptoms have different diagnostic importance, weighted aggregation can be applied as

$$\mu_{\text{Measles}} = \frac{w1 \times 0.7 + w2 \times 0.9 + w3 \times 0.6}{w1 + w2 + w3}.$$

The resulting degree of membership for a disease can be interpreted as a risk score: low values may lead to routine monitoring, medium values to additional testing, and high values to immediate clinical action. This process bridges the abstract mathematical definition of a fuzzy set with concrete medical decision-making.

Zadeh's seminal work pioneered a novel approach to FL rules of inference, characterized by logical operations over membership functions (Zadeh, 1983). These rules, exhibiting an "if–then" structure, have been successfully applied in various domains, including forecasting and medical diagnosis. In the medical context, such rules facilitate the interpretation and processing of patient data to generate clinically relevant outputs (Stoean and Stoean, 2013).

Each fuzzy rule consists of an antecedent (the "if" part) and a consequent (the "then" part). For example, in a clinical setting:

IF temperature is high AND rash is present, THEN the likelihood of measles is high.

Formally, a fuzzy "if–then" rule, also referred to as a fuzzy implication, can be expressed as

$$\text{IF } x = A, \text{ THEN } y = B$$

where A and B are linguistic fuzzy values determined by the corresponding membership functions for variables x and y. For multiple variables $x_1, x_2, \ldots, x_n$, a set of such rules can be aggregated as

$$\text{IF } x_1 = A_1 \text{ AND } x_2 = A_2 \ldots \text{AND } x_n = A_n, \text{ THEN } y = B.$$

Here, $x_1, x_2, \ldots, x_n$ represent the features of the n-dimensional input vector X. This formulation allows complex medical conditions to be described as combinations of multiple symptoms and clinical findings, enabling the system to model uncertainty and support diagnostic reasoning beyond the limits of binary logic.

FL (FL) provides a powerful framework for modeling uncertainty in medical decision-making, bridging human reasoning with computational intelligence. Through processes such as fuzzification, rule inference, and defuzzification, FL systems can represent vague clinical concepts and handle imprecise data effectively (Vyas et al., 2022). This interpretability and flexibility have led to its integration into modern medical decision-support systems and hybrid computational models, such as fuzzy neural networks and IoT-based healthcare architectures (Mohod et al., 2025).

In the domain of medicine, a considerable number of DSSs have been developed. In 1980, a real-time fuzzy control drug delivery system was utilized to regulate blood pressure in patients with open heart surgery. Smets (1983) developed a fuzzy model that employs expected utility theory and fuzzy numbers to optimize decision-making regarding renal transplants. Phuong and Kreinovich (2001) developed a fuzzy system for diagnosing various lung diseases by integrating diagnostic methods

from Eastern and Western medicine based on patient symptoms. Hayward and Davidson (2003) presented a study of a DSS for automating the application of clinical practice guidelines based on fuzzy methods. Beig et al. (2011) designed another model of a FL medical diagnosis control system. This system utilizes FL design, comprising a fuzzifier, an inference engine, a rule base, and a defuzzification process. It is capable of medical diagnosis with five inputs (protein, red blood cells, lymphocytes, neutrophils, and eosinophils) and three outputs. A systematic review titled "Medical Applications of FL Inference Systems" was conducted by Thukral and Bal (2019), encompassing literature from the past decade (2008-2018) on the use of FL in various medical applications and methodologies developed during that time frame. The paper focuses on eight prevalent medical conditions, including heart disease, asthma, liver disease, breast cancer, Parkinson's disease, cholera, dental issues, and diabetes. It explores the potential of implementing FL in different medical domains in the future based on these applications. Matinfar and Golpayegani (2022) developed a fuzzy expert system for the early diagnosis of multiple sclerosis. This system mapped symptoms to fuzzy sets and established rules for predicting the disease. Concurrently, Myrzakerimova et al. (2024) spearheaded research endeavors focusing on advanced mathematical models for disease diagnosis and prediction. Their research led to the development of automated systems based on these models. These systems were designed to reduce subjectivity through computational assessment of imprecision and uncertainty. Additionally, they sought to enhance diagnostic modeling using fuzzy set theory. Building upon these findings, Myrzakerimova and Kolesnikova (2024) created a fuzzy system capable of diagnosing kidney diseases.

While many researchers have explored FL and medical diagnosis, there is a lack of studies addressing the evolution and mapping of this scientific domain. This study dwells on the issues related FL and medical diagnosis in the context of today’s research by identifying the most important lines of research, researchers, and research concentration areas. To explore the trends in the area of fuzzy set theory and medical diagnosis, this study used bibliometric analysis. Bibliometric analysis is a methodology that employs statistical and quantitative techniques to evaluate academic literature. These techniques are designed to identify influential authors, map collaboration networks, and uncover emerging research trends. The application of these techniques provides comprehensive insights into research landscapes and enhances understanding of various domains (Kumar, 2025). The present study employed VOSviewer, a software program designed to facilitate the analysis of bibliometric data. This software offers a comprehensive visualization of literature relevant to the subject of material selection, thereby enabling an in-depth analysis of the associated research trends. The utilization of VOSviewer facilitates a thorough exploration of the evolution of decision-making processes in material selection, leading to the identification of significant trends and connections within the existing literature. The software assists in mapping essential studies, authors, and collaborative efforts, thereby providing a more holistic view of the research landscape. This approach not only illustrates the development of research concepts but also emphasizes the interrelationships among various contributors and institutions (Sahoo et al., 2024). This study aims to answer the overarching question: “What are the main trends and developments in the research on the application of FL in medical diagnosis, and how does this field appear in terms of scientific publications, prevalent topics, leading authors, and future directions?”

To achieve the aim of this paper, the sub-research questions (RQs) are formulated as follows:

RQ1: What is the distribution of research in the field of FL and medical diagnosis?
RQ2: What is the distribution of research in the field of FL and medical diagnosis in different periods of time?
RQ3: Who are the prolific researchers in the field of FL and medical diagnosis?
RQ4: What was the contribution of each university in the field of FL and medical diagnosis?
RQ5: Which journals mainly published FL and medical diagnosis research?
RQ6: What languages are the publishes in the field of FL and medical diagnosis mainly in?
RQ7: What was the distribution of each country in the field of FL and medical diagnosis publishes?
RQ8: What are the most important themes in the field of FL and medical diagnosis?

2. Methodologies

To address the research inquiries and discern patterns in the evolution of FL in medical diagnosis, a comprehensive bibliometric analysis is undertaken, focusing on the most significant scholarly contributions and areas of research concentration. This analytical method, which employs mapping and clustering techniques, consists of three primary phases: data acquisition, data processing, and results extraction (da Silva and de Souza, 2021).

The Scopus database is selected as the singular data repository for this bibliometric study due to its extensive scope, rigorous quality assurance protocols, and comprehensive citation analysis capabilities. Scopus is recognized as the foremost curated repository for abstracts and citations, encompassing a vast collection of scientific journals, conference proceedings, and academic monographs from over 5,000 international publishers across various disciplines (Salleh et al., 2023). Furthermore, Scopus offers advanced profiling for both authors and institutions and is fully compatible with bibliometric visualization instruments such as VOSviewer, thus facilitating efficient data extraction and analysis (Zainuldin and Lui, 2022). In comparison to databases like Web of Science, Scopus provides broader global coverage and a more heterogeneous array of content, thereby enabling a thorough and unbiased perspective of the research landscape (Salleh et al., 2023). The attributes of Scopus render it particularly suitable for executing comprehensive bibliometric analyses in our investigation. Scopus is regarded as one of the largest databases for abstracts and citations. It encompasses journal articles and conference proceedings across diverse fields and offers high compatibility with bibliometric tools such as VOSviewer. Its export functionalities are user-friendly and well-supported for bibliometric mapping. A total of 356 documents are acquired after the application of the search criteria. Although using only one database may introduce some bias, Scopus provides extensive coverage and is widely accepted in bibliometric studies, especially when using network mapping tools like VOSviewer.

For quantitative data analysis and visualization of bibliometric networks, the free-access VOSviewer (version 1.6.17) software is used, and co-authorship and co-words networks are constructed. VOSviewer is intended primarily for analyzing bibliometric networks, and it provides three visualizations: network visualization,

overlay visualization, and density visualization (Skačkauskienė, 2022). The bibliographic data retrieved from Scopus is examined using VOSviewer following its import. The data is exported straight from the Scopus database in CSV format to ensure complete compatibility with VOSviewer. A thesaurus document is meticulously constructed to amalgamate synonymous terminology and to consolidate variations of author names (e.g., 'Smith J.' and 'Smith, John'). The document adhered to VOSviewer's bifurcated format, wherein synonyms are categorized beneath a singular preferred designation. To ensure methodological rigor and transparency, this study followed the systematic mapping process outlined by Kitchenham and Charters (2007). We applied the PICOC framework to define the scope of our research and guide the formulation of search strings:

- Population (P): Studies focusing on FL applications in medical diagnosis
- Intervention (I): Bibliometric analysis of published literature
- Comparison (C): Not applicable (no intervention comparison is made)
- Outcome (O): Identification of publication trends, active contributors, and thematic clusters
- Context (C): Peer-reviewed scientific publications indexed in the Scopus database.

Based on this framework, we develop a structured search query: TITLE-ABS-KEY (fuzzy AND logic AND medical AND diagnosis). The query is applied in the Scopus database in January 2025, without restrictions on publication year, document type, or subject area. All retrieved records containing English titles, abstracts, or keywords are included to ensure linguistic consistency for bibliometric analysis. The extracted data is then analyzed using VOSviewer for network visualization and mapping.

The methodological framework for data collection and analysis is structured according to a predefined research protocol, which ensures transparency and reproducibility while allowing for minor adjustments depending on the specific research questions or stages of the inquiry. Consistent with this protocol, a comprehensive dataset was retrieved from the Scopus database and systematically analyzed using bibliometric mapping techniques. The parameters of the data collection and analytical process are summarized in Table 1.

Table 1. Data collection and analytical protocol

Topic	**Research Questions (Q1-Q8)**
Keywords	"FL" AND "Medical diagnosis" (including related terms identified during query refinement)
Database	Scopus (chosen for its broad coverage of multidisciplinary journals and compatibility with VOSviewer)
Search type	Title, Abstract, Keywords
Document type	Articles, reviews, conference papers
Analysis techniques	Co-occurrence, co-authorship, co-citation, trend analysis
Period	1974 to 2025 (January)

Following this protocol, the bibliometric search was executed in the Scopus database. The retrieved records were exported in CSV format and processed through VOSviewer (version 1.6.17) and Microsoft Excel for data visualization and network analysis. This workflow enabled the construction of co-authorship, keyword, and citation networks to identify research trends and collaborations within the domain. Figure 1 provides an overview of the steps taken in this research.

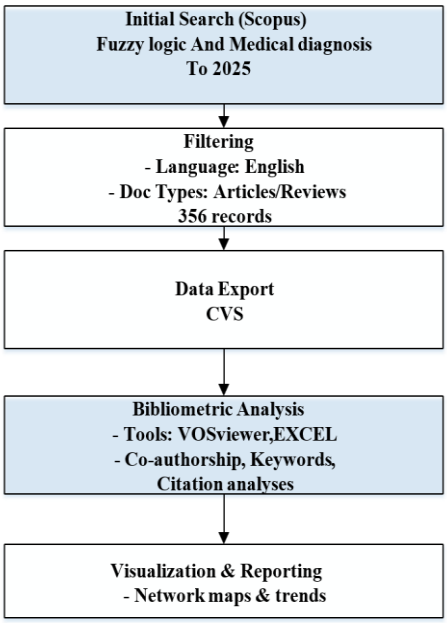


Figure 1. The stages of research

3. Results

In this section, the answers to the previously stated RQs are presented, with the results accompanied by the corresponding tables and graphs as follows.

3.1. What is the distribution of research in the field of FL and medical diagnosis? (RQ1)

As illustrated in Figure 2, the field of FL has been demonstrated to intersect with medical diagnosis. Among the various academic disciplines, computer science has demonstrated the most interest in this concept, with 31% of research, followed by engineering sciences (19%), medicine (13%), and mathematics (11%).

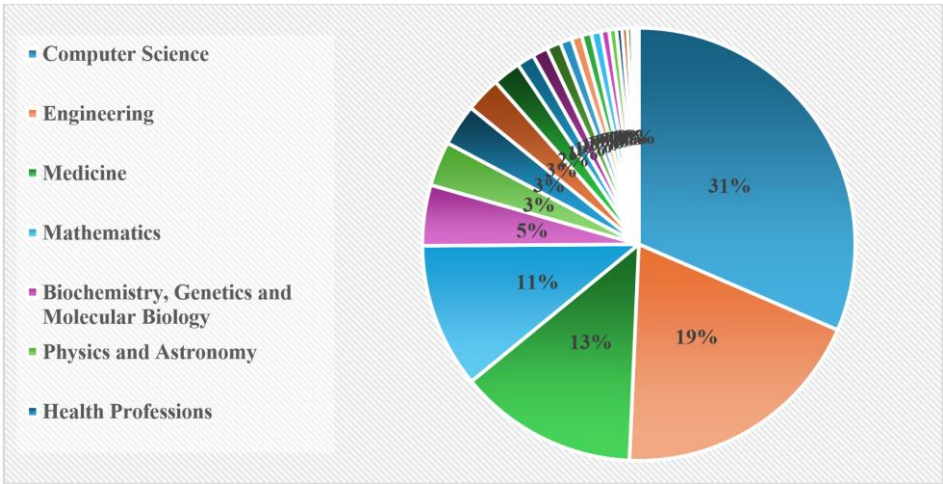


Figure 2. The distribution of research in the field of FL and medical diagnosis

Computer Science (31%): The highest percentage indicates that computer scientists are extensively involved in developing FL algorithms, AI models, and DSSs for medical applications. These contributions likely focus on areas such as expert systems, machine learning, and health informatics.

Engineering Sciences (19%): Engineers, particularly those in biomedical and electrical engineering, contribute to hardware implementations, sensor technology, and medical imaging applications using FL.

Medicine (13%): While medical professionals apply FL in clinical decision-making, disease classification, and personalized treatment strategies, their contribution is lower, possibly due to the field being more application-oriented rather than being theory-driven.

Mathematics (11%): The involvement of mathematicians suggests foundational research on FL theory, the development of new mathematical models, and the improvement of fuzzy inference mechanisms.

3.2. What is the distribution of research in the field of FL and medical diagnosis in different periods of time? (RQ2)

As illustrated in Figure 3, there has been an increasing trend in research studies concerning the application of FL in the field of medical diagnosis. The earliest documented studies in this domain date back to 1974, with two studies being conducted in that year. Notably, the year 2024 stands out as a significant peak in research activity, with 128 studies related to this subject being published. Note that the analysis was conducted in January 2025.

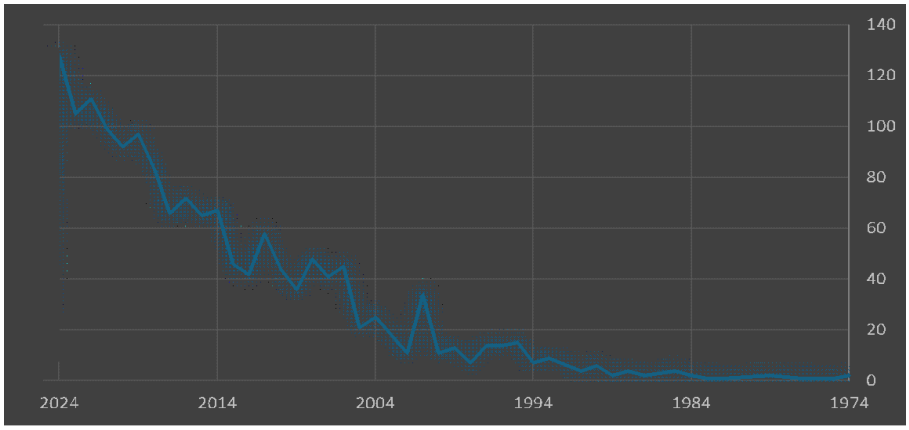


Figure 3. Distribution of research in the field of FL and medical diagnosis

The earliest documented studies on this topic show that researchers began exploring FL applications in medical decision-making nearly five decades ago. The small number of studies (only two) suggests that it is an emerging concept with limited adoption. Over the years, the number of studies has grown, reflecting a rising interest in FL for medical diagnosis. This trend likely corresponds to advancements in computational power, artificial intelligence, and the increasing complexity of medical decision-making.

The sharp rise in publications after 2020 (Figure 3) is closely tied to the global COVID-19 pandemic. Healthcare systems faced immense pressure and uncertainty, prompting the adoption of artificial intelligence (AI) and telemedicine tools. FL played a crucial role in handling imprecise symptom data, enabling diagnostic systems to triage patients based on indicators such as fever, oxygen saturation, and comorbidities. This surge in real-world relevance likely fueled the growing academic interest in FL-based medical applications.

COVID-19 has catalyzed the expansion of telemedicine and highlighted the imperative for uncertainty modeling within the healthcare sector. During the pandemic, there was a global demand for advanced health monitoring and diagnostic frameworks tailored for individuals with severe medical conditions (Rahman et al., 2023). The unprecedented scale of the crisis compelled healthcare systems worldwide to rapidly and substantially reconfigure their service delivery strategies (Temesgen et al., 2020). As a result, the application of FL in healthcare has attracted growing attention since 2020. Intelligent clinical DSS for triage has contributed to improving the quality of care in the emergency departments as well as identifying the challenges they have been facing (Fernandes et al, 2020). FL models can also be used for classifying the risk of medical equipment. Since FL is closer to the way humans think, it is expected to improve the prioritization of devices (Tawfik et al., 2013). FL is used in AI-based diagnostic systems, particularly for predicting heart disease, brain disease, prostate disease, liver disease, and kidney disease (Kaur et al, 2020).

The year 2024 marks a record high in research activity, with 128 studies published. This peak could be attributed to several factors, such as:

• Increased adoption of AI and machine learning in healthcare.
• Growing interest in explainable AI and interpretable decision-making models, where FL plays a key role.
• The expansion of digital health technologies and data-driven diagnostics.
• A possible surge in funding and interdisciplinary collaborations in medical informatics.

3.3. Who are the prolific researchers in the field of FL and medical diagnosis? (RQ3)

A total of 7 researchers have demonstrated remarkable productivity in the domain of FL applied to medical diagnosis, with each researcher having at least 9 studies in this field. Among these researchers, Adlassnig, K.P., has published 17 studies, while Al-Kasasbeh, R.T., has published 15 studies, positions them as the most prolific researchers in this area. The complete list of these researchers and the number of their publications are presented in Table 2.

Table 2. The prolific researchers in the field of FL and medical diagnosis

Author (s)	**Document number**
Adlassnig, K.P	17
Al-Kasasbeh, R.T	15
Hata, Y.	11
Straszecka, E.	11
Singla, J.	11
Obot, O.U.	9
Uzoka, F.M.E.	9

Figure 4 delineates the co-authorship network, constructed using VOSviewer software, among scholars engaged in the exploration of FL within the domain of medical diagnosis. This examination addresses pertinent research inquiries concerning the patterns of collaboration and scientific alliances prevalent in this discipline. By rendering a visualization of the interrelationships and clusters of authors, the figure elucidates principal contributors and their collaborative networks, thereby offering insights into the mechanisms of knowledge dissemination and collective advancement. As depicted in Figure 4, the investigation of authors conducting research in the realm of FL in medical diagnosis has identified 4,697 authors. Of these, 28 authors have disseminated more than five scholarly documents, which are categorized into five distinct clusters (Figure 4). The nodes symbolize authors engaged in collaborative relationships, and the dimensions of the nodes likely represent the significance or frequency of publications or collaborative endeavors. The connections between nodes are represented by lines, indicating instances of co-authorship. Lighter lines between groups denote less

substantial or infrequent collaborations. This analysis underscores that a mere 0.6% of the authors have produced five or more publications in the designated field (28 authors), with only 11 authors having released seven or more documents.

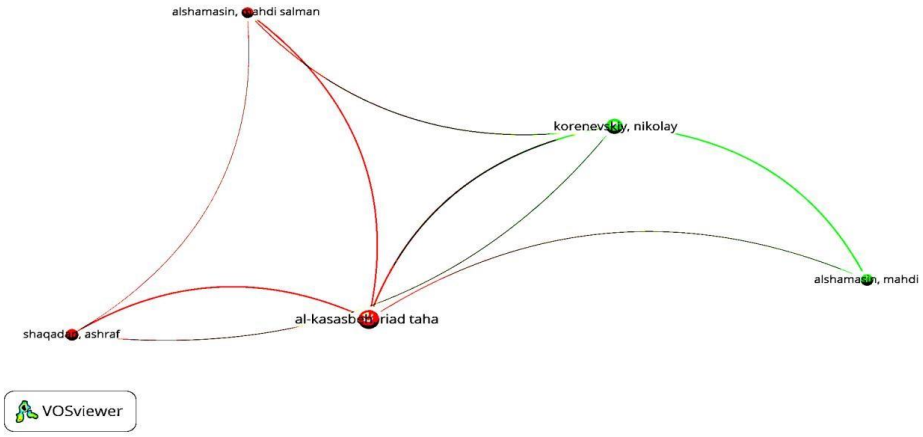


Figure 4. Analysis of authors contribution conducting research on areas of FL in medical diagnosis

3.4. What was the contribution of each university in the field of FL and medical diagnosis? (RQ4)

The following universities have been identified as the most prominent in the field of FL research in disease diagnosis: Lovely Professional University, Southwest State University, and Vellore Institute of Technology. Table 3 presents a comprehensive list of the most active universities in this field and the number of studies they have conducted.

Table 3. The contribution of each university in the field of FL and medical diagnosis

University	Number of studies
Lovely Professional University	24
Southwest State University	21
Vellore Institute of Technology	20
Silesian University of Technology	18
Al-Balqa Applied University	17
Medizinische Universität Wien	14
University of Uyo	13
Amirkabir University of Technology	11
University of Toronto	11
Jadavpur University	11

3.5. Which journals mainly published FL and medical diagnosis research? (RQ5)

The most prominent journals in the domain of FL studies in medical diagnosis include AI in Medicine, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), and Journal of Medical Systems. These journals and the number of their publications in this field are outlined in Table 4.

Table 4. Most important journals in FL and medical diagnosis research

Journals	Number of documents
Artificial Intelligence in Medicine	42
Lecture Notes in Computer Science Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics	40
Journal of Medical Systems	35
Computer Methods and Programs in Biomedicine	33
Computers in Biology and Medicine	32

3.6. What languages are the publishes in the field of FL and medical diagnosis mainly in? (RQ6)

It is important to note that in this study, documents with English titles, keywords or abstracts are searched and reviewed, so documents that are entirely in other languages are not included in the search scope. Although most publications are in English, a few non-English records (e.g., in Chinese, German, Russian, Turkish, etc.) were also identified. Therefore, the inclusion of Figure 5 provides a complete representation of the language diversity observed in the dataset, even though the dominance of English is visually overwhelming.

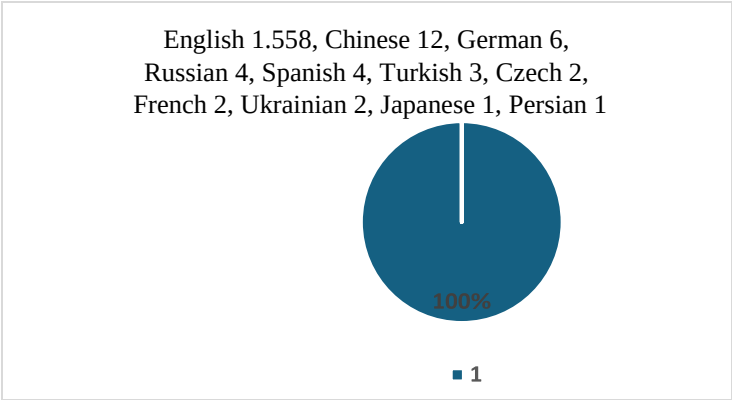


Figure 5. The distribution of studies on FL and medical diagnosis across different languages

3.7. What was the distribution of each country in the field of FL and medical diagnosis publishes? (RQ7)

Figure 6 presents a visual representation of the countries that have demonstrated the most activity in the domain of FL studies in medical diagnosis. These countries have conducted a minimum of 30 studies within this specific field. The three countries that have exhibited the most activity in this domain are India, China, and the United States of America.

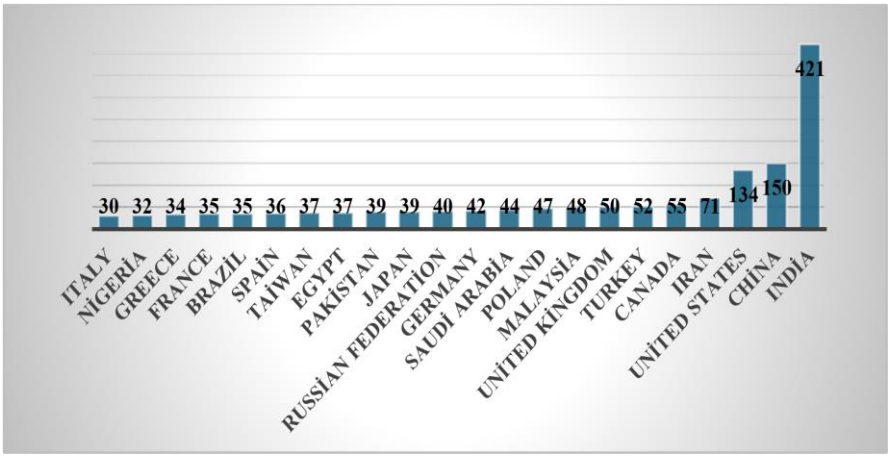


Figure 6. Distribution of each country in the field of FL and medical diagnosis publishes

India and China are leading the way in innovative research on intelligent disease detection, primarily due to significant governmental efforts and substantial investments in digital health infrastructures. The Ayushman Bharat Digital Mission (ABDM) of India, launched in 2021, aims to create an extensive national digital health ecosystem that improves accessibility, affordability, and quality of healthcare through interoperable platforms and real-time data exchange. Building upon India's vast digital public infrastructure, which includes over 1.2 billion Aadhaar digital identities and widespread internet connectivity, this initiative has accelerated the deployment of AI-driven and fuzzy-logic-based technologies in both urban and rural environments (Sharma et al., 2023; Velan et al., 2024).

China's preeminence in intelligent disease detection research is closely linked to its national health initiative, Healthy China 2030, which promotes the integration of digital health tools and AI-based diagnostics to strengthen disease prevention, early detection, and healthcare management (Tan et al., 2017). Consequently, both countries exhibit strong thematic clusters in fuzzy modeling for chronic disease management and diagnostic decision support, reflecting the translation of national policy priorities into measurable research outputs.

Conversely, the United States displays a relatively lower academic publication volume, which may be attributed to the predominance of proprietary, industry-led research activities within private technology and biotechnology companies. This divergence suggests a structural difference in knowledge dissemination, where publicly

funded open-access research dominates in Asian countries, while innovation in the United States often remains confined to industrial settings.

3.8. What are the most important themes in the field of FL and medical diagnosis? (RQ8)

The keywords co-occurrence network helps in the identification of main themes that are focused therein. The clusters of keywords of high relevance can be interpreted as research themes. Out of 10612 keywords, 2010 meet the co-occurrence threshold of 20 times. The highly connected keywords that appeared in FL in medical diagnosis, formed 4 clusters with 14926 links having a TLS of 106432. Figure 7 displays a simplified keyword network. Each cluster is represented by a color representing nodes of keywords. As shown in Figure 7, several keywords, such as FL, diagnosis, humans, disease, fuzzy inference, fuzzy sets and medical images have been the most frequently used in documents related to FL in medical diagnosis.

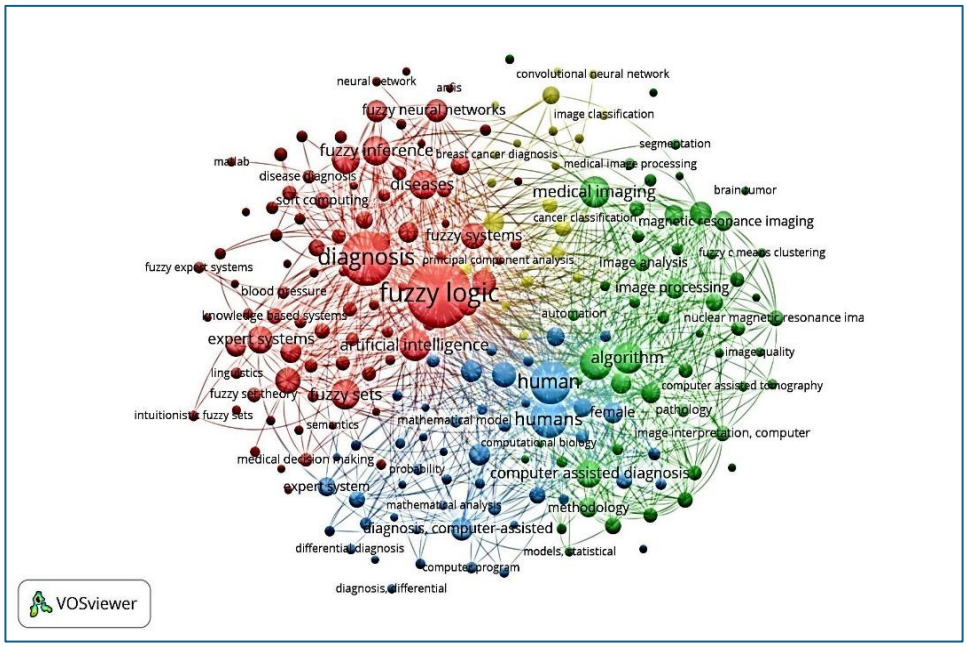


Figure 7. Keywords co-occurrence network in the field of FL and medical diagnosis

The keyword co-occurrence network in Figure 7 shows the main research topics and interrelationships in FL applications for medical diagnosis. Each color-coded cluster represents a specific research focus:

1. Red Cluster (FL in Medical Diagnosis): This cluster is centered around "FL", "diagnosis", and "diseases", indicating that FL is widely applied to medical decision making and disease classification. Other closely related terms include "fuzzy

inference", "fuzzy neural networks", "artificial intelligence", and "expert systems", suggesting a strong link between AI-based fuzzy methods and diagnostic accuracy.

2. Blue cluster (human-centered applications and statistical methods): Keywords such as "human", "computer-aided diagnosis", "statistical models", and "mathematical models" highlight research that focuses on patient-oriented diagnosis and computational methods. This suggests that a significant proportion of studies in this area involve human subjects and statistical tools to improve diagnostic accuracy.

3. Green Cluster (Algorithms and Image Processing in Diagnostics): This cluster contains terms such as "algorithm," "image analysis," "image processing," and "nuclear magnetic resonance imaging," indicating that machine learning and computer vision techniques are being integrated with FL for medical image analysis. These technologies are likely to be used to improve diagnostic accuracy by analyzing radiological and pathological images.

4. Yellow Cluster (Neural Networks and Automation in Diagnostics): The presence of "convolutional neural network (CNN)", "image classification", "medical imaging", and "automation" suggests the growing role of deep learning models and automation in medical image diagnosis. The most cited keywords in the field of FL and medical diagnosis are shown in Table 5.

Table 5. The most cited keywords in the field of FL and medical diagnosis

Cluster1 (85)	**Cluster2 (53)**	**Cluster3 (50)**	**Cluster4 (21)**
Anfis	Algorithm	Accuracy	Biological organs
Artificial intelligence	Algorithms	Adolescent	Breast cancer
Bioinformatics	Automated pattern recognition	Adult	Breast cancer diagnosis
Biomedical engineering	Automation	Aged	Cancer classification
Blood	Biology	Artificial neural network	Cancer diagnosis
Blood pressure	Brain	Clinical article	Classification algorithm
Cardiology	Brain tumor	Clinical decision making	Convolutional neural network
Cardiovascular disease	Breast neoplasms	Clinical decision support system	Convolutional neural networks
Classification	Breast tumor	Computer program	Database, factual
Classification (of information)	Cluster analysis	Controlled study	Deep learning
Classification accuracy	Clustering algorithms	Data analysis	Early diagnosis
Cognitive systems	Comparative study	Data base	Extraction
Computation theory	Computational biology	Decision support system	Feature extraction
Computer aided diagnosis	Computer assisted diagnosis	Decision su	Feature selection
Computer Circuits	Computer assisted tomography	systems, clinical	Fuzzy
Data mining	Computer simulation	Decision su	Genetic algorithm
Data sets	Computerized tomography	techniques	Image classification
Decision making	Diagnostic imaging	Diagnosis, computer assisted	Lung cancer
Decision support system	Echography	Diagnosis, differential	Machin learning
Decision theory	Entropy	Diagnostic accuracy	Neural networks, computer
Decision tree	Fuzzy c means clustering	Diagnostic test accuracy	Support vector machine
Decision trees	Fuzzy clustering	Diagnostic value	
Diabetes mellitus			
Diagnosis			

Disease diagnosis Disease Electrocardiography Expert systems Forecasting Formal logic Fuzzy expert systems Fuzzy inference Fuzzy inference system Fuzzy inference systems Fuzzy expert systems FL Fuzzy neural networks Fuzzy rules Fuzzy set theory Fuzzy sets Fuzzy systems Fuzzy-logic Genetics algorithms Healthcare Heart Hearth disease Hospitals Information science Intelligent systems Internet of things Intuitionistic fuzzy sets Knowledge base Knowledge base systems Knowledge representation Learning algorithms Learning systems Linguistics Machine-learning Mathematical models MATLAB Medical applications Medical computing Medical data Medical Decision making Medical Decision support system Medical diagnosis Medical diagnostics Medical expert system Medical information systems Medical knowledge Medical problems Medicine Membership functions Neural network Neural networks	Fuzzy filters Image analysis Image enhancement Image fusion Image interpretation, computer assisted Image processing Image processing, computer assisted Image quality Image segmentation Magnetic resonance imaging Mammography Medical image processing Medical imaging Methodology Models, statistical Nuclear magnetic resonance imaging Pathology Pattern recognition, automated Positron emission tomography Procedures Reproducibility of results Segmentation Sensitivity and specificity Signal processing Statistical model Textures Three dimensional image Tissue Tomography, X-ray computed Tumors	Differential diagnosis Disease classification Evaluation Expert system Female Fuzzy control Fuzzy system Human Humans Information processing Information retrieval Logic Major clinical study Male Mathematical analysis Mathematical model Medical informatics Medical record Middle aged Model Neural networks (computer) Patient monitoring Prediction Predictive value Priority journal Probability Prognosis Review Risk factor Software	

Optimization Patient treatment Pattern recognition Principal component analysis Problem solving Risk assessment Semantics Soft computing Support vector machine Uncertainty Uncertainty analysis			

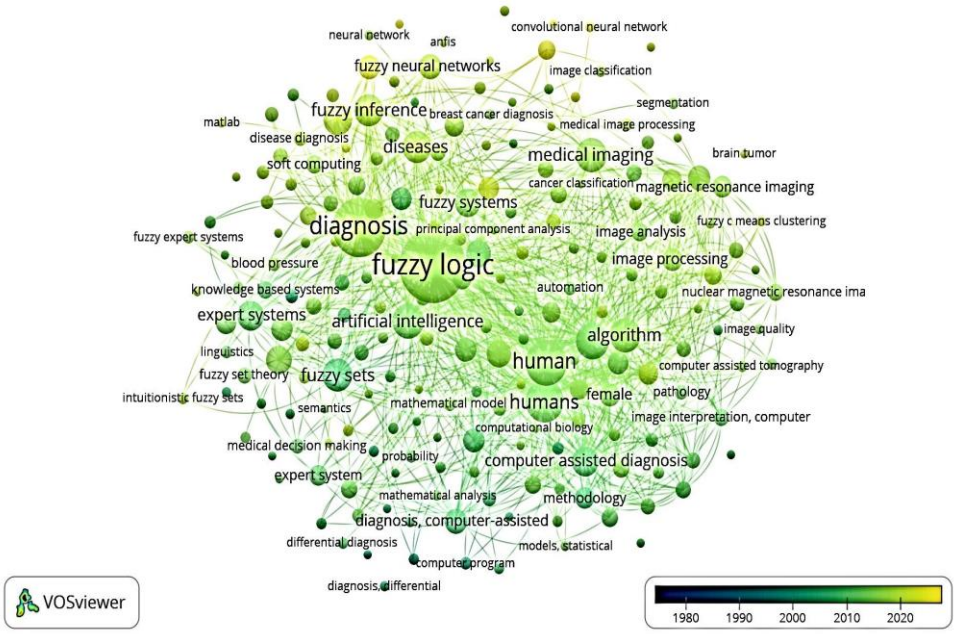


Figure 8. Co-occurrence based on keywords in the overlay visualization in the field of FL and medical diagnosis

Figure 8 shows co-occurrence based on keywords in the Overlay visualization. The color of each item is determined by the average score of the year of publication. The yellow color indicates that these keywords are associated with the post-2020 period. In other words, they have attracted the attention of researchers in recent years, and they can be used in further research. The items related to the year 2000 and earlier (blue color) mean that the keywords associated with these years have been the focus of articles and research in the field of FL in medical diagnosis in the early years. Nowadays, they are less addressed. The fundamental research domains, depicted in darker green and blue

shades, encompass "FL," "diagnosis," "fuzzy inference," "expert systems," and "artificial intelligence." This observation indicates that the initial research endeavors are predominantly centered on the development of FL-based decision-making systems for medical diagnosis. The presence of bright yellow nodes, such as "convolutional neural networks (CNNs)," "image classification," "segmentation," and "deep learning," signifies the latest advancements in the field. This suggests a shift toward AI-driven medical imaging techniques, where neural networks and deep learning are now integrated with FL for more precise medical diagnosis. The presence of "automation," "brain tumor," and "medical image processing" indicates that modern AI techniques are being applied to automate medical image analysis.

Based on this overlay visualization, the most important recent keywords in the field of FL in medical diagnosis are shown in Table 6.

Table 6. The most important recent keywords in the field of FL in medical diagnosis

Keywords	**Avg.pub.year**
Machin-learning	2023.05
Convolutional neural network	2022.61
Deep learning	2022.58

In the Density Visualization map, the concentration of research emphasis and the prevalence of cooccurring keywords within the domain of FL in medical diagnosis are illustrated through a chromatic gradient. Regions characterized by vibrant yellow tones signify the most recurrent and conceptually rich clusters, whereas green and blue zones pertain to terms that are less frequent or in the process of emerging. Fundamental terms such as "FL," "diagnosis," "medical imaging," "human," and "algorithm" are positioned at the nucleus of the research landscape, thereby indicating their pivotal role and considerable importance within the discipline. The elevated density in these sectors signifies the advancement and aggregation of academic research, especially in image-based diagnosis, decision-support algorithms, and the incorporation of AI methodologies. Conversely, less densely populated areas may denote underexamined research prospects or novel interdisciplinary intersections that warrant additional scrutiny. Figure 9 illustrates this concept. Density visualization is a data representation technique that employs color coding to depict the frequency and intensity of keyword occurrences within a given dataset. Areas imbued with bright yellow in the visual representation denote keywords that exhibit high frequency and robust interconnections, while green areas signify keywords of moderate frequency. In contrast, blue/dark regions indicate keywords that manifest with less frequency within the dataset. As depicted in Figure 9, the most luminous regions, centered on "FL," "diagnosis," "human," and "artificial intelligence," signify these as the most impactful and frequently co-occurring keywords. The co-occurrence of "diagnosis" and "FL" at the epicenter of Figure 9 accentuates the importance of these constructs. The peripheral blue regions encompass keywords such as "pathology," "brain tumor," "segmentation," and "image classification," suggesting that while these domains are integral to the research landscape, they are explored with less frequency in comparison to the core subjects.

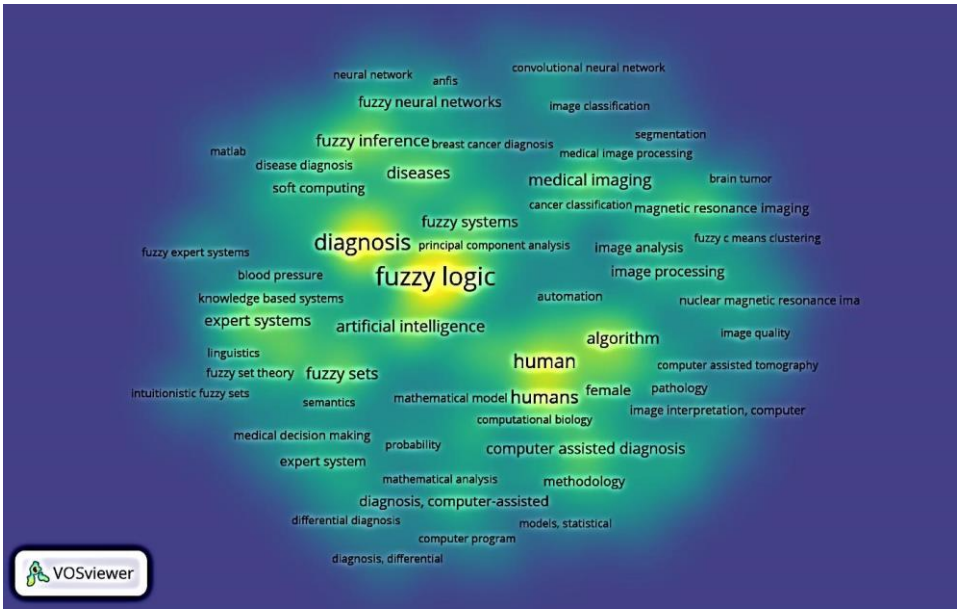


Figure 9. Co-occurrence based on keywords in the density visualization in the field of FL in medical diagnosis

4. Discussions

To provide a deeper understanding of how FL contributes to medical diagnosis, the discussion is structured into two complementary parts. First part explores its clinical relevance and real-world applications, while second one focuses on the computational mechanisms, advantages, and prospects of FL–based diagnostic systems.

4.1. Applications and clinical relevance of FL in medical diagnosis

In fields such as cardiovascular and infectious diseases, there is a growing need for decision-making tools that can manage diagnostic uncertainty. FL-based systems have proven effective in these areas by reducing diagnostic ambiguity and improving clinical decision support, particularly when patient information is incomplete or imprecise. FL's ability to handle uncertainty and vagueness improves decision-making, medical diagnosis, and treatment (Abdalla et al., 2024).

The increasing prevalence of diseases like cardiovascular conditions, cancer, and COVID-19 underscores the growing complexity of medical data and the critical need for personalized treatment approaches. This complexity often leads to significant decision uncertainty in clinical settings. Cardiovascular diseases, for instance, are influenced by a myriad of interconnected genetic, environmental, and lifestyle factors, generating vast amounts of heterogeneous data. Similarly, cancer diagnosis and treatment involve intricate pathways, with patient responses varying widely based on

tumor characteristics and individual physiology. The recent COVID-19 pandemic further highlighted the challenge of managing rapidly evolving, often incomplete, and uncertain clinical information. These diseases are characterized by a confluence of factors: the sheer volume and complexity of patient data, the necessity for highly individualized treatment plans to maximize efficacy and minimize side effects, and the inherent uncertainty in diagnosis and prognosis due to overlapping symptoms or varying disease manifestations.

A key characteristic of FL is its capacity to address the inherent uncertainty and imprecision of medical diagnosis. As a branch of mathematical logic, it is designed to resolve problems involving explicit, imprecise, or approximate reasoning, providing a framework to derive definitive outcomes from uncertain, ambiguous, imprecise, noisy, or incomplete input data (Aslan and Hızıroğlu, 2024). FL is particularly adept at managing complex and uncertain clinical information, often encountered in conditions like cardiovascular diseases and cancer. It allows examining data that cannot be easily categorized, offering a more nuanced understanding of patient health. FL methods have specifically improved clinical decision-making and diagnostic accuracy by reducing misdiagnosis, enabling earlier detection, and supporting more effective treatment planning. For example, in cardiovascular diseases, fuzzy systems can integrate diverse risk factors (e.g., blood pressure, cholesterol levels, lifestyle habits) to provide a nuanced risk assessment, going beyond traditional binary classifications and enabling personalized preventative strategies. In cancer diagnosis, FL can process ambiguous imaging data or biomarker levels to assist clinicians in identifying subtle indicators of malignancy earlier. During the COVID-19 pandemic, fuzzy inference systems were developed to assess disease severity and predict patient outcomes based on a range of symptoms and laboratory findings, providing critical support for resource allocation and treatment prioritization.

The integration of fuzzy inference systems with technologies such as the Internet of Things (IoT) has further enhanced the accuracy and efficiency of disease diagnosis, as demonstrated in applications for COVID-19 and malaria (Ferreira, 2023). These systems can simulate expert decision-making, offering personalized risk assessments and treatment recommendations, which is especially valuable in cardiovascular diseases where individual risk factors vary significantly among patients (Casalino et al, 2018). While precise real-world implementations can be challenging to track due to proprietary development and varied clinical settings, the literature indicates growing use. For instance, FLbased DSSs have been explored for integration into hospital information systems to aid clinicians in realtime diagnosis. During the COVID-19 pandemic, FL models were utilized in telemedicine platforms to help healthcare providers remotely assess patient conditions and provide timely advice, thus contributing to better patient management and reducing the burden on physical healthcare infrastructure. Measurable impacts reported in various studies include improved diagnostic accuracy rates, reduced rates of misdiagnosis, and more timely initiation of appropriate treatments, all contributing to better patient outcomes.

Based on the bibliometric mapping of FL applications in medical diagnosis, the synthesis of thematic clusters and citation trends shows key development directions with significant potential for practical impact. The leading research areas include combining fuzzy reasoning with large-scale, real-world clinical datasets and electronic health records (EHRs). This allows for adaptive and personalized decision support across diverse patient populations (Rachel et al., 2025). Related high-impact domains involve hybrid architectures that mix FL with machine learning and deep learning techniques.

This combination offers advantages in performance and interpretability (Dalkılıç et al., 2025). From a practical standpoint, studies focusing on lightweight fuzzy-based modules for use in resource-limited clinical settings (Menacuer et al., 2025) align well with global healthcare priorities. Addressing the gaps found in the bibliometric analysis, such as dependence on benchmark datasets and lack of explainability in system design, will be crucial. Concentrating on these emerging areas while ensuring regulatory compliance, integrating into clinical workflows, and maintaining transparency can help close the gap between research and practice and enhance the real benefits of FL in medical diagnosis.

FL's increasing application in clinical decision-making stems from its ability to handle the imprecision and uncertainty inherent in medical data. This methodology emulates human cognitive processes, enabling more nuanced and adaptable decision-making frameworks. FL systems are particularly advantageous in areas where establishing precise mathematical models is challenging, such as medical diagnostics and prognosis.

By integrating linguistic variables and expert insights, which are often characterized by imprecision and subjectivity, FL significantly augments diagnostic precision. This approach is highly beneficial for analyzing complex medical data and symptoms that do not fit into binary classifications (Gürsel, 2015). Fuzzy inference systems are used to address the uncertainties in diagnostic data, thereby improving the reliability of medical diagnostics (Tency and Harish, 2024). The FL paradigm has been applied in various contexts, including the diagnosis of infectious diseases (Arji et al, 2019), prediction of allergens (Saravanan and Lakashmi, 2014), diagnosis of Covid-19 (Jayalakshmi et al, 2021), management of cardiac patients (Hussain et al, 2016), and prediction of lung cancer (Aslan and Hızıroğlu, 2024), among numerous other applications.

The purpose of this research is to offer an extensive viewpoint for academics interested in FL for medical diagnosis, which benefits researchers by keeping them informed about key trends in this area to stay current. Since this article used only the Scopus database to review research literature regarding the application of FL in medicine, future research can use other databases such as Web of Science to perform bibliometric analysis. The future is focused on improving interpretability, of these systems, rigorously validating them across various real-world environments, seamlessly incorporating them into clinical workflows, and persistently tackling the ethical aspect of AI in healthcare, while also investigating new data sources and applications.

Although the implementation of FL in the realm of medical diagnosis presents numerous advantages, it is not without its inherent challenges. The interoperability of medical terminologies is imperative for the efficient exchange and analysis of data. FL systems frequently encounter difficulties in assimilating varied medical terminologies, which may result in the emergence of isolated systems and impede the fluid interpretation of medical data. The heterogeneity and variability inherent in medical data, encompassing symptoms, patient history, and demographic factors, present considerable obstacles. Consequently, FL systems must adeptly navigate this complexity to yield precise and timely diagnoses. (Shoaip et al., 2024). Also FL systems can experience a "rules explosion" challenge, resulting in an overwhelming number of rules that hinder the scalability and efficiency of the system. This problem calls for creative solutions, like the single-input rule module, to enhance system parameters and boost diagnostic precision (Zhang and Wen, 2019).

While the quantitative results presented in this study may appear largely descriptive, bibliometric analysis provides a methodological framework that goes well beyond counting publications, authors, or countries. It offers systematic and reproducible metrics to chart the intellectual structure and evolution of FL applications in medical diagnostics, revealing not only the scale of research activity but also knowledge gaps, emerging trends, and strategic research priorities (de Oliveira et al., 2019). Through network-based analyses such as co-authorship mapping, co-citation analysis, and thematic clustering, this study identifies key foundational works that serve as conceptual cornerstones, highlights emerging research frontiers, and uncovers interdisciplinary links that promote the dissemination of knowledge across medical subfields (Dai et al., 2022). Bibliometric analysis thus helps align research efforts with clinical needs, guides effective collaboration strategies, and provides evidence-based insights for policy and funding decisions (McQuire et al., 2024). In doing so, it offers a detailed knowledge map for the field, supports hypothesis generation by identifying research gaps, and enables researchers and institutions to position their work within the broader context of FL's role in advancing medical diagnostics (Akhtar et al., 2024).

4.2. Advances and challenges in FL–based medical diagnostic systems

FL (FL) is an effective method for modeling unclear and ambiguous information. It connects human language and computer decision-making models. Human communication has built-in uncertainty, and FL uses fuzzy set theory to assign meanings to vague traits. In this setup, each language term links to a fuzzy subset. This allows for approximate reasoning in uncertain situations (Shukla et al., 2025). Membership functions, which are often triangular, trapezoidal, or sigmoidal, show degrees of membership on a scale from 0 to 1. Here, 0 indicates no membership and 1 indicates full membership, which helps in making nuanced interpretations (Vyas et al., 2022).

Fuzzy models generally use the Mamdani rule-based inference system, structured in these layers:

1. **Fuzzification:** Numeric input values are changed into fuzzy sets represented by membership functions. This allows inputs to be described in everyday terms with degrees of belonging (Vyas et al., 2022).

2. **Rule Evaluation**: The fuzzy inference engine applies relevant if–then rules to the fuzzified inputs to create fuzzy outputs.

3. **Defuzzification**: These fuzzy outputs are then converted back into clear, actionable results for decision-making.

Medical diagnosis systems often use specialized software called expert system shells. These shells include inference methods like forward chaining, backward chaining, or hybrid approaches to process knowledge encoded as production rules. Object-oriented expert system shells make it easier to integrate with outside clinical databases, allowing real-time data use (Sikchi and Sikchi, 2013). The knowledge base contains definitions of diseases encoded as collections of if-then rules that reflect clinical understanding, such as tuberculosis symptoms structured as discrete production rules. The fact base holds

observed patient data, which the inference engine uses to reason and produce new conclusions.

Medical decision support (MDS) systems have developed over the last fifty years to handle various clinical decisions, using fuzzy methods for their ability to manage uncertainty and ambiguity. Research shows that MDS accuracy increases with the modeling of unclear and changing data, integration of information from multiple patient sources, and improvements in algorithmic reasoning (Wagholikar et al., 2012). Soft computing techniques like FL combined with data mining can accurately predict diseases such as heart attacks by analyzing patient data and identifying the best diagnostic models (Dianirani and Claudia, 2021).

The field benefits from hybrid models that mix FL with other computational intelligence techniques such as Swarm Intelligence, Evolutionary Computing, Neural Networks, and deep learning. These combined frameworks offer scalable, context-sensitive, and personalized healthcare solutions by mimicking complex clinical reasoning processes (Dianirani and Claudia, 2021; Tariq et al., 2024). Recent studies highlight pairing fuzzy systems with IoT-enabled real-time data, which improves system flexibility and resilience across diseases like hypertension, diabetes, and multimorbidity (Mohod et al., 2025).

FL systems make use of membership function design, compositional inference, and new techniques like type-2 fuzzy sets to better represent uncertainty. Various system architectures have been examined through case studies, showing strong agreement with expert clinician decisions, especially in borderline cases where deterministic models do not perform well (Mohod et al., 2025). Improved architectures confirm model accuracy, strength, and adaptability in different clinical situations.

Fuzzy expert systems offer clear and understandable decision-making processes, durability against incomplete or noisy data, and flexibility in diverse healthcare settings. However, they face challenges including scaling rule bases, difficulties with system integration, regulatory issues, and the computational demands of hybrid approaches (Vyas et al., 2022; Wagholikar et al., 2012). Understanding deep hybrid models and ensuring data quality are still crucial concerns.

The growth of hybrid fuzzy-AI techniques, fueled by data-driven optimization and IoT-enabled health monitoring, aims to advance precision medicine. There is a rising need for explainable, clinically appropriate AI tools, and FL—known for its transparency and fit with medical reasoning—serves as an important enabler (Mohod et al., 2025). Future research should focus on increasing dataset diversity, automating rule creation, improving integration with clinical workflows, and navigating regulatory challenges to support real-world use. In the end, fuzzy expert systems are likely to transform into collaborative tools that support patient-centered care alongside clinicians.The employment of fuzzy sets and rules facilitates the consideration of uncertainty and membership degrees, thereby enabling healthcare providers to efficiently interpret and navigate the intricate information present within medical databases (Castaneda et al., 2015). The utilization of FL not only improves the precision of diagnostic findings but also improves their comprehensibility, a development that is vital in the creation of sophisticated decision-making tools that can assist healthcare professionals in the identification and management of various illnesses (Edison, 2023).

5. Limitations

This investigation utilized the Scopus database to procure of bibliometric data, which may have resulted in the omission of pertinent studies that are indexed within alternative databases, such as Web of Science or PubMed. While Scopus is recognized for its extensive coverage, this limitation has the potential to influence the comprehensiveness and representativeness of the dataset. Furthermore, the emphasis on publications that feature English titles, abstracts, or keywords introduces a risk of linguistic bias, which could lead to an underrepresentation of research disseminated in other languages, especially from regions where English is not the primary language. These considerations must be taken into account when analyzing the results, particularly within global contexts. Additionally, the potential biases inherent in co-authorship and keyword clustering can significantly impact the interpretation of research findings, thereby underscoring the necessity for transparency and meticulous analysis in bibliometric investigations. A notable limitation of this bibliometric analysis is the possibility of bias arising from both co-authorship and keyword clustering methodologies. Within co-authorship networks, closely collaborating research entities may induce interdependence among outcomes, possibly leading to an inflated visibility of specific authors or clusters if such variables are not properly addressed. Similarly, keyword clustering may inadequately capture the full spectrum of research themes, as algorithmic categorization may merge articles that share keywords yet exhibit divergent content or perspectives. These methodological constraints highlight the imperative for careful interpretation of co-authorship and keyword clustering results in bibliometric studies (Glänzel and Schubert, 2005; Zupic and Čater, 2015). The search is executed within the Scopus database employing the query TITLE-ABS-KEY (fuzzy AND logic AND medical AND diagnosis). Only documents with titles or abstracts in English are considered for inclusion. No limitations are imposed regarding subject categories. The analysis encompasses peer reviewed journal articles and conference proceedings, while editorials, letters, and book chapters are excluded.

Furthermore, to overcome the aforementioned limitations, future bibliometric investigations may consider integrating data from multiple databases (e.g., Web of Science, PubMed, and IEEE Xplore) to ensure broader coverage and to mitigate potential database-specific biases. Expanding the scope to include non-English publications could also alleviate linguistic bias and improve the representativeness of global research contributions. Beyond bibliometric mapping, combining these analytical findings with real-world clinical datasets would strengthen the practical relevance of FL–based diagnostic models. Future efforts should also emphasize explainability and interpretability in such systems to foster transparency and trust in clinical decision support applications.

6. Conclusions

A review of the Scopus database reveals that research on the application of FL to medical diagnosis has garnered increasing attention over time. Machine learning has emerged as the predominant theme in this domain. China, India, and the United States have emerged as the most active countries in this field, with the vast majority of studies being published in English. The Journal of Artificial Intelligence in Medicine has been

identified as the most prominent journal in this field, and Lovely Professional University has been recognized as the most active university. The most prolific researchers in this field are Adlassnig, K.P. and Al-Kasasbeh, R.T., with computer science demonstrating the most interest in this concept, accounting for 31% of documents. Future research could build upon these trends by exploring hybrid approaches that combine FL with advanced deep learning architectures, thereby improving diagnostic accuracy in complex clinical scenarios. Moreover, expanding applications beyond computer science into biomedical engineering, public health, and personalized medicine could open new interdisciplinary opportunities. The focused research areas, such as hybrid fuzzy and machine learning frameworks, uncertainty management systems, and clinical decision support integration, point toward a trend that seeks to balance diagnostic accuracy with interpretability in various healthcare settings. From these trends, clear directions for future research arise. Future studies should focus on: (i) validating findings in large, real-world settings across various clinical environments and patient groups to improve generalization; (ii) creating lightweight, efficient fuzzy systems that can work in resource-limited settings; (iii) integrating FL models with electronic health records and real-time patient monitoring for better decision-making; and (iv) exploring user interface designs and transparency methods to build clinician trust and encourage use. Following these pathways will connect bibliometric insights with real-world clinical practice, helping advance interpretable AI in medical diagnostics.

Acknowledgments

We sincerely appreciate the reviewers' valuable suggestions and recommendations, which have significantly improved both the scientific quality and the presentation of the paper.

References

Abdalla, A. Y., Abdalla, T. Y., Chyaid, A. M. (2024). Internet of things based fuzzy systems for medical applications: A review. *IEEE Access*, 163883- 163902.

Ahmadi, H., Gholamzadeh, M., Shahmoradi, L., Nilashi, M., Rashvand, P. (2018). Diseases diagnosis using fuzzy logic methods: A systematic and meta-analysis review. *Computer Methods and Programs in Biomedicine*, 161, 145-172.

Akhtar, M. N., Haleem, A., Javaid, M. (2024). Exploring the advent of Medical 4.0: a bibliometric analysis systematic review and technology adoption insights. *Informatics and Health*, **1**(1), 16-28.

Arji, G., Ahmadi, H., Nilashi, M., Rashid, T. A., Ahmed, O. H., Aljojo, N., Zainol, A. (2019). Fuzzy logic approach for infectious disease diagnosis: A methodical evaluation, literature and classification. *Biocybernetics and biomedical engineering*, **39**(4), 937-955.

Aslan, B., Hızıroğlu, O. A. (2024). Prediction of Lung Cancer with Fuzzy Logic Methods: A Systematic Review. *Artificial Intelligence Theory and Applications*, **4**(2), 155-192.

Awotunde, J. B., Matiluko, O. E., Fatai, O. W. (2014). Medical diagnosis system using fuzzy logic. *African Journal of Computing & ICT*, **7**(2), 99-106

Casalino, G., Castellano, G., Castiello, C., Pasquadibisceglie, V., Zaza, G. (2018). A fuzzy rule-based decision support system for cardiovascular risk assessment. In *International Workshop on Fuzzy Logic and Applications*, 97-108, Cham: Springer International Publishing.

Castaneda, C., Nalley, K., Mannion, C., Bhattacharyya, P., Blake, P., Pecora, A., ... Suh, K. S. (2015). Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. *Journal of clinical bioinformatics*, **5**, 1-16.

Charanbir, C. S., Kumar, R., Kumar, S. (2023). Leveraging computer-aided education for enhanced learning: Innovations, benefits, and challenges in the medical sector. *Asian Journal of Engineering and Applied Technology*, **12**(2), 40-49.

Dai, Z., Xu, S., Wu, X., Hu, R., Li, H., He, H., ... Liao, X. (2022). Knowledge mapping of multicriteria decision analysis in healthcare: a bibliometric analysis. *Frontiers in public health*, **10**, 895552.

Dalkılıç, O., Demirtaş, N., Demirtaş, A. (2025). Evaluating Prostate Cancer Diagnosis Using the Adaptive Neural Fuzzy Inference System (ANFIS): A Comparative Analysis of Diagnostic Accuracy. *Turkish Journal of Science and Technology*, **20**(2), 583-593.

da Silva, R. F., de Souza, G. F. M. (2021). Mapping the literature on asset management: A bibliometric analysis. *Journal of Scientometric Research*, **10**(1), 27-36.

de Oliveira, O. J., da Silva, F. F., Juliani, F., Barbosa, L. C. F. M., Nunhes, T. V. (2019). Bibliometric method for mapping the state-of-the-art and identifying research gaps and trends in literature: An essential instrument to support the development of scientific projects. In *Scientometrics recent advances.* IntechOpen.

Dianirani, A. S., Claudia, Z. D. (2021). Fuzzy-based decision for coronary heart disease diagnosis: systematic literature review. *Engineering, Mathematics and Computer Science Journal (EMACS)*, **3**(2), 73-78.

Fernandes, M., Vieira, S. M., Leite, F., Palos, C., Finkelstein, S., Sousa, J. M. (2020). Clinical decision support systems for triage in the emergency department using intelligent systems: a review. *Artificial intelligence in medicine*, **102**, 101762.

Ferreira, F. de Sá (2023). Advancements in medical diagnostics through fuzzy logic: a comprehensive review. *Revista ft*, 40–41.

Glänzel, W., Schubert, A. (2005). Analysing scientific networks through co-authorship, *Handbook of quantitative science and technology research* (pp. 257–276). Springer. https://doi.org/10.1007/1-4020-2755-9_12

Gürsel, G. (2015). Fuzzy logic in healthcare. In *Handbook of Research on Artificial Intelligence Techniques and Algorithms, Advances in Computational Intelligence and Robotics*, 679-707. IGI Global.

Hayward, G., Davidson, V. (2003). Fuzzy logic applications. *Analyst*, **128**(11), 1304-1306.

Jayalakshmi, M., Garg, L., Maharajan, K., Jayakumar, K., Srinivasan, K., Bashir, A. K., Ramesh, K. (2021). Fuzzy logic-based health monitoring system for COVID'19 patients. *Computers, Materials and Continua*, **67**(2), 2431-2447.

Kanchanachitra, C., Lindelow, M., Johnston, T., Hanvoravongchai, P., Lorenzo, F. M., Huong, N. L., ... Dela Rosa, J. F. (2011). Human resources for health in southeast Asia: shortages, distributional challenges, and international trade in health services. *The Lancet*, **377**(9767), 769-781.

Kaur, S., Singla, J., Nkenyereye, L., Jha, S., Prashar, D., Joshi, G. P., ... Islam, S. R. (2020). Medical diagnostic systems using artificial intelligence (ai) algorithms: Principles and perspectives. *IEEE Access*, **8**, 228049-228069.

Kitchenham, B., Charters, S. M. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report.

Kumar, R. (2025). Bibliometric Analysis: Comprehensive Insights into Tools, Techniques, Applications, and Solutions for Research Excellence. *Spectrum of Engineering and Management Sciences*, **3**(1), 45-62.

Matinfar, F., Golpaygani, A. T. (2022). A fuzzy expert system for early diagnosis of multiple sclerosis. *Journal of Biomedical Physics & Engineering*, **12**(2), 181.

McQuire, C., Frennesson, N. F., Parsonage, J., Van der Heiden, M., Troy, D., Zuccolo, L. (2024). Trends in fetal alcohol spectrum disorder research: A bibliometric review of original articles published between 2000 and 2023. Alcohol: *Clinical and Experimental Research*, **48**(10), 1819-1833.

Meyer, A. N., Giardina, T. D., Khawaja, L., Singh, H. (2021). Patient and clinician experiences of uncertainty in the diagnostic process: current understanding and future directions. *Patient Education and Counseling*, **104**(11), 2606-2615.

Mohod, S. B., Ingole, K. R., Raghuwanshi, P., Jain, S. K., Shrivastava, A., Rachel, B. (2025). Next-Gen Medical Intelligence: Fuzzy Logic-Driven Expert Systems For Clinical Decision-Making. *International Journal of Environmental Sciences*, **11**(21s).

Myrzakerimova. A., Kolesnikova. K. (2024). Development of mathematical methods for diagnosing kidney diseases using fuzzy set tools. *Indonesian Journal of Electrical Engineering and Computer Science*. **35**(1), July 2024, pp. 405~417.

Myrzakerimova, A., Kolesnikova, K., Nurmaganbetova, M. (2024). Use of Mathematical Modeling Tools to Support Decision-Making in Medicine. *Procedia Computer Science*, 231, 335-340.

Ohayon, M. M. (1999). Improving decision making processes with the fuzzy logic approach in the epidemiology of sleep disorders. *Journal of Psychosomatic Research*, **47**(4), 297-311.

Pandey, S. R. (2016). Temporal logic-based fuzzy decision support system for diagnosis of rheumatic fever and rheumatic heart disease (Doctoral dissertation, University of Greenwich).

Phuong, N. H., Kreinovich, V. (2001). Fuzzy logic and its applications in medicine. *International Journal of Medical Informatics*, **62**(2-3), 165-173.

Rahman, M. Z., Akbar, M. A., Leiva, V., Tahir, A., Riaz, M. T., Martin-Barreiro, C. (2023). An intelligent health monitoring and diagnosis system based on the internet of things and fuzzy logic for cardiac arrhythmia COVID-19 patients. *Computers in Biology and Medicine*, **154**, 106583.

Rachel, B., Kavali, S., Raghuwanshi, P., Jain, S. K., Choudhary, S., Shrivastava, A.(2025). Revolutionizing Healthcare Informatics With Fuzzy Logic: Smarter Data, Smarter Decisions. *International Journal of Environmental Sciences*, **11**(21s).

Sahoo, S. K., Choudhury, B. B., Dhal, P. R. (2024). A bibliometric analysis of material selection using MCDM methods: trends and insights. *Spectrum of Mechanical Engineering and Operational Research*, **1**(1), 189-205.

Salleh, N. Z. M., Abdullah, M., Ali, A., Faisal, F., Nor, R. M. (2023). Research trends, developments, and future perspectives in brand attitude: A bibliometric analysis utilizing the Scopus database (1944–2021). *Heliyon*, **9**(1).

Saravanan, V., Lakshmi, P. T. V. (2014). Fuzzy logic for personalized healthcare and diagnostics: FuzzyApp—A fuzzy logic based allergen-protein predictor. *OMICS: A Journal of Integrative Biology*, **18**(9), 570-581.

Sharma, R. S., Rohatgi, A., Jain, S., Singh, D. (2023). The Ayushman Bharat Digital Mission (ABDM): making of India's digital health story. *CSI transactions on ICT*, **11**(1), 3-9.

Shoaip, N., El-Sappagh, S., Abuhmed, T., Elmogy, M. (2024). A dynamic fuzzy rule-based inference system using fuzzy inference with semantic reasoning. *Scientific reports*, **14**(1), 4275.

Shukla, A. K., Mehra, P., Muhuri, P. K. (2025). Fuzzy Sets-Based Approaches for Improved Medical Diagnosis: An Analysis and Overview of Major Research Directions. *ACM Computing Surveys.*

Sikchi, S. S., Sikchi, S. (2013). Fuzzy expert systems (FES) for medical diagnosis. *International Journal of Computer Applications*, **63**(11).

Skačkauskienė, I. (2022). Research on management theory: A development review and bibliometric analysis. *Problems and Perspectives in Management*, 335-347

Smets, P. (1983). Fuzzy sets theory for medical decision making. In *Objective Medical Decision making; Systems Approach in Acute Disease*: Eindhoven, The Netherlands, 19–22 April 1983 Proceedings (pp. 7-19). Berlin, Heidelberg: Springer Berlin Heidelberg.

Stoean, R., Stoean, C. (2013). Modeling medical decision making by support vector machines, explaining by rules of evolutionary algorithms with feature selection. *Expert Systems with Applications*, **40**(7), 2677-2686.

Tan, X., Liu, X., Shao, H. (2017). Healthy China 2030: a vision for health care. *Value in health regional issues*, **12**, 112-114.

Tang, H. H., Ahmad, N. S. (2024). Fuzzy logic approach for controlling uncertain and nonlinear systems: a comprehensive review of applications and advances. *Systems Science & Control Engineering*, **12**(1), 2394429.

Tawfik, B., Ouda, B. K., Abd El Samad, Y. M. (2013) A fuzzy logic model for medical equipment risk classification. *Journal of Clinical Engineering*, **38(**4), 185-190.

Tariq, M., Hayat, Y., Hussain, A., Tariq, A., Rasool, S. (2024). Principles and perspectives in medical diagnostic systems employing artificial intelligence (AI) algorithms. *International Research Journal of Economics and Management Studies IRJEMS*, **3**(1).

Temesgen, Z. M., DeSimone, D. C., Mahmood, M., Libertin, C. R., Palraj, B. R. V., Berbari, E. F. (2020). Health care after the COVID-19 pandemic and the influence of telemedicine. In *Mayo Clinic Proceedings*, **95(**9), pp.S66-S68).

Tency, E. L. M., Harish, M. (2024). Applications of Fuzzy Logics in Modern Systems: A Simple Survey. *International Journal of Research Publication and Reviews*, **5**(5), 7598–7600. https://doi.org/10.55248/gengpi.5.0524.1316.

Thukral, S., Bal, J. S. (2019). Medical applications on fuzzy logic inference system: a review. *International Journal of Advanced Networking and Applications*, **10**(4), 3944-3950.

Velan, D., Mohandoss, H., Valarmathi, S., Sundar, J. S., Kalpana, S., Srinivas, G. (2024). Digital health in your hands: A narrative review of exploring Ayushman Bharat's digital revolution. *World Journal of Advanced Research and Reviews*, **23**(3), 1630-1641.

Vyas, S., Gupta, S., Bhargava, D., Boddu, R. (2022). [Retracted] Fuzzy Logic System Implementation on the Performance Parameters of Health Data Management Frameworks. *Journal of Healthcare Engineerin*g, 2022(1), 9382322.

Wagholikar, K. B., Sundararajan, V., Deshpande, A. W. (2012). Modeling paradigms for medical diagnostic decision support: a survey and future directions. *Journal of Medical Systems*, **36**(5), 3029-3049.

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, **8(**3), 338-353.

Zadeh, L. A. (1983). The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy sets and systems*, **11**(1-3), 199-227.

Zainuldin, M. H., Lui, T. K. (2022). A bibliometric analysis of CSR in the banking industry: a decade study based on Scopus scientific mapping. *International Journal of Bank Marketing*, **40**(1), 126.

Zhang, Q., Wen, C. (2019). A Novel Single-Input Rule Module Connected Fuzzy Logic System and Its Applications to Medical Diagnosis. In *Chinese Intelligent Systems Conference* (pp. 357366). Singapore: Springer Singapore.

Zupic, I., Čater, T. (2015). Bibliometric methods in management and organization. *Organizational Research Methods*, 18(3), 429–472.

Received August 14, 2025, revised October 24, 2025, accepted October 27, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, pp. 834–861
http://doi.org/10.22364/bjmc.2025.13.4.05

ReInTa: A Novel Requirements Interdependencies Taxonomy

Mohamad GHARIB[1], Elvin MIRZAZADA[1]

University of Tartu, Estonia

{mohamad.gharib,elvin.mirzazada}@ut.ee

ORCID 0000-0003-2286-2819, ORCID 0009-0003-0195-579X

Abstract. Requirements interdependencies define how requirements relate to and influence one another, playing a critical role in the delivery of robust software systems. Despite their importance, existing literature lacks a unified framework to systematically capture and classify these interdependencies. This paper bridges this gap by introducing ReInTa, a novel taxonomy of requirements interdependencies derived through a Systematic Literature Review (SLR). ReInTa organizes the identified 16 interdependencies into three categories — Intra-type, Cross-type, and Multi-type—addressing gaps in fragmentation and terminology inconsistency. The taxonomy is evaluated against established frameworks, demonstrating broader coverage, including previously under-addressed interdependencies like Continuance, Cooperation, and iCost. A demonstration using an example abstracted from the PHArA-ON project illustrates RelnTa's practical utility in enhancing the Motivational Goal Model (MGM), enabling richer modeling of interactions among functional, quality, and emotional goals. ReInTa provides researchers and practitioners with a standardized and actionable framework for managing complex requirement relationships, ultimately improving system quality and development efficiency.

Keywords: Requirements interdependencies, Requirements taxonomy, Requirements relationships, Requirements modeling, Requirements Engineering

1 Introduction

The Requirements Engineering (RE) community widely acknowledges that most requirements are interrelated Dahlstedt and Persson (2005). These interdependencies define how requirements influence and relate to one another Tabassum et al. (2014). Empirical studies, such as those by Carlshamre et al. Carlshamre et al. (2001), suggest that only a fifth of the requirements are not related to or influenced by any other requirements. Given their importance, requirements interdependencies must be carefully managed Tabassum et al. (2014), as they significantly impact software quality, development

time, and cost Robinson et al. (2003). Failure to properly address these relationships can even contribute to software project failures Noviyanto et al. (2023).

Prior research has proposed diverse approaches for capturing requirements interdependencies (e.g., Dalpiaz et al. (2016); Van Lamsweerde (2001); Miller et al. (2015)) and identified various dependency types, such as requires, refines, similar, and conflicts Gharib et al. (2016). However, these contributions remain fragmented and lack standardization, revealing a critical gap: no unified framework exists to integrate key requirements interdependencies. This fragmentation is further complicated by the absence of standardized terminology and classification criteria Robinson et al. (2003). This complicates the comparison and synthesis of interdependencies, a fundamental aspect of software development. Consequently, researchers lack a common understanding and terminology, leading to potential analytical oversights and incorrect requirements analysis Dahlstedt and Persson (2005). Without a comprehensive, structured taxonomy, the identification, understanding, and application of requirements interdependencies continues to present substantial challenges.

The main objective of this paper is to develop a novel taxonomy for requirements interdependencies that offers a deeper and more comprehensive understanding of these relationships. The taxonomy will enable improved comprehension, analysis, and management of complex interdependencies in modern software systems. To achieve this objective, we have defined the following Research Questions (RQs):

RQ1: What types of requirements interdependencies/relationships exist in the literature?

RQ2: How are interdependencies between requirements utilized across different requirement types?

RQ3: What are the key coverage gaps in existing studies on requirements interdependencies?

This study employs a Systematic Literature Review (SLR) methodology to comprehensively survey existing research and identify the most mature and relevant studies in the field. Through careful analysis of these selected works, we address our defined RQs. Building upon these findings, we develop **ReInTa** (**Re**quirements **In**terdependency **Ta**xonomy), a novel taxonomy designed to systematically identify, classify, and manage requirements interdependencies. As we demonstrate in later sections, ReInTa has significant potential to enhance both software quality and development efficiency by providing structured guidance for handling complex requirement interdependencies.

The rest of this paper is structured as follows: Section 2 establishes the research baseline, followed by the methodology in Section 3. Section 4 presents our Systematic Literature Review (SLR), with results and findings discussed in Section 5. We introduce ReInTa and detail its construction in Section 6. The taxonomy evaluation appears in Section 7, followed by its application to a Motivational Goal Model (MGM) example in Section 8. Section 9 examines threats to validity, while Section 10 concludes the paper and outlines future work.

2 Baseline

Software requirements have evolved through distinct historical stages. Initially, the focus was predominantly on defining software system requirements—the precise specifications of what the software must do Greenspan et al. (1982). By the 1980s, this perspective broadened to explicitly include the operational environment, acknowledging that a system's functionality depends upon its interactions with external entities and surrounding conditions Dubois et al. (1986). An important shift occurred in the 1990s with the incorporation of the social and organizational context, recognizing that software is embedded within complex human structures, goals, and norms Yu (1993). This modern, holistic view ensures that the developed software system is not only technically sound but also viable, valuable, and effectively integrated within its intended ecosystem of stakeholders, policies, and cultural practices.

Requirements do not exist in isolation; they often depend on and influence one another Dahlstedt and Persson (2005). These interdependencies can take several forms, including: *Functional dependencies*, where fulfilling one requirement depends on functionality provided by another Zave and Jackson (1997); Dalpiaz et al. (2016); Van Lamsweerde (2001); Miller et al. (2015); *Temporal dependencies*, where one requirement must be fulfilled before another can be addressed Van Lamsweerde et al. (1995); *Resource dependencies*, where fulfilling one requirement depends on resources provided by another Dalpiaz et al. (2016); and *Conflict dependencies*, where fulfilling one requirement may hinder another Nuseibeh et al. (2000); Dalpiaz et al. (2016).

Capturing and analyzing requirements interdependencies is critical for several reasons, most notably: *Impact analysis*, understanding dependencies helps assess how changes to one requirement may propagate through the system Arnold and Bohner (1996). *Risk mitigation*, unrecognized dependencies can lead to design flaws, integration failures, or project delays Wiegers and Beatty (2013). *Prioritization and trade-offs*, dependency mapping supports informed decision-making when balancing competing constraints Helfert and Herrmann (2002). Finally, *traceability,* clear interdependencies enhance requirements traceability and ensure alignment with system architecture and test cases Dahlstedt and Persson (2005).

Requirements interdependencies play a critical role throughout the entire requirements engineering lifecycle, influencing elicitation, classification, prioritization, and validation phases. Recognizing and managing these interdependencies is therefore essential for effective requirements engineering. In what follows, we discuss each of these phases, followed by how requirements interdependencies are relevant to them.

Requirements elicitation aims at discovering the requirements for the system-to-be through consulting relevant stakeholders, investigating system documentation, and applying domain knowledge Sommerville and Sawyer (1997); Kotonya et al. (1998); Sommerville (2007). During this process, dependencies naturally emerge as stakeholders articulate needs that may conflict or rely on one another Zave and Jackson (1997), necessitating specialized techniques like scenario analysis to uncover these implicit relationships Dahlstedt and Persson (2005). The identification of such dependencies at this early stage proves critical for subsequent development phases.

Requirements classification transforms an unstructured collection of requirements into coherent clusters Sommerville (2007). While multiple classification approaches ex-

ist Aurum and Wohlin (2005), the fundamental distinction between functional requirements (specifying what the system shall do) and non-functional requirements (defining how it should operate) remains paramount Chung et al. (1999). As classification proceeds, interdependencies play a key role in determining appropriate grouping strategies, ensuring logical structuring while preventing artificial fragmentation of related requirements Bass et al. (1997). This facilitates more effective requirements management throughout the development lifecycle.

Requirements prioritization involves ranking requirements based on their relative importance Sommerville (2007); Helfert and Herrmann (2002), guiding decisions about implementation sequencing. This process must carefully consider requirements interdependencies Karlsson (2004), as high-priority features may depend on lower-priority foundational elements Berander (2004). Methods like dependency-aware ranking have been developed to optimize resource allocation while respecting these critical relationships Leffingwell and Widrig (2000).

Finally, *requirements validation* ensures that the final requirements set accurately reflects stakeholder expectations Kotonya et al. (1998); Sommerville (2007). At this stage, interdependencies become particularly important as they determine whether testing should evaluate requirements in isolation or as integrated components. Proper attention to these relationships helps prevent system failures caused by overlooked interactions between requirements Nuseibeh et al. (2000). Please note that requirements validation ensures a requirement set is correct, complete, and consistent with stakeholder expectations, typically through five checks: validity, completeness, consistency, realism, and verifiability Sommerville (2007); Terry Bahill and Henderson (2005). Among these, interdependencies are central to consistency checks (e.g., identifying conflict or exclusion relationships) and completeness checks (e.g., revealing missing prerequisites). They also support realism assessments by exposing cost or technical constraints and aid verifiability by establishing traceable links (e.g., satisfies) for testing. Thus, analyzing interdependencies is a core mechanism, not an ancillary activity, for robust validation.

In summary, requirements interdependencies form a critical dimension throughout the entire RE lifecycle. Their proper identification and management directly contribute to more robust systems, informed decision-making, and ultimately project success. As the discussed phases demonstrate, neglecting these relationships risks compromising system integrity, while their systematic consideration enables more effective requirements engineering practices.

3 Research Methodology

This study follows a three-phase, iterative research methodology, as illustrated in Figure 1. The process begins with a Systematic Literature Review (SLR) structured around three sub-phases, each corresponding to one of our Research Questions (RQs). The second phase synthesizes these findings to develop our taxonomy. Finally, we evaluate and refine the taxonomy. In what follows, the three key phases of the methodology are briefly discussed.

1. **Identify the most relevant studies to answer the RQs via an SLR:** This phase systematically identifies the most mature and relevant studies on requirements inter-

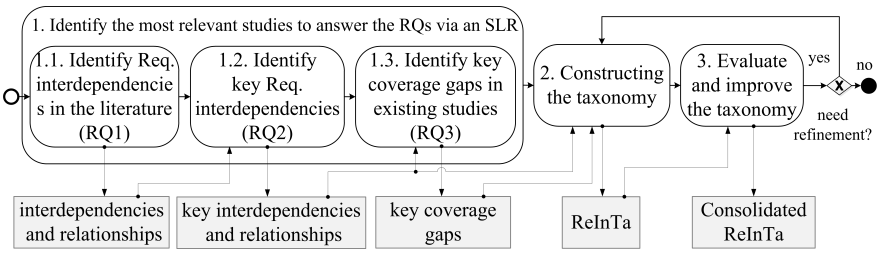


Fig. 1. The methodology process

dependencies to address the RQs. It follows established SLR protocols, including: (1) planning the review, (2) defining search strategies and study selection criteria, and (3) executing data extraction and synthesis. The SLR outcomes provide input for addressing RQ1, RQ2, and RQ3 through three sub-phases that are dedicated to them, as detailed in Sections 4 and 5.

2. **Constructing the taxonomy:** This phase systematically develops the taxonomy by synthesizing two key outputs from the SLR: (1) the identified requirements interdependencies (from RQ2) and (2) the coverage gaps in existing studies (from RQ3). The construction process involves three critical activities: first, identifying and rigorously defining core concepts for capturing requirements interdependencies; second, resolving synonymous terms to ensure conceptual clarity and distinctness; and third, categorizing these refined concepts into logical, high-level groupings that form the structural foundation of the taxonomy. The taxonomy construction is discussed in Section 6.
3. **Evaluate and improve the taxonomy:** This phase evaluates the resulting taxonomy, assessing its internal coherence and consistency, its completeness regarding existing work, and its practical utility. The evaluation involved a comparative analysis of the developed taxonomy against three established requirements engineering frameworks: iStar 2.0 Dalpiaz et al. (2016), KAOS (Knowledge Acquisition in automated specification) Van Lamsweerde (2001), and Emotion-led Modeling language (Motivational Goal Model (MGM)) Miller et al. (2015). This comparison allowed for an assessment of the coverage of the taxonomy and its alignment with existing conceptualizations, leading to minor refinements. To demonstrate the practical applicability of the taxonomy, we applied it to model relationships within a real-world example abstracted from the PHArA-ON (Pilots for Healthy and Active Ageing in Europe) H2020 European Project[1]. This proves the practical relevance of the taxonomy and its ability to represent a wider range of interdependencies compared to existing approaches. The evaluation and demonstration of the taxonomy are presented in Sections 7 and 8, respectively.

Through this systematic, iterative methodology, we ensure the resulting taxonomy achieves three critical objectives: (1) strong theoretical foundations grounded in comprehensive literature analysis, (2) direct mitigation of identified knowledge gaps in cur-

[1] `https://www.pharaon.eu/`

rent requirements engineering practice, and (3) demonstrable practical utility. This guarantees that the final taxonomy is both theoretically sound and practically relevant for improving requirements engineering activities.

4 The SLR to identify the most relevant studies to answer the RQs

To systematically identify the most relevant and mature studies on requirements interdependencies, we conducted a Systematic Literature Review (SLR) following the established guidelines by Kitchenham (2004). As illustrated in Figure 2, our SLR methodology comprises three distinct phases:

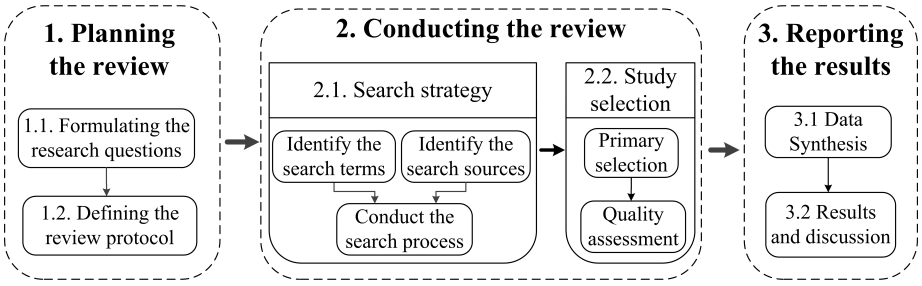


Fig. 2. The Systematic Literature Review Process.

1. **Planning the review:** We formulated the RQs and developed a comprehensive review protocol to ensure methodological rigor.
2. **Conducting the review:** This phase involved (a) identifying appropriate search terms and literature sources, (b) executing the search process, and (c) systematically selecting relevant studies through predefined criteria.
3. **Reporting the results of the review:** We extracted and analyzed detailed information from the selected studies to address each RQ.

The subsequent sections elaborate on each phase in detail.

1. Planning the review. This phase is fundamental for the success of the review, as it establishes the research objectives and the process to conduct the review. This phase encompasses three primary activities:

1.1. Formulating the RQs is a vital activity since RQs serve as the basis for developing the overall approach for a systematic review Kitchenham (2004). Therefore, we formulate three RQs of this research (presented earlier) to identify key requirements interdependencies that have been presented in the literature, as well as the gaps in existing studies on such interdependencies.

1.2. Defining the review protocol. A review protocol specifies the strategy that will be used to search for relevant studies, study inclusion and exclusion criteria, and study selection criteria. In what follows, we discuss how each of these activities was performed.

2. Conducting the review. This phase is composed of two main activities: 1 - search strategy, and 2 - study selection, where each of them is composed of several sub-activities.

2.1. Search strategy. This activity aims to find studies related to the RQs using an objective and repeatable search strategy. The search activity consists of three main sub-activities:

Identify the search terms. We derived the main search terms from the research questions. In particular, we used the Boolean AND to link the major terms, and we used the Boolean OR to incorporate alternative synonyms of such terms. The resulting search terms are:

("Requirements" OR "requirement") AND ("dependency" OR "dependencies") AND ("relation" OR "relations" OR "relationship" OR "relationships")

Identify the literature resources. We have selected several electronic database sources, namely IEEE Xplore, ACM Digital Library, Web of Science, and Scopus, as they index the main scientific publications in the fields of software and requirements engineering.

Conduct the search. We used the search terms to search the selected electronic database sources, and all returned studies were considered.

2.2. Study selection. Using the search terms to search the electronic database sources, returned 265 papers for Scopus, 411 papers for IEEE Xplore, 1052 papers for WoS, and 59 papers for ACM Digital Library, resulting in 1787 papers. After removing duplicate papers, we had 1485 papers. We read the titles and abstracts and skimmed through the rest of the papers to apply the inclusion and exclusion criteria (shown in Table 1). This comprehensive approach ensured that the final selection of papers was highly relevant and contributed to answering the RQs. This resulted in selecting 106 papers.

Table 1. Inclusion and exclusion criteria.

Exclusion criteria	Inclusion criteria
EC1: Papers that are not published in English	**IC1:** Papers related to at least one of the research questions
EC2: Papers that are not peer- reviewed. (i.e., not published in a conference or journal)	**IC2:** Only papers for which the full text is available will be included, ensuring that a thorough review can be conducted.
EC3: If a paper has several versions, only the most complete one is included.	**IC3:** Papers focusing on software/requirements engineering or systems engineering domains to maintain relevance to the field of study.

The 106 papers were fully read, and only papers that contained sufficient information to contribute to satisfying at least three of the quality assessment questions (shown in Table 2) were included. More specifically, the four quality assessment questions were

Table 2. Quality Assessment Questions.

QA Question
Q1. Are the objectives of the proposed work related to eliciting, capturing, modeling, and/or analyzing requirements interdependencies?
Q2. Are the proposed dependencies/relationships clearly defined?
Q3. Does the work propose sufficient concepts to capture dependencies/relationships among requirements?
Q4. Have the dependencies/relationships been applied to/within a project/case study, or justified by appropriate examples?

used to filter 106 papers, requiring studies to be directly relevant to requirements interdependencies (Q1), offer clear definitions (Q2), provide sufficient conceptual depth (Q3), and, preferably, include practical validation or examples (Q4). This resulted in the selection of 22 papers. A simplified representation of the search and selection process is shown in Figure 3. For transparency and further research, a spreadsheet documenting the lists of the selected papers in each step has been made available online[2].

3. Reporting the results. This phase involves summarizing the results, and it consists of two main activities: 1- data synthesis; and 2- results and discussion.

Data synthesis aims at combining the findings of the selected studies in a way that allows answering the RQs. To facilitate this activity, each of the selected papers has been analyzed, and its contribution to the RQs has been summarized. In what follows, we describe how data synthesis was conducted. Data related to *RQ1* were extracted from the list of selected papers, which were analyzed to identify relevant relationships/interdependencies that are shown in Table 3 along with their frequency of appearance in the selected papers. Key relationships/interdependencies are distinguished using **Bold** typeset. These key relationships/interdependencies have been selected based on their frequency of appearance and their importance for capturing relationships/interdependencies among requirements. To answer *RQ2*, we analyzed how each identified key requirement interdependency has been used in the relevant literature. Table 4 includes these key relationships/interdependencies accompanied by a brief description of each of them. Finally, to answer *RQ3*, we first categorized interdependencies into three groups. Then, categorized the selected studies based on their coverage of these three groups (see Table 5). In the following section, we discuss how each *RQ* has been answered.

5 Review results and discussion

In this section, we present and discuss the findings of this review concerning each RQ in its corresponding step, as follows:

RQ1: What types of requirements interdependencies/relationships exist in the literature? To address this RQ, we systematically analyzed the selected literature to identify the most relevant and frequently cited requirements interdependencies. Our analysis

[2] `https://doi.org/10.5281/zenodo.16731599`.

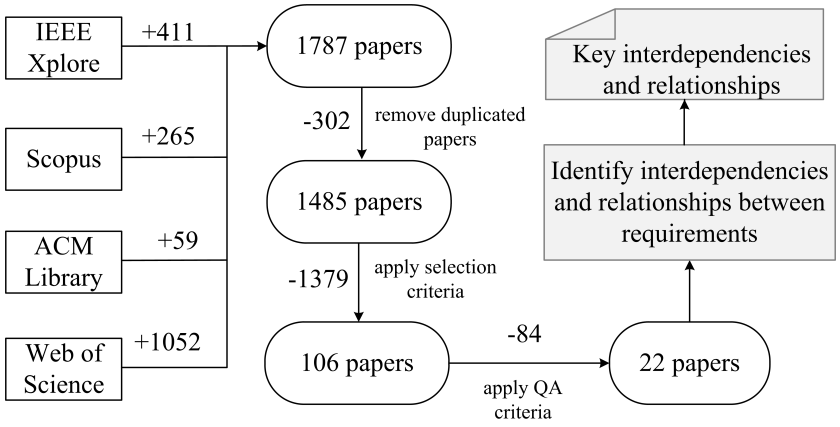


Fig. 3. Paper search and selection process.

revealed 30 interdependencies (shown in Table 3), from which we selected 16 key relationships highlighted in **bold**. The selection criteria were based on: (1) conceptual distinctiveness (removing overlaps with similar interdependencies), (2) frequency of appearance across studies, and (3) their critical role in capturing essential requirement relationships. This selection process ensures our findings focus on the most impactful interdependencies for requirements engineering practice.

The identified dependencies cover diverse aspects of requirements interdependencies as they were derived from studies from different domains. For instance, some dependencies focus on capturing relationships between Functional Requirements (FR), while others capture relationships between Non-Functional Requirements (NFR). The first type of relationship captures the interdependency between what a system must do, and the last captures the relationships among qualities the system strives to achieve. Moreover, some identified dependencies involve an interaction between FR and NFR. This type of relation highlights the interdependency between what a system must do and the qualities or constraints under which the system must operate. Moreover, there were some infrequently identified dependencies such as *Resource-Based Dependencies*, which considers the availability of resources that a certain requirement is dependent on the completion or availability of resources allocated to other requirements.

RQ2: How are interdependencies between requirements utilized across different requirement types? Interdependencies between different requirement types—specifically FRs and NFRs—play an important role in ensuring that the overall system design and development are coherent, efficient, and aligned with both users' needs and system constraints. Through examining the selected papers, we refined and precisely characterized these interdependencies, establishing an unambiguous context of their use that articulates their distinct functions within system specification. Our analysis of how these interdependencies have been employed and defined across the papers enabled us to: (1) disambiguate their conceptual boundaries, (2) clarify their operational roles, and (3) standardize their interpretation. The complete list of key interdependencies is presented

Table 3. Requirements dependencies in the selected papers

	Carlshamre et al. (2001)	**Kulshreshtha et al. (2012)**	**Alzyoudi et al. (2015)**	**Svensson et al. (2010)**	**Sumesh et al. (2019)**	**Tabassum et al. (2014)**	**Guan et al. (2021)**	**Guo et al. (2021)**	**Vasilache and Tanaka (2005)**	**In and Boehm (2001)**	**Gupta and Gupta (2022)**	**Martinez et al. (2019)**	**Soomro et al. (2013)**	**Jaeger and Hoffmann (2008)**	**Egyed and Grünbacher (2004)**	**Sumesh and Krishna (2022)**	**Mylopoulos et al. (2020)**	**Dalpiaz et al. (2016)**	**Van Lamsweerde (2001)**	**Miller et al. (2015)**	**Tong et al. (2018)**	**Monostori et al. (2000)**
And	✓		✓	✓		✓	✓	✓			✓	✓			✓	✓		✓	✓	✓		
Temporal	✓																					
Replace								✓														
Or	✓		✓	✓		✓	✓	✓			✓	✓			✓	✓		✓	✓	✓		
Before			✓																	✓		
Compliance		✓														✓			✓			
Wish																			✓			
Contribution					✓											✓		✓	✓			
Continuance		✓																				
Non-Invertible																						✓
Consequential		✓																				
Conflict			✓						✓				✓		✓	✓			✓		✓	
Concurency									✓													
Cooperation		✓													✓				✓			
CValue	✓																					
Dependence													✓								✓	
ICost	✓					✓				✓												
Social																	✓					
Exclusion								✓	✓													
Requires	✓					✓		✓										✓		✓		
Coordinate													✓								✓	
Contradiction													✓								✓	
Contractual		✓																				
S-Cost										✓												
Cause-Effect									✓													
Time									✓													
Supports			✓					✓				✓							✓		✓	
Abstraction									✓													
Equivalence								✓														
Satisfies								✓										✓	✓			

in Table 4. Further, we have classified these interdependencies based on the type of the involved requirements into the following three top-level categories:

1. **Intra-type (FR-to-FR or NFR-to-NFR) interdependencies:** capture relationships among the system-to-be functionalities/capabilities (FR-to-FR) or among the system-to-be qualities (NFR-to-NFR). These include *(1) FR-to-FR interdependencies:* are used to capture the hierarchies and sequence relationships among the FR of the system-to-be (e.g., *And*, *Or*). *(2) NFR-to-NFR interdependencies:* are used to capture the relationships among the quality attributes of the system-to-be (e.g., *Cooperation*). *(3) FR-to-FR or NFR-to-NFR interdependencies:* capture interactions either among functional capabilities (FR-to-FR) or among quality attributes (NFR-to-NFR) of the system-to-be (e.g., *Conflict*, *Consequential*).
2. **Cross-type (FR-to-NFR and NFR-to-FR) interdependencies:** capture how the capabilities of the system-to-be might influence or get influenced by its qualities (e.g., *Exclusion - (FR-to-NFR)*, *Compliance - (NFR-to-FR)*, *Contribution (FR-to-NFR or NFR-to-FR)*), i.e., they capture how the capabilities of the system-to-be fundamentally intertwines with its qualities.

(3) Multi-type (FR/NFR-to-FR/NFR) interdependencies: describe interdependencies that can be both Intra-type (FR-to-FR or NFR-to-NFR) and cross-type (FR-to-NFR and NFR-to-FR) interdependencies (e.g., *Supports*, *ICost*). Unlike the first two categories, which are confined to either homogeneous (FR/FR, NFR/NFR) or bidirectional (FR/NFR) dependencies, multi-type interdependencies involve dynamic combinations of FR and NFR, often emerging in scenarios where system behaviors and qualities are co-dependent in non-linear ways (e.g., an FR modifying an NFR, which in turn impacts another NFR).

Table 4: Key requirements interdependencies in relevant literature

And refines top-level functional requirement (FR) into several lower-level FRs, where the fulfillment of the top-level FR requires the fulfillment of all lower-level FRs Carlshamre et al. (2001); Alzyoudi et al. (2015); Svensson et al. (2010); Tabassum et al. (2014); Guan et al. (2021); Guo et al. (2021); Gharib et al. (2021).
Or refines top-level FR into several lower-level FRs, where the fulfillment of the top-level FR require the fulfillment of any of the lower-level FRs Carlshamre et al. (2001); Alzyoudi et al. (2015); Svensson et al. (2010).
Before ensures that one FR must be fulfilled before another can be satisfied Alzyoudi et al. (2015); Van Lamsweerde (2001).
Contractual captures the relationship where one requirement necessitates data from another Kulshreshtha et al. (2012). As FR requires data in most existing languages, we limit this interdependency to FR-to-FR.

Cooperation captures a state when two or more requirements collaborate to achieve a shared goal, where the effective execution of these requirements enhances or enables the execution of another requirement Kulshreshtha et al. (2012); Egyed and Grünbacher (2004); Van Lamsweerde (2001). While literature does not explicitly limit cooperation to NFRs, we argue they are uniquely suited to NFRs due to their inherent traits: (1) their satisfaction involves continuous trade-offs rather than binary fulfillment (e.g., balancing security vs. usability), and (2) they produce emergent system qualities through synergy (e.g., reliability and performance jointly enabling resilience). Moreover, And, a FR-to-FR interdependency, can capture how two or more FRs can collaborate to achieve a top level FR.
Conflict captures a state where two requirements (FR-to-FR or NFR-to-NFR) are mutually incompatible, creating contradictions that must be resolved to maintain system coherence and functionality Soomro et al. (2013); Egyed and Grünbacher (2004). In iStar Dalpiaz et al. (2016) "break" between FR-to-NFR has been introduced. However, break is not equivalent to conflict conceptually.
Continuance captures a state where one requirement triggers another, ensuring a logical and orderly development process Kulshreshtha et al. (2012). **Continuance** is applicable to both FRs and NFRs.
Consequential captures a dependency between two requirements (FR or NFR), where modifying or fulfilling one directly necessitates a corresponding adjustment or fulfillment of the other (FR or NFR) to maintain consistency Kulshreshtha et al. (2012).
Compliance captures the conformity of one requirement (usually a FR) to another NFR concerning laws, rules, standards, or policies Kulshreshtha et al. (2012); Sumesh and Krishna (2022).
Exclusion captures a state when the implementation of one NFR is explicitly prohibited or invalidated by a FR, often due to mutual exclusivity or conflicting nature Vasilache and Tanaka (2005).
Contribution captures how requirements collectively enable or hinder the fulfillment of other requirements Sumesh et al. (2019). It operates bidirectionally between FRs and NFRs—capturing how FR implementations influence NFR satisfaction (e.g., caching improving performance) and how NFRs constrain FR (e.g., security mandating encryption). While iStar Dalpiaz et al. (2016) considers NFR-to-NFR contributions, we deliberately exclude these interactions because they are more accurately represented through other interdependency types: Cooperation for synergistic NFR pairs (e.g., scalability and availability enhancing system resilience) and Conflict for competing qualities (e.g., security versus usability).
Satisfies captures a relationship where the implementation of a requirement ensures the fulfillment of the specified objectives or criteria of another requirement Guo et al. (2021); Dalpiaz et al. (2016); Van Lamsweerde (2001), applicable to both FRs and NFRs.
Supports captures a relationship where the successful implementation of one requirement assists the effective achievement of another Alzyoudi et al. (2015); Van Lamsweerde (2001), applicable to both FRs and NFRs.

Requires captures a relationship where the fulfillment of one requirement (FR or NFR) is contingent upon the satisfaction of another requirement (FR or NFR), indicating that one requirement relies on another to function effectively Tabassum et al. (2014); Guo et al. (2021); Dalpiaz et al. (2016); Miller et al. (2015).
ICost captures the impact of implementing one requirement (FR or NFR) on the cost of implementing another requirement (FR or NFR), either increasing or decreasing the expense Carlshamre et al. (2001); Tabassum et al. (2014); In and Boehm (2001).
CValue captures the impact of implementing one requirement (FR or NFR) on the perceived value of another requirement (FR or NFR), influencing development prioritization and the overall value proposition of the system Carlshamre et al. (2001).

RQ3: What are the key coverage gaps in existing studies on requirements interdependencies? We answer this question by comparing the coverage of the selected papers with the three categories of requirements interdependencies. The results presented in Table 5 reveal important insights into how existing studies address the three categories of requirements interdependencies: Intra-type, Cross-type, and Multi-type. Most studies provide partial coverage of Intra-type interdependencies (FR-to-FR or NFR-to-NFR), reflecting their alignment with traditional requirements engineering practices like goal decomposition and conflict detection. For instance, only Carlshamre et al. Carlshamre et al. (2001), Dalpiaz et al. Dalpiaz et al. (2016), and Van Lamsweerde Van Lamsweerde (2001) sufficiently cover these homogeneous relationships. While some studies (e.g., In and Boehm (2001)) offer poor coverage, likely due to their narrower focus on specific interdependency types such as temporal or resource constraints.

Cross-type interdependencies (FR-to-NFR or NFR-to-FR) receive moderate attention in the papers. Despite this, nearly half of the reviewed papers overlook Cross-type relationships entirely. This gap suggests that many approaches still prioritize homogeneous requirement analysis over heterogeneous interactions, limiting their ability to address real-world scenarios where functionality and quality constraints intersect. Multi-type interdependencies (FR/NFR-to-FR/NFR) also received moderate attention, where only a handful of studies (e.g., Tong et al. (2018); Kulshreshtha et al. (2012)) sufficiently address these interdependencies. The scarcity of research in this area highlights a critical gap in requirements engineering, as Multi-type dependencies are essential for modeling dynamic systems where FR and NFR co-evolve. The limited coverage in existing frameworks like iStar Dalpiaz et al. (2016) and KAOS Van Lamsweerde (2001) further validates the need for the comprehensive taxonomy proposed in this paper.

The uneven distribution of coverage across these categories underscores the fragmented nature of current research on requirements interdependencies. While Intra-type relationships are well-established, the lack of standardized approaches for Cross-type and Multi-type dependencies hinders holistic requirements management.

6 Constructing the Taxonomy for Requirements Interdependencies

The construction of the taxonomy was straightforward, as we first identified key interdependencies while addressing *RQ1* (see Table 3) and later categorized them into three

Table 5. A coverage of interdependency categories in selected primary studies.

Paper	Intra-type interd.	Cross-type interd.	Multi-type interd.
Carlshamre et al. (2001)	●	◐	◐
Kulshreshtha et al. (2012)	◐	○	●
Alzyoudi et al. (2015)	◐	◐	◐
Svensson et al. (2010)	◐	◐	○
Sumesh et al. (2019)	◐	◐	○
Tabassum et al. (2014)	◐	○	○
Guan et al. (2021)	◐	○	◐
Guo et al. (2021)	◐	○	○
Vasilache and Tanaka (2005)	◐	○	●
In and Boehm (2001)	○	○	◐
Gupta and Gupta (2022)	◐	◐	◐
Martinez et al. (2019)	●	◐	◐
Soomro et al. (2013)	◐	○	○
Jaeger and Hoffmann (2008)	◐	○	○
Egyed and Grünbacher (2004)	◐	○	◐
Sumesh and Krishna (2022)	◐	◐	○
Mylopoulos et al. (2020)	◐	◐	○
Dalpiaz et al. (2016)	●	◐	◐
Van Lamsweerde (2001)	◐	◐	◐
Miller et al. (2015)	◐	○	○
Tong et al. (2018)	◐	◐	●
Monostori et al. (2000)	◐	○	○
● Sufficiently covered; ◐ Partially covered; ○ Poorly covered			

distinct groups in *RQ2* (see Table 4). The resulting taxonomy is illustrated in Figure 4, which highlights the top-level categories of requirements interdependencies. This section details each interdependency type, clarifies conceptual distinctions between them, and illustrates their application with relevant examples.

1. Intra-type (FR-to-FR or NFR-to-NFR) interdependencies.

And *interdependency* decomposes complex FR into operational sub-requirements. For example, "**FR1.** *The system shall process online payments"* can be refined into "**FR1.1.** *The system shall validate the user's credit card information"* and "**FR1.2.** *The system shall confirm payment completion"*. Fulfilling **FR1** demands the fulfillment of **FR1.1** and **FR1.2**.

Or *interdependency* provides alternative refinements for a FR. For example, "**FR1.** *The system shall authenticate users"* can be refined into "**FR1.1.** *The system shall authenticate users via user name and password"* and "**FR1.2.** *The system shall authenticate users via biometric verification (fingerprint/facial recognition)*". Fulfilling **FR1** demands the fulfillment of **FR1.1** or **FR1.2**, offering design flexibility.

Before *interdependency* ensure that a dependent requirement (**FR2**) cannot be implemented until its prerequisite requirement (**FR1**) has been satisfied. For example, *Before* can enforce the implementation of "**FR1.** *The system shall store user profile data in a database*" before "**FR2.** *The system shall display user profiles on the dashboard*".

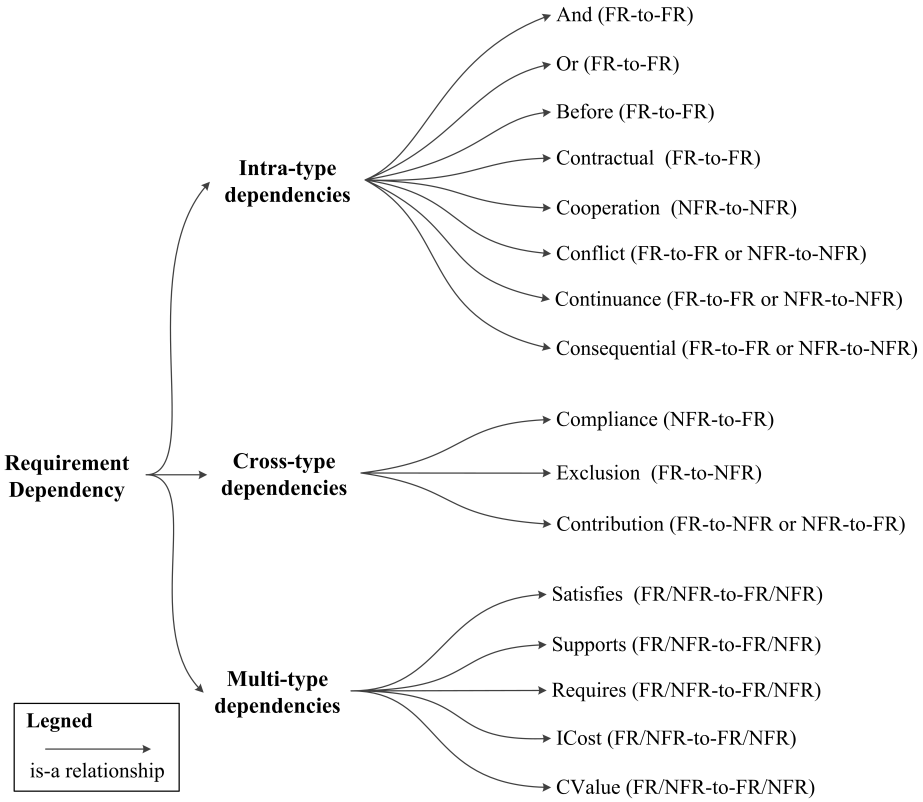


Fig. 4. The Proposed Novel Taxonomy of Requirements Interdependencies.

Contractual *interdependency* mandates data flow between FRs Kulshreshtha et al. (2012). Unlike *Before* that enforces implementation order, *Contractual* ensures data flow. For example we can use this interdependency to contractually obligate "**FR1.** *The payment system shall generate a transaction receipt containing: amount, date, and merchant ID"* to provide specific data to "**FR2.** *The accounting module shall record all transactions using the receipt data (amount, date, merchant ID)"*.

Cooperation *interdependency* becomes handy when two or more NFRs collaborate to achieve a shared goal. For instance, consider a cloud service where "**NFR1.** *Resilience: requires the system to withstand 10x traffic spikes"*, which can be achieved through the Cooperation of "**NFR2.** *Availability: mandating 99.9% uptime during 10x normal load"* and "**NFR3.** *Performance: requiring ¡2s API response under peak loads"*. **NFR2.** and **NFR3.** create a resilient system where the combined effect exceeds their individual contributions, exemplifying the emergent properties characteristic of NFR cooperation.

Conflict *interdependency* can arise in both FRs and NFRs. Among FRs, conflicts occur when implementation needs are mutually exclusive, creating binary incompatibilities that require prioritization or design arbitration. For example, "**FR1.** *The system*

shall delete all user-identifiable data 30 days after account termination (GDPR Article 17)" is clearly conflicting with "**FR2.** *The system shall retain complete transaction audit logs for one year"*. In contrast, NFR conflicts emerge from fundamental quality trade-offs. For example, "**NFR1.** *All communications shall use AES-256 encryption"* and "**NFR2.** *API endpoints shall respond within 100ms for 95% of requests"* are clearly conflicting and cannot be satisfied simultaneously.

Continuance *interdependency* can be applied to both FRs and NFRs, dictating the order of requirements and conditions for successful project completion. For example, "**FR1.** *The system shall allow users to add items to a shopping cart"* triggers "**FR2.** *The system shall calculate and display the total order cost in real-time"*. This ensures that developers cannot implement real-time price updates (**FR2.**) without first building the cart functionality (**FR1.**). Another example, "**NFR1.** *The system shall handle 10,000 concurrent users"* triggers "**NFR2.** *The database shall support connection pooling to manage high traffic.*". Unlike Before that specifies a temporal implementation constraint between FRs, Continuance describes a logical triggering relationship where fulfilling one requirement (e.g., "ensure data privacy") inherently necessitates addressing another (e.g., "implement encryption").

Consequential *interdependency* maintains consistency between requirements (FR or NFR) when one of them changes. For example, if "**FR1.** *The shopping cart shall apply location-based tax calculations during checkout."* is modified to "apply real-time tax API calls", then, "**FR2.** *The system shall store the user's verified shipping address before payment processing."* must consequentially be updated to: "Store shipping address with timestamp verification" (to ensure tax calculations use current rates). Unlike *Continuance*, which dictates a logical trigger for implementation order, *Consequential* ensures mutual adjustment during system evolution.

2. Cross-type (FR-to-NFR and NFR-to-FR) interdependencies.

Compliance *interdependency* guarantees that system implementation aligns with external standards (e.g., GDPR, ISO) or internal governance rules, mitigating legal and operational risks. For example, "**FR1.** *Doctors shall access patient medical histories."* must comply with "**NFR1.** *Patient data access shall comply with the GDPR"*.

Exclusion *interdependency* is essential in cases where specific functionalities are unable to coexist within some system qualities. For example, "**FR1.** *The car shall allow drivers to manually override autonomous controls at any time."* and "**NFR1.** *The system shall prevent all driver actions that could cause unsafe maneuvers."* **FR1.** requires unrestricted manual control, while **NFR1.** demands forced system intervention to block dangerous actions.

Contribution *interdependency* operates bidirectionally between FRs and NFRs, capturing how one requirement enables or hinders another. For example, implementing **FR1.** *The system shall compress image uploads using lossy compression (JPEG)."* positively contributes to "**NFR1.** *The system shall minimize bandwidth usage during file transfers."*.

(3) Multi-type (FR/NFR-to-FR/NFR) interdependencies.

Satisfies *interdependency* indicates that one requirement completely fulfills another objectives or criteria. For example, implementing **FR2.** *The steering control module shall implement triple redundant sensors with voting logic."* regarding an autonomous

vehicle control system satisfies "**NFR1.** *The system shall achieve ASIL-D fault tolerance per ISO 26262."* in an autonomous vehicle system.

Supports *interdependency* indicates partial assistance where one requirement improves another requirement achievement likelihood without guaranteeing it. Unlike *Satisfies* (which ensures complete fulfillment), *Supports* identifies probabilistic enhancements. Consider the following two requirements: "**NFR1.** *The system shall maintain* $\leq$*10ms latency for emergency braking signals."* and "**FR1.** *The controller area network (CAN) bus shall prioritize braking messages over non-critical signals."* **FR1.** supports **NFR1.** by improving response times, but cannot alone guarantee the $\leq$10ms target.

Requires *interdependency* captures essential logical dependencies where one requirement cannot function without another. Unlike *Satisfies* (fulfillment) or *Before* (sequence), *Requires* denotes prerequisite necessity. For example, "**FR1.** *The system shall activate emergency braking when obstacles are detected"* requires "**NFR1.** *The system shall process lidar sensor data in real-time"* to function.

ICost *interdependency* captures how one requirement affects another's implementation cost. For example, "**FR1.** *The system shall record HD video (1080p) from all cameras"* increases the storage cost of "**FR2.** *The system shall store video footage for 30 days.".*

CValue *interdependency* captures how one requirement enhances another's perceived value. For example, "**FR1.** *Let users save favorite restaurants"* increases the value of "**FR2.** *Show personalized recommendations."* by providing important user preference data.

In conclusion, the taxonomy presented here provides a structured set of interdependencies to model complex requirement relationships. It is important to note that these interdependencies are not isolated; they often interact and complement each other within a requirements model. For example, a high-level goal might be achieved not through a simple And/Or refinement, but through the combined effect of a functional requirement and a non-functional requirement, linked by a Satisfies or Requires Cross-type relationship. More complex chains can emerge: for instance, an And-refinement of a goal might lead to a sub-goal that is in Conflict with another requirement, while a separate branch of the refinement Contributes to a key quality goal. Similarly, realizing one requirement might increase the cost (ICost) of another, creating a trade-off. Therefore, the practical power of this taxonomy lies in its ability to be used compositionally, allowing practitioners to capture the intricate networks of dependencies where functional, qualitative, and emotional requirements continuously influence one another to define the complete system behavior.

7 Evaluation

To evaluate the comprehensiveness and relevance of our proposed taxonomy, we conducted a comparative analysis against three well-established requirements engineering frameworks: iStar 2.0 Dalpiaz et al. (2016), KAOS (Knowledge Acquisition in automated specification) Van Lamsweerde (2001), and Emotion-led Modeling language (Motivational Goal Model (MGM)) Miller et al. (2015). The frameworks selected for evaluation—iStar 2.0, KAOS, and MGM—were chosen to provide a comprehensive

assessment across the principal paradigms of goal-oriented requirements engineering. This selection ensures the taxonomy is evaluated against diverse and complementary conceptualizations of requirements relationships. iStar 2.0 was chosen as the representative of social-oriented modeling, with its rich constructs for capturing strategic dependencies and rationales among actors. KAOS was included as the exemplar of formal, system-oriented modeling, providing a benchmark for taxonomic precision through its rigorous goal refinement and formal semantics. Finally, MGM was selected to incorporate the critical, yet often overlooked, human-affective perspective, testing the taxonomy's ability to capture emotional goals and quality-of-experience requirements that are increasingly vital in human-centric systems. By benchmarking against this triad, we demonstrate RelnTa's applicability across the key dimensions of modern requirements engineering: social, formal, and affective.

Our analysis systematically mapped the interdependencies in our taxonomy to related constructs in these reference frameworks, with the complete results presented in Table 6. The evaluation examines the 16 interdependencies defined in the ReInTa taxonomy, systematically analyzing their coverage across three distinct categories of interdependencies: intra-type, cross-type, and multi-type.

For **intra-type dependencies**, iStar 2.0 and KAOS provide robust support for **And/Or** refinements in goal or task decomposition, while MGM only supports basic goal decomposition via **And** without explicit **Or**-refinement. Temporal dependencies like **Before** are uniquely supported by KAOS, whereas **contractual** interdependencies remain unaddressed across all frameworks. **Cooperation** among NFRs can be indirectly modeled in iStar and KAOS but is absent in MGM. **Conflict** dependencies are partially addressed: iStar's **Break** relationship differs conceptually from true conflict, KAOS explicitly supports **conflict**, and MGM handles it implicitly. Notably, **continuance** and **consequential** dependencies, critical for dynamic requirements management, are not covered by any framework, underscoring a significant gap.

In **cross-type dependencies**, **compliance** interdependency is partially modeled in iStar (via quality attributes) and KAOS (as constraints), but MGM lacks explicit support. **Exclusion** is represented in iStar through **Break** links, while KAOS and MGM omit this concept. **Contribution** dependencies are well-supported across all frameworks, though terminology varies (e.g., "Implicit emotional influence" in MGM).

In **cross-type dependencies**, **compliance** interdependency is partially modeled in iStar (via quality attributes) and KAOS (as constraints), but MGM lacks explicit support. **Exclusion** is represented in iStar through **Break** links and in MGM via negative emotional influences, while KAOS omits this concept. **Contribution** dependencies are well-supported across all frameworks, though terminology varies (e.g., "Implicit emotional influence" in MGM).

The evaluation demonstrates that while current frameworks capture subsets of interdependencies in ReInTa taxonomy, their fragmented and inconsistent approaches leave critical gaps, particularly in cross-type and multi-type relationships. This inconsistency in terminology and partial coverage highlights the need for standardization. ReInTa addresses these limitations by offering a unified, comprehensive taxonomy with clearly defined terms and broader coverage.

Table 6. Comparative evaluation of the Taxonomy with existing frameworks.

Dependency - req. classification	**iStar 2.0 Dalpiaz et al. (2016)**	**KAOS Van Lamsweerde (2001)**	**MGM Miller et al. (2015)**
And - (FR-to-FR)	AND-refinement (Goal/Task decomposition)	AND-refinement (Goal/Task decomposition)	Goal (requirements) decomposition
Or - (FR-to-FR)	OR refinement	OR-refinement (Alternative goal refinements)	-
Before - (FR-to-FR)	-	Before	-
Contractual - (FR-to-FR)	-	-	-
Cooperation - (NFR-to-NFR)	Could be modeled as mutual dependencies	Could be modeled via joint agent assignments	-
Conflict - (FR-to-FR or NFR-to-NFR)	"Break" - conceptually, not equivalent to Conflict	Conflict	Conflict can be captured but not explicitly considered
Continuance - (FR-to-FR or NFR-to-NFR)	-	-	-
Consequential - (FR-to-FR or NFR-to-NFR)	-	-	-
Compliance - (NFR-to-FR)	Could be modeled as a Quality with "Make" contributions from compliant tasks	Could be modeled as domain properties or constraints, not as a distinct dependency type.	-
Exclusion - (FR-to-NFR)	Can be represented as a "break" link	-	-
Contribution - (FR-to-NFR or NFR-to-FR)	Contribution	Contribution	Implicit emotional/quality influence
Satisfies - (FR/NFR-to-FR/NFR)	Satisfy	Operationalization	is Implicitly captured only in the context of fulfilling a goal via its sub-goals
Supports - (FR/NFR-to-FR/NFR)	Supports	Supports	-
Requires - (FR/NFR-to-FR/NFR)	NeededBy	Only in goal-to-agent assignments	-
ICost - (FR/NFR-to-FR/NFR)	Could be modeled as Quality (Cost) and contributions	-	-
CValue - (FR/NFR-to-FR/NFR)	Could be modeled as Quality (value) and contributions	Could be modeled as contribution to soft goals (e.g., "MinimizeCost")	Only emotional/quality value contribution

8 Demonstration of Applicability

Following design science principles Hevner et al. (2004); Venable et al. (2012), we evaluate the utility of the RelnTa taxonomy by applying it to the Motivational Goal Model (MGM) framework Miller et al. (2015). This practical application serves to demonstrate RelnTa's relevance and its capacity to extend and enhance existing modeling frameworks. Specifically, we utilize ReInTa to introduce new interdependencies to MGM, thereby enhancing both the robustness and expressiveness of the resulting model. This expansion enables a richer understanding of the interdependencies between functional, quality, and emotional goals, leading to a more nuanced and complete representation of system requirements. This builds on our earlier demonstration that ReInTa encompasses a broader range of requirements interdependencies than MGM (see Table 6).

To demonstrate this, we apply ReInTa to a real-world example derived from the PHArA-ON (Pilots for Healthy and Active Ageing) project[3], an H2020 European initiative aimed at developing smart and active living solutions for Europe's aging population. Figure 5 depicts the requirements model for "Support the well-being of older adults" represented in the MGM language Miller et al. (2015). The original example has been extended to incorporate the complete set of concepts from MGM, including negative emotions and several quality and emotional goals[4]. The model is developed starting with specifying the top-level functional goal, namely "Support the well-being of older adults", which represents the main objective/goal the system needs to achieve. Then, the top-level goal is refined into three sub-functional goals: "Improve Digital Skills", "Participate in the community", and "Provide cognitive stimulation". These sub-goals are characterized by several quality and emotional goals, i.e., these quality and emotional goals are attached to relevant functional goals. For example, the functional goal "Improve Digital Skills" is associated with "Guided", "Adaptive", "Interactive", "Easy to use", "Consideration of skills", and "Helpful", quality goals as well as "Included" emotional goal, and "Overwhelmed" as negative emotions that should be avoided. Finally, the roles responsible for achieving/activating the goals are added to the model, namely: The Older adult in our example.

As illustrated in the figure and established in the preceding analysis, the MGM framework provides only a limited set of interdependencies for capturing the different relationships among requirements. In the rest of this section, we first analyze the interdependencies supported by MGM. Subsequently, we demonstrate how ReInTa's additional interdependencies - not accounted for in MGM - can extend both the scope and depth of MGM's requirements modeling capabilities.

1. Interdependencies supported by ReInTa and MGM. ReInTa encompasses nearly all core MGM interdependencies, and we will discuss how they are covered in the following discussion.

And interdependency is explicitly captured in MGM as shown in the figure, where the top-level functional goal *Support the well-being of older adults* is refined into three lower-level functional goals via And interdependency.

[3] `https://www.pharaon.eu/`

[4] Both emotional and quality goals are widely classified as NFRs

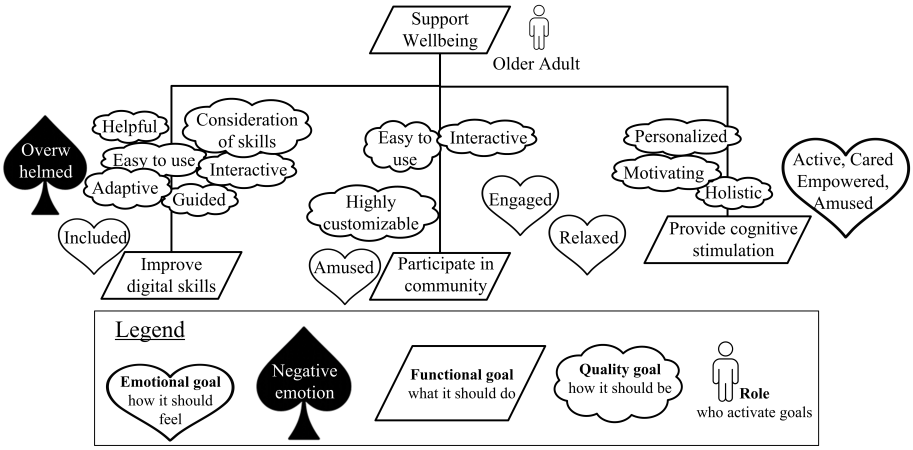


Fig. 5. MGM for Supporting the Well-being of older adults Gharib et al. (2024).

Conflict interdependency is implicitly considered in MGM, as MGM lacks a dedicated construct for such interdependency. For instance, the emotional goal *Included* is conflicting with the negative emotion *Overwhelmed*, which might result from the FR *Improve digital skills* as it requires intensive training or complex interfaces. This could trigger feelings of being overwhelmed (e.g., information overload, rapid skill demands). Another example is the conflict between *Easy to use* and *Highly customizable* as the former prioritizes simplicity, minimal steps, and intuitive interfaces, and the latter introduces complexity (e.g., settings, preferences), potentially overwhelming novice users. Note that the dotted link *Conflict* (represented in Fig. 6) does not exist in MGM since conflict is implicit, and has been added to the figure to facilitate comprehension.

Contribute interdependency can be captured implicitly in MGM as an influence of an emotional or quality goal over functional goals. For instance, both the emotional goal *Included* and the quality goal *Adaptive* contribute to the functional goal *Improve digital skills* in Fig. 6.

Satisfies interdependency In MGM, Satisfies is only captured in the context of fulfilling a goal through its sub-goals. In contrast, ReInTa models Satisfies as a bidirectional relationship between FR and NFR. For example, the functional goal *Implement gamified progress tracking* satisfies the emotional goal *Engaged*, illustrating how an FR can directly fulfill an NFR. This relationship is essential for capturing how FRs/NFRs are realized through other FRs/NFRs. Again, the dotted link *Satisfies* (shown in Fig. 6) does not exist in MGM since satisfies is not captured in this context, and there is no dedicated construct for it.

CValue interdependency can be captured implicitly in MGM as the realization of a goal/requirement can increase the value of another. For example, the quality goal *Easy to use* can increase the value of the emotional goal *Engaged* contributes to the functional goal *Improve digital skills*. Since no dedicated construct existed for this interdependency, we introduced it into the model as a dotted directed relationship, as illustrated in Fig. 6.

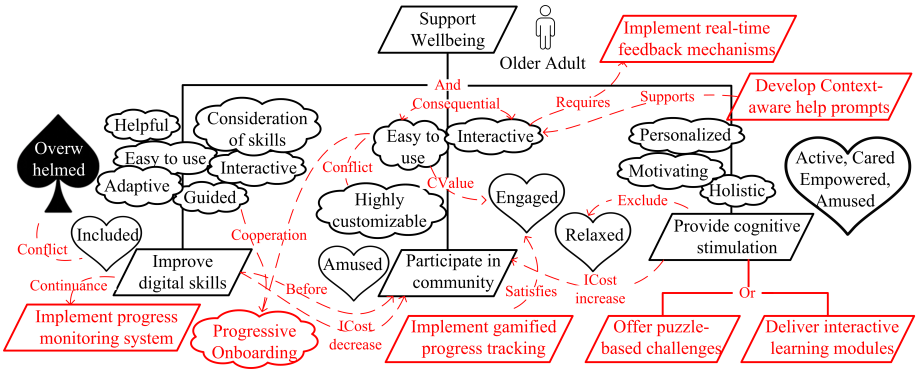


Fig. 6. Extended version of the MGM model for Supporting the Well-being of older adults

2. Interdependencies supported by ReInTa but absent in MGM. ReInTa introduces novel interdependencies not present in MGM. In the following discussion, we demonstrate how these extensions (shown in Fig. 6) enhance both the robustness and expressiveness of MGM.

Or interdependency we extend MGM with this interdependency and provide an example of its usage by refining the functional goal *Provide cognitive stimulation* via or into two sub-functional goals *Offer puzzle-based challenges* and *Deliver interactive learning modules*, i.e., the system could use puzzles OR learning modules to achieve cognitive stimulation. Additionally, to ensure consistency in representation, we enhanced the And refinement construct by explicitly including the term “And”.

Before interdependency we extend MGM with this interdependency and show how it can be used to require the realization of *Improve Digital Skills* before *Participate in the Community* as users must first acquire basic digital skills (prerequisite) to meaningfully engage in community platforms (e.g., forums, social media). E.g., a senior citizen must learn to use a tablet (Improve Digital Skills) before joining an online hobby group (Participate in the Community).

Contractual interdependency is essential for modeling systems where information flow between components must be explicitly defined. However, since MGM lacks any representation of data or information as a first-class concept, Contractual interdependency cannot be meaningfully expressed within the MGM framework.

Cooperation interdependency formalizes how synergistic interactions between requirements collectively achieve higher-level objectives. Considering it in MGM allows capturing how the quality goals *Easy to use* and *Guided* cooperate to fulfill the quality goal *Progressive Onboarding*. Here, first *Easy to use* ensures low entry barriers while *Guided* provides structured learning. Their combined effect enables seamless user acclimation by balancing accessibility with systematic instruction—precisely realizing *Progressive Onboarding* without overwhelming complexity.

Continuance interdependency adopting it in MGM will help uncover FRs that might otherwise be missed by highlighting dependencies between existing needs. For example, the functional goals *Improve digital skills* consequentially trigger *Implement*

progress monitoring system as tracking skill development becomes a necessary supporting requirement. Without this interdependency, the monitoring system might be overlooked, leading to an incomplete solution.

Consequential interdependency considering this interdependency in MGM will ensure that requirements remain logically consistent and operationally viable as they evolve throughout the requirements engineering process. For example, the quality goals *Easy to use* and *Interactive* have a consequential interdependency as adding new interactive features (e.g., drag-and-drop, live previews) requires compensatory ease-of-use measures (e.g., tooltips), while easy to use goals may constrain interaction complexity.

Compliance interdependency can proactively embeds compliance into system design when considered in MGM. For example, the functional goal *Participate in community* must comply with the quality goal *GDPR-Compliant processing* as user-generated content contains personal information and might pose a threat to users' privacy.

Exclusion interdependency considering this interdependency in MGM will pose a strict constraint on the inclusion of some NFRs that can compromise the fulfillment of FRs. For instance, the functional goal *Provide cognitive stimulation* fundamentally excludes the emotional goal *Relaxed*, as stimulating mental engagement (requiring active attention) directly contradicts the passive, stress-free state implied by relaxation. This interdependency serves as a vital design guardrail, preventing incompatible requirements from being implemented simultaneously.

Supports interdependency considering this interdependency in MGM will provide essential insights into how requirements might influence each other. For instance, the functional goal *Develop context-aware help prompts* supports the quality goal *Interactive*, as delivering real-time, adaptive help enhances interactivity.

Requires interdependency is particularly vital in complex systems where requirements have cascading dependencies. Considering it in MGM, will force explicit acknowledgment of technical and operational prerequisites during design. For example, the quality goal *Interactive* requires the functional goal *Implement real-time feedback mechanisms* since *Interactive* cannot be fulfilled without implementing real-time feedback features.

ICost interdependency considering this interdependency in MGM will capture the impact of implementing one requirement on the cost of implementing another one. For example, the functional goal *Improve Digital Skills* decreases the implementation cost of the functional goal *Participate in the Community* as users with stronger digital skills (from the first FR) require fewer moderation resources (reduced admin overhead) and simpler community interfaces, to mention a few.

This demonstration shows that RelnTa meaningfully extends MGM by introducing critical new interdependencies (e.g., Continuance, Cooperation, iCost) while fully supporting MGM's original relationships. Using the PHArA-ON case study, we illustrate how ReInTa captures subtle yet essential requirement interactions—such as timing constraints (Before), cost impacts (iCost), and mutual exclusions—that MGM cannot express. By enabling systematic analysis of trade-offs, compliance, and value delivery, ReInTa offers a practical upgrade to MGM for modeling complex systems.

9 Threats to validity

This section discusses potential threats to the validity of the study. Following Runeson et al. Runeson and Höst (2009), we categorize these threats into four types:

1. Construct validity addresses whether the study accurately measures the intended concepts and relationships Runeson and Höst (2009). A key threat in this regard is *systematic error*, which could emerge during the design or execution of our SLR. To minimize this risk, we employed multiple safeguards: First, we grounded our review protocol in established SLR methodologies, ensuring methodological rigor. Second, we maintained strict adherence to the protocol throughout all review phases, including the formulation of the search strategy, study selection, and data extraction. Third, we implemented peer verification, where two researchers independently validated key decisions to reduce subjective bias. These measures collectively enhance the reliability of our findings by aligning operational procedures with theoretical constructs.

2. Internal validity examines whether unaccounted factors may have influenced the study's outcomes, ensuring the observed effects accurately reflect the true relationships under investigation Runeson and Höst (2009). A critical threat in this regard is publication bias, a well-documented phenomenon where studies with statistically significant or positive results are more likely to be published than those with null or negative findings Kitchenham (2004). To mitigate this threat, we considered: (1) explicit inclusion of studies reporting failed or neutral results about interdependency applications (see Table 1 criteria); and (2) quality assessment using four questions (Table 2) that evaluated studies regardless of their outcome directionality. These measures help ensure our findings reflect the true spectrum of evidence in requirements interdependency research.

3. External validity examines the generalizability of our findings to broader contexts in requirements engineering Runeson and Höst (2009). A primary threat to external validity is *study completeness* - the practical impossibility of identifying every relevant study in the field. While exhaustive coverage remains theoretically unattainable, we systematically search across four major digital libraries (IEEE Xplore, ACM DL, Scopus, and Web of Science) using rigorously developed search strings; and (3) predefined inclusion/exclusion criteria (Table 1) to ensure methodological consistency. These measures collectively enhance the likelihood that our taxonomy reflects the true diversity of requirements interdependencies in both research and practice, strengthening confidence in its applicability beyond the immediate study context.

4. Reliability threats evaluate the consistency and reproducibility of our study's methodology and findings. The study's methodology, including search terms, sources, inclusion and exclusion criteria, and quality assessment questions, is fully documented[5]. This transparency ensures that other researchers can replicate the review and expect to obtain similar results.

[5] `https://doi.org/10.5281/zenodo.16731599`

10 Conclusions and future work

This paper presented ReInTa, a comprehensive taxonomy for requirements interdependencies developed through a systematic literature review and demonstrated its applicability through practical example. The taxonomy addresses critical gaps in existing frameworks by standardizing terminology, expanding coverage to include previously unaddressed relationships, and demonstrating practical utility through its extension of the Motivational Goal Model. By organizing 16 interdependencies into three clear categories (intra-type, Cross-type, and Multi-type), ReInTa provides researchers and practitioners with a more complete framework for understanding and managing complex requirement relationships. The comparative evaluation against established frameworks (iStar 2.0, KAOS, and MGM) confirmed ReInTa's superior coverage. The application to the PHArA-ON project showcased ReInTa's ability to enhance requirements modeling, especially for systems involving emotional and quality goals. These contributions represent a significant step forward in requirements engineering, offering both theoretical foundations and practical applications for managing interdependencies throughout the software development lifecycle.

Several promising directions emerge for extending this research. First, large-scale empirical studies across different industrial domains would help validate ReInTa's generalizability and identify potential adaptations for specific contexts. Second, developing tool support to automate dependency identification and analysis could significantly enhance practical adoption. Third, the taxonomy could be extended to address emerging challenges in AI-driven systems and IoT environments, where requirements interdependencies may exhibit unique characteristics. Finally, research could explore incorporating probabilistic measures to assess dependency strength and impact, enabling more informed decision-making. These future efforts would further strengthen ReInTa's position as a comprehensive framework for requirements interdependency management.

Acknowledgments

This work was supported by the Estonian Research Council grant "Developing human-centric digital solutions" (TEM-TA120).

References

Alzyoudi, R., Almakadmeh, K., Natoureah, H. (2015). A Probability Algorithm for Requirement Selection In Component-Based Software Development, *Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication*, Association for Computing Machinery (ACM), pp. 1–6.

Arnold, R., Bohner, S. (1996). *Software Change Impact Analysis (Practitioners)*. `http://www.amazon.com/Software-Change-Impact-Analysis-Practitioners/dp/0818673842`

Aurum, A., Wohlin, C. (2005). Requirements engineering: Setting the context, *Engineering and Managing Software Requirements*, Springer, pp. 1–15.

Bass, L., Clements, P. C., Kazman, R. (1997). *Software Architecture in Practice*.

Berander, P. (2004). Using students as subjects in requirements prioritization, *Proceedings - 2004 International Symposium on Empirical Software Engineering, ISESE 2004*, IEEE, pp. 167–176.

Carlshamre, P., Sandahl, K., Lindvall, M., Regnell, B., Natt och Dag, J. (2001). An industrial survey of requirements interdependencies in software product release planning, *Proceedings of the IEEE International Conference on Requirements Engineering*, IEEE, pp. 84–91.

Chung, L., Nixon, B., Yu, E., Mylopoulos, J. (1999). Non-Functional Requirements in Software Engineering, *Conceptual modeling: Foundations and applications* p. 472.

Dahlstedt, Å. G., Persson, A. (2005). Requirements interdependencies: State of the art and future challenges, *Engineering and Managing Software Requirements* pp. 95–116.

Dalpiaz, F., Franch, X., Horkoff, J. (2016). iStar 2.0 Language Guide, *Technical report*. `http://arxiv.org/abs/1605.07767`

Dubois, E., Hagelstein, J., Lahou, E., Ponsaert, F., Rifaut, a. (1986). A knowledge representation language for requirements engineering, *Proceedings of the IEEE* **74**(10), 1431–1444.

Egyed, A., Grünbacher, P. (2004). Identifying requirements conflicts and cooperation: How quality attributes and automated traceability can help, *IEEE Software* **21**(6), 50–58.

Gharib, M., Falco, M., Nijboer, F., Tinga, A. M., D'Agostini, S., Rovini, E., Fiorini, L., Cavallo, F., Taveter, K. (2024). Dealing with Emotional Requirements for Software Ecosystems: Findings and Lessons Learned in the PHArA-ON Project, *in*, *Proceedings of the 18th International Conference on Research Challenges in Information Science (RCIS24)*, Cham: Springer Nature Switzerland, pp. 99–114. `https://link.springer.com/10.1007/978-3-031-59465-6_7`

Gharib, M., Giorgini, P., Mylopoulos, J. (2021). COPri v.2 — A core ontology for privacy requirements, *Data and Knowledge Engineering* **133**, 101888. `https://linkinghub.elsevier.com/retrieve/pii/S0169023X2100015X`

Gharib, M., Salnitri, M., Paja, E., Giorgini, P., Mouratidis, H., Pavlidis, M., Ruiz, J. F., Fernandez, S., Siria, A. D. (2016). Privacy Requirements: Findings and Lessons Learned in Developing a Privacy Platform, *the 24th International Requirements Engineering Conference, RE 2016*, IEEE, pp. 256–265.

Greenspan, S., Mylopoulos, J., Borgida, A. (1982). Capturing more world knowledge in the requirements specification, *Proc. 6th international conference on Software Engineering*, IEEE Computer Society Press, pp. 225–234. `http://dl.acm.org/citation.cfm?id=800254.807765`

Guan, H., Cai, G., Zhao, C. (2021). An Automatic Approach to Extracting Requirement Dependencies based on Semantic Web, *The 8th International Conference on Dependable Systems and Their Applications*, IEEE, pp. 414–420.

Guo, W., Zhang, L., Lian, X. (2021). Putting software requirements under the microscope: Automated extraction of their semantic elements, *Proceedings of the IEEE International Conference on Requirements Engineering*, pp. 416–417.

Gupta, A., Gupta, C. (2022). A novel collaborative requirement prioritization approach to handle priority vagueness and inter-relationships, *Journal of King Saud University - Computer and Information Sciences* **34**(5), 2288–2297. `https://www.sciencedirect.com/science/article/pii/S1319157819310511`

Helfert, M., Herrmann, C. (2002). Proactive Data Quality Management for Data Warehouse Systems - A Metadata based Data Quality System, *4th International Workshop on Design and Management of Data Warehouses (DMDW 2002)*, Vol. 2002, pp. 97–106. `http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-58/herrmann.pdf`

Hevner, A. R., March, S. T., Park, J., Ram, S. (2004). Design science in information systems research, *MIS Quarterly: Management Information Systems* **28**(1), 75–105.

In, H., Boehm, B. W. (2001). Using WinWin Quality Requirements Management Tools: A Case Study, *Annals of Software Engineering* **11**(1), 141–174.
https://link.springer.com/article/10.1023/A:1012547320602

Jaeger, M. C., Hoffmann, A. (2008). Assessing Relations between Non-Functional Requirements, *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI)*, Vol. P-122, pp. 451–456.
https://dl.gi.de/bitstreams/4e5d24b2-05c8-45f2-8872-9c2c1cfc403a/download

Karlsson, L. (2004). Requirements prioritisation: an experiment on exhaustive pair-wise comparisons versus planning game partitioning, *"8th Internation Conference on Empirical Assessment in Software Engineering (EASE 2004)" Workshop - 26th International Conference on Software Engineering* (Ease), 145–154.
http://link.aip.org/link/IEESEM/v2004/i920/p145/s1&Agg=doi

Kitchenham, B. (2004). Procedures for Performing Systematic Reviews, *Technical Report 2004*.

Kotonya, G., Sommerville, I., Kotonya, G. (1998). *Requirements engineering: processes and techniques*, 1st edn, Wiley Publishing.

Kulshreshtha, V., Boardman, J., Verma, D. (2012). The emergence of requirements networks: The case for requirements inter-dependencies, *International Journal of Computer Applications in Technology* **45**(1), 28–41.
https://www.inderscienceonline.com/doi/abs/10.1504/IJCAT.2012.050130

Leffingwell, D., Widrig, D. (2000). *Managing software requirements: a unified approach*, Addison-Wesley Professional.

Martinez, G. G., Carpio, A. F. D., Gomez, L. N. (2019). A model for detecting conflicts and dependencies in non-functional requirements using scenarios and use cases, *Proceedings 45th Latin American Computing Conference, CLEI*, pp. 1–8.
https://ieeexplore.ieee.org/abstract/document/9073957/

Miller, T., Pedell, S., Lopez-Lorca, A. A., Mendoza, A., Sterling, L., Keirnan, A. (2015). Emotion-led modelling for people-oriented requirements engineering: The case study of emergency systems, *Journal of Systems and Software* **105**, 54–71.
https://www.sciencedirect.com/science/article/pii/S0164121215000667

Monostori, L., Viharos, Z. J., Markos, S. (2000). Satisfying various requirements in different levels and stages of machining using one general ANN-based process model, *Journal of Materials Processing Technology* **107**(1-3), 228–235.

Mylopoulos, J., Amyot, D., Logrippo, L., Parvizimosaed, A., Sharifi, S. (2020). Social dependence relationships in requirements engineering, *CEUR Workshop Proceedings*, Vol. 2641, pp. 55–60.

Noviyanto, F., Razali, R., Nazree, M. Z. A. (2023). Understanding requirements dependency in requirements prioritization: a systematic literature review, *International Journal of Advances in Intelligent Informatics* **9**(2), 249.
http://ijain.org/index.php/IJAIN/article/view/1082

Nuseibeh, B., Easterbrook, S., Russo, A. (2000). Leveraging inconsistency in software development, *Computer* **33**(4), 24–29.

Robinson, W. N., Pawlowski, S. D., Volkov, V. (2003). Requirements Interaction Management, *ACM Computing Surveys* **35**(2), 132–190.

Runeson, P., Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering, *Empirical Software Engineering* **14**(2), 131–164.

Sommerville, I. (2007). *Software engineering*, pearson Education limitd.

Sommerville, I., Sawyer, P. (1997). *Requirements engineering: a good practice guide*, John Wiley & Sons, Inc.

Soomro, S., Hafeez, A., Shaikh, A., Musavi, S. H. A. (2013). Ontology based requirement interdependency representation and visualization, *Communications in Computer and Information Science*, Vol. 414, Springer Verlag, pp. 259–270.
https://link.springer.com/chapter/10.1007/978-3-319-10987-9_24

Sumesh, S., Krishna, A. (2022). Sensitivity Analysis of Conflicting Goals in the i* Goal Model, *Computer Journal* **65**(6), 1434–1460.
https://academic.oup.com/comjnl/article-abstract/65/6/1434/6132358

Sumesh, S., Krishna, A., Subramanian, C. M., Murtagh, F. (2019). Game Theory-Based Reasoning of Opposing Non-functional Requirements using Inter-actor Dependencies, *Computer Journal* **62**(11), 1557–1583.
https://academic.oup.com/comjnl/article-abstract/62/11/1557/5288329

Svensson, R. B., Regnell, B., Aurum, A. (2010). Towards modeling guidelines for capturing the cost of improving software product quality in release planning, *ACM International Conference Proceeding Series*, pp. 20–23.

Tabassum, M. R., Siddik, M. S., Shoyaib, M., Khaled, S. M. (2014). Determining interdependency among non-functional requirements to reduce conflict, *International Conference on Informatics, Electronics and Vision*, IEEE, pp. 1–6.

Terry Bahill, A., Henderson, S. J. (2005). Requirements development, verification, and validation exhibited in famous failures, *Systems Engineering* **8**(1), 1–14.

Tong, Z. X., Su, X. H., Ding, X., Lin, J. X. (2018). Optimizing software development requirements based on dependency relations, *Journal of Information Hiding and Multimedia Signal Processing* **9**(4), 1014–1030.

Van Lamsweerde, A. (2001). Goal-oriented requirements engineering: A guided tour, *Requirements Engineering, 2001. Proceedings. Fifth IEEE International Symposium on*, IEEE, pp. 249–262.

van Lamsweerde, A., Darimont, R., Massonet, P. (1995). Goal-directed elaboration of requirements for a meeting scheduler: problems and lessons learnt, *Proceedings of the IEEE International Conference on Requirements Engineering*, Published by the IEEE Computer Society, pp. 194–203.

Vasilache, S., Tanaka, J. (2005). Bridging the gap between analysis and design using dependency diagrams, *Third ACIS International Conference on Software Engineering Research, Management and Applications, SERA 2005*, Vol. 2005, pp. 407–414.
https://ieeexplore.ieee.org/abstract/document/1563190/

Venable, J., Pries-Heje, J., Baskerville, R. (2012). A comprehensive framework for evaluation in design science research, *International Conference on Design Science Research in Information Systems*, Springer, pp. 423–438.

Wiegers, K., Beatty, J. (2013). *Software requirements*.

Yu, E. S. K. (1993). Modeling organizations for information systems requirements engineering, *Requirements Engineering, 1993., Proceedings of IEEE International Symposium on*, IEEE, pp. 34–41.

Zave, P., Jackson, M. (1997). Four dark corners of requirements engineering, *ACM Transactions on Software Engineering and Methodology (TOSEM)* **6**(1), 1–30.

Received August 3, 2025 , revised September 30, 2025, accepted November 5, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, pp. 862–884
https://doi.org/10.22364/bjmc.2025.13.4.06

Practical Assessment of the SSH Services' Transition to Post-Quantum Cryptography

Simona ZAVACKĖ, Linas BUKAUSKAS

Institute of Computer Science, Vilnius University, Vilnius, Lithuania

simona.zavacke@mif.stud.vu.lt, linas.bukauskas@mif.vu.lt

ORCID 0009-0007-5910-5658, ORCID 0000-0002-9781-9690

Abstract. Quantum computing poses a critical threat to classical public-key cryptography, driving the adoption of post-quantum cryptography (PQC). While PQC performance has been extensively benchmarked in TLS, empirical studies on the Secure Shell (SSH) protocol remain limited. This paper evaluates the integration of PQC and hybrid (classical+PQC) mechanisms into SSH using an OQS-OpenSSH 8.9 prototype with liboqs support. We measured handshake latency, session size, and Secure Copy Protocol (SCP) transfer performance across standardized algorithms (CRYSTALS-Kyber, CRYSTALS Dilithium, Sphincs+), Falcon, and hybrid configurations. Experiments on heterogeneous legacy hardware under realistic LAN conditions were validated through Wireshark captures and protocol checklists. The results show that Falcon yields the smallest handshake sizes, Dilithium offers balanced and stable performance, Kyber scales predictably with NIST security levels, and Sphincs+ incurs significant overhead. Hybrid modes add limited cost while retaining classical trust anchors. These findings extend SSH PQC research beyond handshake analysis, offering deployment-oriented guidance for quantum-safe migration.

Keywords: post-quantum cryptography, hybrid PQC, transition to PQC, SSH, ML-KEM, Kyber, ML-DSA, SLH-DSA, Falcon, NIST security levels

1 Introduction

The rapid development of quantum computing represents a credible threat to existing asymmetric cryptographic methods. Schemes such as RSA, DSA, and Elliptic Curve Cryptography (ECC) rely on number-theoretic problems whose security assumptions collapse in the presence of quantum adversaries. In response to this, the National Institute of Standards and Technology (NIST) finalized the first Post-Quantum Cryptography (PQC) standards on August 13, 2024. These standards, published as FIPS 203, FIPS 204, and FIPS 205, are designed to protect against attacks from quantum computers (Jacob W. S. Schneider, 2024; National Institute of Standards and Technology, 2024e,a):

- ML-KEM (FIPS 203), a module lattice-based key encapsulation mechanism derived from CRYSTALS-*Kyber* algorithm is for key exchange (National Institute of Standards and Technology, 2024b);
- ML-DSA (FIPS 204), a module lattice-based digital signature algorithm based on CRYSTALS *Dilithium* algorithm (National Institute of Standards and Technology, 2024c);
- SLH-DSA (FIPS 205), a stateless hash-based digital signature scheme derived from *Sphincs+* algorithm, as a backup in case ML-DSA is compromised (National Institute of Standards and Technology, 2024d).

Additional candidates, such as Falcon, remain under evaluation for inclusion in future standards (e.g., as FIPS 206, FN-DSA). Falcon is a lattice-based digital signature algorithm that leverages Fast Fourier Transforms (FFT) to optimize the efficiency and speed of key generation, signing, and verification (Prest et al., 2020). While its compact signatures and high verification performance make it an attractive option, Falcon's reliance on discrete Gaussian sampling introduces implementation complexity and potential side-channel risks. These factors, along with the need for further cryptanalysis and validation of secure, constant-time implementations, are among the reasons why Falcon has not yet completed the full standardization process (National Institute of Standards and Technology, 2022b; Pierre-Alain et al., 2024).

To ensure continuity of trust during the transition to PQC, some organizations and researchers have explored hybrid cryptographic constructions, which combine classical and PQC components to provide security guarantees (Bindel et al., 2019; Crockett et al., 2019; Rossi, 2021). For example, such approaches are reflected in X.509 hybrid certificate prototypes (Nina et al., 2019), hybrid TLS 1.3 key exchanges (Stebila et al., 2021), and IETF drafts on hybrid signature mechanisms (Ounsworth and Pala, 2024).

Despite significant progress in developing PQC algorithms and advancing hybrid approaches, the most significant challenge is ensuring that new cryptographic algorithms can be integrated seamlessly into existing systems, applications, and protocols. The key difficulty lies in validating the practical feasibility of PQC schemes: behavior under real operational conditions, compatibility with existing protocols, the magnitude of their computational burden, and their sensitivity to subtle integration flaws. Real-life experimentation is essential to assess not only the computational cost of PQC primitives but also their resilience to implementation-specific pitfalls such as timing leaks, resource exhaustion, or protocol-level incompatibilities. Without such empirical grounding, recommendations for PQC migration remain speculative and potentially misleading (Ott et al., 2019; Alnahawi et al., 2021).

Running a pilot project to evaluate the performance and security of new algorithms in real-world scenarios is a critical step in understanding how these PQC cryptographic systems behave under operational conditions. Real-world performance evaluation enables the identification of the most suitable algorithm variant and the fine-tuning of performance parameters (Mosca, 2018; Kreutzer et al., 2018; Barker et al., 2021; Joseph et al., 2024). This process helps validate the practical feasibility of PQC schemes and supports a smoother migration to a quantum-safe state.

The objective of this research is to evaluate the integration of PQC and its hybrid implementation into the Secure Shell (SSH) protocol. By developing a functioning

prototype based on OpenSSH and the liboqs framework, this study investigates the performance, compatibility, and practical deployment considerations of PQC algorithms as ML-KEM, ML-DSA, SLH-DSA, Falcon, and selected hybrid configurations.

The contributions of this work are threefold:

1. Integrating and evaluating PQC and its hybrid implementation in OpenSSH using liboqs.
2. Measuring handshake size and latency for various algorithm combinations in a controlled but realistic environment;
3. Analyzing the trade-offs between security level, session size, and connection time.
4. Recommending optimal algorithm configurations for practical PQC adoption in SSH.

This paper is structured as follows. Section 2 reviews the relevant literature and previous SSH-related PQC evaluations. Section 3 details the prototype environment, algorithms, and measurement methodology. Section 4 presents and interprets the performance data, followed by deeper analysis in Section 5. Section 6 summarizes conclusions and outlines directions for future work.

2 Related Work

PQC research has focused mainly on design and formal security analysis, while research on their practical implementation in protocols remains limited. Prior work on PQC integration in protocols is richest in TLS, where macro-benchmarks and even protocol variants (e.g., KEMTLS) quantify handshake time trade-offs and message size growth (Paquin et al., 2019; Schwabe et al., 2020; Stebila et al., 2021). TLS benchmarks reveal non-trivial PQC impact on handshake latency:

– Paquin et al. (2019) presents one of the first comprehensive integrations of PQC into mainstream TLS libraries. They integrated Kyber and Dilithium into OpenSSL and BoringSSL (TLS 1.3), experimented on virtual machines the client and server processes. Results demonstrated that while post-quantum public-key operations dominate median TLS handshake latency on fast, reliable links, on lossy networks ($\geq$3–5% packet loss), larger PQC message sizes become the primary bottleneck—and as application payloads increase, the relative cost of the PQC handshake rapidly diminishes.
– Schwabe et al. (2020) propose KEMTLS, removing signature rounds in TLS, and show optimized KEM-only handshakes cut latency by up to 30 %.

Although our focus is on SSH, we review PQC benchmarking in TLS because the two protocols share cryptographic primitives and migration concerns. TLS studies provide valuable baselines and methodological lessons, while also highlighting the distinctive aspects of SSH. In particular, TLS results emphasize handshake latency in short-lived connections, whereas SSH's session-oriented design raises different challenges such as sustained throughput and multi-client concurrency. Reviewing TLS benchmarks, therefore, helps situate our contribution within the broader landscape of PQC deployment studies.

SSH has credible but comparatively fewer empirical studies, mostly handshake-oriented. In particular, the handshake latency, session size, and throughput impacts of PQC remain underexplored outside of controlled laboratory settings:

- Crockett, Paquin, and Stebila (Crockett et al., 2019) demonstrated how PQC and hybrid key exchange could be integrated into SSH (and TLS) as proof-of-concept, using liboqs primitives and adapting OpenSSH code paths. Their emphasis was on design feasibility, not full-scale performance benchmarking.
- Sikeridis et al. (2020b) implemented hybrid ECDHE+PQC key exchanges (Kyber, Dilithium, SPHINCS+) in both TLS 1.3 and OpenSSH, demonstrating that PQC integration incurs handshake latency overheads ranging from just 0.5 % (Kyber) up to 50 % (SPHINCS+) and shows that simple TCP initial-window tuning can halve the perceived slowdown. These findings suggest that both cryptographic and network-level considerations are critical in PQC deployment. This research provided valuable early measurements of PQC integration into SSH, but their focus remained primarily on handshake costs, not sustained session or file transfer performance.

Formal analyses of hybrid SSH protocol correctness and security under hybrid models have been treated by:

- Crockett et al. (2019), who implemented and discussed protocol-level changes for hybrid PQC key exchange in OpenSSH 7.9.
- Tran et al. (2024), using symbolic methods (CafeOBJ) to verify that classical and PQC components interoperate securely.
- Kampanakis et al. (2020) represents one of the first standardization-oriented efforts to define how SSH can negotiate and fall back safely between classical and PQC key exchange algorithms. Their design patterns inform the hybrid architecture evaluated in this paper.

Our work complements these efforts by extending the analysis to full SSH sessions on real hardware, including file transfer workloads and hybrid configurations. Whereas most prior studies combine performance metrics with CPU profiling or formal analysis often limited to single-client PQC implementations in virtualized settings, this work focuses exclusively on session-level performance (handshake latency, session size, and throughput). We evaluate PQC in a uniform OpenSSH environment under realistic client–server loads, deployment–focused recommendations. In doing so, this study contributes empirical evidence that helps close the "SSH in the wild" gap by:

- Incorporating Falcon into SSH performance benchmarking.
- Testing on heterogeneous, legacy hardware representative of operational infrastructures.
- Quantifying handshake size and latency across PQC and hybrid configurations.
- Evaluating large file transfers to capture end-to-end SSH workload performance.
- Providing a balanced assessment of algorithm choices, weighing operational efficiency against required security assurance levels.

Table 1 summarizes how our work extends prior research.

Table 1: Comparison of selected SSH-related PQC studies

Study	Algorithms	Highlights
Crockett et al. (2019)	Kyber, Dilithium, SPHINCS+	First hybrid SSH in OpenSSH 7.9. No Falcon. Limited performance analysis.
Sikeridis et al. (2020a)	NIST Round 2/3	Latency testing under PQC; observed 0.5–50% delay; TCP tuning mitigates impact.
Stebila et al. (2021)	Kyber + ECC	Designed hybrid KEM groups for TLS 1.3; proposed hybridization strategies.
This work	Kyber, Dilithium, Falcon, SPHINCS+	Full SSH implementation on real hardware. Includes Falcon. Benchmarked handshakes and file transfers (SCP). Performance is evaluated in terms of security level.

3 Methods

We designed a replicable prototype method for the realistic operational conditions of SSH services and developed a prototype to empirically evaluate and identify the most suitable PQC algorithms and hybrid implementations for SSH-based communication. SSH works as an application and establishes automated connections.

3.1 Sandbox network architecture

To simulate real-world operating conditions, an evaluation test environment was created using legacy equipment representing a typical small business environment. Figure 1 illustrates the sandbox topology, where the client and server are interconnected through a Local Area Network. Experimental setup:

- Client: Ubuntu 20.04.5 LTS, Intel i5-4570S CPU (4 cores, 2.9 GHz), 16 GB RAM.
- Server: Ubuntu 22.04.1 LTS, AMD A8-5500 APU with Radeon HD Graphics (4 cores, 2.14 GHz), 16 GB RAM.
- Network infrastructure: TP-Link TL-SF1005D desktop switch (100 Mbps Ethernet), realistically limiting bandwidth. Network speed validated at 94 Mbps via `iperf`.

The prototype leveraged OpenSSH 8.9 (Git tag V_8_9_P1), integrating PQC through OQS-OpenSSH, a fork designed explicitly for quantum-resistant cryptography using liboqs version 0.7.2. libraries[1]. Initially developed for isolated environments (where client-server interactions were confined to a single virtual machine), the prototype required modifications to the SSH configuration files to enable automated connections between client and server instances running on separate physical devices. SSH was configured explicitly over IPv4, and TCP window scaling extensions were enabled to reflect common operational practices for optimized network data transmission.

[1] The prototype was made in 2022 November, OQS-OpenSSH snapshot 2022-08 was used`https://github.com/open-quantum-safe/openssh`. The old names of algorithms are used, as they were tested at that time, and there are some minor differences with standardised versions.

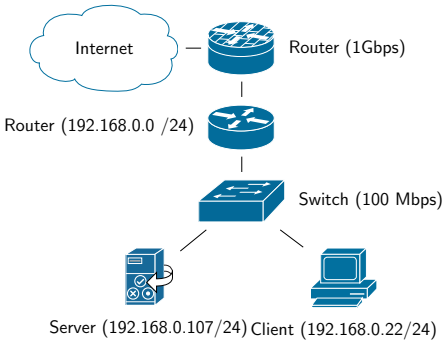


Fig. 1: Prototyping SSH: network test-bed

3.2 Algorithms under test

For comparative benchmarking, the prototype first establishes a *classical* cryptographic baseline. This baseline uses Ed25519 digital signatures[2] in combination with Elliptic Curve Diffie–Hellman (ECDH) key exchange across three standardized NIST recommended curves: `ecdh-nistp256`, `ecdh-nistp384`, and `ecdh-nistp521`.[3]

Table 2: SSH session performance with classical algorithms: Ed25519 signatures and ECDH key exchange

Key Exchange with Ed25519	**Time (ms)**	**Size (bytes)**	**NIST Security**
ECDH-nistp256	459	34,540	Level 1
ECDH-nistp384	475	34,700	Level 3
ECDH-nistp521	465	34,942	Level 5
Average	**466**	**34,727**	–

As shown in Table 2, this configuration achieves an average SSH handshake size of **34,727 bytes** and an average handshake latency of **466 ms**. These values serve as the primary reference point for evaluating the performance and practicality of PQC and hybrid implementations.

[2] Ed25519 is a high-performance Edwards-curve signature scheme designed for fast verification, short keys, and strong security guarantees (Bernstein et al., 2012). For more: `https://ed25519.cr.yp.to`, accessed 2025-08-05.

[3] ECDH is supported in SSH using the NIST recommended prime field curves `nistp256`, `nistp384`, and `nistp521`, implemented in OpenSSH as `ecdh-nistp256`, `ecdh-nistp384`, and `ecdh-nistp521`. RFC 5656 specifies all three for use in SSH key exchange (IETF SSH Working Group, 2009), while NIST SP 800-57 Part 3 designates `nistp256` and `nistp384` as mandatory-to-implement for federal systems, with `nistp521` available as an additional high-security option (Barker et al., 2015).

In PQC-based experiments, the ECDH is replaced by the *Kyber* Key Encapsulation Mechanism (KEM). Table 3 lists the six Kyber parameter sets evaluated. While all generate the same 32-byte shared secret, their key and ciphertext sizes differ as per the NIST security level. `Important note:` Kyber-90s is not standardized by NIST and is included here for comparative performance analysis only. For hybrid implementations, Kyber is combined with the corresponding ECDH parameters[4] to retain classical security alongside PQC resistance.

Table 3: PQC Kyber parameter sets

Parameter set	Public key size (bytes)	Secret key size (bytes)	Ciphertext size (bytes)	NIST Security	Hybrid with
Kyber512	800	1632	768	Level 1	ecdh-nistp256
Kyber512-90s	800	1632	768	Level 1	ecdh-nistp384
Kyber768	1184	2400	1088	Level 3	ecdh-nistp521
Kyber768-90s	1184	2400	1088	Level 3	ecdh-nistp256
Kyber1024	1568	3168	1568	Level 5	ecdh-nistp384
Kyber1024-90s	1568	3168	1568	Level 5	ecdh-nistp521

Authentication in the SSH prototype is performed via PQC digital signatures, as summarized in Table 4, and is tested:

- **Falcon** (lattice-based): offers the smallest signature size among PQC candidates, making it efficient in bandwidth-constrained environments.
- **Dilithium** (lattice-based): provides balanced key sizes and signatures.
- **Sphincs+** (hash-based): delivers strong security, but incurs the largest signature sizes, which may approach SSH packet fragmentation limits.

Table 4: PQC digital signature algorithm parameters

Parameter set	Public key size (bytes)	Secret key size (bytes)	Signature size (bytes)	NIST Security	Hybrid with
Falcon-512	897	1281	690	Level 1	ecdsa-nistp256
Falcon-1024	1793	2305	1330	Level 5	ecdsa-nistp521
Dilithium2-AES	1312	2528	2420	Level 1	ecdsa-nistp256
Dilithium3	1952	4000	3293	Level 3	ecdsa-nistp384
Dilithium5-AES	2592	4864	4595	Level 5	ecdsa-nistp521
Sphincs+ -128f	32	64	17088	Level 1	ecdsa-nistp256
Sphincs+ -192f	48	96	35664	Level 3	ecdsa-nistp384

[4] NIST elliptic curves follow the FIPS 186-5 specification National Institute of Standards and Technology (2023), with curve identifiers such as `nistp256` indicating key length in bits.

From a performance–security trade-off perspective, a higher NIST security level improves long-term resistance but typically increases the size of public keys, ciphertexts, or signatures (National Institute of Standards and Technology, 2016).

These baselines and PQC variants allow us to directly observe how key length, ciphertext size, and signature size translate into measurable SSH handshake performance. By aligning each algorithm with its corresponding NIST security level, these experiments enable the quantification of how the required security margin translates into real-world SSH handshake performance, providing an evidence-based basis for selecting algorithm variants for transition to PQC.

3.3 Evaluation Methodology

We perform repeated Black-Box trials measuring: (i) *Handshake latency* (client connect to finished key exchange and authentication); (ii) *Session-establishment size* (bytes on the wire); and (iii) *SCP transfer* behavior for files of 1 MB, 10 MB, 100 MB, 200 MB and 1024 MB.

Network analysis utilized Wireshark to capture and thoroughly quantify session-related data exchanges during SSH session establishment. For capturing precise timing and event sequencing to ensure comprehensive session validation the detailed SSH protocol steps checklist was created. The primary metrics utilized for evaluating each SSH connection include:

- **Handshake Latency:** The total duration from the initiation of an SSH connection request by the client to the successful completion of the cryptographic handshake, directly influencing user-perceived responsiveness.
- **Session Establishment Data Size:** The volume of network traffic data exchanged during the handshake phase. Lower data volumes reduce network load and enhance scalability, especially in bandwidth-limited environments.
- **Security Level:** Assessed based on NIST's standardized categorization, ranging from Security Level 1 (equivalent to AES-128) to Security Level 5 (equivalent to AES-256) (National Institute of Standards and Technology, 2016). This criterion ensures a balanced evaluation of algorithms, considering both efficiency and relative security assurance.
- **Success criteria.** No downtime; no noticeable performance degradation; and overhead acceptable for typical LAN conditions.

Limitations: Despite efforts toward realism, the sandbox configuration may not fully capture all potential complexities present in large-scale enterprise networks. Results should thus be interpreted with caution concerning generalizability beyond similar organizational contexts.

4 Results

The measurements were collected under the same network, hardware, and OpenSSH and liboqs configurations described in Section 3. Each signature algorithm was tested in both *PQC-only* and *Hybrid* configurations, with handshake session size (in bytes) and handshake time (in milliseconds) recorded for every KEM and signature pairing.

4.1 Performance data: Hybrid implementation

The results obtained from combinations of Hybrid implementations over an SSH session are presented. Table 5 presents data on the size of SSH session establishment in bytes, while Table 6 displays the session time results measured in milliseconds (ms).

Classical *Ed2519* was also combined with all hybrid Kyber to compare against classical ECDH in the SSH protocol. Together, these metrics enable a direct, evidence-based comparison of the bandwidth and latency implications of different Hybrid configurations.

Table 5: Hybrid algorithms combinations: session size (bytes)

Hybrid combinations results (size)	ecdsa-nistp256-Falcon512	ecdsa-nistp521-Falcon1024	ecdsa-nistp384-Dilithium3	ecdsa-nistp256-Dilithium2aes	ecdsa-nistp521-Dilithium5aes	ecdsa-nistp256-Sphincs+-128f	ecdsa-nistp384-Sphincs+-192f	Ed25519
ecdh-nistp256 Kyber512	42384	49958	54942	48730	62048	72130	112076	34892
ecdh-nistp384 Kyber768	42972	50864	55644	49564	62742	72840	112984	35396
ecdh-nistp521 Kyber1024	44172	51860	56786	50714	64016	73784	113532	36546
ecdh-nistp256 Kyber512-90s	**42204**	49826	55146	48738	62122	72014	112290	34636
ecdh-nistp384 Kyber768-90s	42972	50594	56238	50100	63352	73442	113528	35734
ecdh-nistp521 Kyber1024-90s	44502	52256	57256	51184	64552	74386	**114464**	36810

Table 6: Hybrid algorithms combinations: session time (milliseconds)

Hybrid session time (ms)	ecdsa-nistp256-Falcon512	ecdsa-nistp521-Falcon1024	ecdsa-nistp384-Dilithium3	ecdsa-nistp256-Dilithium2aes	ecdsa-nistp521-Dilithium5aes	ecdsa-nistp256-Sphincs+-128f	secdsa-nistp384-Sphincs+-192f	Ed25519
ecdh-nistp256 Kyber512	400	425	482	396	440	401	496	429
ecdh-nistp384 Kyber768	413	416	496	410	474	427	505	406
ecdh-nistp521 Kyber1024	439	404	460	382	425	422	476	415
ecdh-nistp256 Kyber512-90s	**347**	452	464	391	444	423	477	384
ecdh-nistp384 Kyber768-90s	369	453	536	428	503	482	577	421
ecdh-nistp521 Kyber1024-90s	426	477	508	389	483	481	**584**	435

The results of the Hybrid implementation (without *Ed25519*) show the following statistics: the average session size is **64188 bytes**, with an average time of **450 ms**. The medians are 56238 bytes and 440 ms.

4.2 Performance data: PQC implementation

Results of the combination of PQC algorithms over an SSH session are presented.

Table 7: PQC algorithms combinations: session size (bytes)

PQC session size (byte)	Falcon512	Falcon1024	Dilithium3	Dilithium2aes	Dilithium5aes	Sphincs+ -128f	Sphincs+ -192f
Kyber512	**41082**	48350	53614	47468	60358	70926	110748
Kyber768	41786	48930	54450	48172	60930	71498	111452
Kyber1024	43160	49942	55124	49102	62058	72692	111986
Kyber512-90s	41098	48308	53630	47476	60300	70926	110616
Kyber768-90s	41810	49012	54862	48708	61532	72224	111980
Kyber1024-90s	43120	50470	55858	49704	62528	73154	**112844**

Table 8: PQC algorithms combinations: session time (milliseconds)

PQC session time (ms)	Falcon512	Falcon1024	Dilithium3	Dilithium2aes	Dilithium5aes	Sphincsharaka 128fsimple	Sphincsharaka 192frobust
Kyber512	387	390	377	380	385	384	433
Kyber768	404	381	372	383	376	396	423
Kyber1024	390	395	**344**	397	381	363	423
Kyber512-90s	366	376	350	349	379	391	404
Kyber768-90s	388	399	386	410	408	407	**496**
Kyber1024-90s	367	428	390	404	431	456	480

Table 7 presents the size data (in bytes) for SSH session establishment, while Table 8 consists of the session's time results (in milliseconds). The statistical mean indicates that the average session size is **62714** bytes, with an average time of **396** ms. The medians are 54450 bytes and 390 ms.

4.3 Results of file transfer test

This subsection presents the results of SSH file transfers using the Secure Copy Protocol (SCP)[5] in PQC and Hybrid implementations where *Kyber512* operates in conjunction with three signature algorithms: *Sphincs+-128f* (abbreviated as *Sphincs*), *Dilithium2aes* (abbreviated as *Dilithium*) and *Falcon512* (abbreviated as *Falcon*). Files of 1MB, 10MB,

Table 9: File transfer results: Hybrid implementation

Hybrid mode	*File size (MB)*	*Transfer session time (ms)*	*Transfer session size (bytes)*	*Real size (bytes) of file*	*Time without SSH session establishment (ms)*	*Transfer session without SSH establishment and file size (bytes)*
Sphincs	1	435	1170980	1048576	12	50390
Sphincs	10	1187	11059856	10485760	764	502082
Sphincs	100	9354	109957252	104857600	8931	5027638
Sphincs	200	18206	220004306	209715200	17783	10217092
Dilithium	1	397	1148100	1048576	6	50786
Dilithium	10	1209	11043348	10485760	818	508850
Dilithium	100	9254	110009828	104857600	8863	5103490
Dilithium	200	18177	219947906	209715200	17786	10183968
Falcon	1	407	1140840	1048576	60	50060
Falcon	10	1195	11032106	10485760	848	504142
Falcon	100	9269	110003418	104857600	8922	5103614
Falcon	200	18190	219928354	209715200	17843	10170950

Table 10: File transfer results: PQC implementation

Pure PQC mode	*File size (MB)*	*Transfer session time (ms)*	*Transfer session size (bytes)*	*Real size (bytes) of file*	*Time without SSH session establishment (ms)*	*Transfer session without SSH establishment and file size (bytes)*
Sphincs	1	397	1167318	1048576	6	47816
Sphincs	10	1194	11058012	10485760	803	501326
Sphincs	100	9235	110010038	104857600	8844	5081512
Sphincs	200	18205	219933726	209715200	17814	10147600
Dilithium	1	380	1144066	1048576	31	48014
Dilithium	10	1194	11041954	10485760	845	508718
Dilithium	100	9222	110012526	104857600	8873	5107450
Dilithium	200	18185	219929410	209715200	17836	10166734
Falcon	1	384	1136294	1048576	18	46620
Falcon	10	1194	11036235	10485760	828	509377
Falcon	100	9232	109987610	104857600	8866	5088912
Falcon	200	18165	219939392	209715200	17799	10224192

100MB, 200MB, and 1GB were transmitted from the client to the server.

Notably, transferring a 1GB file took the same duration (1 minute and 29 seconds) and the speed of sessions was 11,2 MB/s across all combinations of algorithms in both PQC and Hybrid implementations, making it less relevant for exploring variations. The SSH SCP test results are presented in Table 9 (Hybrid implementation) and Table 10 (PQC implementation), detailing the times and sizes of file transfers.

5 Discussion

5.1 SSH handshake performance: PQC vs Hybrid vs Classical

Figure 2 compares the performance of SSH session establishments in terms of both time and size across three cryptographic configurations: PQC, Hybrid, and classical ECC only. While the classical ECC configuration achieves the smallest session establishment size, its handshake latency is not lower than that of PQC. Both PQC and Hybrid configurations show performance variability depending on the specific key exchange and signature algorithm combinations. Figure 3 presents the same dataset as box plots to highlight

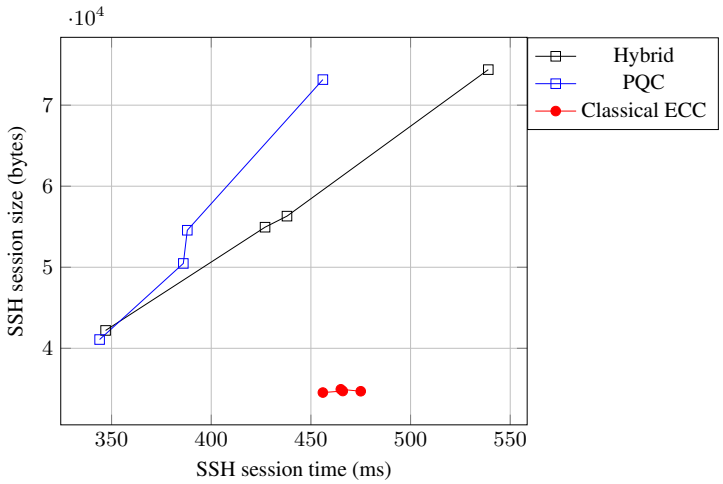


Fig. 2: SSH session establishment time and size for PQC, Hybrid, and Classical ECC configurations.

the statistical distribution of session sizes (left panel) and times (right panel). Each box plot shows the minimum, first quartile, median (represented by a vertical line), third quartile, and maximum values, with outliers indicated as circular markers. From a size perspective, Classical ECC appears as an outlier relative to PQC and Hybrid results due to its significantly smaller handshake sizes. In contrast, handshake time values for Classical ECC fall within the overall spread of Hybrid results. Handshake size results for

[5] SCP is a secure method within SSH for transferring files between a client and a server.

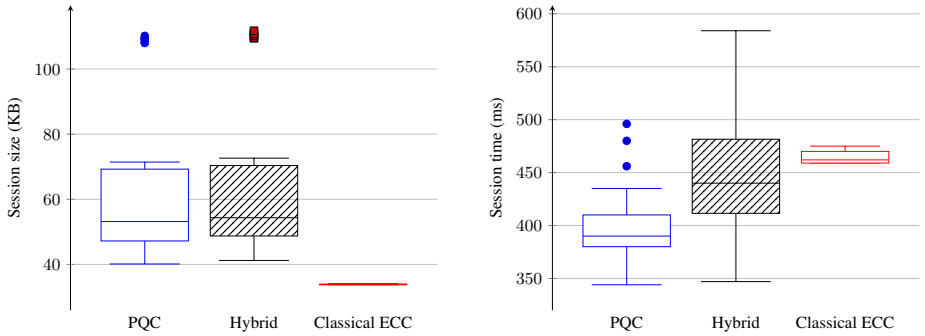


Fig. 3: Box plots of SSH session size and time for PQC, Hybrid, and Classical ECC configurations.

Hybrid and PQC are broadly comparable, showing no significant differences in median or quartile values. However, session time data reveals more variability:

- The PQC time box plot is relatively compact, suggesting that all tested PQC algorithm combinations achieve similar handshake latencies.
- The Hybrid time box plot is more dispersed, indicating that algorithm choice has a larger influence on latency in Hybrid configurations.

Outlier detection using the interquartile range (IQR) method identified *Sphincs+-192f* as a high-latency outlier. Given its deviation from the main distribution in both PQC and Hybrid cases, this parameter set is excluded from subsequent prototype candidate analysis to maintain a focus on representative, performant configurations.

5.2 SCP: Aggregate Efficiency

Following the handshake performance evaluation (Subsection 5.1), we extended the analysis to full SSH SCP file transfers to assess how key exchange and signature algorithm choices influence the established end-to-end session efficiency. As SCP operates over the TCP transport protocol, additional TCP headers and acknowledgement packets inherently generate extra traffic, increasing the overall size of file transfer sessions. This overhead was quantified by subtracting both the initial SSH session establishment size and the raw file size from the total file transfer session size.

We first examined the proportional relationship between file transfer session size and completion time to identify any irregularities or performance anomalies. Across all tested algorithms, no malfunctions or unexpected outliers were observed; all results fell within expected variance bounds, indicating stable and predictable performance for every implemented variant.

Figure 4 compares average SCP time and size across signature algorithms for payloads of 1 MB, 10 MB, 100 MB and 200 MB. At larger file sizes, the choice of cryptographic algorithm has less impact on transfer time; however, Falcon tends to be the most size-efficient, while Dilithium is marginally faster in Hybrid mode.

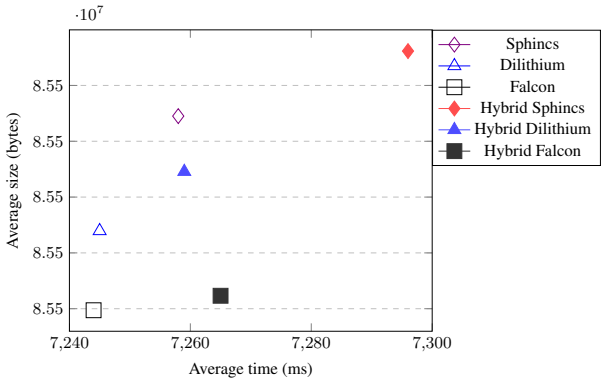


Fig. 4: Average SCP time vs. size across 1 MB, 10 MB, 100 MB and 200 MB

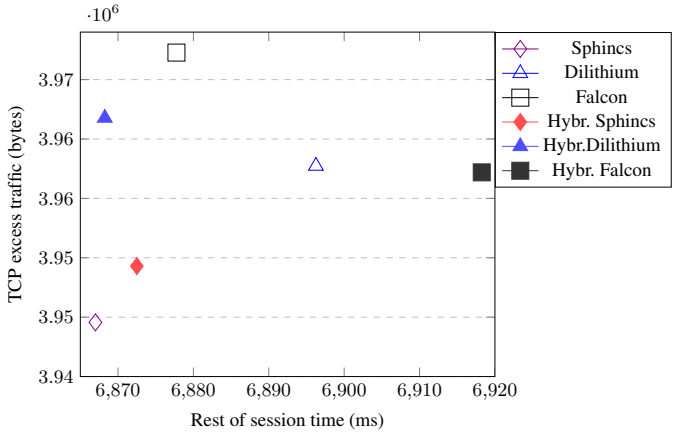


Fig. 5: Residual SCP session metrics: TCP overhead traffic (session size minus SSH session size and file size) vs. residual transfer time (excluding handshake).

Figure 5 presents the averaged residual session metrics, where the Y-axis represents TCP overhead (in bytes) and the X-axis shows the residual transfer time (excluding SSH session establishment). This residual phase isolates the sustained data transfer component, independent of handshake costs.

Across both PQC and its hybrid implementations, *Sphincs* consistently exhibits the smallest residual size and shortest residual duration. However, this apparent efficiency is misleading: *Sphincs* remains the slowest overall and produces the largest total session size due to its exceptionally large signatures.

In PQC-only mode, *Dilithium* and *Falcon* differ in transfer time by only ~ 1 ms, yet *Falcon* achieves smaller total session sizes. In Hybrid configurations, *Dilithium* is marginally faster, while *Falcon* maintains superior size efficiency.

5.3 KEM analysis: Kyber vs Kyber-90s in SSH Handshakes

Kyber has two variants of the main pseudorandom function (PRF) options and other symmetric sub-primitives:

- Standard Kyber $\rightarrow$ all symmetric sub-primitives are instantiated with Keccak-based FIPS 202 functions (XOF = SHAKE-128 (for generating the public matrix), PRF / KDF = SHAKE-256, H = SHA3-256, G = SHA3-512). This design ties Kyber tightly to the SHA-3 family.
- Kyber-90s $\rightarrow$ PRF based on AES-256/SHA-2 in counter mode (AES-CTR), designed for faster performance when hardware AES acceleration (AES-NI) is available. The XOF is replaced with AES-256 in CTR mode, the PRF is replaced with AES-256 in CTR mode keyed differently, H and G are instantiated with SHA-2 functions (SHA-256 and SHA-512) instead of SHA-3 (Avanzi et al. (2021)).

NIST excluded Kyber-90s from the ML-KEM standard (National Institute of Standards and Technology, 2022a) to simplify adoption by selecting a single, SHAKE/SHA-3 based variant, though AES-based designs may still be relevant in niche or constrained deployments.

To isolate the impact of the KEM choice from the large performance variance introduced by different signature schemes (e.g., compact Falcon vs large SPHINCS+), we computed the *average* SSH handshake session size and latency for each Kyber parameter set, aggregated across all tested PQC signatures. These results, shown in Table 11, focus solely on the contribution of the KEM implementation to overall performance and Security Levels.

Observed trends. The aggregated results confirm that **handshake session size is virtually identical** between Kyber and Kyber-90s at the same security level, size differences are minimal (<1%), meaning storage and transmission overhead between the two variants is practically the same.

Latency favors Kyber-90s at Level 1, but reverses at Levels 3 and 5. Since both variants share identical public key and ciphertext lengths, differing only in their internal pseudorandom function (SHAKE vs AES-256-CTR). Kyber-90s does not outperform Kyber at higher security levels in latency and shows measurable variation:

- **Level 1 (Kyber512):** Kyber-90s is $\approx 4.4\%$ faster.

Table 11: Average SSH handshake size and time for Kyber vs Kyber-90s across all tested signatures.

KEM Parameter Set	Average Size (bytes)		Average Time (ms)		NIST Security
	Kyber	Kyber-90s	Kyber	Kyber-90s	
Kyber512	61,792	61,765	390.9	373.6	Level 1
Kyber768	62,460	62,875	390.7	413.4	Level 3
Kyber1024	63,438	63,954	384.7	422.2	Level 5

– **Level 3 (Kyber768):** Kyber-90s is $\approx 5.8\%$ slower.
– **Level 5 (Kyber1024):** Kyber-90s incurs $\approx 9.7\%$ slowdown.

Benchmarks show mixed results — sometimes Kyber is faster at larger parameter sets, sometimes Kyber-90s wins, depending on whether SHA-2/AES acceleration is present and on implementation. There are published implementations of Kyber-90s optimized for constrained devices (ESP32) demonstrating that 90s can be very competitive at L1 on some platforms (Segatz and Hafiz (2025)). The observed Kyber-90s are faster than Kyber, which is plausibly caused by the presence of CPU instructions that accelerate AES. Kyber-90s does not outperform Kyber at higher security levels, suggesting that larger key and ciphertext sizes negate AES performance advantages — consistent with the Kyber designers' assertion (Avanzi et al. (2019)) that the AES-based variant is only preferable when AES is accelerated in hardware.

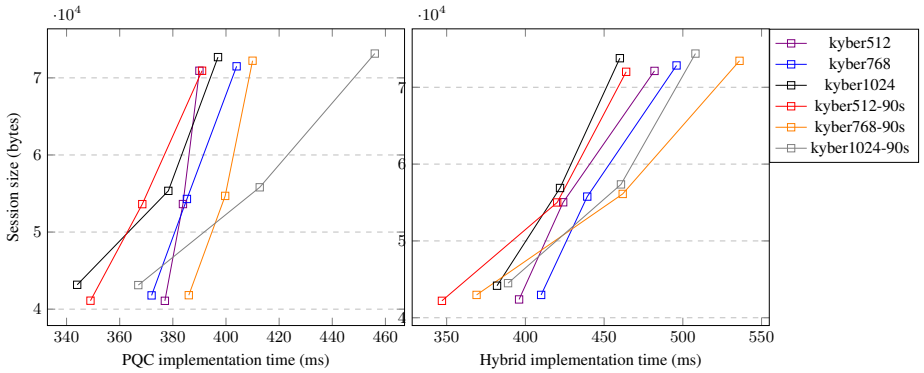


Fig. 6: Kyber parameters of SSH session time and size: PQC vs Hybrid

Figure 6 presents the minimum, average, and maximum data for both time and size, distinguishing between PQC implementations on the left and their hybrid implementations on the right of every Kyber parameter set. To sum up, Kyber's contribution to session size and time remained consistent, exhibiting linear scaling with security level and showing low variance across signature pairings. When combined with classical ECC in hybrid handshakes, Kyber added only a minimal size overhead compared to PQC mode.

5.4 Digital Signature Algorithms Analysis

To complement the key encapsulation mechanism results, we analyzed the performance characteristics of the PQC digital signature algorithms (Falcon, Dilithium, Sphincs+) used in our SSH handshake experiments. The goal was to quantify their impact on overall handshake efficiency and to identify trade-offs relevant to post-quantum migration strategies.

Table 12 summarizes the average handshake size and time for each signature algorithm, computed across all Kyber parameter sets. This averaging smooths out the effect of KEM size scaling and isolates the performance contribution of the signature component.

Observed Trends. From the aggregated results, several clear patterns emerge:

- **Handshake size is dominated by signature size.** Sphincs+ variants have the largest signatures (17 KB for -128f, 36 KB for -192f), making them significantly larger than lattice-based schemes. Falcon and Dilithium are much more practical for SSH, with moderate signature sizes (0.7– 46 KB) and handshake times (<500 ms for hybrid and <400 ms PQC).
- **Handshake time differences are smaller than size differences.** While Falcon consistently shows good size efficiency, its handshake times are not significantly faster than Dilithium, and the highest security Level5 Dilithium5-AES slightly outperforms Falcon1024 in speed (2 ms).
- **Hybrid vs PQC impact is modest.** In PQC, introducing a classical ECC component into the handshake increases average handshake size by ≈1.4–3.5% depending on algorithm, so size overhead for hybrid handshakes is relatively small. Latency impact is more pronounced than size overhead for hybrid PQC handshakes:
  - Falcon: + 15–43 ms (3.9 – 10.9 %)
  - Dilithium: + 12–91 ms (3.1 – 32.7 %)
  - Sphincs+: + 39–76 ms (9.8 – 17.2 %)
- **Sphincs+ paradox.** Although Sphincs+ shows competitive residual (post-handshake) SCP performance in Section 5.2, its handshake cost is the highest in both size and time, making it less suitable for frequent connection establishments in bandwidth-limited environments.

Figure 7 illustrates the performance data of all tested digital signature algorithms in terms of size and time. The left side displays the results from the PQC implementation and Classical ECC, while the right side highlights the results from the Hybrid implementation. From the interpretation of Figure 7, the fastest time results are located further left in the plot.

Security-Level Perspective. Falcon-512 and Dilithium2-AES offer NIST Level 1 security, while Falcon-1024 and Dilithium5-AES reach Level 5. Dilithium3 is Level 3. The higher-security variants predictably increase handshake sizes and times, but the growth is far more pronounced for Sphincs+, whose large signature sizes scale sharply with security level.

Table 12: Average SSH handshake size and time for post-quantum signature algorithms (averaged over all Kyber variants).

Signature Algorithm	**Average Size (bytes)**		**Average Time (ms)**	
	PQC	Hybrid	PQC	Hybrid
Falcon-512	42,009	43,201	384	399
Falcon-1024	49,169	50,864	395	438
Dilithium2-AES	48,438	49,838	387	399
Dilithium3	54,590	56,002	370	491
Dilithium5-AES	61,284	63,352	393	461.5
Sphincs+-128f	71,903	73,009	400	439
Sphincs+-192f	111,604	113,146	443	519

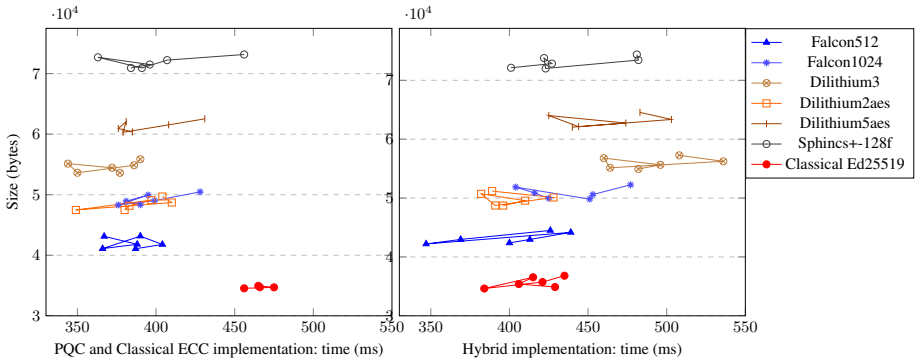


Fig. 7: Digital signature algorithms parameters of SSH session time and size: PQC vs Classical ECC vs Hybrid

Practical Takeaways. From an operational deployment perspective:

- *Falcon* is the best candidate for minimizing handshake size, which is valuable for low-bandwidth or high-latency links.
- *Dilithium* variants offer competitive speed and simpler implementation properties, making them suitable when size is less critical.
- *Sphincs+* should be reserved for scenarios prioritizing conservative, hash-based security over performance, given its handshake overhead.

These results highlight that, in the context of SSH, the choice of signature algorithm can have a larger impact on handshake size than the choice between Kyber variants for KEM, emphasizing the need to consider signature size in PQC migration strategies.

5.5 NIST Security Levels Relevance to PQC Implementation

National Institute of Standards and Technology (2016) defined *security levels* for PQC algorithms as part of its standardization process. Secure communication protocols such as SSH typically operate at the equivalent of Level 1. For example, RSA-2048 or ECDSA P-256 aligns with AES-128 security and is sufficient for most enterprise and short-to-medium data retention uses (IETF PQC Working Group, 2024). Levels 3–5, by contrast, offer higher security margins suitable for long-term protection of sensitive or archival data.

This study evaluated the performance of PQC algorithms and provides recommendations for PQC migration covering security levels:

- **Kyber512 combination with Falcon512 or Dilithium2-AES:** Level 1 — fastest performance and smallest sizes; adequate for most enterprise SSH deployments with typical data retention requirements.
- **Kyber768 combination with Dilithium3:** Level 3 — stronger security at a moderate performance cost.
- **Kyber1024 combination with Dilithium5-AES or Falcon1024:** Level 5 — maximum standardized security strength; introduces significant size overhead with only marginal handshake time impact.

From an operational perspective, selecting the security level for SSH should balance performance impact against the required security margin. For short-lived SSH sessions, lower levels can maximize throughput, whereas archival or compliance-driven environments may justify the higher costs of Level 3 or above.

6 Conclusions and Future Work

This study conducted a comprehensive empirical evaluation of post-quantum cryptographic algorithms in their pure form and in hybrid combinations with classical Elliptic Curve Cryptography within SSH protocol handshakes and file transfer sessions. By systematically measuring handshake latency, session size, and sustained transfer performance across multiple NIST security levels, we established a **dataset linking cryptographic parameter choices** to real-world operational outcomes.

These baselines and PQC variants allow us to directly observe how key length, ciphertext size, and signature size translate into measurable SSH handshake performance. By aligning each algorithm with its NIST security level, it is possible to quantify how the required security margin affects real-world SSH performance. This provides an evidence-based **justification for algorithm selection in post-quantum migration strategies**, ensuring that the security gain is achieved by selecting the most efficient combination of algorithms.

The results show that the PQC digital signature algorithm signature size is relevant to SSH session size; a correlation is observed between SSH sessions and signature size. No significant correlation was observed between the PQC algorithm's key sizes and SSH session time performance:

- **Falcon** consistently achieves the smallest handshake sizes and remains competitive in latency, making it suitable for bandwidth-sensitive environments.
- **Dilithium** offers stable performance and strong standardization support, with marginal speed advantages in Hybrid configurations.
- **Sphincs+**, while providing conservative hash-based security, imposes significant size and latency costs and is best reserved for scenarios prioritizing long-term cryptographic resilience over efficiency.
- **Kyber** exhibits consistently low variance in handshake times across NIST levels, making it predictable and practical for deployment planning.
- Hybrid modes preserved classical trust anchors at only modest performance cost, making them practical for transitional deployments.

From an operational perspective, the choice of *security level* and *signature scheme* has a more substantial performance impact. Table 13 summarizes recommended PQC algorithm pairings for SSH migration. Overall, our analysis suggests that Falcon paired

Table 13: Recommended PQC algorithm choices for SSH migration

Scenario	**Recommendation**	**Rationale**
Bandwidth-limited	Falcon-512 + Kyber512	Level 1 security; Smallest handshake
Long-term security	Falcon-1024 + Kyber1024	Level 5; Balanced speed and size
Balanced migration	Dilithium3 + Kyber1024	Level 3 & 5; Best speed / size trade-off
High-trust Hybrid	ecdh-nistp512-Kyber1024 wtih Ed25519	Level 5 PQC + Classical trust anchor
Conservative assurance	Sphincs+-128f/192f + Kyber768	Hash-based; Highest performance cost

with Kyber provides favorable trade-offs between size and performance. However, the adoption of Falcon requires careful consideration of side-channel resistance and implementation robustness, given its reliance on Gaussian sampling. Importantly, the experimental results indicate that **the most effective combination is Dilithium3 with Kyber1024**. While this pairing does not achieve the smallest session size, it delivers the shortest session time and ensures a balanced security profile by combining NIST levels 3 and 5.

Future research should extend these experiments beyond controlled laboratory conditions to more realistic settings. Evaluating PQC-enabled SSH in wide-area and high-latency networks would clarify how packet loss and congestion influence handshake and session overheads. Multi-client scalability and server-side load analysis are also critical for assessing performance in enterprise and cloud deployments. Finally, future work should examine the integration of PQC into public key infrastructures, including hybrid certificates, and assess compliance implications under regulatory frameworks such as the European Union NIS2 and the Cyber Resilience Act.

References

Alnahawi, N., Wiesmaier, A., Grasmeyer, T., Geißler, J., Zeier, A., Bauspieß, P., Heinemann, A. (2021). On the state of post-quantum cryptography migration, *INFORMATIK 2021*, Gesellschaft für Informatik, Bonn, pp. 907–941.

Avanzi, R., Bos, J., Ducas, L., Kiltz, E., Stehlé, D., Schwabe, P. et al. (2021). CRYSTALS-Kyber algorithm specification (round 3), *Technical report*, PQCRystals. `https://pq-crystals.org/kyber/data/kyber-specification-round3-20210131.pdf`.

Avanzi, R. et al. (2019). Crystals-kyber round 2 specification, `https://pq-crystals.org/kyber/data/kyber-specification-round2.pdf`. Accessed: 2025-08-09.

Barker, E., Dang, Q., Roginsky, A. (2015). Recommendation for key management: Application-specific key management guidance, *NIST Special Publication 800-57 Part 3 Revision 1*, National Institute of Standards and Technology.
`https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-57pt3r1.pdf`

Barker, W., Polk, W., Souppaya, M. (2021). Getting ready for post-quantum cryptography: Exploring challenges associated with adopting and using post-quantum cryptographic algorithms. Last accessed: 2024-10-13, `https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04282021.pdf`.

Bernstein, D. J., Duif, N., Lange, T., Schwabe, P., Yang, B.-Y. (2012). High-speed high-security signatures, *Cryptographic Hardware and Embedded Systems – CHES 2011*, Vol. 6917 of *Lecture Notes in Computer Science*, Springer, pp. 124–142.
`https://ed25519.cr.yp.to/ed25519-20110926.pdf`

Bindel, N., Brendel, J., Fischlin, M., Goncalves, B., Stebila, D. (2019). Hybrid key encapsulation mechanisms and authenticated key exchange, *in* Ding, J., Steinwandt, R. (eds), *Proc. 10th International Conference on Post-Quantum Cryptography (PQCrypto) 2019*, Vol. 11505 of *LNCS*, Springer, pp. 202–226.

Crockett, E., Paquin, C., Stebila, D. (2019). Prototyping post-quantum and hybrid key exchange and authentication in tls and ssh, *Cryptology ePrint Archive* . Access online: `https://ia.cr/2019/858`.

IETF PQC Working Group (2024). Post-quantum cryptography for engineers, IETF Draft (draft-ietf-pquip-pqc-engineers-06), Section 11. Accessed: 2025-08-10.

IETF SSH Working Group (2009). Elliptic curve algorithm integration in the secure shell (ssh) transport layer, RFC 5656.
`https://datatracker.ietf.org/doc/html/rfc5656`

Jacob W. S. Schneider, P. S. (2024). NIST Releases Three Post-Quantum Cryptography Standards | Insights | Holland & Knight — hklaw.com, `https://www.hklaw.com/en/insights/publications/2024/08/nist-releases-three-post-quantum-cryptography-standards`. [Accessed 13-10-2024].

Joseph, D., Misoczki, R., Manzano, M. (2024). Transitioning organizations to post-quantum cryptography, *Nature* **605**, 237–243.

Kampanakis, P., Stebila, D., Friedl, M., Hansen, T., Sikeridis, D. (2020). Post-quantum public key algorithms for the secure shell (ssh) protocol, IETF Internet-Draft, draft-kampanakis-curdle-pq-ssh-00. Available at IETF Datatracker.
https://datatracker.ietf.org/doc/html/draft-kampanakis-curdle-pq-ssh-00

Kreutzer, M., Niederhagen, R., Waidner, M., Gespräch, E. (2018). Next generation crypto. Last accessed: 2024-04-14, https://www.sit.fraunhofer.de/fileadmin/dokumente/studien_und_technical_reports/EberbacherBroschuere_prefinal_V10.pdf?_=1520946028.

Mosca, M. (2018). Cybersecurity in an era with quantum computers: Will we be ready?, *IEEE Security Privacy* **16**(5), 38–41.

National Institute of Standards and Technology (2016). Security (evaluation criteria), https://csrc.nist.gov/projects/post-quantum-cryptography/post-quantum-cryptography-standardization/evaluation-criteria/security-(evaluation-criteria). Accessed: 2025-08-10.

National Institute of Standards and Technology (2022a). Official comment on the selection of crystals-kyber as the nist post-quantum cryptography standard, https://csrc.nist.gov/csrc/media/Projects/post-quantum-cryptography/documents/selected-algos-2022/official-comments/crystals-kyber-selected-algo-official-comment.pdf. Accessed: 2025-08-09.

National Institute of Standards and Technology (2022b). Official comment on the selection of falcon as a nist post-quantum cryptography candidate, NIST PQC Standardization Project. Accessed: 2025-09-07.
https://csrc.nist.gov/csrc/media/Projects/post-quantum-cryptography/documents/selected-algos-2022/official-comments/falcon-selected-algo-official-comment.pdf

National Institute of Standards and Technology (2023). Fips pub 186-5: Digital signature standard (dss), https://doi.org/10.6028/NIST.FIPS.186-5. Accessed: 2025-08-09.

National Institute of Standards and Technology (2024a). Announcing issuance of federal information processing standards (fips) 203, module-lattice-based key-encapsulation mechanism standard; fips 204, module-lattice-based digital signature standard; and fips 205, stateless hash-based digital signature standard, *Federal Register* **89**(157), 66052–66057.

National Institute of Standards and Technology (2024b). FIPS 203: ML-KEM – module lattice-based key encapsulation mechanism, *Federal Information Processing Standards Publication 203*, U.S. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD.
https://doi.org/10.6028/NIST.FIPS.203

National Institute of Standards and Technology (2024c). FIPS 204: ML-DSA – module lattice-based digital signature algorithm, *Federal Information Processing Standards Publication 204*, U.S. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD.
https://doi.org/10.6028/NIST.FIPS.204

National Institute of Standards and Technology (2024d). FIPS 205: SLH-DSA – stateless hash-based digital signature algorithm, *Federal Information Processing Standards Publication 205*, U.S. Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD.
https://doi.org/10.6028/NIST.FIPS.205

National Institute of Standards and Technology (2024e). Post-quantum cryptography: Publications, https://csrc.nist.gov/Projects/post-quantum-cryptography/publications. Accessed July 26, 2025.

Nina, B., Johannes, B., Luca, G., Tobias, S., Johannes, W. (2019). X.509-compliant hybrid certificates for the post-quantum transition, *Journal of Open Source Software* **4**(40), 1606. Last accessed: 2024-04-14, https://doi.org/10.21105/joss.01606.

Ott, D., Peikert, C., participants, o. w. (2019). Identifying research challenges in post quantum cryptography migration and cryptographic agility. Last accessed: 2024-04-23, https://arxiv.org/abs/1909.07353.
https://arxiv.org/abs/1909.07353

Ounsworth, M., Pala, M. (2024). Composite signatures for use in internet pki, ietf, 8 june 2024, during conference PQCrypto2021. Last accessed: 2024-04-14, https://www.ietf.org/id/draft-ounsworth-pq-composite-sigs-07.html.

Paquin, C., Stebila, D., Tamvada, G. (2019). Benchmarking post-quantum cryptography in tls, *Technical Report 2019/1447*, IACR Cryptology ePrint Archive.
https://eprint.iacr.org/2019/1447

Pierre-Alain, F., Jeffrey, H., Paul, K., Vadim, L., Thomas, P., Thomas, P., Thomas, R., S., W. G., Whyte, Z. Z. (2024). Falcon: Fast-fourier lattice-based compact signatures over ntru, https://www.di.ens.fr/~prest/Publications/falcon.pdf. [Accessed 12-09-2025].

Prest, T., Ducas, L., Lepoint, T., Lyubashevsky, V., Schwabe, P., Stehlé, D., Zaverucha, G. (2020). Falcon: Fast-fourier lattice-based compact signatures over ntru, https://falcon-sign.info. Submission to NIST PQC Standardization Project, Round 3.

Rossi, M. (2021). Pqc transition - anssi views, during conference PQCrypto2021. Last accessed: 2024-04-14, http://pqcrypto2021.kr/download/program/PQC_transition_in_France.pdf.

Schwabe, P., Stebila, D., Wiggers, T. (2020). Post-quantum tls without handshake signatures, *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS '20)*, ACM, pp. 1234–1243.

Segatz, F., Hafiz, M. I. A. (2025). Efficient implementation of crystals-kyber key encapsulation mechanism on esp32, *arXiv preprint arXiv:2503.10207v1* . https://arxiv.org/abs/2503.10207v1.

Sikeridis, D., Kampanakis, P., Devetsikiotis, M. (2020a). Assessing the overhead of post-quantum cryptography in tls 1.3 and ssh, *Proceedings of the 16th International Conference on Emerging Networking Experiments and Technologies*, CoNEXT '20, Association for Computing Machinery, New York, NY, USA, p. 149–156.
https://doi.org/10.1145/3386367.3431305

Sikeridis, D., Kampanakis, P., Devetsikiotis, M. (2020b). Assessing the overhead of post-quantum cryptography in tls 1.3 and ssh, *Proceedings of CoNEXT '20*, Barcelona, Spain, pp. 1–7.

Stebila, D., Fluhrer, S., Gueron, S. (2021). Hybrid key exchange in tls 1.3, IETF Internet-Draft, draft-ietf-tls-hybrid-design-03. Last accessed: 2025-09-07.
https://datatracker.ietf.org/doc/html/draft-ietf-tls-hybrid-design-03

Tran, D. D., Ogata, K., Escobar, S., Akleylek, S., Otmani, A. (2024). Formal analysis of post-quantum hybrid key exchange ssh transport layer protocol, *IEEE Access* **12**, 1672–1687.

Received February 13, 2025 , revised September 28, 2025, accepted November 5, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, pp. 885–893
https://doi.org/10.22364/bjmc.2025.13.4.07

Predicting Student Performance on a Novel Moodle Dataset Using GRU Time Series Model

Martins SNEIDERS[1], Evalds URTANS[2], Amjed ABU SAA[3]

[1] Riga Technical University, Riga, Latvia
[2] Ventspils University of Applied Sciences, Ventspils, Latvia
[3] Ajman University, UAE

martins.sneiders@gmail.com, evalds.urtans@venta.lv, amjed.abusaa@ku.ac.ae

ORCiD 0009-0006-7961-0822, ORCID 0000-0001-9813-0548, ORCID 0009-0009-4765-6107

Abstract. The prediction of student grades and behavior on online learning platforms is essential for enhancing teaching and learning outcomes. Traditional methods frequently neglect the temporal dynamics inherent in student activity. This study presents a novel Moodle dataset, available on Kaggle, comprising over one million student activity logs. It aims to develop a machine learning time series model designed to predict student grades and behavior, specifically within the context of small universities characterized by limited student populations and course sizes. Moreover, the research seeks to construct a highly accurate predictive model, achieving an accuracy of 83% for semester-long courses. The investigation systematically evaluates the impact of various factors—including course type, passing grade, predictive features, time-splitting schemes, supervised learning algorithms, and performance metrics—on prediction accuracy. The results indicate that GRU (gated recurrent units) models are particularly effective in forecasting student performance in courses that feature active and continuous student engagement. These findings may assist higher education institutions in identifying student performance determinants and establishing an early warning system for potential academic challenges.

Keywords: machine learning, time series, gru, moodle

1 Introduction

Recent advances in machine learning have opened new avenues for analyzing educational data, with time series models emerging as a promising approach for understanding student behavior over time. This study explores the application of a machine learning time series model, specifically one based on gated recurrent units (GRU) to predict student performance using data from Moodle, one of the most popular open-source learning management systems (LMS) worldwide.

A key motivation for this work is the challenge posed by traditional prediction methods, which often fail to capture the dynamic and sequential nature of student interactions. In educational settings, especially in small universities where class sizes are limited, data tend to accumulate slowly, and the temporal patterns of student activity can be critical for early intervention. Moreover, while many institutions have started to implement machine learning models within their LMS, the reliance on black-box approaches raises concerns about prediction reliability and fairness—issues that must be addressed to support sound pedagogical decision-making.

Moodle provides a rich environment for this investigation. Its widespread use and integrated analytics functionalities make it an ideal platform for developing and testing predictive models. Although recent Moodle releases (from version 3.4 onward) support in-platform ML (Machine Learning) model creation, few studies rigorously evaluate these approaches as shown in Tagharobi and Simbeck (2022) and Bognár, Fauszt and Váraljai (2021). This paper not only examines the effectiveness of a GRU-based time series model in forecasting student performance but also discusses the broader implications of integrating machine learning within educational systems. Consistent with prior findings in Daugule et al. (2022), slowly accumulating semester-level data in small cohorts can limit model reliability and timeliness.

Moodle is one of the most popular open-source Learning Management Systems (LMS) in the world with millions of users. Although since the release of Moodle 3.4, it is possible to create ML models within the LMS system, very few studies have been published. Using these models as black boxes poses serious risks of getting unreliable predictions and false alarms as shown in Bognár, Fauszt and Nagy (2021).

Online learning platforms implement ML in their learning analytics functionalities, offering educators predictions on student progress and allowing early interventions. However, these predictive systems are prone to the same fairness problems. If educators use the prediction results in their approach or grading decisions, this can have a major impact on student success. Therefore, educational institutions must assess the fairness of learning analytics before implementing it like shown in Tagharobi and Simbeck (2022).

This work contributes to the growing body of research on educational data analytics by demonstrating how time series models can leverage the rich, sequential data available in LMS platforms like Moodle to provide early warnings and support interventions that enhance student success.

2 Related work

A range of classical machine learning studies has been conducted using engineered features for study course performance prediction. Research utilizing interaction logs, attendance records, quiz submissions, and demographic data demonstrates that when features are meticulously chosen and validated, tree-based models and linear classifiers prove effective in previous work done by Evangelista (2021), Bhutto et al. (2020), Duch et al. (2024). These studies highlight the importance of data quality, the significance of feature importance, and the advantages offered by multi-source attributes. Across institutional datasets, Random Forest often emerges as a strong baseline with

key predictors spanning demographics, prior performance, course/instructor attributes, and general factors as shown in Saa et al. (2019). Early-risk benchmarking at 6–60% of course progress reports Random Forest as strongest among algorithms while emphasizing the trade-off between timeliness and accuracy when combining institutional and Moodle logs achieving 84.32% at 20% of course progress and 91.78% at 60% of course progress as listed in Tamada et al. (2021). Also, analysis of Moodle activity confirms a strong link between online engagement patterns and academic achievement, guiding compact and informative feature sets as shown in Shrestha and Pokharel (2021).

Models designed for sequential analysis of temporal behavior have been studied extensively. To effectively capture the sequence and timing of events, RNN-based frameworks such as GRU/LSTM, along with hybrid models like CNN-RNN, have been investigated. These often outperform non-sequential baselines in works like He et al. (2020), Baniata et al. (2024), Aljaloud et al. (2022), Abbaspour et al. (2020), Yin et al. (2023). These approaches can leverage detailed temporal patterns but must be applied with attention to issues of sparsity and the length of the sequences.

Alternative models focus on aspects such as fairness, dependability, and data sparsity as demonstrated in Bognár, Fauszt and Nagy (2021). For instance, a study investigates the critical factors for generating reliable predictions within Moodle, accentuating elements like the size of the predictor matrix, temporal divisions, and evaluation metrics, alongside an examination of the bias–variance dilemma. The issue of data sparsity is particularly pronounced in smaller institutions. It is demonstrated that predictability is significantly influenced by course structure and assessment designs, with self-assessments and quizzes acting as more robust indicators, as evidenced in Kaensar and Wongnin (2023). At the course level, analyses of Moodle indicators—cognitive depth and social breadth—reveal that file/URL interactions constitute weak signals, while self-assessment tests enhance prediction accuracy, and the calculation of quiz indicators proves challenging, as illustrated in Fauszt et al. (2021).

In contrast to previous research, this study presents two main contributions: the release of a new Moodle dataset with over one million logs tailored for small universities, and the benchmarking of a GRU-based temporal model across fine-grained and grouped-grade configurations to tackle scarce long-tail labels. Citations are consistently formatted in author-year style without article titles.

3 Methodology

The main contribution of this research is the Moodle dataset that was obtained from a University of Latvia, enabling the creation of a novel dataset as described in this paper. In accordance with the confidentiality agreement, the university's Moodle data—installed on the institution's server—was provided to the author. Due to institutional constraints and data availability, obtaining additional semesters of data for this analysis was impossible. As a result, the models developed here are based solely on activity and assessment data from one semester. Despite this limitation, the current dataset offers valuable insights into student engagement and performance within the observed semester, and the published data can serve as a benchmark for future research.

After a review of Moodle documentation and publications, the following database tables were selected for use:

1. `mdl_logstore_standard_log` – contains records of various user activities, such as system logins, chapter accesses, answer submissions, and the posting of assessments to students.
2. `mdl_grade_grades` – stores data related to student grades.
3. `mdl_grade_items` – establishes the link between grades and corresponding course information.

Subsequent to data processing, the dataset was published as CSV files:

1. Comprehensive grade data is stored in the file `udk_moodle_all_grades.csv`.
2. Log data is stored in the file `udk_moodle_log.csv`[4].
3. These files are publicly available in the repository, with associated statistics presented in Table 1.

Table 1: Overview of dataset

Metric	udk_moodle_all_grades.csv	udk_moodle_log.csv
Samples	20317	1259411
Columns	id, timemodified, userid, courseid, finalgrade, itemtype	id, timecreated, eventname, action, target, userid, courseid, other
Labels	‘itemtype‘: ("category", "course", "manual", "mod"), ‘userid‘, ‘courseid‘	‘eventname‘, ‘action‘, ‘target‘, ‘userid‘, ‘courseid‘

Analyzing the distribution of grades shown in Table 2, it can be estimated that the majority of students received a "pass" grade, but there are 287 failed grades. A grade of "-1" means that the student did not receive a grade in this subject, and there are 761 such entries.

Table 2: Distribution of grades in dataset

Grade	-1	0	1	2	3	4	5	6	7	8	9	10
Number of samples	761	154	53	23	56	97	162	281	662	1500	2098	1143

An analysis of the grade types presented in Table 3 reveals that the grades are grouped by type. Specifically, the "course" type represents the final grade, while the other types correspond to various assessments administered during the semester. Notably, most of these assessments align with the "mod" type, which indicates that they are generated by different Moodle plugins. To evaluate the prevalence of semester grades,

[4] https://www.kaggle.com/datasets/martinssneiders/moodle-grades-and-action-logs

the column "CWG" (courses with grades) was incorporated into the analysis. The results indicate that the majority of courses are conducted without semester assessments in Moodle; rather, only final grades are recorded, with merely approximately 3.5% of the courses incorporating grading during the semester.

Table 3: Distribution of dataset mark types

	category	**course**	**manual**	**mod**	**CWG**
graded	371	6229	118	7078	216
with ungraded	562	6990	151	12614	255

An analysis of the log file data reveals that the majority of entries were generated by less than 5% of the users and were concentrated in only a few courses. This observation indicates that only a limited number of courses are actively utilizing the Moodle system, possibly those related to information technology; however, this conclusion cannot be definitively verified given the constraints imposed by the anonymized data.

The model employs the Gated Recurrent Unit (GRU) architecture to facilitate efficient time-series modeling while simultaneously conserving memory resources, as shown in Gao and Glowacka (2016).

The hyperparameters of the model include a hidden dimension size of 512, a learning rate of 1e-3, a batch size of 16, and a single-layer GRU model. The model is designed to process Moodle logs by feeding sequential activity data represented as features, followed by a final fully connected layer used for classification into 11-grade categories, with input data being padded for variable length sequences. The training involves the Adam optimizer with a loss function that incorporates class weights to address the class imbalance, which is defined as an alpha coefficient scaled by the inverse frequency of each label in the dataset. Three aggregate representations are computed for each sequence—namely, the element-wise maximum, mean, and the final GRU output—whose concatenation forms a 1536-dimensional vector that is fed into a fully connected layer mapping to 11 output classes, after which a softmax activation is applied to yield a probability distribution over discrete grade categories.

4 Results

An analysis of the dataset facilitated the accurate prediction of student grades, although certain limitations were identified. Specifically, these predictions were achievable solely for a subset of the accessible data. Further scrutiny disclosed that the Moodle platform is actively utilized by fewer than 5% of courses; thus, reliable grade predictions can only be realized for courses that employ continuous grading throughout the semester. This constraint introduces selection bias and restricts the external validity of our findings to Moodle-enabled courses with continuous assessment. Furthermore, the grade distribution is imbalanced, with several grade values being sparsely represented, as observed in Figure 1 and Figure 2. The test dataset was created using a fixed random seed and 20% random sampling without overlapping time sequences.

In the initial iteration of the model, interim assessments conducted during the semester were excluded based on the presumption that they were infrequently utilized by teaching staff, leading to a model founded solely on student activities. However, this approach proved inadequate. Consequently, intermediate evaluation outcomes were subsequently incorporated, and the model was trained exclusively on courses where such evaluations occur. Although this enhances internal consistency, it further restricts the target population to courses with regular intermediate evaluations. Final scores were not analyzed independently but were computed utilizing the highest, lowest, and average values. Preliminary tests revealed that analyses predicated on the highest grade did not yield significant insights. Consequently, emphasis was redirected towards the lowest grade, thereby enabling the prediction of failing students. This decision is supported by the available data, which indicates a predominance of high grades (beginning from 8) in conjunction with a sufficient number of low grades. Accordingly, only those students and courses for which assessments were conducted regularly throughout the instructional process were included in the forecast. Consequently, courses in which students merely accessed reading materials and instructors solely recorded final results were excluded. The forecast input data could then be configured to include or exclude intermediate results as necessary.

As illustrated by a histogram Figure 1, the model demonstrated minor deviations between the predicted and actual marks, culminating in an accuracy of 83%. Nevertheless, this aggregate accuracy warrants cautious interpretation due to class imbalance and low counts per grade. Metrics specific to each grade level, along with ordinal error, offer a more robust evaluation of the model's goodness-of-fit compared to accuracy alone in this context.

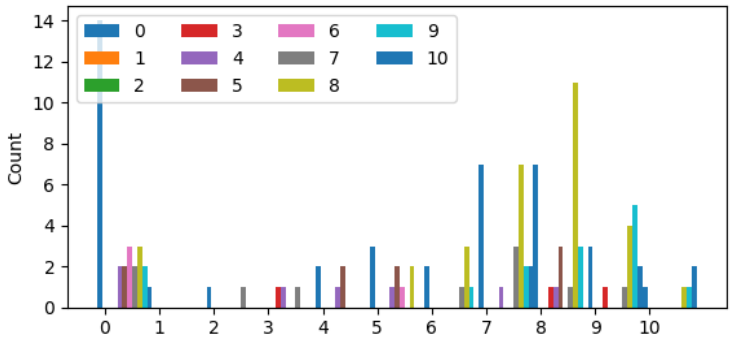


Fig. 1: Histogram of predicted grades indicated by colors and ground truth grades indicated by the X-axis

Additionally, experiments were conducted using the same dataset while excluding grade information and relying solely on student activities. These experiments achieved an accuracy of 80% in predicting final student outcomes, as demonstrated by a hitogram in Figure 2. This result suggests that although intermediate assessment data improves

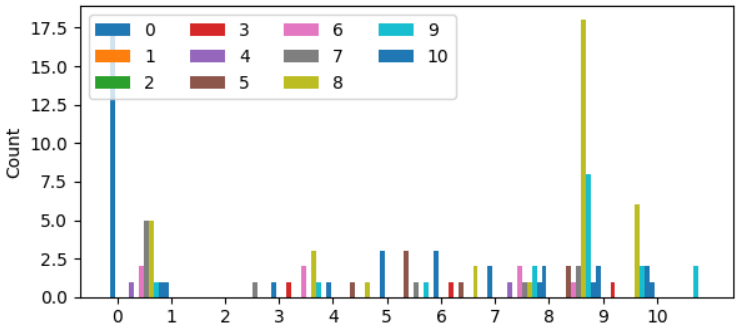


Fig. 2: Histogram of predicted grades indicated by colors and ground truth grades indicated by the X-axis excluding intermediate assessment data

predictive accuracy, a reasonably reliable prediction can still be obtained based solely on user activity data.

5 Further research

In light of the findings of this study, several avenues for future work are evident. One promising area of exploration is the investigation of alternative machine learning models, specifically diverse implementations of Transformer models. Beyond treating grades as an 11-class softmax problem, future work should reformulate the task using ordinal methods—such as cumulative link models or CORN—or as direct regression on $[0, 10]$ with post-hoc rounding. Given skewed and sparsely populated categories, it is also recommended to construct the model with aggregated grade groups (e.g., 0; 1–2; 3–4; 5–6; 7–8; 9–10) and compare performance against the ordinal/regression formulations.

Given that the Moodle platform is presently deployed in less than 5% of courses, it is imperative for future research to focus on enhancing its integration within instructional workflows and evaluating how enriched platform signals impact the modeling capacity and accuracy. To fortify external validity and assess domain shift, it is recommended to broaden the dataset to encompass additional semesters and at least one supplementary institution, utilizing methodologies such as train-on-A/test-on-B and train-on-previous-semester/test-on-next-semester. Privacy-preserving collaboration among universities should be investigated through federated learning to prevent raw data exposure while potentially enhancing generalization.

Evaluation should extend beyond accuracy to include error, agreement, and calibration: report mean absolute error (MAE) of grade points, Quadratic Weighted Kappa, and calibration metrics such as the Brier score and expected calibration error (ECE). To address class imbalance, compare cost-sensitive objectives, focal loss, and class-balanced sampling. Finally, to mitigate sparsity at the course level, investigate transfer and meta learning on “dense” courses and fine-tuning on “sparse” ones—as well as semi-supervised approaches to leverage unlabeled or partially labeled data.

6 Conclusions

This study demonstrates that a compact GRU-based time–series model can forecast student outcomes on a real Moodle deployment from a small university with competitive accuracy, provided that courses exhibit regular, logged engagement and intermediate evaluations. Using a new public dataset comprising 1,259,411 interaction logs and 20,317 grade records. In terms of coverage and adoption, only a small fraction of courses systematically use Moodle for continuous assessment, with 6,990 course-level (final) grade records observed, of which 6,229 are graded and 761 are ungraded. Courses with regular semester assessments account for approximately 3.5% relative to the recorded course finals, and usage is highly skewed with fewer than 5% of users and a handful of courses generating the majority of logs. Regarding the outcome distribution and imbalance, among the 6,229 graded course finals, high marks in the range of 8–10 make up 76.1%, while low marks between 0–3 are rare at 4.6%. Depending on the pass threshold set by institutions, the proportion of fails varies, highlighting the relevance of the long-tailed label distribution for both class weighting and ordinal-aware evaluation. In terms of predictive performance and error shape, including intermediate assessment signals enables the GRU model to achieve 83% accuracy on the test set, with activity logs alone leading to 80% accuracy. errors tend to be small and ordinally local, with a high proportion of predictions falling within one grade point of the true grade, though very low grades suffer from noisier estimates due to limited test samples, which emphasizes the necessity for imbalance-aware training and reporting. Despite sparse adoption and skewed outcomes paired with careful scope limitations, calibration, and ordinal-aware evaluation, such models can underpin practical, low-latency early-warning systems for universities.

References

Abbaspour, S., Fotouhi, F., Sedaghatbaf, A., Fotouhi, H., Vahabi, M., Lindén, M. (2020). A comparative analysis of hybrid deep learning models for human activity recognition, *Sensors (Basel, Switzerland)* **20**.
https://api.semanticscholar.org/CorpusID:222255551

Aljaloud, A. S., Uliyan, D. M., Alkhalil, A., Elrhman, M. A., Alogali, A. F. M., Altameemi, Y. M., Altamimi., M., Kwan, P. (2022). A deep learning model to predict student learning outcomes in lms using cnn and lstm, *IEEE Access* **10**, 85255–85265.
https://api.semanticscholar.org/CorpusID:251378155

Baniata, L. H., Kang, S., Alsharaiah, M., Baniata, M. H. (2024). Advanced deep learning model for predicting the academic performances of students in educational institutions, *Applied Sciences* .
https://api.semanticscholar.org/CorpusID:268136123

Bhutto, E. S., Siddiqui, I. F., Arain, Q. A., Anwar, M. (2020). Predicting students' academic performance through supervised machine learning, *2020 International Conference on Information Science and Communication Technology (ICISCT)* pp. 1–6.
https://api.semanticscholar.org/CorpusID:218468275

Bognár, L., Fauszt, T., Nagy, G. (2021). Analysis of conditions for reliable predictions by moodle machine learning models, *Int. J. Emerg. Technol. Learn.* **16**.

Bognár, L., Fauszt, T., Váraljai, M. (2021). The impact of online quizzes on student success, *Int. J. Emerg. Technol. Learn.* **16**.

Daugule, I., Kapenieks, A., Timsans, Z. (2022). Use of knowledge acquisition surface to monitor and assess students' success, *Int. J. Emerg. Technol. Learn.* **17**, 109–125.
https://api.semanticscholar.org/CorpusID:251124300

Duch, D., May, M., George, S. (2024). Enhancing predictive analytics for students' performance in moodle: Insight from an empirical study, *Journal of Data Science and Intelligent Systems* .
https://api.semanticscholar.org/CorpusID:272935088

Evangelista, E. D. L. (2021). A hybrid machine learning framework for predicting students' performance in virtual learning environment, *Int. J. Emerg. Technol. Learn.* **16**.

Fauszt, T., Bognár, L., Sándor, Á. (2021). Increasing the prediction power of moodle machine learning models with self-defined indicators, *Int. J. Emerg. Technol. Learn.* **16**.

Gao, Y., Glowacka, D. (2016). Deep gate recurrent neural network, *Proceedings of the Asian Conference on Machine Learning (ACML)*, Vol. 63, pp. 1–16.

He, Y., Chen, R., Li, X., Hao, C., Liu, S., Zhang, G., Jiang, B. (2020). Online at-risk student identification using rnn-gru joint neural networks, *Inf.* **11**, 474.
https://api.semanticscholar.org/CorpusID:225166054

Kaensar, C., Wongnin, W. (2023). Analysis and prediction of student performance based on moodle log data using machine learning techniques, *Int. J. Emerg. Technol. Learn.* **18**, 184–203.
https://api.semanticscholar.org/CorpusID:258881108

Saa, A. A., Al-Emran, M., Shaalan, K. F. (2019). Mining student information system records to predict students' academic performance, *International Conference on Advanced Machine Learning Technologies and Applications*.

Shrestha, S., Pokharel, M. (2021). Educational data mining in moodle data, *International Journal of Informatics and Communication Technology* **10**, 9–18.

Tagharobi, H., Simbeck, K. (2022). Introducing a framework for code based fairness audits of learning analytics systems on the example of moodle learning analytics, *International Conference on Computer Supported Education*.

Tamada, M. M., Giusti, R., de Magalhães Netto, J. F. (2021). Predicting student performance based on logs in moodle lms, *2021 IEEE Frontiers in Education Conference (FIE)* pp. 1–8.
https://api.semanticscholar.org/CorpusID:245422775

Yin, C., Tang, D., Zhang, F., Tang, Q., Feng, Y., He, Z. (2023). Students learning performance prediction based on feature extraction algorithm and attention-based bidirectional gated recurrent unit network, *PLOS ONE* **18**.
https://api.semanticscholar.org/CorpusID:264487735

Received September 7, 2023 , revised September 15, 2025, accepted November 13, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, pp. 894–918
https://doi.org/10.22364/bjmc.2025.13.4.08

AnROM: A Methodology and Comparative Study of Custom Android Systems

Naser AlDuaij

Department of Computer Science, Kuwait University, Kuwait

naser.alduaij@ku.edu.kw

ORCID 0000-0002-0028-5902

Abstract. As official mobile operating system support becomes increasingly restrictive, users turn to custom operating systems, or custom ROMs, to prolong device usability, enhance performance, and regain control over privacy. Despite their popularity, the custom ROM ecosystem relies heavily on anecdotal evidence, lacking a standardized methodology to evaluate the significant trade-offs in performance, privacy, and security. This paper introduces AnROM, a novel framework that replaces this ambiguity with empirical analysis by providing a reproducible methodology to identify, rank, and evaluate widely used custom ROMs on a consistent hardware and software baseline. The framework's evaluation spans system resource usage, application responsiveness, and default privacy and security configurations, revealing that the choice of a custom ROM involves a complex set of compromises. The findings show some ROMs deliver superior performance at the cost of weaker privacy defaults, while others exhibit significant and counter-intuitive behaviors, such as minimalist systems generating excessive network traffic. The primary contribution of this work is the AnROM framework itself, which serves as an essential tool for quantifying these critical trade-offs and empowering both users and developers to make more informed decisions in the custom Android ecosystem.

Keywords: Operating Systems, Mobile Systems, Mobile Security, Mobile Privacy, Performance Evaluation, Reproducible Methodology, Android, Custom ROMs

1 Introduction

Mobile systems are essential to modern life, with a majority of the global population relying on smartphones (Howarth, 2023; Oberlo, 2023; Wise, 2023). These smartphones are manufactured and supported by hundreds of companies (PhoneArena, 2025; StarDust, 2025). Each smartphone model or brand may run a specific mobile operating system or a set of versions of a mobile operating system. This leads to significant mobile

operating system fragmentation (StarDust, 2025), and users are left with a multitude of choices when selecting a smartphone.

Mobile operating systems typically follow one of the two models: closed source, such as iOS (Apple Inc., 2025), or open source, such as Android (Google, 2025a). Despite the diversity in mobile system hardware, only a small number of operating systems are in widespread use (BankMyCell, 2025). The most popular mobile operating systems, iOS and Android, combined have almost all of the mobile operating system market share (BankMyCell, 2025; Counterpoint Technology Market Research, 2023). Closed source systems limit user customization, while Android-based systems, despite being open source in principle, are often tightly coupled with proprietary software (MUO, 2025). Users are thus constrained by manufacturer-enforced software policies and limited to preloaded features, short support cycles, and predefined privacy settings. Choosing a smartphone with a mobile operating system such as Android or iOS restricts the user to certain preloaded software, predetermined performance and battery life, and a limited support cycle, with devices typically reaching end-of-life after two years (Lanxon, 2022). Users have limited control over the personalization of their smartphone and limited control, even through permissions, over their privacy (Norton Labs, 2021).

Mobile systems and smartphones are known to contain a significant amount of personal and professional user data (European Data Protection Supervisor, 2025; Consumer Reports Inc., 2025). Users are often asked by the mobile operating system and applications to share information or enable specific device permissions relevant to the user (Norton Labs, 2021). The average user has over fifty mobile applications on their device with half of those used on a regular basis (Threat Intelligence, 2018; Fussell, 2022). Furthermore, almost half of the top mobile applications request personally identifiable data (Threat Intelligence, 2018). As such, there have been growing concerns about privacy and security on mobile systems in recent years (Klosowski, 2022). Legal action has not fully deterred the collection and abuse of personal user data (Klosowski, 2022). Legal measures such as the General Data Protection Regulation law (GDPR) (European Commission, 2025), implemented in 2018, have not fully protected users and their personal data in practice (Burgess, 2022; Wodinsky, 2022; Saqr, 2022). When confronted with this information and the fact that their personal information is being exposed through software, the majority of users show concern and many are not willing to share their personal data for any reason (KPMG LLP, 2023). Users face a constant dilemma: they depend on smartphones in daily life but feel compelled to limit usage due to privacy concerns (Bian et al., 2021).

To regain control, users and developers have turned to alternative mobile operating systems. There is a growing community of developers creating alternative mobile operating systems by using and repurposing the Android open source mobile operating system (AOSP, 2025a). These developers can modify the Android kernel and the Android framework, enabling them to remove, modify, or add more features. They can also add new devices or extend the life of a device that has reached end-of-life. These modifications can be released to the public and their purpose can be a more lightweight, performant, secure, private, or even simplistic version of Android (King, 2019). Additionally, developers can remove proprietary services such as Google Apps and Google

Play Services (Android Developers, 2024), which contain closed source Application Program Interfaces (APIs) for certain functionalities for applications such as location tracking or notifications (Vonau, 2021). Some versions that developers created not only remove Google Apps but replace Google Play Services entirely with a fully open source API (microG Team, 2025).

Over fifty custom Android mobile operating systems exist (GearJail, 2023). Many of those available online have been quite popular, with some supporting over four hundred devices and receiving over a million and a half active installs (The LineageOS Project, 2023; jhenrique09, 2023). Many of these custom-developed Android versions claim better performance, enhanced privacy, or improved usability, but there is little independent evaluation of these claims (King, 2019). Users often rely on informal community opinions, which may lack rigor or objectivity.

AnROM introduces a structured and reproducible methodology for evaluating popular Android custom operating systems. This includes a novel, publicly sourced ranking mechanism for identifying widely used custom operating systems, classification criteria based on device support and maintenance status, and a thorough empirical evaluation. The evaluation controls for hardware and Android version to ensure fair comparisons, and it includes tests for performance, privacy defaults, system resources, and security configurations.

Throughout this paper, the term "custom ROM" refers to a modified version of the Android operating system, as further detailed in Section 3. The term AnROM, in contrast, does not refer to an operating system but to the novel methodology and evaluation framework introduced in this work. AnROM provides the structured, reproducible process for selecting, evaluating, and comparing custom ROMs which is the central contribution of this work.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 discusses the architecture behind Android and what constitutes a custom ROM, Section 4 describes the AnROM framework and methodology, Section 5 presents the results, Section 6 reviews the findings and implications of this study, Section 7 discusses threats to validity, and finally, Section 8 presents the conclusion and future work.

2 Related Work

Prior work has compared eight Android versions based on added features, user experience, processing, security, and memory management without establishing a software or hardware baseline (Kaur et al., 2017). Several research studies (Singh, 2014; Holla and Katti, 2012; Nirmala and Sathya, 2015; Mohini et al., 2013) discuss the security features of Android. Android provides permissions and process isolation at the kernel level, sandboxing of applications, secure inter-process communication, application signing, and application and user-based permissions. This body of work also discusses the security shortcomings of these features.

Installing custom ROMs presents both advantages and disadvantages. Several studies (Gupta et al., 2015; Jha, 2023; Parekh, 2022; Anwer Basha et al., 2017; Sasi Kumar et al., 2018) discuss the benefits of rooting and installing custom ROMs. Another study (Manjrekar and Bhati, 2016) also lists custom ROMs. Using custom ROMs, the

latest Android versions can be installed, additional software can be added to improve performance, a user may easily back up all data, some default applications that use many resources can be removed, and a developer or user might be able to achieve more customization and personalization of Android. However, the disadvantages can be alarming; rooting and installing a custom ROM can allow for installing applications from different sources. Additionally, applications might not work correctly, a device might not boot properly, some security issues or bugs might be introduced, and in some cases, warranty for the device might be voided. Some studies (Rahul et al., 2014) detail how to port and install custom ROMs.

Additional studies create their own custom ROM for a specific purpose. For example, a specific study (Charan et al., 2014) customizes Android for running on different embedded devices and claims that the ease of portability is due to the reliance on the Linux kernel for Android. Another study (Sharma and Nimawat, 2019) creates a custom Android ROM to improve memory performance, showing that memory and battery usage can be improved with a customized and optimized version of Android. Similarly, one study (Kanthed and Yadav, 2017) creates a custom Android ROM to improve memory usage. A different study (Shreyas, 2020) creates a custom Android ROM to improve performance and user interface of Android and then compares against other non-Android mobile operating systems with respect to memory management. Finally, one study (Rajput et al., 2018) creates a custom ROM to improve user interface and extend the support life of lower-end Android devices.

Multiple survey studies exist. For example, one survey study (Okediran et al., 2014) compares different mobile operating systems and application development platforms but does not elaborate on different versions of mobile operating systems or custom ROMs. Additionally, another survey study (Suleman et al., 2020) compares custom Android ROMs from different Android versions. There is no methodology or reasoning behind choosing the eight custom ROMs aside from claiming they are popular. The comparison only shows the difference in ROM size, boot-up time, memory usage, ROM performance, and power savings. The comparison does not explain why the differences exist in these properties and does not note that these differences might be attributed to the different Android versions used by these custom ROMs. This makes the comparison potentially misleading, as the differences observed could be artifacts of the underlying Android version rather than the custom ROMs themselves.

Methodological work exists on Android. A methodological study (Barrera et al., 2010) develops a methodology for the empirical analysis of permission-based security models, using Android apps as a case study. The methodology highlights potential points of improvement for the Android permission model and is applicable beyond Android permissions.

In summary, existing work has addressed security features, rooting benefits, and isolated ROM projects, but lacks a unified, methodologically sound, and reproducible study of widely used custom ROMs. AnROM fills this gap by selecting popular ROMs based on transparent ranking, standardizing evaluation across uniform hardware and software versions, and comparing real-world privacy, performance, and security configurations.

3 Background

3.1 Android Architecture

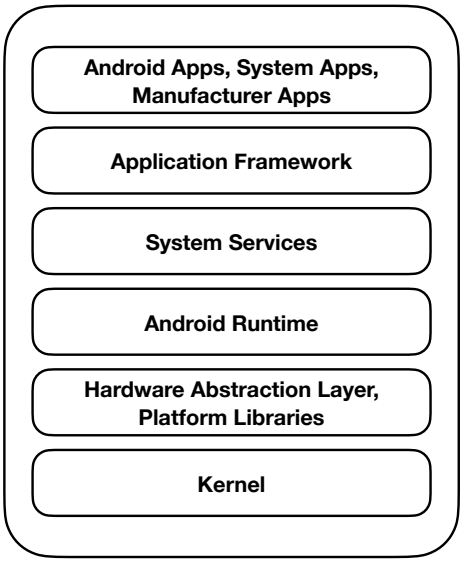


Fig. 1: Android Top-Down Architecture

To understand how developers create their own versions of the Android mobile operating system, we first examine the Android architecture. Figure 1 illustrates the top-down view of the Android architecture. Android ships with stock applications, system applications, and in some cases, device manufacturer applications. Users can also download third-party applications through the Google Play Store (Google, 2025b).

These applications run on the application framework, which provides the API. The Android framework provides access to system services such as location or sensors. These system services run as a single instance to initialize and manage access to hardware and databases.

At a lower level, the Android Runtime (ART) is the application runtime environment used by Android applications and some system services. The Hardware Abstraction Layer (HAL) manages access to the hardware and the native platform libraries provide all the native APIs for Android. Finally, the kernel, based on Linux, runs with device drivers and the core operating system components (AOSP, 2023b).

By default, stock Android ships with Google Apps and Google Play Services (Vonau, 2021). Google Play Services requires licensing from Google and contains proprietary software for some background services such as location tracking, payment processing, and integrated advertising. However, Google Apps and Play Services are not required to be able to run Android on a device. Android Open Source Project provides Android without Google Apps or Play Services (AOSP, 2023b; Android Developers, 2024).

3.2 Android-Based Operating Systems

Developers can use the Android Open Source Project as a base to create their own custom versions, allowing them to add, remove, or modify features. Developers can add support for new devices or prolong support for devices that are no longer supported in stock Android (AOSP, 2023b). The modified components can be strictly related to the collection of shipped applications or can be as involved as containing kernel changes.

Android framework and API, system services, and platform libraries can all be modified to serve the purpose of the modified Android version. The end product is made available by the developers as a custom firmware or more widely known as an Android or custom ROM, a term derived from Read-Only Memory. The actual product is not a ROM but a type of flash memory with a system image of Android including the modifications along with any applications (PCMag, 2025).

The Android Open Source Project does not include Google Play Services since the services are proprietary and require licensing from Google (AOSP, 2025c). Some developers include Google Play Services in their custom ROM, while others do not. Some developers also include open source alternatives for Google Apps and Play Services (Vonau, 2021; microG Team, 2025; The Open GApps Team, 2025). These architectural and licensing decisions significantly affect both functionality and privacy, making the ROM's design choices an important area of study.

4 Methodology

The **AnROM Framework** is structured as a process with multiple stages:

- **Stage 1: Candidate Selection**: AnROM first identifies custom ROMs based on objective, publicly available popularity metrics to ensure a relevant and unbiased selection.
- **Stage 2: Standardized Environment**: All selected custom ROMs are installed on identical hardware with a consistent software baseline to isolate operating system performance and eliminate hardware variables.
- **Stage 3: Reproducible Evaluation**: AnROM employs a suite of quantitative benchmarks to measure key performance indicators, including system resource utilization, application responsiveness, and boot times. To ensure the process is transparent and reproducible, all tests are conducted using publicly available tools and clearly defined procedures, allowing other researchers to validate the findings.

The remainder of this section details the application of Stage 1, Stage 2, and Stage 3 of this framework, while the evaluation of the selected custom ROMs and hardware is presented in Section 5.

4.1 Candidate Selection

The current number of custom ROMs far exceeds fifty ROMs (GearJail, 2023). These include ROMs released by smartphone manufacturers, software companies, and com-

munity developers. To conduct a proper *comparative* study, two factors are considered when choosing custom ROMs: (1) The custom ROM must support multiple devices and multiple brands. Custom ROMs that are exclusively released only for a specific brand (Alibaba Cloud, 2023; OPPO, 2023; emteria GmbH, 2023; Huawei Device Co., Ltd., 2023; Amazon.com, Inc., 2023; vivo, 2023; Indus OS, 2023; Xiaomi, 2023; ZTE, 2023; Samsung Electronics Co., Ltd., 2023; OnePlus, 2023; SHIFT GmbH, 2023; Smartisan Technology Co., Ltd., 2023) are not considered. Table 1 lists some stock ROMs exclusive to specific well-known smartphone brands. These ROMs are typically proprietary stock ROMs tailored to their specific branded devices. Additionally, since these custom ROM developers focus on a specific list of hardware, they might be specifically optimized to run on these devices. (2) Defunct or no longer supported custom ROMs are not considered. Only currently active ROM projects that support at least five different devices are considered.

Brand	Custom ROM
Alibaba	AliOS (Alibaba Cloud, 2023)
Amazon	Fire OS (Amazon.com, Inc., 2023)
Huawei/Honor	EMUI (Huawei Device Co., Ltd., 2023)
OnePlus	OxygenOS (OnePlus, 2023)
Oppo	ColorOS (OPPO, 2023)
Samsung	One UI (Samsung Electronics Co., Ltd., 2023)
Vivo	Funtouch OS (vivo, 2023)
Xiaomi	MIUI (Xiaomi, 2023)
ZTE	MyOS (ZTE, 2023)

Table 1: Examples of Brand-Exclusive Custom ROMs

Custom ROMs that do not support more than four devices are usually exclusive, focusing on the same brand, or they are created by a very small number of community developers that target a small number of users. Removing these brand-specific ROMs, non-active projects, and device-limited ROMs, fewer than fifty ROMs remain.

Since AnROM results and findings should reflect the ROMs with the most users, a list of the most recognized and used ROMs was created. However, there is no standard or recognized rankings of custom ROMs. AnROM develops a novel methodology for ranking items in a category. To identify top custom ROMs, AnROM utilizes Google Search (Google, 2023) results and curated lists from relevant technology publications and websites. Therefore, to focus on the most popular custom ROMs that are widely used, Google Search results and lists of top custom ROMs from technology-related websites are used as the inclusion criteria.

The search queries included the name of the custom ROM along with the phrase "ROM" including the quotes. This filters out any non-ROM-related queries since some custom ROMs have generic names. The list was refined to include only custom ROMs that are not brand-specific and have more than one hundred thousand search results. To verify these results, the list is cross-referenced with lists of top custom ROMs from technology websites (GearJail, 2023; Patel, 2023; Cawley, 2020; Time News, 2022; Regmi,

2022; Ghosh, 2023; Chapman, 2022; Congleton, 2023; Abhijeet, 2023; Mithran, 2022; Hazarika, 2021b; Sha, 2023).

Table 2 shows 24 ROMs that matched the criteria as of March 30, 2023. These ROMs also generally matched the top custom ROMs listed on technology websites. AOSP, GrapheneOS, and CopperheadOS mostly target Google devices, they were kept in this list for thoroughness. The custom ROMs in Table 2 are sorted by the number of Google Search results. The *Purpose* column defines the purpose of the custom ROM as listed by the developers of the custom ROM. Table 2 also lists the latest Android version the custom ROM is based on along with the total number of devices supported and the total number of Android 13 devices supported.

ROM (as of March 30, 2023)	Google Search Results	Purpose	Latest Version	Num of Devices	Android 13 Devices
LineageOS (The LineageOS Project, 2023)	1,550,000	Customization	Android 13	420	92
AOSP (AOSP, 2023b)	1,260,000	Development	Android 13	36	9
Paranoid Android (Paranoid Android, 2023)	678,000	Customization	Android 13	47	30
Descendant (Descendant, 2023)	541,000	Customization	Android 12	75	0
PixelExperience (jhenrique09, 2023)	541,000	Customization	Android 13	138	85
Evolution X (Haruka LLC, 2023)	492,000	Customization	Android 13	73	59
Replicant (Replicant, 2023)	385,000	Free Software	Android 11	8	0
crDroid (crDroid Android, 2023)	344,000	Customization	Android 13	173	63
Resurrection Remix OS (Resurrection Remix OS, 2023)	318,000	Customization	Android 10	88	0
Xtended (Xtended, 2023)	260,000	Customization	Android 13	50	31
GrapheneOS (Only Pixels) (GrapheneOS, 2023)	220,000	Privacy/Security	Android 13	11	11
CarbonROM (CarbonROM, 2023)	218,000	Customization	Android 11	51	0
Havoc-OS (Prasal, 2023)	190,000	Customization	Android 12.1	106	0
OmniROM (OmniROM, 2023)	190,000	Customization	Android 13	60	10
Kali NetHunter (OffSec Services Limited, 2023)	174,000	Testing	Android 13	82	2
SparkOS (SparkOS, 2023)	174,000	Customization	Android 13	18	18
/e/ e Foundation (e.foundation, 2023)	173,000	Privacy	Android 12.1	210	0
MoKee (MoKee Open Source Project, 2023)	171,000	Customization	Android 11	98	0
YAAP (Mohan, 2023)	157,000	Customization	Android 13	11	7
dotOS (dotOS, 2023)	135,000	Customization	Android 12.1	71	0
ArrowOS (ArrowOS, 2023)	118,000	Customization	Android 13	102	25
Potato Project (Potato Open Sauce Project, 2023)	107,000	Customization	Android 13	74	3
CopperheadOS (Only Pixels) (Copperhead Limited, 2023)	104,000	Privacy/Security	Android 13	18	6
BlissRoms (BlissLabs, 2023)	103,000	Privacy/Security	Android 13	33	24

Table 2: Popular Custom ROMs

4.2 Standardized Environment

For a true comparison, the same software version and the same hardware are used for testing. Focusing on the list in Table 2, the most popular custom ROMs, with more than three hundred thousand Google Search results, that support the latest Android version are chosen. Any custom ROMs that only support a certain brand such as Google devices are removed.

Refining this further, Table 3 lists the five most popular ROMs that support the latest Android version. Note that LineageOS is the successor of CyanogenMod (CyanogenMod open-source community, 2023). Also, AnROM uses the Plus edition of the Pix-

elExperience, which includes more features, since it is much more popular than the regular edition given the number of downloads.

ROM	Google Search Results	Latest Version	Num of Devices	Android 13 Devices
LineageOS (The LineageOS Project, 2023)	1,550,000	Android 13	420	92
Paranoid Android (Paranoid Android, 2023)	678,000	Android 13	47	30
PixelExperience (jhenrique09, 2023)	541,000	Android 13	138	85
Evolution X (Haruka LLC, 2023)	492,000	Android 13	73	59
crDroid (crDroid Android, 2023)	344,000	Android 13	173	63

Table 3: Most Popular Custom ROMs

In addition to maintaining the testing for custom ROMs across the same Android version, the same hardware across all custom ROMs is required for a true comparison. Each custom ROM supports a specific set of devices. Searching all the custom ROMs for devices supported with the latest Android, only three devices are collectively supported by every custom ROM listed in Table 3: the OnePlus 7 Pro, 7T, and 7T Pro.

To ensure reliable performance metrics and account for measurement variance, all dynamic tests, such as application launch times and benchmark scores, were executed at least three times for each custom ROM after a clean boot. The average values of these runs were reported in the results.

4.3 Reproducible Evaluation

To facilitate a comprehensive and reproducible evaluation of the AnROM framework, a series of metrics was defined to assess each custom ROM's default state and its dynamic performance. These metrics are listed in Table 4. The evaluation is structured into two distinct phases. The first phase focuses on a static analysis of the custom ROM's default configuration. This involves quantifying the initial software package by inventorying pre-installed applications and establishing a baseline resource footprint by measuring storage, idle memory, CPU usage, and running processes. The privacy and security defaults are also inspected by examining enabled connectivity services (e.g., Bluetooth, NFC), active permissions, and the status of core features like location services and device encryption. Static information about kernel configuration and Android properties should also be examined. Finally, the camera's output should also be compared in terms of quality and file output.

Following the baseline analysis, and after a system reboot to ensure a standardized starting state for all custom ROMs, the second phase consists of dynamic testing to measure performance under typical workloads. This phase includes measuring application launch times to gauge user-perceived responsiveness and executing a suite of synthetic benchmarks to quantitatively assess CPU, memory, and disk performance. To evaluate efficiency, battery consumption is measured after a sustained workload. Finally, network traffic is monitored immediately after boot to detect any unsolicited data

Test/Query	Details
Applications Package	Does it have pre-installed official Google Apps, alternatives, or none?
Pre-installed Applications	How many applications were initially installed?
Processes	How many processes after starting the system?
Storage	How much space is the ROM taking?
Memory	How much RAM is being used while idle?
CPU	What CPU governor is used? How much CPU is being used while idle?
Connectivity	What is enabled by default (Bluetooth, NFC, Nearby Share)?
Permissions	What are the permissions in use so far?
Location	Is Google Services location disabled by default?
Encryption	Is the device encrypted by default?
Camera	What is the quality of the camera? Does it differ across ROMs?
Kernel	What kernel/compiler version is being used? What is the kernel config?
Android Properties	What are the default Android properties?
Reboot Device	Reboot the device after the above tests/queries
Application Launch Time	How long does it take to launch a certain application
Performance Tests	Run benchmarks for CPU, memory, disk, and GPU
Battery Life	Run benchmarks 3 times then after 15 minutes measure the battery
Networking	Check the amount of data going in and out after the reboot

Table 4: AnROM Study Questions And Evaluation Metrics

transmissions, providing further insight into the custom ROM's default privacy practices. The complete set of metrics, detailed in Table 4, provides a holistic view of each custom ROM's characteristics.

5 Evaluation

For evaluation, various tests and studies were run using the same device and hardware to get comparable results across ROMs. The OnePlus 7 Pro (guacamole, GM1913, Qualcomm Snapdragon 855 Octa-core, 12GB RAM, 256GB storage, 4G LTE) was used. A valid LTE-enabled SIM card was installed. Table 5 shows the custom ROMs and their latest versions, based on Android 13, used for testing. In addition to the chosen most popular custom ROMs, the original or stock ROM, OxygenOS, which ships with the OnePlus 7 Pro was included and all the relevant automated updates were applied. The latest version available for the OnePlus 7 Pro is OxygenOS 12.1 which is based on Android 12.1 and not Android 13. The stock ROM was included as a usability comparison of what users have available to them by default versus the range of the latest custom ROMs available. Note that all custom ROMs require a base firmware of OxygenOS 12.1 except Paranoid Android, which requires the OxygenOS 11 base firmware to be installed.

To prepare the device and install custom ROMs, developer options and USB debugging were both enabled. After a clean install of the stock ROM or any custom ROM, a newly created Google Account was used with the default configurations for setting up the new device but with a few exceptions: a WiFi network was used and data was not copied from previous devices, and for Google Services, backups, location, sending usage and diagnostic data, and network scanning were all disabled. Additionally, a PIN, Google Assistant, and Google Pay were not set up. The device was then left charging and idle for a few minutes until any updates running in the background were completed.

ROM	Version
LineageOS	20
Paranoid Android	Topaz Stable 1
PixelExperience	13.0
Evolution X	7.6.2
crDroid	9.3
OxygenOS	12.1

Table 5: ROM Versions Tested

After the device was set up for each ROM and stabilized, several tests and queries were conducted. Table 4 shows a list of tests and queries. Results were obtained either through the graphical user interface of Android using the Settings application or through the Android Debug Bridge (*adb*) tool (Android Developers, 2023a) using the command-line with *top*, *procfs*, *ifconfig*, and *logcat*.

The type of applications package, used or installed, was noted. In some cases, the default Google Apps are already installed. In other cases, some open source or free alternatives to Google Apps are installed. Note that not all ROMs are released with applications packages, some are released without Google Apps or any alternatives. The number of applications initially installed and the number of processes running were also noted. The amount of storage space, RAM, and CPU initially used while idle were all measured. Additionally, the CPU governor is noted. As a privacy check, connectivity technologies such as Bluetooth, NFC, and Google Nearby Share, were checked to see if they were enabled by default. Additionally, the number of permissions in use by all the applications was verified. The default settings for location and encryption are also noted. Since the majority of users care about camera quality (Sharma, 2021), images across ROMs were taken from the same position and similar angles with the same lighting and compared using an image quality assessment algorithm.

Next, the kernel version, kernel configuration file, and Android properties were all queried. The kernel *config* file specifies what kernel features are to be compiled or included. The Android properties (AOSP, 2025b) are system-wide resources that are shared across all processes. Configuration information can be stored in these Android properties. For example, it stores the build date, kernel version, the status of the security module, SELinux (SELinux, 2024), and any other configurations set up by the ROM developers.

After completing these queries, the fully charged device was rebooted and disconnected from the charger. The launch time was then measured, defined as the amount of time in milliseconds that elapses between launching the process and finishing drawing the corresponding activity on the screen (Android Developers, 2023b). The PassMark PerformanceTest benchmark application (version 10.2.1001) (PassMark Software, 2023b) was relaunched five times, without running the benchmark, and launch times, that were logged by Android, were averaged. The PassMark benchmarks test the CPU, memory, disk, and GPU performance. The PassMark benchmarks were run three times and the results were averaged, the benchmark results were not reported to PassMark servers. After running the benchmarks, the battery percentage drop was measured

fifteen minutes later. Finally, the number of networking packets received and transmitted during those fifteen minutes were recorded. The networking connections were also investigated to see which connections were responsible for the data exchange, if any.

Table 6 shows some of the study results. Most ROMs used Google Apps. LineageOS and crDroid did not include Google Apps but provided alternative options to be installed by the user. The number of pre-installed applications varied, ROMs using Google Apps had around 26-29 applications. Since LineageOS recommends MindTheGapps (MindTheGapps, 2023) and crDroid recommends NikGapps (core) (NikGApps, 2023), they had a slightly lower number of applications pre-installed, around 22-24. Stock ROM had a much larger number of applications since it included OnePlus applications and some additional applications provided by OnePlus for OxygenOS. The number of processes running after booting the ROMs was within the 745-895 range, depending on the pre-installed applications and background services for each ROM. Storage used differed significantly between the ROMs. Storage usage is divided between system storage and application storage. crDroid and Evolution X use up less than 6GB of storage. Stock ROM uses the most storage due to the number of applications and additional system applications. Memory usage varied for all ROMs, with stock ROM being the most optimized and crDroid using the most RAM.

Test/Query	LineageOS	Paranoid Android	PixelExperience	Evolution X	crDroid	Stock ROM
Applications Package	MindTheGapps	GApps	GApps	GApps	NikGapps	GApps
Pre-installed Applications	24	29	26	27	22	42
Processes	856	745	861	868	848	895
Storage: Total Used	22GB	17GB	17GB	5.1GB	3.9GB	25.2GB
Storage: Applications	0.63GB	1.3GB	1.9GB	1.1GB	0.06GB	3.64GB
Memory	3424MB	1908MB	2243MB	1723MB	4122MB	270MB
CPU (% used)	6%	7%	17%	14%	12%	15%
Connectivity	Not Nearby	All	Not Nearby	None	None	Not Bluetooth
Permissions (text)						
Location	Off	Off	Off	Off	Off	Off
Encryption	On	On	On	On	On	On
Camera (Figure 2)						
Kernel Version	4.14.180	4.14.305	4.14.309	4.14.180	4.14.311	4.14.180
Kernel Compiler (Clang)	14.0.6	16.0.2	14.0.6	14.0.6	14.0.6	10.0.7
Android Properties (text)						
Reboot Device						
Application Launch Time	129ms	118ms	107ms	160ms	131ms	201ms
Performance Tests (text)						
Battery Life (% used)	4%	3%	3%	3%	3%	4%
Networking (Table 8)						

Table 6: AnROM Study Results

Even though all ROMs used the same *schedutil* CPU governor, CPU usage can significantly vary even within the same ROM. CPU usage was measured as the aggregate utilization across all eight cores, where a value of 100% signifies the full load of a single core, making the theoretical maximum 800%. CPU usage while idle was less than 15% for four custom ROMs, with PixelExperience and stock ROM using 15% or higher. For connectivity, Evolution X and crDroid had all technologies off by default. LineageOS and PixelExperience only had Nearby Share disabled by default. Stock ROM only had

Bluetooth disabled by default. Paranoid Android was the only ROM with all technologies enabled by default. Google location services were disabled by default for all ROMs. Encryption was enabled by default for all ROMs. Even though all ROMs, except stock ROM, were Android 13, not all ran the same kernel. Only LineageOS, Evolution X, and stock ROM shared the same kernel version. Only LineageOS and Evolution X shared both, the same kernel version and the same kernel compiler version.

After rebooting the device, the average application launch time was measured. PixelExperience was the fastest followed by Paranoid Android. The slowest launch times were Evolution X and stock ROM. Stock ROM had the highest number of processes running and applications pre-installed, which might contribute to the slower launch time. After running the benchmark tests, battery percentages only decreased by 3% for most ROMs. Stock ROM decreased by 4% due to either a different Android version being used, a higher number of running processes, or both. LineageOS battery percentage also decreased by 4% because it had a noticeably higher default brightness than the rest of the ROMs.

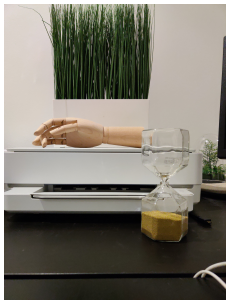
For the pre-installed applications, the majority of permissions were requested and given by the ROMs. Permissions such as SMS, media, phone, notifications, Nearby Share, audio, microphone, location, contacts, camera, call logs, and calendar are available and given to most applications.

Figure 2 shows the camera images taken from the same position and similar angle with the same lighting for all ROMs. All ROMs produced images in the same JPEG format with identical resolution and similar Exif data. Most ROM images were around 2.00 megabytes, except stock ROM which was 3.25 megabytes, suggesting differences in image compression or post-processing.

To quantitatively assess camera performance, the BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) algorithm was selected (Mittal et al., 2012). BRISQUE was chosen as it is a widely-accepted No-Reference Image Quality Assessment model that correlates strongly with human perception of image quality (The MathWorks Inc., 2025). The results, shown in Figure 2, where a lower score indicates better image quality, show that the stock ROM produced the highest quality image, while the custom ROMs were significantly lower in quality, though relatively close to one another. These measurable differences can be attributed to software-level choices, such as low-level image compression values, camera application processing, or underlying device drivers.

For kernel configuration options, there are a total of 2078 unique options across all tested ROMs. The vast majority of these configuration options are the same for all of the ROMs. Some are only used for certain ROMs, such as ROM-specific options. An interesting configuration option that is worth mentioning is the configuration option responsible for the default type of I/O scheduler. All ROMs, except Paranoid Android, use the completely fair queuing scheduler. Paranoid Android uses the NOOP I/O scheduler instead. The difference in schedulers would affect the performance and give the user a different usability experience. Note that the choice of kernel configuration options may expose the user to different types of security and privacy risks.

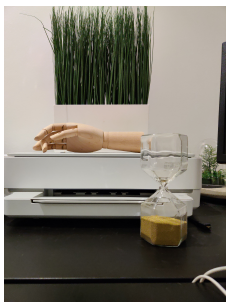
All ROMs include support for SELinux, as verified in the kernel configuration file. SELinux runs in one of three modes: Disabled, permissive, or enforcing. Permissive does not fully enforce SELinux but logs any operation rather than denying it. Enforcing


LineageOS (2.050MB)
BRISQUE 42.59


Paranoid Android (2.038MB)
BRISQUE 42.87


PixelExperience (2.052MB)
BRISQUE 41.97


Evolution X (2.046MB)
BRISQUE 42.79


crDroid (2.080MB)
BRISQUE 40.70


Stock ROM (3.250MB)
BRISQUE 15.25

Fig. 2: Camera Image Quality (BRISQUE Score: Lower Is Better)

actually enforces SELinux. By definition, the enforcing mode is the most secure and any other mode is a security risk (Hazarika, 2021a). Any custom ROM could choose to use the same kernel version but without SELinux enabled, thus, it is essential to review kernel configurations for every custom ROM rather than simply relying on the kernel version.

For Android properties, there are a total of 1327 unique properties across all tested ROMs. Similar to kernel configuration options, many of these Android properties are the same across all ROMs. Some are different across ROMs, such as the *dev.mnt* which includes the mount point name, but do not actually affect performance nor user experience. A few other Android properties are specific to a ROM.

Table 7 lists some of the Android properties across ROMs. These Android properties are listed because their values differ across ROMs and they also might affect the performance or security of the system as a whole. The first set of Android properties is related to the Android Runtime, ART or Dalvik, which is the managed runtime used by applications and some system services (AOSP, 2023c). The specific Android properties listed for the runtime are related to memory usage and utilization by applications. These default values can affect the performance and behavior of Android applications. For example, heap max free defines the maximum amount of free memory allowed af-

ter a garbage collection event. This affects the frequency of the virtual machine garbage collection.

Android Properties	LineageOS	Paranoid Android	PixelExperience	Evolution X	crDroid	Stock ROM
dalvik.vm.heapmaxfree	32m	56m	32m	32m	32m	16m
dalvik.vm.heapminfree	8m	8m	8m	8m	8m	4m
dalvik.vm.heapstartsize	16m	24m	16m	16m	16m	16m
dalvik.vm.heaptargetutilization	0.5	0.42	0.5	0.5	0.5	0.75
ro.surface_flinger.set_touch_timer_ms	4000	1500	4000	4000	4000	4000
ro.boot.type	sdebug	normal	sdebug	sdebug	sdebug	sdebug
ro.system.build.type	userdebug	user	user	user	userdebug	user
ro.debuggable	1	0	0	0	0	0
ro.build.version.security_patch	2023-03-05	2023-02-05	2023-03-05	2023-02-05	2023-03-05	2022-12-05

Table 7: Android Properties Comparison

Stock ROM has the lowest heap max memory allowed, making garbage collection more frequent. Paranoid Android has the highest heap max memory allowed. The rest of the ROMs are equal at 32 megabytes. Heap min free defines the minimum amount of free memory reserved after a garbage collection event. Stock ROM reserves 4 megabytes while the rest of the ROMs use 8 megabytes. Heap start size defines the default size of the heap for an application, with Paranoid Android having the highest at 24 megabytes and the rest of the ROMs use 16 megabytes. The higher the heap start size, the better performance is expected as applications will initially have more memory. Finally, the heap target utilization defines the target heap size using the heap max and heap min free parameters. Most ROMs have the same utilization ratio of half. Paranoid Android sets a slightly lower target utilization rate, given their higher heap values. Stock ROM has a much higher target utilization of 0.75 since it has a lower heap min free value (Peng, 2023; Rajendran and Banerjee, 2021). A garbage collector optimization study (Rajendran and Banerjee, 2021) of these parameters shows a significant performance improvement with graphics and application launch times when changing these heap parameters. Any custom ROM assigning values to these Android properties is thus directly affecting application performance and user experience.

Table 7 lists a touch timer for the *SurfaceFlinger*, which composes and sends buffers to the display (AOSP, 2023f). The set touch timer Android property allows setting the time, in milliseconds, after which the refresh rate is reduced if no touches occur (AOSP, 2023e). For example, if 1.5 seconds pass without any touches from the user, the refresh rate for the display is set to a reduced rate for the case of Paranoid Android. For the rest of the ROMs, the default is 4 seconds. This setting can affect the performance and battery consumption of the device.

The boot and system build type Android properties define the type of boot option and build. User builds are generally more limited in terms of debugging and include fewer features and code. Debug builds contain more debugging in terms of features such as logging and code. Additionally, some ROMs rely on the boot and build type to allow for more privileged options to be used. Note that some of these properties are generally read-only and changing them does not change the behavior of the system.

When the *debuggable* Android property is set to one, debugging for all applications is enabled (AOSP, 2023d). Additionally, this property is required to allow root access through adb (AOSP, 2023a). LineageOS is the only ROM to have this property enabled by default. While this property is useful for debugging and testing, it is a security risk for the system (Android Developers, 2025).

Finally, the security patch Android property is a read-only value that lists the version of the security patch installed. This shows the latest security patch installed for each ROM. From Table 7, the security patches installed for some ROMs are older than others. Installing the latest security patches is essential to maintain a secure system and each ROM should provide the latest security patches from mainstream Android (Thomas, 2021).

Table 8 shows the network data transmission of the ROMs through WiFi. Cellular data usage was negligible since WiFi was enabled. Some ROMs did not have significant data received or transmitted. LineageOS and stock ROM received data mostly from Google Play Services. crDroid had the most network activity, mostly from Google Play Services as well.

ROM	RX Packets	TX Packets
LineageOS	11200	2870
Paranoid Android	3864	2584
PixelExperience	1959	1515
Evolution X	1203	443
crDroid	90205	12113
Stock ROM	12031	5177

Table 8: WiFi Network Data Transmission of ROMs

Table 9 presents the numerical results of the AnROM evaluation, showing the percentage difference of each custom ROM compared to the stock ROM baseline. For performance and resource metrics, a negative percentage indicates an improvement (e.g., lower memory usage, faster launch times), while a positive percentage indicates worse performance.

The results presented in Table 9 provide a clear, quantitative answer to this study's central research questions and demonstrate the main achievement of this work. The data reveals that the choice of a custom ROM involves significant and often non-obvious trade-offs. For example, while crDroid offers a remarkable 84.5% reduction in total storage used and a 98.4% reduction in application storage footprint compared to the stock ROM, this efficiency comes at the cost of a staggering 1426.7% increase in idle memory consumption. Similarly, PixelExperience provides the fastest application launch times (46.8% faster than stock), but it does so with a 13.3% increase in idle CPU usage. Therefore, the primary achievement of this research is twofold: first, it provides a concrete, numerical quantification of these critical performance and resource trade-offs; and second, it validates the AnROM framework itself as a methodology that successfully uncovers these data-driven insights, moving the evaluation of custom operating systems from anecdotal recommendations to empirical analysis.

Numerical Test/Query	Stock ROM	LineageOS	Paranoid Android	PixelExperience	Evolution X	crDroid
Pre-installed Applications	42	-42.9%	-31.0%	-38.1%	-35.7%	-47.6%
Processes	895	-4.4%	-16.8%	-3.8%	-3.0%	-5.3%
Storage: Total Used	25.2GB	-12.7%	-32.5%	-32.5%	-79.8%	-84.5%
Storage: Applications	3.64GB	-82.7%	-64.3%	-47.8%	-69.8%	-98.4%
Memory	270MB	+1168.1%	+606.7%	+730.7%	+538.1%	+1426.7%
CPU (% used)	15%	-60.0%	-53.3%	+13.3%	-6.7%	-20.0%
Camera (image size)	3.25MB	-36.9%	-37.3%	-36.9%	-37.0%	-36.0%
Camera (BRISQUE)	15.25	+179.3%	+181.1%	+175.2%	+180.6%	+166.9%
Application Launch Time	201ms	-35.8%	-41.3%	-46.8%	-20.4%	-34.8%
Battery Life (% used)	4%	0.0%	-25.0%	-25.0%	-25.0%	-25.0%
Networking (WiFi data RX)	12031 Pkts	-6.9%	-67.9%	-83.7%	-90.0%	+649.8%
Networking (WiFi data TX)	5177 Pkts	-44.6%	-50.1%	-70.7%	-91.4%	+134.0%

Table 9: A Numerical Comparison of Custom ROMs to the Stock ROM Baseline

Finally, Figure 3 shows the test results from running the PassMark benchmark tests. All results are normalized to the stock ROM and higher results are better. Although the stock ROM is based on Android 12.1 rather than Android 13, all PassMark performance scores were normalized against it to illustrate the difference a user might experience when moving from the latest default ROM shipped with the device to the most popular custom ROMs currently available for it. PassMark runs various tests for CPU, memory, disk, 2D graphics, 3D simple test, and a 3D complex test. The OpenGL ES using Unity test did not run successfully on any ROM including the stock ROM since there is a known issue with this PassMark test for some devices (PassMark Software, 2023a).

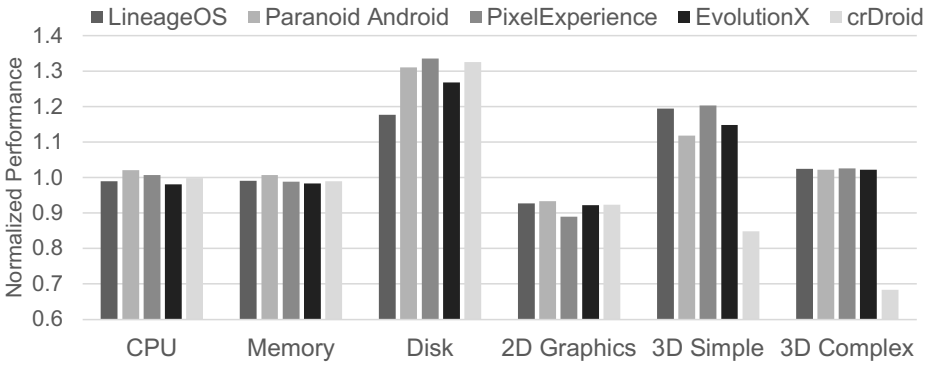


Fig. 3: PassMark Benchmarks (Higher Is Better)

CPU and memory results in Figure 3 show a negligible difference between all ROMs. For the disk results, Paranoid Android, PixelExperience, and crDroid performed similarly with the best results. All of these three ROMs use a kernel version higher than 4.14.300. LineageOS and Evolution X performed differently but with worse results compared to the three custom ROMs. They both use the same kernel version and the difference could be attributed to Evolution X using less storage and memory than

LineageOS. For 2D graphics, all ROMs performed worse than stock ROM. The difference could be attributed to stock ROM using a manufacturer customized Android 12.1, with the rest of the ROMs using Android 13. Additionally, PixelExperience performed slightly worse compared to other ROMs due to possibly using a uniquely different kernel version. For 3D graphics, the 3D simple test shows varying results with most performing better than stock ROM. crDroid performed significantly worse than stock ROM and the rest of the custom ROMs. Finally, the 3D complex test had very similar results across all ROMs except for crDroid which performed the worst. Even though crDroid has higher CPU usage and has the highest RAM usage compared to all ROMs, the difference in performance is most likely due to resource management and a different codebase.

6 Discussion

Custom ROMs offer users a compelling alternative to stock mobile operating systems, but the user experience involves significant, measurable trade-offs. While aesthetic differences in design and themes are a matter of preference, the choice of pre-installed applications has a quantifiable impact. For instance, crDroid and LineageOS ship with nearly half the number of applications as the stock ROM (a 47.6% and 42.9% reduction, respectively), offering a minimalist starting point. This contrasts with ROMs like Paranoid Android, which includes only 31.0% fewer apps, providing a more feature-rich but less lean initial experience.

Beyond the user interface, the AnROM findings prove that performance differences are not marginal but substantial. No single custom ROM provides a universal performance benefit; instead, they present a complex balance of gains and regressions. For example, PixelExperience delivers the fastest application launch times (46.8% reduction compared to stock), a clear advantage for user-perceived responsiveness. However, this comes at the cost of a 13.3% increase in idle CPU usage, which can impact battery life. Conversely, Evolution X and crDroid show remarkable storage efficiency, reducing the total space used by 79.8% and 84.5%. This efficiency is offset by a staggering increase in idle memory consumption, with crDroid using 1426.7% more RAM than the stock ROM, highlighting a critical performance trade-off between storage and memory.

In terms of privacy and security, AnROM shows that some custom ROMs provide a more granular permissions model that users need to manually enable through a menu. The quantitative data reveals counterintuitive and critical insights: a leaner application footprint does not guarantee better privacy. For example, crDroid, despite its minimal app count, exhibited the highest idle network activity, transmitting 134.0% more data than the stock ROM. This raises significant questions about potential undisclosed background processes. In contrast, Evolution X offered the best out-of-the-box privacy posture in this regard, with a 91.4% reduction in transmitted data. Furthermore, while not all security aspects are easily benchmarked, the inconsistency in security patch levels noted in AnROM analysis confirms that a custom ROM is not an automatic solution for security; it is a choice that requires user diligence.

These measurable differences in performance and privacy stem from deliberate, albeit often undocumented, choices by the developers. Variations in kernel versions, com-

piler versions, compiler flags, and default Android property settings have direct, quantifiable consequences. The dramatic increases in idle memory usage seen in LineageOS (1168.1%) and crDroid (1426.7%), for example, are likely the result of specific kernel configurations designed to prioritize other aspects of performance. This underscores the critical need for custom ROM developers to provide greater transparency. Users should not have to discover a four-figure percentage increase in memory consumption on their own; it should be a documented design choice.

Ultimately, AnROM contributes a replicable methodology for transforming this complex landscape into a field of empirical analysis. Unlike prior work, which often compares ROMs under inconsistent hardware and software baselines, AnROM standardizes the evaluation to provide fair, reproducible comparisons. Its core achievement is demonstrating that the real-world trade-offs in performance, privacy, and resource management can be quantified. As the custom ROM ecosystem evolves, a rigorous, data-driven framework like AnROM is essential for empowering users and holding developers accountable.

7 Threats to Validity

While the AnROM framework establishes a reproducible methodology, the findings of this initial study are subject to specific threats that can be addressed in future work.

First, the evaluation was conducted on a single hardware platform. Performance metrics, particularly those related to graphics and disk I/O, are tightly coupled with the hardware technology of the device. The trade-offs observed in this study may manifest differently across other hardware platforms. Future work should extend AnROM to a matrix of devices with varying hardware profiles.

Second, the performance analysis relied on synthetic benchmarks and tests. While useful for standardized comparison, these do not fully capture the complexity of real-world usage patterns. Future work could incorporate more sophisticated, scripted workloads that simulate common user behaviors like web browsing, mobile gaming, and multi-application task switching to provide a more complete performance profile.

Third, this study represents a snapshot in time. Custom ROMs are continuously updated, and the performance and security status of a given version may change over time. A longitudinal study, tracking the evolution of a select group of custom ROMs across several updates, would provide valuable insights into developer maintenance practices and the long-term stability of their performance profiles.

Finally, to address error and reliability more formally, future studies should execute each test a greater number of times to establish statistical significance and report results with confidence intervals and standard deviations. This would provide a more robust assessment of performance variance both within and between custom ROMs.

8 Conclusion

Custom Android operating systems offer users a compelling alternative to vendor-controlled mobile operating systems by enabling device longevity, enhanced customization, and, potentially, stronger privacy and performance. However, the diversity among

these custom Android operating systems, ranging from development practices to privacy configurations, makes it difficult for users to make informed choices. This paper introduced AnROM, a methodology and evaluation framework designed to fill that gap. AnROM provides a reproducible way to rank and select widely used custom operating systems and to evaluate them on a consistent hardware and software baseline. Through performance benchmarking, privacy inspection, and security analysis, AnROM uncovers notable differences between custom Android operating systems that are often invisible to users or poorly documented by developers. For example, some custom ROMs achieve a dramatic reduction in storage use but exhibit a massive increase in idle memory consumption and unsolicited network activity. Others deliver considerably faster application launch times but with higher idle CPU usage.

AnROM findings demonstrate that no single custom ROM consistently outperforms others across all dimensions. Performance, privacy, and security trade-offs vary significantly and depend on factors such as included applications, kernel settings, and update practices. As such, users must be cautious and informed when adopting a custom Android operating system, and developers should aim to improve transparency and maintenance practices. Future work could extend AnROM to include additional custom Android operating systems, a broader range of devices, as well as longitudinal studies to assess how these custom Android operating systems evolve over time. Further automation of testing, integration of battery life metrics, and crowd-sourced user feedback could enhance the depth and reach of the evaluation process.

AnROM encourages a more informed, privacy-conscious, and performance-aware Android user community. AnROM provides a foundational methodology for future research, representing a critical first step in replacing ambiguity with empirical evidence and enabling a new class of rigorous, comparative studies in mobile operating systems.

Acknowledgments

The author would like to thank Kuwait University and its College of Science for their support and facilities.

References

Abhijeet (2023). 20+ Best Custom ROMs for Android, https://www.droidthunder.com/best-custom-roms-android/.

Alibaba Cloud (2023). AliOS, https://alios.cn/.

Amazon.com, Inc. (2023). Fire OS Overview — Amazon Fire TV, https://developer.amazon.com/docs/fire-tv/fire-os-overview.html.

Android Developers (2023a). Android Debug Bridge (adb) — Android Developers, https://developer.android.com/tools/adb.

Android Developers (2023b). App startup time — Android Open Source Project, https://developer.android.com/topic/performance/vitals/launch-time.

Android Developers (2024). Google Play services — Google Developers, https://developers.google.com/android.

Android Developers (2025). android:debuggable — Security — Android Developers, https://developer.android.com/privacy-and-security/risks/android-debuggable.

Anwer Basha, H., Priyanga, K., Bhuwaneshuwaran, V., Bhanu Prakash, J. (2017). Custom rom, *International Journal of Innovative Research in Science and Technology* **4**(7), 43–45.

AOSP (2023a). adb_debug.prop — Android Code Search, https://cs.android.com/android/platform/superproject/+/android-13.0.0_r41:system/core/rootdir/adb_debug.prop;l=8.

AOSP (2023b). Android Open Source Project, https://source.android.com/.

AOSP (2023c). Android Runtime (ART) and Dalvik — Android Open Source Project, https://source.android.com/docs/core/runtime.

AOSP (2023d). debugger.html — Android Code Search, https://cs.android.com/android/platform/superproject/+/android-13.0.0_r41:dalvik/docs/debugger.html;l=44.

AOSP (2023e). Multiple refresh rate — Android Open Source Project, https://source.android.com/docs/core/graphics/multiple-refresh-rate.

AOSP (2023f). SurfaceFlinger and WindowManager — Android Open Source Project, https://source.android.com/docs/core/graphics/surfaceflinger-windowmanager.

AOSP (2025a). Android community and contacts — Android Open Source Project, https://source.android.com/docs/setup/community.

AOSP (2025b). Configuration overview — Android Open Source Project, https://source.android.com/docs/core/architecture/configuration.

AOSP (2025c). Frequently asked questions — Android Open Source Project, https://source.android.com/docs/setup/about/faqs.

Apple Inc. (2025). iOS - Apple, https://www.apple.com/ios.

ArrowOS (2023). ArrowOS, https://arrowos.net/.

BankMyCell (2025). Android vs. Apple Market Share: Leading Mobile OS (2025), https://www.bankmycell.com/blog/android-vs-apple-market-share/.

Barrera, D., Kayacik, H. G., van Oorschot, P. C., Somayaji, A. (2010). A methodology for empirical analysis of permission-based security models and its application to android, *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS '10, pp. 73–84.

Bian, B., Ma, X., Tang, H. (2021). The supply and demand for data privacy: Evidence from mobile apps, *SSRN Electronic Journal* pp. 1–80.

BlissLabs (2023). BlissRoms, https://blissroms.org/.

Burgess, M. (2022). How GDPR Is Failing, https://www.wired.co.uk/article/gdpr-2022.

CarbonROM (2023). Home - CarbonROM, https://carbonrom.org/.

Cawley, C. (2020). The 5 Best Custom Android ROMs Still Worth Trying, https://www.makeuseof.com/best-custom-android-roms/.

Chapman, G. (2022). 10 Best Free Custom ROM OS for Android, https://www.digitbin.com/best-custom-rom-android/.

Charan, K. V., Sharmila, S. P., Manjunath, A. S. (2014). Customizing aosp for different embedded devices, *International Conference on Computing for Sustainable Global Development*, pp. 259–264.

Congleton, N. (2023). The 6 Best Android ROMs of 2023, https://www.lifewire.com/best-android-roms-4777481.

Consumer Reports Inc. (2025). What Personal Data Stays on a Phone? - Consumer Reports, https://www.consumerreports.org/cell-phones-services/what-personal-data-stays-on-your-phone--a5605449074/.

Copperhead Limited (2023). Secure Android - CopperheadOS - Copperhead, https://copperhead.co/android/.

Counterpoint Technology Market Research (2023). Global Smartphone OS Market Share: Android vs. iOS, https://www.counterpointresearch.com/global-smartphone-os-market-share/.

crDroid Android (2023). crDroid.net - increase performance and reliability over stock Android for your device, https://crdroid.net/.

CyanogenMod open-source community (2023). CyanogenMod — Android Community Operating System, https://web.archive.org/web/20161225043707/https://www.cyanogenmod.org/.

Descendant (2023). Descendant, https://descendant.me/.

dotOS (2023). dotOS — HomePage, https://www.droidontime.com/.

e.foundation (2023). /e/OS - e Foundation - deGoogled unGoogled smartphone operating systems and online services - your data is your data, https://e.foundation/e-os/.

emteria GmbH (2023). emteria.OS — Embedded Android, https://emteria.com/emteria-os.

European Commission (2025). Data Protection, https://commission.europa.eu/law/law-topic/data-protection_en.

European Data Protection Supervisor (2025). Mobile devices — European Data Protection Supervisor, https://edps.europa.eu/data-protection/data-protection/reference-library/mobile-devices_en.

Fussell, S. (2022). The Most Important Things to Know About Apps That Track Your Location, https://time.com/6209991/apps-collecting-personal-data/.

GearJail (2023). Custom ROM List, https://gearjail.neocities.org/pda-os.

Ghosh, N. (2023). 11 Best Custom ROMs for Android in 2022, https://androidblog.org/best-custom-roms/.

Google (2023). Google, https://www.google.com/.

Google (2025a). Android - Secure & Reliable Mobile Operating System, https://www.android.com/.

Google (2025b). Android Apps on Google Play, https://play.google.com/.

GrapheneOS (2023). GrapheneOS: the private and secure mobile OS, https://grapheneos.org/.

Gupta, M., Bhardwaj, A., Garg, L. (2015). Custom rom's in android, *International Journal of Computer Science and Information Technologies* **6**(2), 1874–1875.

Haruka LLC (2023). Evolution X, https://evolution-x.org/.

Hazarika, S. (2021a). Here's why you should be wary of installing anything that sets SELinux to permissive, https://www.xda-developers.com/permissive-selinux-dangers-exploits/.

Hazarika, S. (2021b). Most popular custom ROMs for Android in 2023, https://www.xda-developers.com/most-popular-custom-roms-android/.

Holla, S., Katti, M. M. (2012). Android based mobile application development and its security, *International Journal of Computer Trends and Technology* **3**(3), 486–490.

Howarth, J. (2023). How Many People Own Smartphones (2023-2028), https://explodingtopics.com/blog/smartphone-stats.

Huawei Device Co., Ltd. (2023). EMUI 13 - HUAWEI Global, https://consumer.huawei.com/en/emui/.

Indus OS (2023). App Store India — Indian Android App Store — Indus OS, https://www.indusos.com/.

Jha, M. M. (2023). Custom roms on android devices, *International Journal of Innovative Science and Research Technology* **8**(7), 1893–1894.

jhenrique09 (2023). PixelExperience, https://pixelexperience.org/.

Kanthed, C., Yadav, Y. (2017). Building custom rom using aosp and improving ram usage in it, *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* **2**(5), 145–149.

Kaur, R., Singh, J., Kochhar, N., Singh, A., Kaur, M. (2017). Evaluation and analysis of android operating systems, *International Journal for Research in Applied Science and Engineering Technology* **5**(12), 2379–2382.

King, B. (2019). 12 Reasons to Install a Custom Android ROM, `https://www.makeuseof.com/tag/6-reasons-you-need-using-custom-rom/`.

Klosowski, T. (2022). How Mobile Phones Became a Privacy Battleground - and How to Protect Yourself, `https://www.nytimes.com/wirecutter/blog/protect-your-privacy-in-mobile-phones/`.

KPMG LLP (2023). Corporate data responsibility: Bridging the consumer trust gap, `https://kpmg.com/us/en/articles/2023/bridging-the-trust-chasm.html`.

Lanxon, A. (2022). Is It Safe to Use an Old or Used Phone? Here's What You Should Know, `https://www.cnet.com/tech/mobile/is-that-old-used-refurbished-android-phone-safe-use-what-you-should-know-security/`.

Manjrekar, S., Bhati, R. (2016). Custom rom - a prominent aspects of android, *International Journal of Advanced Research in Computer Engineering and Technology* **5**(5), 1590–1593.

microG Team (2025). microG Project, `https://microg.org/`.

MindTheGapps (2023). MindTheGapps: Download Custom Google Apps Package for Android, `https://mindthegapps.com/`.

Mithran, A. (2022). 13 Best Android Custom ROMs For 2022 That You Must Try, `https://fossbytes.com/android-custom-roms/`.

Mittal, A., Moorthy, A. K., Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain, *IEEE Transactions on Image Processing* **21**(12), 4695–4708.

Mohan, A. (2023). YAAP - Yet Another AOSP Project, `https://yaosp.dev/`.

Mohini, T., Srivastava, A. K., Nitesh, G. (2013). Review on android and smartphone security, *Research Journal of Computer and Information Technology Sciences* **1**(6), 12–19.

MoKee Open Source Project (2023). MoKee ROM, `https://www.mokeedev.com/en/`.

MUO (2025). Is Android Really Open-Source? And Does It Even Matter?, `https://www.makeuseof.com/tag/android-really-open-source-matter/`.

NikGApps (2023). NikGApps - Custom Google Apps Package!, `https://nikgapps.com/`.

Nirmala, T., Sathya, M. (2015). Android based mobile application development and its security, *International Journal of Trend in Research and Development* **Special Issue**(PCIT-15).

Norton Labs (2021). How much private information was gathered from my phone?, `https://www.nortonlifelock.com/blogs/norton-labs/private-information-gathered-phone`.

Oberlo (2023). How Many People Have Smartphones? [Jan 2023 Update] — Oberlo, `https://www.oberlo.com/statistics/how-many-people-have-smartphones`.

OffSec Services Limited (2023). Get Kali — Kali Linux, `https://www.kali.org/get-kali/#kali-mobile`.

Okediran, O., Arulogun, T., Ganiyu, A., Oyeleye, A. (2014). Mobile operating systems and application development platforms: A survey, *International Journal of Advanced Networking and Applications* **6**(1), 2195–2201.

OmniROM (2023). OmniROM, `https://omnirom.org/`.

OnePlus (2023). OnePlus OxygenOs - OnePlus (United States), `https://www.oneplus.com/us/oxygenos`.

OPPO (2023). ColorOS 12 — OPPO Global, `https://www.oppo.com/en/coloros/`.

Paranoid Android (2023). Paranoid Android, `https://paranoidandroid.co/`.

Parekh, M. R. (2022). Custom android rom's, *International Research Journal of Engineering and Technology* **9**(3), 501–504.

PassMark Software (2023a). PassMark Android Benchmarks General Test Info, https://www.androidbenchmark.net/cpu_test_info.html.

PassMark Software (2023b). PassMark PerformanceTest - Apps on Google Play, https://play.google.com/store/apps/details?id=com.passmark.pt_mobile.

Patel, P. (2023). 5 De-Googled Android-based Operating Systems to Free Your Smartphone from Google and other Big Tech, https://itsfoss.com/android-distributions-roms/.

PCMag (2025). Definition of Android ROM — PCMag, https://www.pcmag.com/encyclopedia/term/android-rom.

Peng, H. (2023). CN110175075A - Android system Memory Optimize Method and device - Google Patents, https://patents.google.com/patent/CN110175075A/en.

PhoneArena (2025). All Phone Brands - PhoneArena, https://www.phonearena.com/phones/manufacturers.

Potato Open Sauce Project (2023). PotatoHub, https://www.potatoproject.co/.

Prasal, A. (2023). Havoc-OS, https://havoc-os.com/.

Rahul, P., Kr. Das, R., Anand, R. R. (2014). Rooting of android devices and customized firmware installation and its calibre, *International Journal of Scientific Engineering and Technology* **3**(5), 553–556.

Rajendran, J., Banerjee, B. (2021). Optimize Garbage Collector Parameters for High Memory in Chrome* OS-ARC++, https://01.org/blogs/jaishank/2021/optimize-garbage-collector-parameters-high-memory-chrome-os-arc.

Rajput, P., Koraganti, V., Champaty, B. (2018). Custom rom, *International Journal of Advances in Science Engineering and Technology* **6**(4), 140–143.

Regmi, K. (2022). Best Custom ROMs for Android You Should Try in 2022, https://www.xtechkr.com/best-custom-roms-for-android/.

Replicant (2023). Replicant, https://replicant.us/.

Resurrection Remix OS (2023). Resurrection Remix OS — Get Resurrected, https://resurrectionremix.com/.

Samsung Electronics Co., Ltd. (2023). One UI — Apps - The Official Samsung Galaxy Site, https://www.samsung.com/global/galaxy/apps/one-ui/.

Saqr, M. (2022). Is gdpr failing? a tale of the many challenges in interpretations, applications, and enforcement, *International Journal of Health Sciences* **16**(5), 1–2.

Sasi Kumar, P., Karthi, A., Surya, S., Dinesh Kumar, S. (2018). Android rooting and custom rom, *International Journal of Innovative Research in Science and Technology* **4**(9), 1–3.

SELinux (2024). SELinux Project, https://github.com/SELinuxProject.

Sha, A. (2023). 15 Best Custom ROMs for Android You Can Install, https://beebom.com/best-custom-roms-android-phones/.

Sharma, A. (2021). We asked, you told us: Here's how much cameras influence your phone purchase, https://www.androidauthority.com/smartphone-camera-poll-results-1204074/.

Sharma, A., Nimawat, S. (2019). Customizing lineage for different embedded devices, *International Conference on Advanced Computing Networking and Informatics*, pp. 389–395.

SHIFT GmbH (2023). ShiftOS Downloads, https://downloads.shiftphones.com/.

Shreyas, S. (2020). Developing custom rom based on android using aosp, *International Journal for Research in Applied Science and Engineering Technology* **8**(8), 709–714.

Singh, R. (2014). An overview of android operating system and its security features, *International Journal of Engineering Research and Applications* **4**(2), 519–521.

Smartisan Technology Co., Ltd. (2023). Smartisan OS / OS-6.x, https://www.smartisan.com/os.

SparkOS (2023). SparkOS - Home, https://spark-os.live/.

StarDust (2025). The Mini-Guide to Smartphone Fragmentation, https://www2.stardust-testing.com/en/the-mini-guide-to-smartphone-fragmentation.

Suleman, M., Zhong, X., Sun, Y. (2020). Empirical research and auxiliary tool for custom android roms, *International Symposium on Computer Engineering and Intelligent Communications*, pp. 14–18.

The LineageOS Project (2023). LineageOS - LineageOS Android Distribution, https://lineageos.org/.

The MathWorks Inc. (2025). Train and Use No-Reference Quality Assessment Model, https://www.mathworks.com/help/images/train-and-use-a-no-reference-quality-assessment-model.html.

The Open GApps Team (2025). The Open GApps Project, https://opengapps.org/.

Thomas, D. (2021). Check Your Android Security Patch Level to See if You're Protected Against the Latest Vulnerabilities, https://android.gadgethacks.com/how-to/check-your-android-security-patch-level-see-if-youre-protected-against-latest-vulnerabilities-0384695/.

Threat Intelligence (2018). Mobile Privacy: What Do Your Apps Know About You?, https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence/mobile-privacy-apps.

Time News (2022). The best Android ROMs for all types of phones and how to install them, https://time.news/the-best-android-roms-for-all-types-of-phones-and-how-to-install-them/.

vivo (2023). vivo Funtouch OS 13 - Tailored for Newest Android 13 — vivo EU, https://www.vivo.com/eu/funtouch.

Vonau, M. (2021). Going Google-less: How to install a custom Android ROM with no Google apps or services, https://www.androidpolice.com/2021/03/27/install-use-custom-rom-no-google-apps/.

Wise, J. (2023). Smartphone Statistics 2023: How Many People Have Smartphones?, https://earthweb.com/smartphone-statistics/.

Wodinsky, S. (2022). The Hidden Failure of the World's Biggest Privacy Law, https://gizmodo.com/gdpr-iab-europe-privacy-consent-ad-tech-online-advertis-1848469604.

Xiaomi (2023). Miui 14, https://www.mi.com/global/miui.

Xtended (2023). Xtended - Custom ROM Redefined!, https://project-xtended.org/.

ZTE (2023). ZTE Unveiled New Consumer Devices and its Full-Scenario Intelligent Ecosystem 2.0 at MWC 2023, https://ztedevices.com/en-gl/news/2023/03/zte-unveiled-new-consumer-devices-and-its-full-scenario/.

Received June 5, 2025 , revised October 8, 2025, accepted November 14, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, pp. 919–932
https://doi.org/10.22364/bjmc.2025.13.4.09

Dictionary Attack with Transformed Russian Words using QWERTY Keyboard Layout

Lea MÜLLER[1], Aušrius JUOZAPAVIČIUS[2], Volodymyr OKHRIMCHUK[3], Stefan SÜTTERLIN[1,4]

[1] Albstadt-Sigmaringen University, Albstadt, Germany
[2] General Jonas Žemaitis Military Academy of Lithuania, Vilnius, Lithuania
[3] Korolov Zhytomyr Military Institute, Zhytomyr, Ukraine
[4] Østfold University College, Halden, Norway

muellerlea@hs-albsig.de, ausrius.juozapavicius@lka.lt, okhrimchuk84@ukr.net, suetterlin@hs-albsig.de

ORCID 0009-0009-6538-8046, ORCID 0000-0002-8852-8605, ORCID 0000-0001-7518-9993, ORCID 0000-0002-4337-1296

Abstract. Despite the known vulnerabilities of passwords, the username-password combination remains the most widely used authentication method. Many users still choose simple, memorable passwords, making them susceptible to dictionary attacks. These attacks are especially effective when using target-specific wordlists. This paper introduces a novel wordlist tailored to Russian-speaking users who may type passwords using the QWERTY layout while writing in Russian, leading to seemingly random character strings. Based on this assumption, a dictionary of transformed Russian words was compared with one million unique Russian passwords. The analysis revealed that around 1% of the passwords exactly matched transformed entries, and an additional 6% partially matched, supporting the effectiveness of this new wordlist approach.

Keywords: Dictionary Attack, Passwords, Password Security, Transformed Passwords, Keyboard Layout

1 Introduction

Notwithstanding the vulnerability of passwords to a variety of attacks (Bošnjak et al., 2018; Alkhwaja et al., 2023; Klein, 1990), username and password remain the most prevalent authentication method globally (Statista, 2024; Wash and Rader, 2021; Taneski et al., 2019). While the necessity for complex passwords that are difficult to guess is widely acknowledged (Wash and Rader, 2021), a considerable number of users opt for straightforward passwords that can be easily recalled (Bryant and Campbell, 2006;

Woods and Siponen, 2018; Taneski et al., 2019; Shen et al., 2016), such as simple keyboard patterns (Shen et al., 2016; Klein, 1990; Ur et al., 2015), dictionary words (Shen et al., 2016; Klein, 1990), or a combination of a word and a basic pattern of numbers (Wash and Rader, 2021; Ur et al., 2015).

The potential for exploitation of a given password varies depending on its length, complexity, and the specific attack method employed. Examples of the various types of attacks include exhaustive search or dictionary attacks. An exhaustive search, or brute force attack, involves systematically testing all possible combinations of characters for an unknown password (Bryant and Campbell, 2006). This type of attack is therefore not a suitable option for long passwords (Alkhwaja et al., 2023; Bryant and Campbell, 2006). Dictionary attacks employ wordlists for the purpose of guessing passwords (Alkhwaja et al., 2023; Yan et al., 2000). This method poses a risk in instances where users select passwords that can be found in dictionaries or other wordlists, such as those resulting from previous password breaches. Although this type of attack is less time-consuming than an exhaustive search (Bryant and Campbell, 2006), it is not feasible for passwords that are not included in the wordlist used in the attack (Alkhwaja et al., 2023; Bryant and Campbell, 2006). However, a combination of different approaches can be used to optimise password guessing attacks.

This paper will focus on a dictionary attack using a novel wordlist based on the transformation of Russian words through the substitution of Cyrillic letters with the corresponding characters as found on a QWERTY keyboard layout. The aim of this paper is to ascertain whether the proposed method of selecting transformed Russian words as passwords is a common practice among Russian-speaking users and, if so, whether it could be employed to enhance dictionary attacks by utilising target-specific wordlists. The efficacy of this approach is evaluated through the use of Russian words transformed through a simple substitution of Cyrillic letters with the characters they share a key with on the QWERTY keyboard layout. However, it is conceivable that the same approach can be applied to further keyboard layouts.

The paper is structured as follows: Section 2 will describe the concept of dictionary attacks and provide examples of potential enhancements to this attack type. Section 3 will put forth a novel approach for dictionary attacks against Russian-speaking users. The construction of the wordlist in this attack is based on the assumption that Russian-speaking users select a Russian word as their password and type it in accordance with the Russian keyboard layout, despite having their keyboard configured to the QWERTY layout, thereby creating seemingly random patterns of characters. This assumption is tested against a list of the one million most frequent passwords used by a Russian-speaking audience. Section 4 will present a summary of the findings and offer an outlook on potential future work.

2 Dictionary Attacks

Dictionary attacks represent a category of attack employed for the purpose of guessing unknown passwords or usernames. This is achieved through a systematic process of testing words contained in a dictionary or wordlist (Alkhwaja et al., 2023; Yan et al., 2000). Dictionary attacks are based on the premise that passwords chosen by users are

susceptible to being easily guessed. This is particularly the case when words contained in a dictionary are used as a password. Additionally, as many users reuse the same or similar passwords for multiple purposes (Wash and Rader, 2021; Bryant and Campbell, 2006; Woods and Siponen, 2018; Taneski et al., 2019; Wash et al., 2016), dictionary attacks utilising password lists from previous security breaches can be employed to rapidly deduce a password (Bryant and Campbell, 2006; Taneski et al., 2019).

The selection of secure passwords is a fundamental aspect of information security, yet users continue to rely on easily memorable passwords (Bryant and Campbell, 2006; Woods and Siponen, 2018; Taneski et al., 2019; Shen et al., 2016). Although the use of dictionary words may safeguard a password from being brute-forced (assuming a sufficiently lengthy word is selected), this approach leaves the password vulnerable to dictionary attacks.

Consequently, users seek methods to circumvent the use of words that precisely match dictionary entries while simultaneously striving for a password that is memorable and straightforward to recall. For example, one frequently employed strategy is to append digits to a dictionary word (Wash and Rader, 2021; Ur et al., 2015), resulting in a password such as *password123*.

The approach presented in this paper is based on the premise that Russian-speaking users may select a Russian word as their password and type it in accordance with the Russian keyboard layout, despite having their keyboard configured to the QWERTY keyboard layout. This approach is discussed in detail in Section 3, along with a consideration of how it may be applied in the context of dictionary attacks.

2.1 Optimisation of Dictionary Attacks

Although dictionary attacks, which involve using all entries in a dictionary to guess passwords, are effective when a password matches an existing entry exactly (Alkhwaja et al., 2023), they are less successful when users make minor alterations to their passwords, such as adding an additional character (Wash and Rader, 2021; Ur et al., 2015). Consequently, dictionary attacks can be enhanced by combining them with other types of attacks. The following section outlines some of the approaches that can be employed to improve the efficacy of dictionary attacks.

The optimisation of a dictionary attack through the integration of an exhaustive search methodology entails the utilisation of words from a pre-defined wordlist, which are employed to guess passwords. In contrast to a traditional dictionary attack, this process involves the combination of these words with additional characters, thereby expanding the search space. Many users opt for passwords that comprise a word in conjunction with additional characters, such as letters, numbers, or a year (Wash and Rader, 2021; Rinn et al., 2015; Ur et al., 2015). To illustrate, a combination of a dictionary attack with an exhaustive search methodology would not only try the word *password* but also combinations of *password* with other characters, such as *password1* or *password12*.

Some users create passwords by repeating or concatenating words (Bošnjak et al., 2018; Bryant and Campbell, 2006; Shen et al., 2016; Klein, 1990). Consequently, dictionary attacks can be optimised to not only repeat a dictionary entry once, but to test repetitions of dictionary words, such as *passwordpassword*, or to combine two or more

words from a dictionary, such as *hellopassword*. Consequently, the combination of dictionary attacks with an exhaustive search methodology renders passwords comprising a word and additional characters, repeated words, or concatenated words susceptible to relatively straightforward exploitation.

An additional method for enhancing the efficacy of dictionary attacks is the utilisation of target-specific wordlists. To illustrate, a German dictionary could be employed in a dictionary attack against German-speaking users. Additional customisation is possible should the attacker have access to supplementary personal information regarding the target (Bryant and Campbell, 2006; Klein, 1990; Taneski et al., 2019), such as the names or dates of birth of family members (Shen et al., 2016; Rinn et al., 2015). This can be achieved through the use of open-source intelligence (OSINT).

This paper puts forth a method for optimising dictionary attacks through the utilisation of a novel type of target-specific wordlist. This new approach is predicated on the substitution of the Cyrillic alphabet with Latin letters and a select set of special characters. Section 4 will address the potential applications of this approach in enhancing the efficacy of dictionary attacks.

3 Dictionary Attack with Transformed Russian Words using QWERTY Keyboard Layout

3.1 Premise

In addition to the aforementioned approaches for optimising dictionary attacks, a novel approach is presented and its efficacy is evaluated in the following section.

In the area of password usage and dictionary attacks, research is primarily focused on an English-speaking audience (Wang et al., 2019), while languages that use non-ASCII characters have been less extensively investigated. The approach presented in this article is based on the premise that Russian-speaking users may input their passwords according to the Russian keyboard layout, while having their keyboard configured to the QWERTY layout. Typing a word in this manner would result in a seemingly random sequence of characters. To test this assumption, the Russian language was selected for analysis. The rationale for this decision is threefold. Firstly, Russian is a widely used language (Zeidan, 2023b,a). Secondly, it does not use the QWERTY or a similar keyboard layout (Wilcock and Dempsey, 2024; Unicode, 2021). Thirdly, all Cyrillic letters are mapped to a single ASCII character on the keyboard (compare Figure 1).

As there are multiple keyboard layouts for Cyrillic characters, the most prevalent layout was selected for this study. According to Chumachenko and Burkov, the ЙЦУКЕН (JCUKEN) keyboard layout, a layout specifically adapted for the Russian language, is the most popular Cyrillic layout (Chumachenko and Burkov, 2023). This is also the default Russian keyboard layout available in different operating systems, such as Windows (Wilcock and Dempsey, 2024), ChromeOS (Unicode, 2021) or Ubuntu. The Russian keyboard layout, according to (Wilcock and Dempsey, 2024; Unicode, 2021), is illustrated in Figure 1. In this illustration, the QWERTY keyboard layout is presented in black letters, with the Russian layout superimposed in red. When typing

Fig. 1. QWERTY keyboard layout with Russian keyboard layout superimposed in red letters.

a Russian word according to the Russian layout with the keyboard configuration set to QWERTY, a simple substitution of Cyrillic letters for Latin letters, as well as some special characters, will occur. This implies that each Russian letter is precisely matched to a single letter or special character within the QWERTY keyboard layout. To illustrate, the Cyrillic letter А is mapped to the Latin letter *F*.

Table 1 illustrates the correspondence between the Cyrillic letters of the Russian keyboard layout and the characters of the QWERTY keyboard layout, namely the characters by which they are substituted. Accordingly, the Russian word for *password*, пароль, would be substituted with the seemingly random sequence of characters *gfhjkm*.

Table 1. Cyrillic letters and the characters by which they are substituted (QWERTY).

Cyrillic	ё	й	ц	у	к	е	н	г	ш	щ	з	х	ъ	ф	ы	в	а	п	р	о	л	д	ж	э	я	ч	с	м	и	т	ь	б	ю
QWERTY	‘	q	w	e	r	t	y	u	i	o	p	[	]	a	s	d	f	g	h	j	k	l	;	’	z	x	c	v	b	n	m	,	.

3.2 Preparation

In order to assess the efficacy of this novel approach, a list of the 50,000 most prevalent Russian words, obtained from (Hingston, 2018), was transformed in accordance with the aforementioned substitution. This entailed the replacement of the Cyrillic characters with the corresponding characters on the QWERTY keyboard layout. The result was a list of 50,000 seemingly random sequences of Latin letters and special characters.It should be noted that in this step all Cyrillic letters were transformed to lowercase Latin letters and no uppercase letters were used.

The first part of this study compares the list of transformed words with a list of the one million most frequently used passwords among Russian-speaking users (Sharsi1, 2019) to identify all exact matches between a transformed Russian word and an entry in the password list. The second part of the study compares the same two lists, this

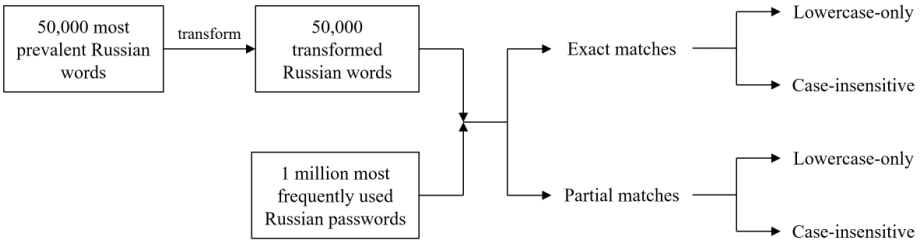


Fig. 2. The steps of the analysis.

time considering partial matches. In the following, the term 'exact match' is used to describe transformed Russian words that are present in the password list in the same form as stated, without the addition of any characters or numbers. To illustrate, the transformed word *gfhjkm* is only deemed an exact match if it occurs in the password list identically. *gfhjkm123*, however, would not be regarded as an exact match. In contrast, 'partial matches' may encompass additional characters. Consequently, both *gfhjkm* and *gfhjkm123* would be identified as a match in this instance.

In the search for both exact and partial matches, lowercase-only and case-insensitive matches were sought in two stages. As the list comprising the 50,000 transformed Russian words was created exclusively using lowercase letters, in the initial phase of each analysis, only lowercase matches are taken into consideration. In the second step of each analysis, all matches regardless of case are searched for. To illustrate, in the lowercase-only search, only *gfhjkm* would be considered a match, whereas in the case-insensitive search, versions such as *Gfhjkm* and *GFHJKM* would additionally be considered a match. The steps of the analysis are illustrated in Figure 2.

To reduce the number of false positives, Russian words comprising a length of less than five characters were excluded from the analysis. In this context, a false positive refers to a situation where a word is identified as a match, despite it being a random sequence or one selected for a different purpose, such as keyboard patterns. Given the prevalence of such instances with shorter words, words comprising four characters or less were excluded. This resulted in the exclusion of 1,565 exact matches with a length of less than five characters when considering lowercase-only matches, and the exclusion of 1,900 case-insensitive matches.

3.3 Analysis

To evaluate the aforementioned premise, the list of 50,000 transformed Russian words was compared with a list of the one million most frequently used passwords among Russian speakers, retrieved from (Sharsi1, 2019). As indicated in the GitHub repository, this is a list of common passwords used by Russian speakers. The data set was uploaded in 2019 and was compiled from a number of data breaches, with the passwords ordered according to the frequency of occurrence. In addition to the passwords, a count is provided for each password. The most frequently used password is *qwerty*, with a count of 8,226,502, and the least frequently used passwords have a count of 65.

In the first step of the analysis, only exact matches were considered. In a subsequent step, partial matches were also taken into consideration. To minimise the number of false positives, only those matches with a minimum length of five characters were considered.

3.3.1 Searching for Exact Matches To perform a comparative analysis between the 50,000 transformed Russian words and the list of the one million most frequently used Russian passwords, the logic illustrated in Listing 1.1 was used. Given two sets, one containing the transformed Russian words and one containing the Russian passwords, computing the intersection of these two sets yields all exact matches. Matches comprising four or fewer characters were discarded.

This yielded 9,580 exact matches, representing approximately 0.96% of the one million most frequent passwords used by Russian speakers.

Listing 1.1. Pseudocode showing the logic for finding all exact matches of a transformed Russian word with a Russian password.

```
temp = set_intersection(transformed Russian words, Russian passwords)
common words = set()
for word in temp:
  if length(word)>=5:
    add word to common words
```

As the 9,580 matches comprise solely those passwords consisting of lowercase letters, the analysis was repeated with a case-insensitive comparison. To achieve this, all Russian passwords were converted to lowercase and subsequently compared with the 50,000 transformed Russian words. This implies that, in addition to lowercase passwords such as *gfhjkm*, uppercase passwords such as *Gfhjkm* or *GFHJKM* were also identified as matches. This search yielded 12,448 results, representing approximately 1.24% of the one million most frequent passwords used by Russian speakers. This number also includes the 9,580 lowercase-only matches.

3.3.2 Searching for Partial Matches In addition to exact matches, the lists of transformed Russian words and one million passwords were compared to identify instances where a transformed Russian word forms a part of a password in the password list.

The logic used to identify all passwords in the list of the one million most frequently used Russian passwords that at least partially consist of a transformed Russian word is shown in Listing 1.2. For each of the transformed Russian words with five or more characters, it was tested whether it is part of any of the Russian passwords.

The comparison yielded 72,950 results, which also included all exact matches. Upon exclusion of the 9,580 previously identified exact matches, the remaining set comprises 63,370 partial matches, representing approximately 6.34% of the one million most frequently used Russian passwords.

Listing 1.2. Pseudocode showing the logic for finding all partial matches, i.e., matches where a transformed Russian word is part of a Russian password.

```
transformed words = transformed Russian words
```

```
passwords = Russian passwords

subwords = set()
for transformed word in transformed words:
  if length(transformed word)>=5:
    for password in passwords:
      if transformed word is part of password:
        add "transformed word | password" to subwords
```

As with the search for exact matches, analysis was repeated with a case-insensitive comparison. To achieve this, all Russian passwords were converted to lowercase and then compared with the 50,000 transformed Russian words once more. Again, the logic presented in Listing 1.2 was used. This yielded 81,031 results, which also included all exact matches. Upon exclusion of the 12,448 previously identified exact matches, the remaining set comprises 68,583 partial matches, representing approximately 6.86% of the one million most frequently used Russian passwords.

As a common approach among many users is to create a password by adding numbers at the beginning or end of a word, such as a birth year or a simple pattern like *123*, the list with partial matches was subsequently filtered for passwords that consist of a transformed Russian word with preceding or following numbers.

The logic used to identify these matches is presented in Listing 1.3. This assumes that all partial matches have been identified according to Listing 1.2, implying that the transformed word that matches a password is already known. Initially, all numerical characters are removed from the password. A match is deemed to be present if the transformed Russian word and the password devoid of numerals are found to be identical. In this instance, exact matches, defined as passwords that corresponded precisely to a transformed Russian word prior to the removal of numbers, were not deemed to be matches.

The comparison yielded 21,355 lowercase-only and 23,245 case-insensitive results, indicating that approximately 2.32% of the one million unique passwords were a combination of a transformed Russian word with preceding or following numbers, such as *gfhjkm1977*.

Other observed passwords were combinations of several transformed Russian words (for example, the transformed words *rfrjqnj* and *gfhjkm* were combined to form *rfrjqnjgfhjkm*) or repeated stringing together of the same transformed Russian word (for example, *gfhjkm* was repeated to form *gfhjkmgfhjkm*).

Listing 1.3. Pseudocode showing the logic for filtering all partial matches that are a combination of a transformed Russian word with added numerals. In this case, the matching transformed word is already known, as all partial matches have been identified previously.

```
def remove_numbers(string):
  return string with numbers removed

passwords = passwords that consist in part of a transformed word

matches with numerals = set()
for password in passwords:
```

```
if remove_numbers(password) equals transformed word and password does
    ↪ not equal transformed word:
  add "transformed word | password" to matches with numerals
```

3.4 Results

Table 2 presents the results of the comparative analysis of the 50,000 most frequently used Russian words, transformed in accordance with the proposed methodology, and the one million most frequently used passwords by Russian speakers.

Table 2. Results as absolute values and as percentages of the 1 million most frequenlty used Russian passwords. It should be noted that case-insensitive matches also include all lowercase-only matches.

	Lowercase-only	Case-insensitive
Exact match	9,580 (0.96%)	12,448 (1.24%)
Partial match (all)	63,370 (6.34%)	68,583 (6.86%)
Partial match (added numbers)	21,355 (2.14%)	23,245 (2.32%)
Total	72,950 (7,30%)	81,031 (8.10%)

A comparison between the 50,000 transformed Russian words and the one million passwords yielded 9,580 exact matches and 63,370 partial matches, when only passwords composed of lowercase letters were considered. When passwords comprising both lowercase and uppercase letters were considered, 12,448 exact matches and 68,583 partial matches were identified. An exact match was defined as a transformed Russian word with a minimum length of five characters that corresponded exactly to an entry in the password list, whereas a partial match was defined as a transformed Russian word that was part of an entry in the password list. Among the partial matches, 23,245 were instances where a transformed Russian word with digits added was used as a password.

Table 2 additionally illustrates the proportion of the one million most frequently used Russian passwords that align with the proposed scheme. While only a minor proportion of the passwords (approximately 1.24%) are an exact match for a transformed Russian word, a larger proportion (approximately 6.86%) consist of a transformed Russian word in part.

It is important to note that only unique passwords were considered in this analysis, and the frequency with which these passwords are used was not taken into account. The presented percentages do not necessarily reflect the actual prevalence of users employing this method. Rather, they represent the proportion of unique passwords within the list of the one million most frequently used Russian passwords that correspond to this method.

Figure 3 shows the distribution of the number of exact matches, both in lowercase-only and case-insensitive categories, across a range of password lengths. The lengths of exact matches ranged from five characters (due to the exclusion of matches with four or

fewer characters from the analysis in order to reduce the number of false positives) to 21 characters.

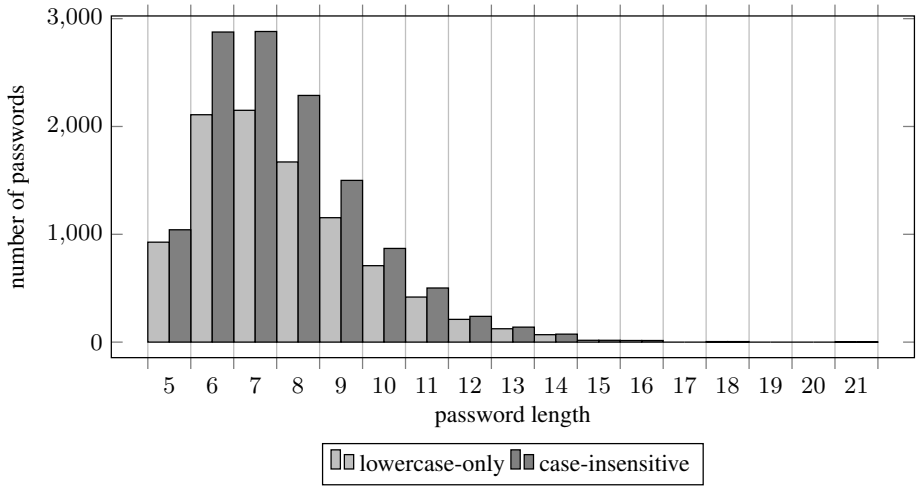


Fig. 3. Distribution of exact matches across different password lengths. Lowercase-only: passwords that only use lowercase letters. Case-insensitive: passwords that use lowercase and/or uppercase letters.

4 Discussion

A novel approach to the creation of target-specific wordlists for dictionary attacks against Russian-speaking users has been presented. This approach is predicated on the assumption that Russian speakers may opt for a Russian word as their password and input it in accordance with the Russian keyboard layout, despite having their keyboard configured to the QWERTY layout. This would result in a seemingly random sequence of Latin letters and special characters.

In order to determine whether this method is employed by Russian speakers, a list comprising the 50,000 most frequently used Russian words was transformed in accordance with the aforementioned scheme and subsequently compared with a list of the one million most frequently used passwords by Russian-speaking users. In order to reduce the number of false positives, only transformed Russian words comprising a minimum of five characters were considered.

The comparison of transformed Russian words with the one million most frequently used Russian passwords revealed that this approach is employed by a subset of Russian-speaking users. Approximately 1% of the passwords were an exact match with one of the transformed Russian words, while an additional approximately 6% of the passwords at least partially consisted of a transformed Russian word.

While the passwords that represented exact matches could be guessed using a traditional dictionary attack, which simply tests the words contained in a wordlist, the passwords that partially consist of transformed Russian words could be attacked by combining a dictionary attack with an exhaustive search approach. One such approach would be to add additional numbers at the beginning or end of a word. The findings suggest that utilising a dictionary of transformed Russian words could facilitate more effective dictionary attacks against Russian-speaking users. Furthermore, they illustrate how the deployment of target-specific wordlists can enhance the efficacy of dictionary attacks.

The presented technique introduces a wordlist that has not previously been considered in the context of dictionary attacks. The proposed approach represents a novel contribution to the field of dictionary attacks. Rather than merely considering existing wordlists for target-specific attacks, such as using a German dictionary for an attack on German-speaking targets, it suggests an entirely new set of words. This is achieved by substituting Cyrillic letters with Latin letters and some special characters.

This reinforces the necessity for users to select robust passwords and demonstrates that circumvention strategies, such as the one outlined in this paper, will inevitably result in the generation of weaker passwords that are susceptible to specific forms of attack.

4.1 Threats to Validity

This section provides a structured analysis of the threats to internal, external, and construct validity.

Internal validity refers to the extent to which the findings can be attributed to the premise of this work, namely that transformed Russian words appear in password lists because users type Russian words while having their keyboard configured to the QWERTY layout. One potential threat arises from alternative explanations for the observed matches. Some matched sequences may originate from simple keyboard patterns or random character strings, rather than from the hypothesised behaviour. Even after excluding strings of fewer than five characters to reduce accidental matches, longer coincidental matches may still occur. A second threat relates to the quality of the password list used in this study. The dataset comprising the one million most commonly used Russian passwords was compiled from publicly available breaches in 2019, and its accuracy cannot be verified independently. Noise, bias or artefacts in the dataset may influence the number and nature of detected matches, thereby affecting internal validity. A third threat stems from the incomplete coverage of the Russian lexicon. The present study relies on a list of the 50,000 most frequent Russian words, which does not include names, slang, domain-specific terms and less common words often used in passwords. This may lead to an underestimation of the prevalence of the hypothesised behaviour. Additionally, reported rates are sensitive to analytical choices, such as the minimum length of five characters and how case is handled. These analytical decisions introduce additional uncertainty into the interpretation.

External validity refers to the extent to which the results can be generalised beyond the specific data and conditions of this study. One threat that affects external validity is temporal generalisability. The password list is based on data that was leaked and

collected prior to 2019, so it may not accurately reflect present-day behaviour, websites, or demographics. Password behaviour and security practices may have changed since then. Therefore, the prevalence of the studied phenomenon in contemporary settings may differ. Another threat concerns population validity. The dataset only includes Russian-speaking users from unspecified services that have been affected by historical breaches. It does not necessarily reflect the broader population of Russian-speaking users or subgroups such as mobile users, different age groups, or users of specific platforms. However, concerns about population validity affect the majority of studies analysing password behaviour based on leaks, as analyses are limited by the availability of lists of leaked passwords. In this case, it was essential to employ a password list for a specific user group, Russian-speakers, which complicated the search for password lists further. Generalising the findings to other keyboard layouts and Cyrillic languages poses another threat. While the present study focuses on the Russian JCUKEN keyboard layout, other layouts map Cyrillic characters differently. The same substitution patterns cannot be assumed to apply across languages or layouts. Although this study relies on publicly available datasets and provides pseudocode, its results may be contingent on the specific list of leaked passwords and the analytical choices employed.

Construct validity refers to whether the operationalisation used in the study accurately captures the behaviour of Russian-speaking users entering passwords using the QWERTY keyboard layout. A primary threat to this arises from the definition of a match. In the present study, matches were defined as exact or partial occurrences of transformed Russian words within leaked passwords. However, these matches may not always reflect the deliberate selection of a Russian word typed while having the keyboard configured to the QWERTY layout. Some matches may be the result of accidental substrings or other password selection habits unrelated to the behaviour under analysis. A second threat is the limitation to a list of 50,000 Russian words. Because this wordlist does not contain a comprehensive set of words, such as names, slang or domain-specific terms that users may select as their passwords, the prevalence of the hypothesised behaviour may be incorrectly represented. These threats may lead to both false negatives (missed legitimate instances) and false positives (matches that do not reflect the behaviour under analysis). A further threat arises from preprocessing decisions, such as case normalisation, the exclusion of short passwords and the criteria for partial matches. These decisions influence how the construct is instantiated in the analysis. As partial matches include passwords that merely contain the transformed sequence anywhere, the operationalisation may at times be broader than the underlying construct. Finally, as the study infers behaviour purely from leaked passwords, there is no direct behavioural validation available. Consequently, the study provides an inferred approximation rather than a directly measured behavioural phenomenon.

4.2 Additional Limitations and Future Work

In addition to the threats to the validity of this research, as discussed in Section 4.1, investigating one language and keyboard layout constitutes a primary limitation of the research. While analysing additional languages and keyboard layouts would significantly strengthen the claim of a novel type of target-specific wordlist presented in this article,

no such analysis was conducted. This was due to the fact that analyses of password behaviour are dependent on the availability of password lists. In this case, password lists of a specific user group (i.e., of a specific language) are required, which further limits their availability. Most password lists do not contain passwords from a single user group characterised by their language. Furthermore, not all languages and keyboard layouts are suitable for the presented approach, as an exact mapping from the language-specific keyboard layout to the QWERTY layout is required. It can generally be assumed that the approach can be applied to all languages and keyboard layouts where the language-specific characters map exactly to one letter or special character on the QWERTY keyboard layout. However, this assumption is subject to the condition that users exhibit the hypothesised behaviour when selecting passwords. Due to the limited scope of this article, we encourage further research analysing password lists of specific languages.

The aim of this study was to ascertain whether the proposed methodology of selecting transformed Russian words is employed by Russian-speaking users. However, only a relatively limited set of Russian passwords, comprising one million unique entries, was subjected to analysis. Moreover, the number of matches identified merely reflects the number of distinct passwords that adhere to this scheme. It may not accurately reflect the proportion of users who employ this strategy when creating passwords.

Further tests could be conducted using a larger set of Russian passwords to enhance the reliability of the findings. An investigation into the frequency of the passwords could provide insight into the prevalence of this scheme among Russian-speaking users. Furthermore, the password list employed in this study is from 2019, thus the utilisation of a more recent dataset could also enhance the reliability of the findings.

There are numerous alternative keyboard layouts in use beyond the QWERTY layout, including Ukrainian, Kazakh, or Belarusian (Chumachenko and Burkov, 2023). To determine whether this same approach to selecting transformed words as passwords is also utilised by other audiences, further analysis could be conducted on additional password datasets to confirm or refute the generalisability of the approach.

References

Alkhwaja, I., Albugami, M., Alkhwaja, A., Alghamdi, M., Abahussain, H., Alfawaz, F., Almurayh, A., Min-Allah, N. (2023). Password cracking with brute force algorithm and dictionary attack using parallel programming, *Applied Sciences* **13**(10), 5979.

Bošnjak, L., Sreš, J., Brumen, B. (2018). Brute-force and dictionary attack on hashed real-world passwords, *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1161–1166.

Bryant, K., Campbell, J. (2006). User behaviours associated with password security and management, *Australasian Journal of Information Systems* **14**(1).

Chumachenko, O., Burkov, A. (2023). Development of an efficient ukranian keyboard layout using a genetic algorithm, *Electronics and Control Systems* **2**(76), 35–39.

Hingston (2018). 50000-russian-words-cyrillic-only, GitHub Repository. https://github.com/hingston/russian/blob/master/50000-russian-words-cyrillic-only.txt

Klein, D. V. (1990). Foiling the cracker: A survey of, and improvements to, password security, *Proceedings of the 2nd USENIX Security Workshop*, pp. 5–14.

Rinn, C., Summers, K., Rhodes, E., Virothaisakun, J., Chisnell, D. (2015). Password creation strategies across high- and low-literacy web users, *Proceedings of the Association for Information Science and Technology* **52**(1), 1–9.

Sharsi1 (2019). stat_russkiwlst_top_1m, GitHub Repository. https://github.com/sharsi1/russkiwlst/blob/master/stat_russkiwlst_top_1M.txt

Shen, C., Yu, T., Xu, H., Yang, G., Guan, X. (2016). User practice in password security: An empirical study of real-life passwords in the wild, *Computers & Security* **61**, 130–141.

Statista (2024). Authentication methods deployment worldwide 2023. https://www.statista.com/statistics/1441144/companies-authentication-methods-deployment-status-worldwide/

Taneski, V., Heričko, M., Brumen, B. (2019). Systematic overview of password security problems, *Acta Polytechnica Hungarica* **16**(3), 143–165.

Unicode (2021). Layouts: Russian (ru). https://www.unicode.org/cldr/charts/40/keyboards/layouts/ru.html

Ur, B., Noma, F., Bees, J., Segreti, S. M., Shay, R., Bauer, L., Christin, N., Cranor, L. F. (2015). "i added'!'at the end to make it secure": Observing password creation in the lab, *Eleventh symposium on usable privacy and security (SOUPS 2015)*, pp. 123–140.

Wang, D., Wang, P., He, D., Tian, Y. (2019). Birthday, name and bifacial-security: Understanding passwords of chinese web users, *28th USENIX Security Symposium (USENIX Security 19)*, USENIX Association, pp. 1537–1555.

Wash, R., Rader, E. (2021). Prioritizing security over usability: Strategies for how people choose passwords, *Journal of Cybersecurity* **7**(1).

Wash, R., Rader, E., Berman, R., Wellmer, Z. (2016). Understanding password choices: How frequently entered passwords are re-used across websites, *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pp. 175–188.

Wilcock, J., Dempsey, P. (2024). Russian keyboard. https://learn.microsoft.com/en-us/globalization/keyboards/kbdru

Woods, N., Siponen, M. (2018). Too many passwords? how understanding our memory can increase password memorability, *International Journal of Human-Computer Studies* **111**, 36–48.

Yan, J., Blackwell, A., Anderson, R., Grant, A. (2000). The memorability and security of passwords – some empirical results, *Technical Report UCAM-CL-TR-500*, University of Cambridge, Computer Laboratory.

Zeidan, A. (2023a). languages by number of native speakers. https://www.britannica.com/topic/languages-by-number-of-native-speakers-2228882

Zeidan, A. (2023b). languages by total number of speakers. https://www.britannica.com/topic/languages-by-total-number-of-speakers-2228881

Received July 28, 2025 , revised November 26, 2025, accepted December 11, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, pp. 933–957
https://doi.org/10.22364/bjmc.2025.13.4.10

ColorMEF: A Novel Transformer Based Multi-Exposure Fusion Model

Matiss LOCANS, Evalds URTANS

Riga Technical University, Riga, Latvia

matiss.locans@edu.rtu.lv, evalds.urtans@rtu.lv

ORCID 0000-0001-9813-054, ORCID 0000-0001-9813-0548

Abstract. This research presents a novel multi-exposure fusion model termed ColorMEF. In contrast to conventional approaches that rely primarily on image luminance data, ColorMEF integrates chromatic color information to augment the image fusion process. The incorporation of chromatic information enables ColorMEF to outperform existing models, as substantiated by evaluation metrics such as SSIM, DISTS, and VSI. Furthermore, the model is systematically trained using an end-to-end framework that optimizes the MEF-SSIM function based on the input data. ColorMEF achieves state-of-the-art performance on 4 of 5 conventional full-reference metrics: SSIM 0.9181 (+0.0092 versus the best prior), GMSD 0.0645 (-0.0043, $\downarrow 6.3\%$), DISTS 0.9266 (+0.0144), and VSI 0.9857 (+0.0025); on VIF ColorMEF is slightly lower (0.5224 vs 0.5394 for IFCNN). On the paper-introduced FSMEF-SSIM metric, ColorMEF ranks *second* (0.9435 vs 0.9465 for MEF-Net). These results indicate that coupling the chroma and luma during fusion improves structural fidelity and perceived quality across diverse scenes.

Keywords: Machine learning, Multiple exposure fusion, Vision transformers, Guided filtering, Chromatic colors

1 Introduction

The technique of Multiple Exposure Fusion (MEF) involving disparate exposure images represents a relatively nascent challenge within the domain of computer vision. By applying this technique with low dynamic range (LDR) images, it is possible to acquire high dynamic range (HDR) images, which contain more information than individual images and make the processing of said information much simpler because it is within a singular image space.

To fuse multiple exposure images, it is necessary to acquire them. It can be done by taking multiple LDR images with different exposure times. By changing the exposure time of the camera sensor, the amount of light to which the sensor is exposed changes.

The increased exposure time translates into longer light exposure, which translates into much brighter images. This can cause certain objects in captured images to be underexposed, causing them to appear much darker and indistinguishable from their surroundings, or overexposed, causing them to appear too bright, creating the same effect of underexposed areas but in the opposite direction.

Since LDR images can fail to capture all of the scene information in a singular image, by taking multiple LDR images with different exposure times and fusing them together, it is possible to attain a singular fused HDR image, which contains most visible features from individual exposures in a single image.

There are a multitude of solutions for the MEF task, spanning classical and deep learning methods alike, with deep learning methods getting better and outperforming their classical counterparts. However, most MEF methods still struggle with similar limitations. Most methods utilize color transformations of images to acquire luminance information from the said images. While luminance information is fused together by using the proposed deep learning architectures, color information is often relegated to being fused by a weighted sum operation. Although this fusion makes sense, as luminance contains much more structural and contrast information than colors do, there remains an unexplored avenue of color fusion with deep learning methods as well.

We propose ColorMEF, a transformer-based multi-exposure fusion model, trained in a self-supervised manner. The overall structure of the model is akin to an encoder-decoder network, and it is trained by using previously used multi-exposure fusion specific metrics for self-supervision, to learn input image feature fusion. The model works in an end-to-end manner, using the segmentation approach to generate weight maps for input exposure images. The weight maps are then applied on top of the input exposures, and they are fused by summing values.

Unlike other solutions, color information, which can be described as chromatic colors, is also fused inside the model. We adapt a transformer-based encoder architecture for global feature extraction from the image, as well as a CNN skip connection for small local feature extraction. We used parallel transformer units to achieve this goal, as experimental results show that color information can affect how luminance is weighed, affecting the final resulting image.

2 Related work

Classical MEF typically relies on multi-scale pyramid fusion (saliency weighted Laplacian or ratio of Gaussians pyramids),exposure-weighted blending (well-exposedness, local contrast, and saturation weights), gradient- or edge-preserving schemes (bilateral/guided filtering) to avoid halos, and optimization formulations that penalize seams and artifacts. These methods are fast and data-free, but struggle with misalignment/ghosting and often treat color as a passenger variable, blending it with the same scalar weights chosen for luminance; chroma-specific artifacts (bleeding, false color in saturated regions) are common.

In recent years, numerous deep learning-based multi-exposure fusion (MEF) solutions have emerged. The first deep learning-based MEF approach was proposed by (Prabhakar et al., 2017) named DeepFuse utilizing CNN features for image fusion. Un-

like classical algorithms, DeepFuse fuses images directly; however, it only fuses luminance information, while color information is fused using the weighted sum (Prabhakar and Babu, 2016). Various other solutions adapt this general approach to MEF by fusing luminance information directly and returning it from the model, while changing the underlying architecture to further try to improve fusion results. PGMI (Zhang, Xu, Xiao, Guo and Ma, 2020) utilizes individual branches for different exposures compared to DeepFuse (Prabhakar et al., 2017), while also utilizing pathwise transfer blocks to exchange information between them. However, this proposes a limitation, locking PGMI to be able to fuse only two images at a time. Other solutions simply use established CNN architectures, e.g. DenseNet (Huang et al., 2017) and fuse multiple exposure images by concatenating luminance information together in channel dimension, returning already fused image outputs directly from the model (Xu, Ma, Jiang, Guo and Ling, 2020), (Xu, Ma, Le, Jiang and Guo, 2020)

The previously mentioned approach has some drawbacks. Trusting the model to provide a complete MEF output is suboptimal. This means that models have to be trained on a large amount of data to be able to produce such output. Some of existing deep learning-based approaches are more similar to classical algorithms, generating weight maps in order to perform MEF (Ma et al., 2020), (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020). When weight maps are generated, the input images are used to generate the fused image by weighing them and then summing the resulting weighted images. This also allows the solutions to be manipulated with weights, for example, MEF-Net (Ma et al., 2020) uses two differing resolutions in the MEF task. Fusion weights are acquired from low-resolution images and, in conjunction with guided filtering (He et al., 2010), weights can be upscaled to higher resolutions, allowing for the fuse of higher resolution images.

Other existing MEF solutions also include the use of GAN (Goodfellow et al., 2014), such as MEF-GAN (Xu, Ma and Zhang, 2020). MEF-GAN in particular is set apart from other solutions by the fact that it fuses RGB images, instead of luminance information from YCbCr images. It is also one of the first models to use an attention mechanism; more specifically, it uses a self-attention mechanism similar to (Wang et al., 2017).

Transformer (Vaswani et al., 2017) architectures have proved to be powerful in a multitude of tasks. Naturally, their strength has also been preserved for vision tasks, e.g., vision transformers (ViT) (Dosovitskiy et al., 2020). One of the more recent solutions, TransMEF (Qu et al., 2021), utilizes the transformer architecture to extract global feature information. This lets the model overcome the receptive field issues that many CNN models have. However, due to the nature of ViT image patching, while they are very proficient at processing global feature patches, they lack the ability to discern the small features within the patches themselves. Because of this, a CNN-based local feature extractor can be used to compensate for these deficiencies. TransMEF (Qu et al., 2021), in particular, is not trained for MEF in a direct way, rather it is trained as a feature extractor. The MEF component is operationalized through a fusion rule, which, although it facilitates the simultaneous amalgamation of multiple exposure images, concurrently detracts from the overall performance and the resultant outcomes.

As mentioned previously, MEF datasets are hard to come by. One of the most prominent data sets used is SICE (Cai et al., 2018), which contains multiple scenes with generated ground truths. For those who wish to use different data for image fusion, it is necessary to obtain MEF data via other means. Another aspect of training data acquisition is the usage of other data sets in conjunction with data augmentation to distort the original image for input and use the original as ground truth (Qu et al., 2021). This can allow the use of large-scale natural image datasets, but the performance might be inhibited because distortion effects are not sufficient to model different exposure levels like they are in originally created images. In the absence of ground-truth data, a self-supervised training approach can be utilized for model training. One of the most popular MEF optimization functions, MEF-SSIM (Ma et al., 2015), has been used as an optimization parameter to train models self-supervised. MEF-SSIM is utilized to process input data and detect image regions of highest contrast, which then are used to train models for the MEF task, avoiding the need for ground truth images.

3 Methodology

The contribution of this work is color-guided luminance fusion that instead of fusing Y and then blending chroma, jointly processes Y and (Cb, Cr) in a full-parallel transformer, including a cross-attention branch that lets chroma steer which luminance structures to trust. Compared to brightness-only MEF, this reduces false structure selections near saturated colors and under white-balance shifts. Compared to post-hoc colorization/blending, it avoids chroma-luma inconsistencies by learning them together. Architecturally, we pair a DPT-style global transformer with CNN skip connections and per-channel deep guided filters for high-res weight upsampling. In this section, we take a look at the proposed model architecture and explain the choices made in the architecture and the philosophy behind them. We also look at the data used in the training and the training process itself.

3.1 Dataset

For training and validation, we use a closed source dataset, which contains about 7600 multi-exposure scenes. Each scene consists of three distinct exposure images, namely dark exposure b_0, middle exposure b_1, and bright exposure b_2. All images are of size 1667×1250 and correspond to a 4:3 aspect ratio. It is important to mention that this large-scale dataset does not contain ground truth images.

For testing and comparison with other methods, we use the SICE (Cai et al., 2018) dataset, which contains 589 multi-exposure scenes. We used the data split provided by the author of the dataset to determine the use of images for model testing. We take the validation split for the model testing. Three exposure images from each scene were sampled, so the model had limited available information during testing. The availability of ground truths in SICE comes as a boon, allowing us to utilize different full reference metrics in order to quantify model performance. However, since (Cai et al., 2018) state that reference images are generated using existing MEF and HDR solutions, this data set can only be used to compare the structural information of images primarily, leaving color information comparisons as a secondary objective.

3.2 Model

The proposed model ColorMEF, as mentioned before, takes inspiration from other previously mentioned methods, as well as different solutions from other computer vision subfields. It is publicly available in open source code [1]. The general approach for the MEF task is borrowed from (Ma et al., 2020). We first downsample the input sequence and make low-resolution weight-map predictions. Using guided filtering (He et al., 2010), we acquire high-resolution weight maps, which are then used for the generation of output exposure by multiplying them by the high-resolution input exposure sequence and adding the results. However, this general process is applied to all three image channels and not exclusively to luminance. The proposed model processes luminance and chromatic color information jointly and separately at different times. Also, chromatic colors are processed together since they represent a 2D point upon a color space. Our reasoning is that processing them jointly is easier for the model since it can fuse both values in conjunction, rather than fusing each of them separately. ColorMEF is trained end to end, optimizing the MEF-SSIM (Ma et al., 2015) criterion on top of high-resolution images.

ColorMEF adheres to the weight-map fusion paradigm as outlined in (Ma et al., 2020). Initially, low-resolution weights are estimated and subsequently upsampled to high resolution through the employment of deep guided filters (DGFs). The fusion of each channel is accomplished via weighted summation. Distinctly from preceding studies, we perform the fusion *luminance and chroma* within the network itself. A DPT-like U-shaped transformer (Ranftl et al., 2021) is utilized to extract global features, while shallow CNN skip pathways are employed to capture local details. A fully-parallel transformer block (Touvron et al., 2022) is implemented, comprising self-attention on Y, combined self-attention on (Cb, Cr), and cross-attention between the two, in order to achieve chroma-guided luminance. Prior to weight prediction, feature refinement is carried out through dictionary convolution units (DCUs) (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020).

3.2.1 Feature extraction block Weight map prediction can be defined as a segmentation task, since weight maps can be interpreted as segments of original images, where information evaluation is performed. Therefore, it makes sense to utilize segmentation networks for this approach. We use a dense prediction transformer (DPT) (Ranftl et al., 2021) to extract global image characteristics. This transformer is U-shaped and uses ViT (Dosovitskiy et al., 2020) for its encoder part, while CNN blocks are used in the decoder to rebuild images. DPT uses reassembly to reconstruct low-resolution image representations in reassembled blocks. These reconstructions are then used to build segmentation maps for each exposure back to the same low-resolution inputs.

In addition to using DPT for global feature extraction, we use parallel transformer blocks (Touvron et al., 2022) in our architecture. The complete parallel transfomer block consists of three MHA units and three feedforward networks. The visual representation of this block can be viewed in Fig 1 on page 938. This is where color information comes in. The image information is divided into two categories: luminance

[1] https://github.com/scrayish/ColorMEF

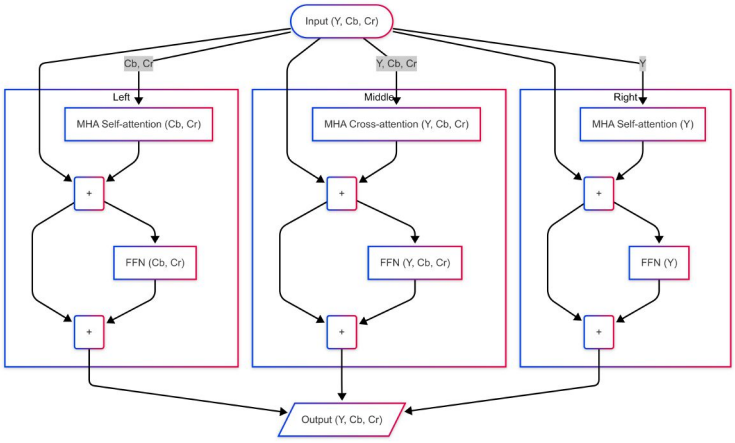


Fig. 1: Visual representation of full parallel transformer block used within DPT for extracting global features from luminance and chrominance information

or Y and chrominance or Cb or Cr. In the full parallel transformer block, we apply self-attention to each of these two categories separately. The third MHA is used for cross-attention between luminance and chrominance. Because luminance information between multiple exposures is more distinguishable than chrominance information, by utilizing cross-attention, it ought to be possible to guide chrominance information in some way to achieve better image fusion results. However, in experiments, regardless of the direction of cross-attention, chrominance information has a profound impact on luminance fusion results, since all of the information is deeply intertwined, therefore changing the overall fused image appearance.

As mentioned above, ViTs are good for global information acquisition, since they do not suffer from receptive field problems as CNNs. However, they are limited when it comes to local features. In a similar fashion to TransMEF (Qu et al., 2021), we use a skip connection to extract local feature information. Our CNN feature module is much more classical in its approach, sporting a normalization and activation function for each convolution layer, of which there are three. While chromatic and luminance information is deeply intertwined inside the DPT, we keep them separate for low-level feature extraction, each information type being processed by its own skip connection.

Also similar to TransMEF (Qu et al., 2021), an enhance block is used to merge the global transformer features with the local features of the CNN module. However, the enhance block uses dictionary convolutional units (DCU) (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020) instead of regular convolutions. This allows us to refine segmentation map results by using a closer link to input information. Also, since chromatic information has a profound influence on luminance information fusion results, it is possible to wrestle back some of the luminance independence by using DCUs. We postulate that the superior clarity of luminance information, in terms of structural and contrastive attributes, compared to chromatic channel information, underpins this phenomenon. The enhancement block head is used to compress the feature information and gain a singu-

lar weight map for each exposure image. The same as skip connections, these enhance blocks are separate for luminance and chrominance information.

3.2.2 Guided image filtering After feature extraction, we applied guided image filtering (He et al., 2010) to enhance the weight maps obtained and upsample them to high resolution to create a high-resolution fusion image. Unlike the MEF-Net approach (Ma et al., 2020), we use a deep guided filter (DGF) (Wu et al., 2018). The deep guided filter utilizes convolutions in order to enhance the weight maps. We also utilize separate DGFs for each information channel, since the information between luminance and chrominance is very distinct. We also separate chrominance back in the Cb and Cr channels and apply a separate DGF to them to reduce the potential production of color artifacts, where a singular DGF is used for both color channels. The complete visual structure scheme of ColorMEF can be viewed in Fig 2 on page 939

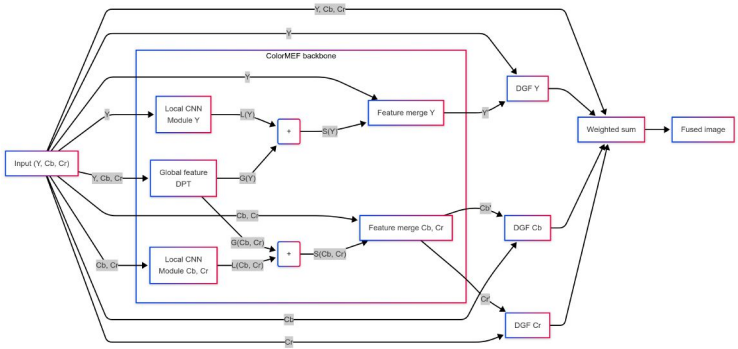


Fig. 2: ColorMEF model full schematic, including guided filtering modules and all necessary post-processing steps. All of these steps happen during model forward pass

3.2.3 Model training The model is trained on a proprietary dataset. A subset of 250 images is randomly selected for validation, while the remaining images comprise the training set. The training process utilizes the MEF-SSIM (Ma et al., 2015) criterion in its entirety. To prevent potential artifacts during training, each image channel, namely Y, Cb, and Cr, is processed separately using the MEF-SSIM criterion. The three loss components are assigned equal weightage ($static\_scale\_coeff = 1.0$, $chroma\_to\_luma\_coeff = 1.0$), and their summation is employed for backpropagation. The model is trained employing a single NVIDIA RTX A6000 GPU with 48 GB of video memory over approximately 200 epochs, with each epoch lasting approximately 2.5–3 hours. Under these conditions, the cumulative wall-clock training time is approximately 21 to 25 days. The optimization process is conducted using the Adam optimizer, starting with an initial learning rate of 1e-4, following a StepLR schedule that reduces the learning rate by 0.9 every 10 epochs (i.e., from 1.0e-4 to approximately

1.22e-5 by epoch 200). A smoothed running loss is calculated using an Exponential Moving Average (EMA) with a beta parameter of 0.9.

Input data are prepared at high and low resolution, similar to MEF-Net (Ma et al., 2020). For high resolution, we use the regular image resolution of 1667×1250 and for low resolution, we use 1280×960. This is a much higher resolution than is normally used for training. Also, since ViT is used for global image information extraction, the low resolution of the model is fixed. In order to train at different resolutions, model configuration should be updated to support said resolution. We also employ learnable positional embeddings, which lock the low resolution further in. It may be possible to train more dynamic-size images if the original transformer positional embedding proposed by (Vaswani et al., 2017) were used.

The training script employs randomized resizing with a high dimension of 1250 and a low dimension of 960. All images are transformed into the YCbCr color space, and the channels are split to facilitate fusion processes. Three exposure levels—dark, middle, and bright—are fused, with explicit supervision applied to both the luma (Y) and chroma (Cb/Cr) channels. During the inference and evaluation phases, the predicted RGB images and exposure weight maps for each sample are saved. The model utilizes a transformer-based architecture with an embedding dimension of 512, employs 8 attention heads, and has an enhancement depth of 3, adopting an identity readout mechanism. Padding is handled through PyTorch reflection padding, with an alternative option for circular panorama padding available. Training and evaluation are conducted sequentially over each epoch, with gradient computations enabled exclusively during training.

Higher resolution also means that more data are required. Although it is possible to train our model using the SICE (Cai et al., 2018) dataset, it is simply far too small for this kind of resolution. This is because a smaller resolution allows for efficient image cropping. By increasing the low-resolution model, image crops become larger, thus individual crops contain less and less unique information per crop, leading to degrading training. Another important aspect of transformer models is the general need for larger datasets than other network architectures. Insufficient data amounts can cause aggressive tiling artifacts in fused images, which the model learns to smooth out when larger data amounts are provided.

3.3 Full Structure MEF-SSIM

In addition to a new model architecture, we also introduce a variation of the MEF-SSIM (Ma et al., 2015) quality measurement metric. Although MEF-SSIM proves itself to be a formidable choice for self-supervised training, it mostly focuses only on the best contrastive information of the input images. Although this works in most cases, there may be cases where additional information from other exposures might be necessary to gain a more balanced fusion output. For this purpose, we propose the Full-Structure (FS) MEF-SSIM.

Algorithm 1 FSMEF-SSIM

1: Variance calculation $\sigma_{y_i}^2$ for input images $y_i \in Y$
2: Variance calculation σ_x^2 for fused image x
3: Covariance calculation σ_{xY} between mean values of fused image μx and input images μY
4: Quality map calculation between input images Y and fused image x with formula

$$\frac{2\sigma_{xy} + C2}{\sigma_x^2 + \sigma_Y^2 + C2}$$

5: Based on varinace calculation σ_Y^2 image segments are grouped into multiple sampling maps $k_i \in K$
6: Sampling maps are used to sample MEF-SSIM results $s_i \in S$ and sampled maps are multiplied with scaling coefficients and summed for final FSMEF-SSIM score

$$\sum_{i=1}^{n} s_i k_i, s \in S; k \in K$$

Fundamentally, this measurement works the same as MEF-SSIM. The contrast regions are calculated and ranked from the best to the worst, based on the returned values. However, with FSMEF-SSIM, all structural information is taken into account by weighting it. In this work, it is adjusted to work with three exposure images ($n = 3$), so weights for each structural segment ranging from best to worst are 0.9, 0.09 and 0.01, respectively. We still highly weigh the best regions with a much smaller contribution from the smallest structural regions. This is done because smaller contrast regions might contain large amounts of similar information with little to no detail, resulting in inferior fusion performance. But it can allow one to smooth out transitions between aggressive structural regions. If lower structural regions are weighed much higher, the final fusion image can become homogeneous, which is an undesirable result, since it can make the image look foggy and lose detail.

This kind of approach can help with the fusing of images with larger exposure time differences. Applying FSMEF-SSIM to an exposure stack of images with a smaller exposure time gap will not give large improvements in results, as the images already are spaced close enough to yield good enough detail coverage for successful fusion.

4 Results

As mentioned above, we used the SICE (Cai et al., 2018) dataset validation split as images to test ColorMEF. We avoid using MEF-SSIM for comparison, as it was used as an optimization criterion during model training. Instead, we utilize ground-truth images and we employ full-reference metrics in order to quantify our solution.

We also use other solutions, MEF-Net (Ma et al., 2020), TransMEF (Qu et al., 2021), CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020) and IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020) for comparison. We used the trained weights provided by the author and source code to replicate selected fusion images using their

solutions. We only make small adjustments to make the code work. No changes have been made to the parameters or the model architecture that could affect the results.

4.1 Quantitative results

In this subsection, we take a look at quantitative results. We employ five different metrics for the determination of image quality: SSIM (Wang et al., 2004), VIF (Sheikh and Bovik, 2006), GMSD (Xue et al., 2013), VSI (Zhang et al., 2014), and DISTS (Ding et al., 2020). Each metric looks at image structure quality, although each does it in a different way. We find it valuable to evaluate the quality of the image based on different calculation approaches.

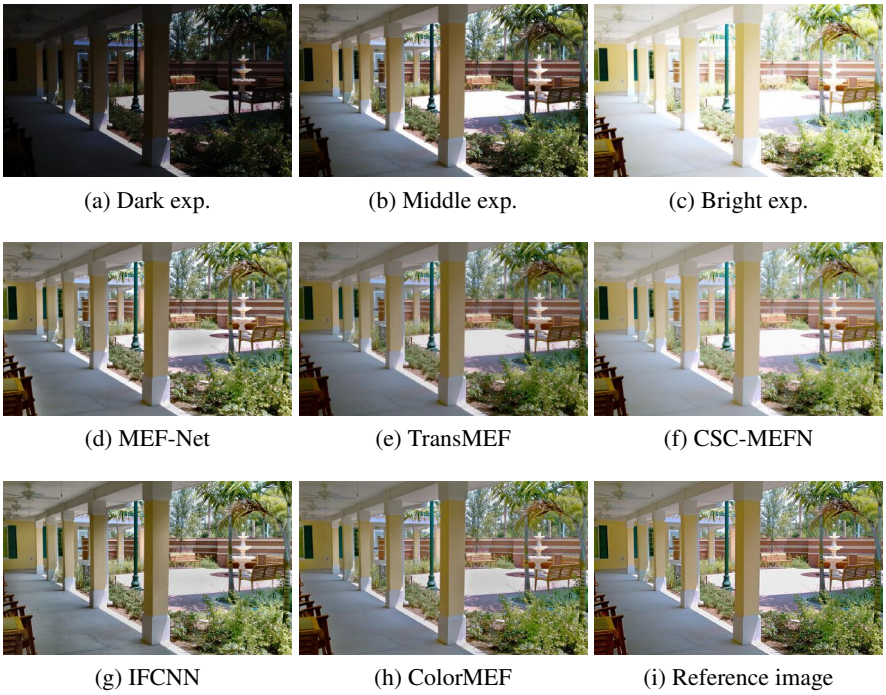
Table 1: Average metric result of all SICE validation test set for all calculated metrics of all tested models. The best results for each metric are marked in bold, while the second-best result is underlined. The star notation indicates the model and metric introduced in this paper

MEF method	MEFNET	TransMEF	CSC-MEFN	IFCNN	ColorMEF*
FSMEF-SSIM*	**0.94646**	0.91000	0.93124	0.88927	0.94348
SSIM	0.90890	0.84143	0.88284	0.87787	**0.91807**
VIF	0.52473	0.44590	0.47646	**0.53935**	0.52240
GMSD	0.06884	0.09076	0.08044	0.08716	**0.06451**
DISTS	0.91214	0.90273	0.90772	0.89346	**0.92656**
VSI	0.98314	0.97694	0.98067	0.97809	**0.98568**

1 contains all average metric values for each quality evaluation criterion tested. We can see that the colorMEF average metric values are best for four of five quality metrics. VIF is the only metric where IFCNN outperforms ColorMEF. Although ColorMEF achieves best results in a multitude of metrics, it is worth mentioning that all models, apart from TransMEF achieve very competitive metric evaluation. This means that all of the models are comparable and can fuse images well to an extent. We also reiterate that (Cai et al., 2018) used existing MEF solutions as well as HDR solutions for ground truth generation. In this way, it is possible to maximize structural information from images, as well as contrast. However, certain details remain subjective, such as coloring, saturation, hue, etc. For more precise results, a qualitative evaluation is also necessary.

4.2 Qualitative results

As seen in Fig 3 on page 943, ColorMEF achieves a balanced fusion between all three exposures. It is not as aggressive as IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), which mixes shadows and small details on the ground very aggressively. The chairs in the bottom left corner are also well fused, while MEF-Net (Ma et al., 2020) produces a light halo around them. While the reference image sports a high color saturation, ColorMEF tones colors a bit down, while remaining fairly vibrant. It also does not



(a) Dark exp. (b) Middle exp. (c) Bright exp.

(d) MEF-Net (e) TransMEF (f) CSC-MEFN

(g) IFCNN (h) ColorMEF (i) Reference image

Fig. 3: ColorMEF result comparison with other deep learning based MEF solutions. (a) - (c) is the input sequence, (d) - (h) are the model fused images, and (i) is the reference image for this fusion sequence

.

suffer from much darkening as with TransMEF (Qu et al., 2021), or too much brightness as CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020).

Fig 4 on page 944 is another example sequence of fusion. In this specific exposure sequence. Although this example uses much different exposure times, it is simpler because you are exposed indoors. It contains some details, which are conveniently clustered.

Inspecting the resulting images, it is apparent that all the models are good enough to fuse this image. By comparing with the reference image, we also deduce that certain models can fuse features better than the reference image contains. In this picture, MEF-Net (Ma et al., 2020) and ColorMEF have fairly similar results. TransMEF (Qu et al., 2021) is darker than others, while CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020) is brighter. IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020) contains very aggressive details, which yields a high structure score, but can degrade the quality of the image, making it look messy.

Apart from influencing the fusion of luminance information, ColorMEF's approach of fusing colors also allows it to correct colors depending on input exposures and input scene. Although not as saturated as reference image colors, ColorMEF can reduce the

(a) Dark exp. (b) Middle exp. (c) Bright exp.

(d) MEF-Net (e) TransMEF (f) CSC-MEFN

(g) IFCNN (h) ColorMEF (i) Reference image

Fig. 4: ColorMEF result comparison with other deep learning based MEF solutions. (a) - (c) is the input sequence, (d) - (h) are the merged images of the model, and (i) is the reference image for this fusion sequence

.

blandness of images, which makes them look flat, and can impact the ease of postprocessing. The fused image also looks warmer than other solution images.

One caveat to including colors inside the network is their influence on luminance information fusion. While we employ cross-attention in a way that colors should follow luminance, the opposite effect could be observed at times. By increasing color importance and influence, it could be possible to gain more saturation and vibrance at the cost of detailing. This avenue has yet to be explored properly.

To gain additional insight on the qualitative performance of the models, we collect the mean opinion score through a survey. We surveyed 47 respondents and asked them to evaluate 9 image sets. Each set contains the fused image of each mdodel. We do not provide a reference image in this comparison, since our aim is to get an opinion on each model performance. The mean quality score (MOS) for each individual sequence and the average score of all 9 sequences can be viewed in 2.

We asked each image to be evaluated on a scale of 1 to 5, where 1 is a very low quality image and 5 is very high quality image. Based on observations from MOS, it is

Table 2: Mean opinion score for each compared model generated image inside provided image sequence. The top score for each image is marked in bold, while the second score is underlined

	MEF-Net	TransMEF	CSC-MEFN	IFCNN	ColorMEF
Sequence 1	3.59574	2.14894	2.44681	**3.91489**	3.63830
Sequence 2	2.89362	1.68085	2.70213	**3.93617**	3.44681
Sequence 3	3.65957	2.14894	2.89362	**4.08511**	3.55319
Sequence 4	3.46809	2.31915	3.12766	3.51064	**3.70213**
Sequence 5	3.48936	2.55319	3.02128	**4.34043**	3.46809
Sequence 6	2.97872	2.38298	3.00000	**3.97872**	3.19149
Sequence 7	3.57447	2.78723	2.36170	**4.38298**	3.14894
Sequence 8	3.70213	2.85106	2.53191	**4.21277**	3.06383
Sequence 9	3.46809	3.23404	2.27660	**4.12766**	3.38298
Average	3.42553	2.45626	2.70686	**4.05437**	3.39953

apparent that IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020) has received the highest evaluation score of 4.05. This can be explained by its aggressive fusion, which makes the resulting images highly detailed. However, it can also cause harsh artifacts in coloring, as was observed in our own qualitative analysis.

Both ColorMEF and MEF-Net (Ma et al., 2020) are very close in second and third place, differing by approximately 0.03 MOS, which is a very small margin. The similar evaluation scores between these two models make sense, as was observed in qualitative result analysis. However, both methods trail IFCNN by more than 0.6 MOS, which is a substantial margin. Especially since both ColorMEF and MEF-Net scored very highly in quantitative results when comparing fused images to reference images.

Both TransMEF (Qu et al., 2021) and CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020) scored lower in the survey. In qualitative analysis, it was observed that these two models produced the smoothest images out of all five models tested. Their fusion images contained the least amount of small detailing, which could be the determining factor for lower MOS values in comparison to other models, which had a higher amount of small detailing.

In summary, small details and image sharpness can be compiled as the two most important aspects for image quality estimation in people testing. While coloring can impact the quality of an image, it seemingly does not play as large a role in image quality evaluation. In qualitative result analysis, we concluded that IFCNN and MEF-Net images are less saturated than those generated by other models, but score higher MOS because of the higher detail. Although ColorMEF scored the highest in most quantitative evaluations, it is trailing in the middle behind the previously mentioned models.

5 Further research

MEF is a method for obtaining HDR images by combining multiple LDR images. However, there may be some pitfalls to this approach. After MEF, the resulting image is usu-

ally still an LDR image, despite containing more details. Sometimes, users may prefer to perform additional post-processing, such as tone mapping.

ColorMEF is different from other models in that it fuses color information within itself along with luminance information. As mentioned in Section 3, we used MEF-SSIM (Ma et al., 2015) to calculate the loss of the color channel. Although it seems to work in this case, it may be suboptimal to use a structural metric to combine the color information. In an optical inspection of the YCbCr channels of an image, it is apparent that the Y channel contains the most information out of the three channels. Because it is equivalent to a grayscale image, this channel has the most pronounced structural and contrast information. While color channels also contain some contrast and structure, it is much weaker, as well as the mean value of all color information in channel is much more towards the middle of value range than in luminance. A wider theoretical knowledge of color spaces and color information can prove useful in further improving color fusion within neural networks.

When it comes to MEF, each additional image taken adds to the total computational cost of fusing images together. Although most of deep learning-based MEF solutions focus on two-image fusion, those images have to be fairly close to each other in terms of exposure times to fuse images successfully without excessive artifacts. Deep-learning models tend to have a strong bias for structural details given their training regimen. It is this bias that can lead to large artifacts in fused images, as models try to utilize much of the information from one image or the other, causing distinct artifact regions in fused images to degrade their quality. Different methods should be explored, which could let models truly fuse information between images, instead of doing hard-headed segmentation and fusion of most prominent feature areas, if there is a necessity to fuse images with larger difference of exposure times in order to keep image count lower.

6 Conclusions

In this paper, we proposed a novel deep learning method for MEF named ColorMEF. By utilizing and fusing image information within the model itself, we can acquire more balanced images, with reduced artifacts, and are better colors than their fused counterparts using other solutions, despite being trained with much larger resolution, which can provide a larger space for errors. Our outperform other model fused images based on objective metrics that use a reference photo. It showed an improvement of 25% using the GMSD metric, 4% using the SSIM and DISTS metrics, and outperformed other methods in 4 of 5 metrics. A novel model provides higher-quality images without additional hardware requirements at higher resolutions. We also give directions for further research, which could bring even more improvements to color fusion methods when applied together with deep learning-based solutions.

References

Cai, J., Gu, S., Zhang, L. (2018). Learning a deep single image contrast enhancer from multi-exposure images, *IEEE Transactions on Image Processing* **27**(4), 2049–2062.

Ding, K., Ma, K., Wang, S., Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 2567–2581.
https://api.semanticscholar.org/CorpusID:215785896

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale, *ArXiv* **abs/2010.11929**.
https://api.semanticscholar.org/CorpusID:225039882

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., Bengio, Y. (2014). Generative adversarial nets, *NIPS*.
https://api.semanticscholar.org/CorpusID:1033682

He, K., Sun, J., Tang, X. (2010). Guided image filtering, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 1397–1409.
https://api.semanticscholar.org/CorpusID:1264129

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K. Q. (2017). Densely connected convolutional networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ma, K., Duanmu, Z., Zhu, H., Fang, Y., Wang, Z. (2020). Deep guided learning for fast multi-exposure image fusion, *IEEE Transactions on Image Processing* **29**, 2808–2819.

Ma, K., Zeng, K., Wang, Z. (2015). Perceptual quality assessment for multi-exposure image fusion, *IEEE Transactions on Image Processing* **24**, 3345–3356.
https://api.semanticscholar.org/CorpusID:4828378

Prabhakar, K., Babu, R. V. (2016). Ghosting-free multi-exposure image fusion in gradient domain, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* pp. 1766–1770.
https://api.semanticscholar.org/CorpusID:8764582

Prabhakar, K., Srikar, V. S., Babu, R. V. (2017). Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs, *2017 IEEE International Conference on Computer Vision (ICCV)* pp. 4724–4732.
https://api.semanticscholar.org/CorpusID:216738

Qu, L., Liu, S., Wang, M., Song, Z. (2021). Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning, *ArXiv* **abs/2112.01030**.
https://api.semanticscholar.org/CorpusID:244799167

Ranftl, R., Bochkovskiy, A., Koltun, V. (2021). Vision transformers for dense prediction, *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* pp. 12159–12168.
https://api.semanticscholar.org/CorpusID:232352612

Sheikh, H. R., Bovik, A. C. (2006). Image information and visual quality, *IEEE Transactions on Image Processing* **15**, 430–444.
https://api.semanticscholar.org/CorpusID:3716103

Touvron, H., Cord, M., El-Nouby, A., Verbeek, J., J'egou, H. (2022). Three things everyone should know about vision transformers, *ArXiv* **abs/2203.09795**.
https://api.semanticscholar.org/CorpusID:247594673

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need, *NIPS*.
https://api.semanticscholar.org/CorpusID:13756489

Wang, X., Girshick, R. B., Gupta, A. K., He, K. (2017). Non-local neural networks, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 7794–7803.
https://api.semanticscholar.org/CorpusID:4852647

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity, *IEEE Transactions on Image Processing* **13**, 600–612.
https://api.semanticscholar.org/CorpusID:207761262

Wu, H., Zheng, S., Zhang, J., Huang, K. (2018). Fast end-to-end trainable guided filter, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 1838–1847.
https://api.semanticscholar.org/CorpusID:3936783

Xu, H., Ma, J., Jiang, J., Guo, X., Ling, H. (2020). U2fusion: A unified unsupervised image fusion network, *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Xu, H., Ma, J., Le, Z., Jiang, J., Guo, X. (2020). Fusiondn: A unified densely connected network for image fusion, *AAAI Conference on Artificial Intelligence*.
https://api.semanticscholar.org/CorpusID:213637621

Xu, H., Ma, J., Zhang, X.-P. (2020). Mef-gan: Multi-exposure image fusion via generative adversarial networks, *IEEE Transactions on Image Processing* **29**, 7203–7216.
https://api.semanticscholar.org/CorpusID:220470749

Xu, S., Zhao, Z., Wang, Y., Zhang, C., Liu, J., Zhang, J. (2020). Deep convolutional sparse coding networks for image fusion, *ArXiv* **abs/2005.08448**.
https://api.semanticscholar.org/CorpusID:218673456

Xue, W., Zhang, L., Mou, X., Bovik, A. C. (2013). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, *IEEE Transactions on Image Processing* **23**, 684–695.
https://api.semanticscholar.org/CorpusID:478859

Zhang, H., Xu, H., Xiao, Y., Guo, X., Ma, J. (2020). Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 12797–12804.

Zhang, L., Shen, Y., Li, H. (2014). Vsi: A visual saliency-induced index for perceptual image quality assessment, *IEEE Transactions on Image Processing* **23**, 4270–4281.
https://api.semanticscholar.org/CorpusID:2995883

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., Zhang, L. (2020). Ifcnn: A general image fusion framework based on convolutional neural network, *Inf. Fusion* **54**, 99–118.
https://api.semanticscholar.org/CorpusID:199677411

A Appendix

In this appendix, we show additional images of ColorMEF testing such as Fig 5 on page 949, Fig 6 on page 950, Fig 7 on page 951, Fig 8 on page 952, Fig 9 on page 953, Fig 10 on page 954, Fig 11 on page 955, Fig 12 on page 956. Fig 13 on page 957 and comparison with the models mentioned previously. Specifically, we show image collages presented in the survey to determine MOS (mean opinion score) for each model-fused output. The image collages are the same as those given to the respondents, and each collage image is numbered, where numbering means that the model used to make the output image. As discussed in section Section 3.1, we used the SICE dataset (Cai et al., 2018) for model evaluation, specifically, we used validation images from the proposed data split.

Fig. 5: Collage of fused images for SICE (Cai et al., 2018) sequence 46 (sequence 1 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

Fig. 6: Collage of fused images for SICE (Cai et al., 2018) sequence 62 (sequence 2 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

Fig. 7: Collage of fused images for SICE (Cai et al., 2018) sequence 56 (sequence 3 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

Fig. 8: Collage of fused images for SICE (Cai et al., 2018) sequence 102 (sequence 4 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

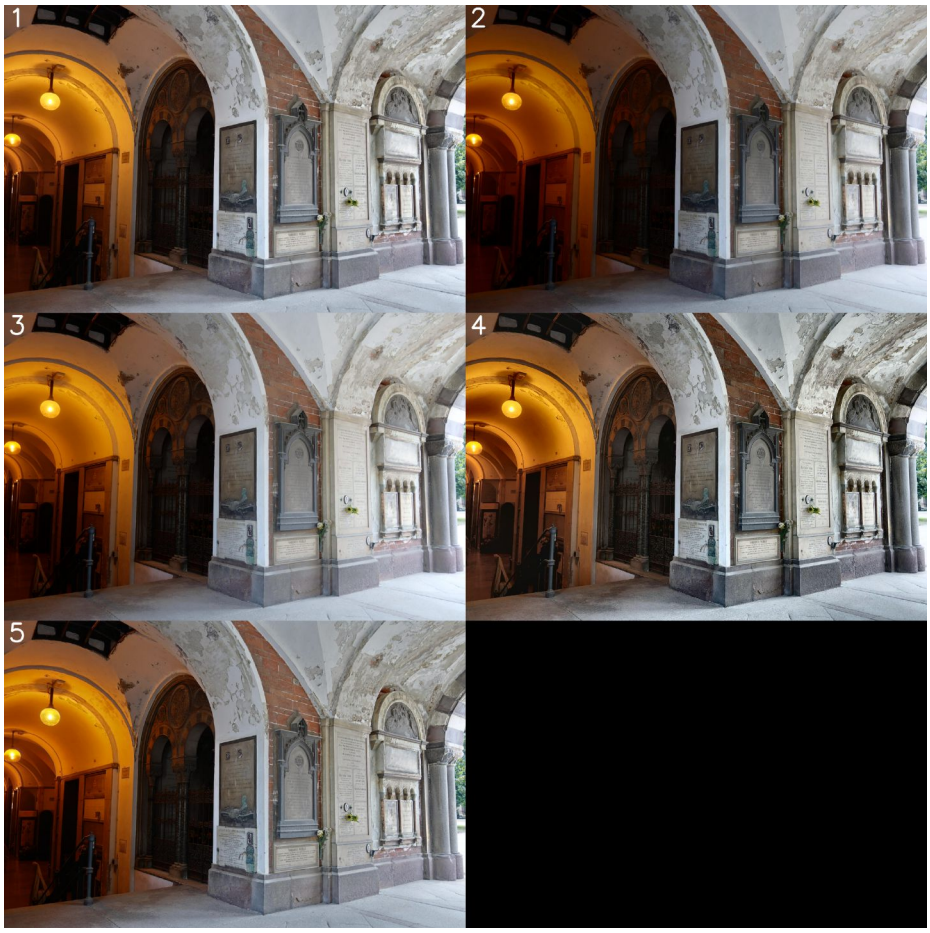


Fig. 9: Collage of fused images for SICE (Cai et al., 2018) sequence 28 (sequence 5 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

Fig. 10: Collage of fused images for SICE (Cai et al., 2018) sequence 78 (sequence 6 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

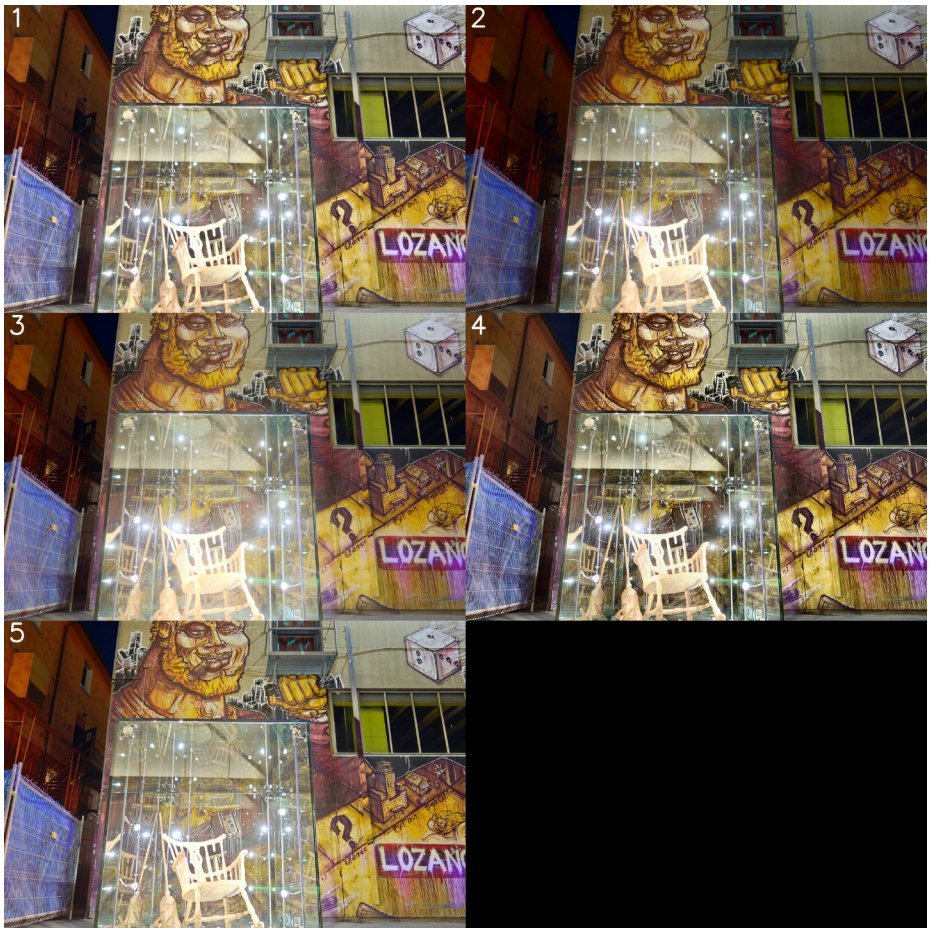


Fig. 11: Collage of fused images for SICE (Cai et al., 2018) sequence 58 (sequence 7 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

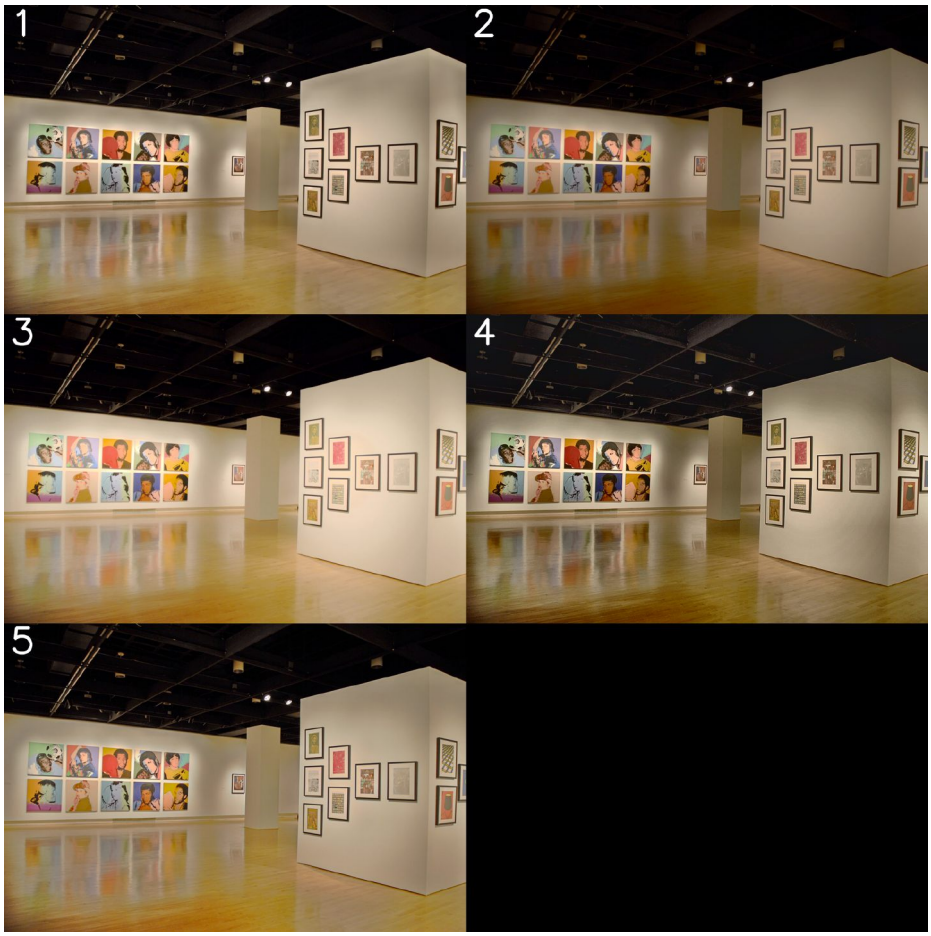


Fig. 12: Collage of fused images for SICE (Cai et al., 2018) sequence 57 (sequence 8 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

Fig. 13: Collage of fused images for SICE (Cai et al., 2018) sequence 52 (sequence 9 in 2). Images acquired using models by numbers: 1 - MEF-Net (Ma et al., 2020), 2 - TransMEF (Qu et al., 2021), 3 - CSC-MEFN (Xu, Zhao, Wang, Zhang, Liu and Zhang, 2020), 4 - IFCNN (Zhang, Liu, Sun, Yan, Zhao and Zhang, 2020), 5 - Ours (ColorMEF)

Received February 26, 2025 , revised September 15, 2025, accepted December 4, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, 958-993
https://doi.org/10.22364/bjmc.2025.13.4.11

Team Resistance Dynamics through a Dual-Pathway Framework for Successful AI Integration

Muhammad A. ALDHAHERY

College of Computing and Information Technology - Shaqra, Shaqra University, Saudi Arabia

maldhahery@su.edu.sa

ORCID 0009-0006-8843-8374

Abstract. This study utilized Partial Least Squares Structural Equation Modelling (PLS-SEM), to examine team resistance successful implantation of AI. The mediation analysis reveals that Digital Skills of Sales Team (DS), Top Management Support (TM), Organizational Readiness (OR), and Innovation Culture (IC) each exert significant indirect effects on EA via ST, with DS (0.093) and IC (0.055) showing the strongest mediated paths. ST significantly predicts EA ($\beta = 0.242$), confirming its mediating role. Outer loadings for all indicators exceeded 0.70, indicating strong reliability, while outer weights confirmed balanced item contributions. OR had the largest total effect (0.395) on EA. Four interaction terms were tested to examine whether ST moderates the effects of DS, TM, OR, and IC on EA. None of the moderating paths were statistically significant (t-values < 1.96), though TM and ST ($\beta = 1.551$) showed a borderline effect. The findings offer practical insights for digitally transforming sales operations.

Keywords: Artificial Intelligence, Digital Skills, Management Support, Sales lifecycle

1. Introduction

AI refers to the simulation of human intelligence processes such as learning, reasoning, and self-correction by machines (techniques/algorithms), particularly computer systems algorithms (Russell and Norvig, 2022; Liu et al., 2025). In the realm of business and sales, AI includes technologies like predictive analytics, which can identify fluctuations in sales over time and forecast future trends (Bughin et al., 2017). The sales lifecycle, spanning from lead generation to customer retention, has traditionally relied on human judgment and interpersonal skills (Moncrief and Marshall, 2005). With the adoption of Customer Relationship Management (CRM) systems in the early 2000s, the sales domain began undergoing digital transformation. From 2015 onward, advancements in cloud computing, big data, and accessible machine learning frameworks accelerated AI adoption (Brynjolfsson and McAfee, 2017). Today, AI plays a strategic role in automating tasks that are repetitive, analyzing the behaviour of consumers, personalizing interactions, and forecasting trends in sales. According to Boppana (2023), platforms such as Salesforce

Einstein, HubSpot AI, and Microsoft Dynamics 365 have vastly improved the way in which businesses manage leads, evaluate performance, and make decisions based on data information. According to Albahri et al. (2023), applications range from chatbots which are used for customer engagement to deep learning algorithms that are used to predict customer churn would eventually dominate all businesses.

Despite these potential advances, AI integration across the entire sales process (sales lifecycle stages) remains inconsistent. “The sales cycle is the series of predictable phases required to sell a product or service. It consists of all the customer-facing steps from prospecting to closing and follow-up” (Moncrief and Marshall, 2005). Sales lifecycle processes are best conceptualized as lifecycles that involve identifiable and measurable steps, allowing for performance tracking and optimization. Typically, it involves 7 steps namely: "Lead Generation", "Lead Qualification", "Needs Assessment / Discovery", "Presentation / Proposal", "Objection Handling", "Closing the Sale" and "Post-Sale Engagement / Relationship Management". Lead Generation stage identify the potential customers who may be interested in the product/service. Bughin et al. (2017) associate this stage with predictive analytics and data mining techniques for AI integration. Lead Qualification assess which leads have the potential to become actual customers based on predefined criteria. Needs Assessment dwell on the discovery and understanding the customer's needs, pain points, and purchasing drivers. Davenport and Ronanki (2018) emphasize the complexity of AI supporting this phase due to human judgment requirements. Presentation / Proposal are offering a direction that tailored product or solution that addresses the lead’s needs. Objection Handling dwells on responding to concerns, doubts, or pushbacks from the prospective customer. It is identified as an underutilized phase for AI (Fountaine et al., 2019). Closing the Sale is the finalizing terms, addressing negotiations, and completing the transaction. The Post-Sale Engagement / Relationship Management ensure customer satisfaction, providing support, and encouraging upselling or repeat business. Olan et al. (2022) and Jarrahi (2018) note that AI is least implemented in this stage due to emotional and contextual complexity. While early stages such as lead generation, scoring, and customer targeting, benefit from well-developed AI tools, later phases like objection handling, relationship building, and upselling remain underutilized (Fountaine et al., 2019; Olan et al., 2022). This is largely due to the complexity of human interactions required in these tasks, which AI tools currently struggle to replicate.

Several challenges hinder the comprehensive adoption of AI in sales. These include limited technical skills among sales teams, resistance to change, and a lack of trust in AI-driven insights—particularly in high-stakes decisions like pricing or contract negotiation, where the black-box nature of some models creates hesitancy (Mikalef and Gupta, 2021; Jarrahi, 2018). Additionally, many AI tools operate in silos, lacking interoperability with core systems such as CRMs and ERPs, which restricts data integration and strategic decision-making. Few standardized frameworks exist to assess the performance of AI across different sales stages, making it difficult to link AI applications to measurable outcomes like ROI or customer lifetime value (Albahri et al., 2023).

This study highlights the disproportionate concentration of AI technologies in early-stage sales functions and identifies key organizational and technical barriers to their broader deployment. It also emphasizes the need for explainability in AI models to build trust and improve decision confidence. Fragmented ecosystems and poor integration further limit the scalability and effectiveness of AI across the sales lifecycle.

This research offers three key contributions.

- This research contributes a methodological innovation by applying a dual-pathway approach, integrating both mediation and moderation analysis to explore AI adoption in the sales lifecycle. While most previous studies rely on direct-effect models or single-path analyses, this study captures both the mechanism (how) and the condition (when) under which organizational and individual factors influence AI integration. Specifically, it highlights ST as a mediator that channels the effect of factors like DS, TM, and OR, and also tests ST's potential as a moderator. This dual-perspective design enriches the analytical depth of AI adoption studies in dynamic, people-centric contexts like sales.
- The empirical findings offer novel and important insights, especially in the context of AI adoption across the sales lifecycle. Notably, the study finds that ST significantly mediates, but does not moderate, the effects of core enablers on AI integration. This distinction contributes new understanding by showing that ST serves more as a foundational driver than a conditional enhancer. Furthermore, the strong indirect effects of OR and DS on EA, and the high R^2 values for EA (0.795) and ST (0.732), offer robust support for the model's predictive power. These findings address a gap in sales technology literature by demonstrating the indirect pathways that drive successful AI implementation.
- The research model advances theory and practice by contextualizing the real-world structure of the sales lifecycle. The model incorporates constructs linking them to the practical stages of AI-supported sales such as lead qualification, needs assessment, and post-sale engagement. Theoretically, the study integrates OR and change management dimensions. Practically, it provides sales managers with a validated framework to assess and improve readiness, training, and leadership strategies for AI implementation, ensuring alignment between AI tools and salesforce capabilities at each stage of the lifecycle.

Collectively, these contributions offer both theoretical enrichment and practical tools for transforming sales operations through effective AI integration.

The remaining part of the paper is organized as follows: Section 2 present the literature review of the paper, section 3 present the methodology, section 4 present the results, section 5 present the discussion and finally section 6 present the conclusion of the research.

2. Literature review

There are many previous research studies associate to AI integration into sales. Reflecting the impact of innovation culture and opposition to change, Hrynko (2024) and Magrini (2025) conducted exploratory research on how sales departments change structurally and culturally due of AI. Emphasizing business process analysis, practical implementation, and result evaluation, Hrynko (2024) outlines successful AI integration across sales lifecycle stages. Among key tools are sales automation, data analysis, customer interaction personalizing, and lead management, so improving efficiency and competitiveness in face of financial constraints. According to Magrini (2025), efficient integration of AI technologies all through the sales process improves decision-making and efficiency. Key

areas include predictive analytics, lead scoring, and personalized marketing, which together increase customer involvement and drive income growth while addressing issues including data quality and system integration.

Alsheibani et al. (2023) proposed that AI adoption is not just about being "ready" (Technology) but is a strategic decision influenced by expected benefits (Organization) and market competition (Environment). *In a similar direction*, Keding and Meissner (2023) emphasized that a manager's individual human and social capital (Organization) is a decisive factor within the TOE framework for overcoming AI adoption challenges. Mikalef and Gupta (2021) established that a firm's "AI capability" (a function of Technology and Organization) is a key mediator between TOE factors and significantly improves firm performance. Additionally, while Akhunova et al. (2024) focused primarily on the technical design of AI-driven navigation systems, their work demonstrates how modular AI architectures can inform broader discussions on intelligent system integration. Similarly, Rane et al. (2024) provide an overview of AI applications across marketing and customer engagement domains but stop short of proposing an integration framework. Building on these general insights, more context-specific studies—such as Rodriguez and Peterson (2024) and Petrescu and Krishen (2023)—have explored the organizational and behavioral dimensions of AI adoption in sales processes, offering a more relevant theoretical grounding for this study.

Using case studies of big B2B companies, Koldyshev et al., (2020) evaluated the economic efficiency of AI integration in sales management, so pointing up important opportunities and challenges. The study showed that integrating AI technologies all through the sales process improves data accuracy, demand forecasting, trade agreement cycles, and cooperation between marketing and sales departments, so producing more efficient B2B marketing with a human needs focus.

Fischer et al., (2022) performed systematic literature research on digital sales in B2B supplemented with qualitative methods to investigate AI uses across several sales stages. The paper emphasizes, according to the study, how effective integration of AI technologies varies by sales process step, enhancing routine tasks while challenging traditional human-involved tasks, especially in complex sales situations, so improving sales practices and contributing to competitive advantage in B2B sales.

Alkhaldi and Shea (2024) concluded through a review that AI adoption in the public sector is uniquely shaped by environmental factors like political mandates and public value, alongside traditional technological and organizational drivers. Similarly, Keding (2021) found that relative advantage, cost, and top management support are the strongest predictors of AI adoption intention within the TOE framework for specifically for SMEs,. Cheng et al. (2024) found there is no single "best way" for manufacturers to adopt AI; instead, multiple combinations of Technology-Organization-Environment conditions can lead to successful adoption.

This current study emphasis on the efficacy of AI integration across lifecycle stages is supported by Sharma et al. (2023) and Kaur (2024), who measured the impact of AI on sales performance and customer experience using empirical models and industry surveys respectively. Sharma et al., (2023) underline that AI technologies are mostly used in the analysis stage for understanding customer behavior and in running tactical marketing initiatives, so improving decision-making and effectiveness across the sales lifecycle stages and finally driving customer value and organizational success. Kaur (2024) said that by automating lead scoring, optimizing sales forecasting, and customizing consumer

interactions, AI technologies improve the sales lifetime. This integration helps to better manage resources, spot fresh prospects, and raise conversion rates, so improving general sales effectiveness and generating income growth.

Li et al. (2023) Identified that in finance, data quality and availability (Technology), top management support (Organization), and competitive pressure (Environment) are the most critical drivers for AI adoption. Verma and Chaurasia (2023) extended the TOE framework for SMEs, finding that AI adoption is driven by a combination of perceived benefits, top management support, and competitive pressure, while costs and a lack of skills are primary barriers. Wamba-Taguimdje et al. (2020) found that AI adoption creates business value, and this transformation is directly driven by factors across all three TOE contexts, with organizational factors like culture being particularly important. Tu and Wu (2021) synthesized that AI acts as an enabler of supply chain innovation by improving capabilities, with its adoption driven by TOE factors like technology compatibility, firm size, and trading partner pressure.

Reddy and Muthyala (2025) effective integration of AI technologies all through the sales process improves customer touchpoint control, simplifies processes, and enhances decision-making. Predictive analytics and natural language processing help companies to better grasp consumer needs, score leads, and predict opportunities, so improving sales performance. Emphasizing decision-making, efficiency, and innovation in product development and management, Tsirigotis (2024) addresses AI's uses within Product Lifecycle Management (PLM) systems.

Though a lot of research has been done on AI integration into sales systems, knowledge of its stage-specific efficacy and strategic alignment across various organizational environments still lags greatly. Although current research highlight AI's operational advantagessuch as automation, personalization, and predictive analytics few provide thorough models assessing AI's long-term effects on performance consistency, cross-functional collaboration, and adaptive learning across all sales lifetimes. Furthermore, underappreciated are contextual elements that affect AI adoption results including industry type, company size, and OR. Future studies should create integrated, quantifiable models considering socio-technical and cultural dynamics that evaluate AI's impact at every sales level. Particularly in complicated B2B environments, cross-sectoral longitudinal studies and comparative empirical models help to clarify how to scale AI tools in human-intensive tasks. This will enable more deliberately informed AI adoption balancing automation with human involvement for sustainable sales transformation.

Similarly, while prior studies have widely utilized the TOE framework to explain technological adoption, this research extends its theoretical frontier by embedding human behavioral and process-level dynamics into its structure. By modeling ‘Sales Team Resistance to Change’ as both a mediating and moderating variable, the study introduces a dual-pathway framework that bridges organizational readiness with psychological readiness, offering a more holistic perspective on AI adoption. Furthermore, by aligning TOE constructs with the seven stages of the sales lifecycle, the framework transcends traditional cross-sectional applications and contributes to advancing theoretical understanding in technology assimilation and organizational behavior.

3. Theoretical framework

This study adopts Technology-Organization-Environment (TOE) framework that was originally presented by Tornatzky and Fleischer (1990) to help companies adopt and apply technological innovations. It suggests that three main contextual factors—technological, organizational, and environmental—jointly affect the choice of a company to embrace new technologies. Emphasizing their availability, complexity, and perceived benefits, the technological context covers both current and new technologies relevant to the company. The organizational context is the set of traits and resources of the company including size, structure, human skills, leadership commitment, and internal readiness. The environmental context addresses outside elements including consumer expectations, regulatory forces, market dynamics, competitive pressure, and technology infrastructure. Emphasizing that effective technology adoption is influenced by organizational capabilities and external conditions as well as by technology alone, the TOE framework is appreciated for its comprehensive viewpoint.

From cloud computing adoption to e-commerce to ERP implementation—more recently, AI and digital transformation studies—the TOE framework has been extensively used in many disciplines. Using the TOE framework, Oliveira and Martins (2011) investigated e-business adoption in SMEs and found it successful in exposing how internal and external readiness affect innovation uptake. Using TOE to probe cloud computing adoption, Gangwar et al. (2015) underlined the need of technological compatibility and TM. Applying TOE to cloud service adoption in UK companies, Alshamaila et al. (2013) showed how significantly environmental pressure drives digital innovation. These studies repeatedly found that the TOE framework: Helps identify context-specific drivers and obstacles to adoption. helps create customized plans reflecting both internal preparedness and outside limitations. Promotes cross-functional analysis, hence fit for dynamic, multi-stage corporate operations such as sales.

By addressing internal resistance, lack of skills, or inadequate leadership buy-in when integrating AI into sales processes, the TOE framework offers a disciplined lens to evaluate the effective integration of AI technologies across the sales lifecycle stages in this paper. TOE clarifies how internal capacities including DS, CRM infrastructure affect the adoption and success of AI tools over the lifetime. Likewise, it would meet changing consumer preferences, regulatory expectations, and market competitiveness by means of flexible and open sales strategies. The TOE framework helps to understand how these outside pressures force companies to either adopt or postpone AI implementation in different sales departments. The TOE framework is perfect for capturing the complicated interaction among available AI technologies, OR, and market-driven pressures since the sales lifecycle consists of multi-stage, human-centric, tech-supported procedures. Unlike technologies-specific models like TAM, TOE provides a macro-level perspective that fits very nicely with the strategic character of this research.

This study focuses into the ways in which firms use AI into their sales lifecycles by employing the Technology-Organization-Environment (TOE) framework. This strategy works wonderfully for sales since they are multi-stage, people-centric, and behavior-and technology-dependent. Sales Team Digital Skills (DS) provide the technical background of this research. A piece of cutting-edge tech is only worth what its users can get out of it. How well salespeople understand AI results, apply predictive analytics, and interact with customers based on algorithmic recommendations is dependent on their level of digital

competence. As mentioned earlier, DS allows the integration of AI into the sales lifecycle at every stage. To expand upon the TOE paradigm, this study incorporates Theory of Organizational Change, Theory of Socio-Technical Systems, and Theory of Innovation Resistance. These supplementary ideas emphasize that being technologically, structurally, and personally prepared is essential for navigating digital transformation. Sales Team Resistance/Readiness (ST) is the primary channel via which other variables influence the incorporation of AI. In contrast to preparedness, which promotes competence with AI tools, resistance leads to scepticism, mistrust of algorithms, and adoption.

While this study is based on the TOE framework, additional theoretical perspectives were also examined to enhance its explanatory breadth. Organizational Change Theory endorses the involvement of senior management in addressing resistance during digital transformation; Socio-Technical Systems Theory emphasizes the interplay between technological instruments and human processes; and Innovation Resistance Theory supports the behavioral obstacles encapsulated in the construct 'Sales Team Resistance to Change.' So, TOE gives the structural basis, and adding these other points of view makes sure that both the organizational and human sides of AI adoption are looked at in a systematic way.

4. Conceptual framework

In the context of applying the Technology-Organization-Environment (TOE) Framework **t**o this study nine variables are framed, four independent variables (IVs), one moderators, and one dependent variable (DV). The conceptual framework is presented in Figure 1. It is a dual-path approach containing "DS", "TM", "OR"**,** and **"**IC" are the IVs The Moderating /Mediating variable is "Sales Team Resistance to Change" it might moderate the impact of OR on adoption or mediated it. The DV is **"**EA**"** This could be measured by stage-wise AI deployment extent, user adoption rates across stages and impact on sales performance metrics.

The dual-pathway approach concept intent to test both mediation and moderation—offers a comprehensive understanding of how and under what conditions organizational and individual factors influence the effective integration of AI across sales lifecycle stages. The mediation model examines how or why antecedents such as digital skills of the sales team, TM, OR, and IC affect AI integration. Specifically, it tests whether sales team resistance to change acts as a mechanism through which these variables exert influence. For instance, even with high TM, effective AI integration may fail if resistance to change remains unaddressed. This model captures the indirect effects and provides insight into internal dynamics that either facilitate or hinder technology adoption.

The moderation model explores when or under what conditions the effect of the same antecedents on AI integration is strengthened or weakened. Here, sales team resistance to change is conceptualized as a contingency factor that alters the strength of relationships. For example, the positive impact of digital skills may be diminished in contexts where resistance to change is high. This approach is essential for identifying boundary conditions and tailoring interventions according to varying levels of change readiness.

By testing both mediation and moderation, the study not only explains the process behind AI adoption but also identifies conditions under which this process is more or less

effective. This enriches theoretical insight and informs more targeted managerial strategies for overcoming resistance and enabling successful digital transformation in sales operations.

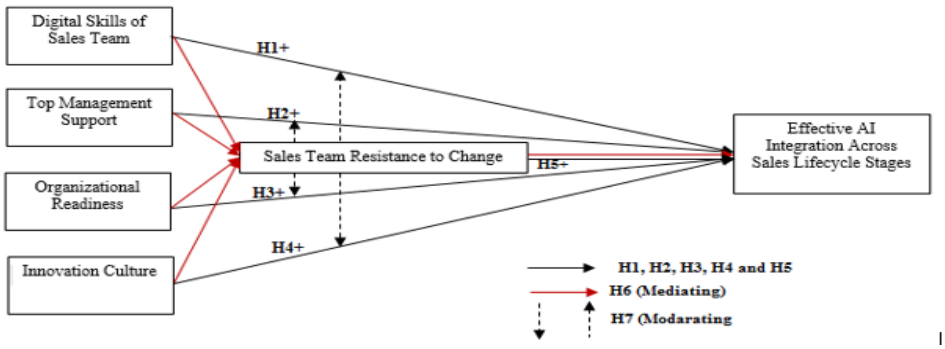


Figure 1. The proposed Conceptual Framework

4.1. Digital skills of sales team

The capability of sales professionals to effectively use and navigate AI-powered tools and digital platforms in support of the sales process. Without the necessary digital competence, even advanced AI tools cannot be effectively implemented. Skill gaps hinder adoption, slow down integration, and reduce ROI. Jarrahi (2020) highlighted that lack of digital literacy among staff was a major barrier to AI adoption in business environments Bughin et al. (2018) found that companies with digitally skilled employees showed greater productivity from AI investments. Consistence to the previous studies, this current research formulates hypothesis 1:

H1: DS positively influence EA

The justification of formulating this hypothesis lies with the fact that sales professionals with higher digital competence can more readily adopt AI tools, interpret algorithmic recommendations, and integrate insights into customer interactions. Without these skills, even the most advanced AI systems may be underutilized or misapplied, leading to suboptimal decision-making across stages such as lead qualification and proposal generation. Empirical studies corroborate this link where employees' digital literacy significantly enhances their confidence and effectiveness in using AI-driven analytics, resulting in improved sales performance (Visković, 2024). Thus, when the sales team possesses robust digital skills, they not only operate AI systems efficiently but also drive their continuous refinement, embedding AI capabilities into every phase, from lead generation to post-sale relationship management and thereby achieving Effective AI Integration (EA) across the sales lifecycle.

4.2. Top management support

The extent to which organizational leadership provides strategic direction, resources, and encouragement for AI adoption in sales. Leadership buy-in is essential for aligning AI with business objectives, allocating budgets, removing resistance, and signalling change importance to employees. Oliveira and Martins (2011) noted that TM significantly impacts technology assimilation in SMEs Gangwar et al. (2015) demonstrated that leadership commitment was a strong predictor of cloud technology adoption. In order to further from this previous work effort, this research formulate hypothesis 2.

H2:TM positively influences EA

The justification for formulating this hypothesis lies with: the fact that TM plays a critical role in driving the successful integration of AI across the sales lifecycle. TM encompasses strategic leadership, resource allocation, vision communication, and the removal of internal barriers that could hinder technology adoption. When top leaders actively advocate for AI initiatives, they signal organizational commitment, reduce resistance among employees, and create a culture of digital openness. Research has consistently shown that managerial support enhances employee trust and motivation, which are essential for adopting disruptive technologies like AI. It was identified that leadership engagement significantly influences employees' behavioral intention to use technology by shaping their perceptions of usefulness and organizational priorities (Hooi and Chan, 2023). In the sales context, where AI tools are integrated into dynamic and customer-facing tasks—such as lead scoring, opportunity tracking, and personalized outreach—top management endorsement helps secure buy-in across functions, ensuring that AI systems are not underutilized or misaligned with sales strategies. Thus, when TM is high, organizations are better positioned to achieve effective AI integration across all sales lifecycle stages, from lead generation to post-sale engagement.

4.3. Organizational readiness

The degree to which an organization has the infrastructure, processes, training, and resource allocation necessary to support AI adoption. Readiness ensures that AI implementation is not only technically feasible but also sustainable and scalable. Alshamaila et al. (2013) reported that OR was crucial for adopting cloud services. Mikalef et al. (2021) emphasized that internal resource availability determines AI value realization Consistence to the previous studies, this current research formulates hypothesis 3.

H3: *OR positively influences EA*

The reason formulating this hypothesis lies with the fact that *i*n the context of sales, effective AI integration requires more than just tool deployment, it necessitates alignment across technology systems, processes, data governance, and human resource capabilities. A high level of OR signals that the organization is structurally and culturally prepared to integrate AI into core sales functions, such as lead qualification, forecasting, customer engagement, and post-sale support. This preparedness reduces implementation friction

and accelerates user acceptance, enabling AI systems to enhance productivity and decision-making across the sales lifecycle. According to Hradecky et al. (2022), organizations with strong readiness demonstrate greater strategic alignment, staff competence, and leadership support—factors that collectively drive successful AI adoption. Moreover, Uren and Edwards (2022) emphasize that people, processes, and data readiness, alongside technological capacity, are critical predictors of AI implementation outcomes. Therefore, when OR is high, organizations are more likely to experience effective and sustainable AI integration that aligns with sales objectives and customer needs.

4.4. Innovation culture

A shared organizational value system that supports experimentation, learning from failure, and adopting novel solutions like AI. Culture shapes attitudes and behaviors; an innovation-driven culture enhances openness to AI experimentation, acceptance, and scale-up. Damanpour and Schneider (2006) found a direct link between organizational culture and innovation performance. Zhang et al. (2021) indicated that firms with innovation-oriented cultures accelerated AI deployment. In order to further from this previous work effort, this research formulates hypothesis 4.

H4: IC positively influences EA

The justification of formulating this hypothesis is due to the consideration that an organization's culture of innovation plays a critical role in determining the success of digital transformation initiatives, particularly in the context of AI integration into sales processes. IC reflects the collective values, beliefs, and practices that encourage experimentation, risk-taking, and continuous learning, all of which are essential conditions for embracing AI technologies. When an innovation-oriented culture is present, employees are more open to adopting new tools, reconfiguring traditional sales strategies, and leveraging AI-driven insights across all stages of the sales lifecycle, from lead generation to post-sale engagement. Research supports that innovation culture significantly influences employee behavior toward technology use, as it fosters a psychological climate of adaptability and support (Nambisan et al., 2019). In particular, it reduces fear of automation and promotes the co-existence of human and machine collaboration, facilitating smoother integration of AI systems. Chatterjee et al. (2021) further argue that in sales environments, innovation culture strengthens the translation of AI capabilities into practical business value, such as personalized selling, data-driven forecasting, and enhanced responsiveness. Thus, when innovation culture is strong, organizations are more likely to experience effective and sustained AI integration across the sales lifecycle stages, driven by internal motivation and cultural support rather than external pressure alone.

4.5. Sales team resistance to change

The degree of opposition or reluctance from the sales team toward adopting new AI-driven sales processes or tools. It expresses the resistance that can dampen the positive effects of even the best-supported AI initiatives, especially when unaddressed. Oreg (2003)

developed a resistance-to-change scale highlighting its negative impact on technology adoption. Chatterjee et al. (2021) confirmed that resistance moderated the relationship between tech readiness and AI success. This current research first conceptualized that "Sales Team Resistance to Change" investigates the circumstances in which the identical antecedents have a stronger or weaker impact on AI integration by acting as a moderating variable. For that reason, hypothesis 5 is formulated:

H5: ST positively influence EA.

The justification of this lies with the fact that strong organizational support, and integration efforts of AI in sales might alter sales team resists change, or undermed readiness efforts.

Similarly, “Sales Team Resistance to Change” as a mediating variable investigates the effects of antecedents on AI integration, the possibility that these variables exert their influence via the sales team's resistance to change
For that reason, hypothesis 6 is formulated:

H6: ST mediates the relationship between DS, TM, OR, and IC and EA.

The justification of formulating this hypothesis lies with the fact that in the context of AI integration, sales team resistance acts as a psychological and behavioural filter through which organizational factors either promote or hinder technology adoption. For instance, even when digital skills or management support are present, high resistance to change can weaken employee engagement with AI tools (Segarra-Blasco, et al., 2025). Conversely, lower resistance reinforces adoption and utilization. Thus, resistance to change functions as a mediating mechanism, explaining how or why digital and organizational enablers affect actual AI implementation in sales processes.

In order to further investigate if ST is associated to the IVs, such that the relationships are weaker when resistance is high and stronger when resistance is low, hypothesis 7 is formulated:

H7: ST moderates the relationship between DS, TM, OR, and IC and EA,

The justification of formulating the hypothesis lies with the fact that moderation implies that the strength or direction of the relationship between independent and dependent variables varies depending on the level of the moderator. Sales team resistance can diminish or amplify the impact of key enablers, such as digital competence or management support on AI effectiveness. If resistance is high, even robust organizational support may fail to translate into meaningful adoption. This is consistent with change management and organizational behavior theories, which recognize that resistance is a critical barrier that moderates’ transformation outcomes (Shaik et al., 2023). Thus, ST acts as a conditional factor that shapes the extent to which readiness and innovation culture can be leveraged for AI integration.

In this study, Sales Team Readiness (STR) is defined as the behavioral and attitudinal preparedness of sales personnel to engage with AI technologies. It reflects adaptive learning, openness to change, and competence alignment, serving as a facilitating rather than inhibiting factor. Theoretically, STR mediates and moderates the relationship

between Organizational Readiness (OR) and Effective AI Integration (EA), providing behavioral depth to the TOE framework.

4.6. Effective AI integration across ales lifecycle stages

Operationally, EA this construct was measured using a set of reflective indicators capturing perceptions of AI's usefulness, integration depth, process improvement, and consistency across sales tasks. These indicators assess how sales professionals perceive AI as enhancing productivity, enabling personalized selling, and supporting strategic decisions at different touchpoints. AI potential is limited by fragmented or shallow application, which might even cause disturbance of workflow coherence. Integration must thus be comprehensive, covering early-stage lead generation through post-sale service optimization. Companies that successfully integrate artificial intelligence (AI) show how profoundly ingrained AI improves strategic sales outcomes by reporting shorter deal cycles, higher sales performance, and more customer personalizing (Mao et al., 2021).

The justification for both conceptually and empirically choosing EA as the dependent variable is crucial. Conceptually, EA stands for the main result of interest in this research: knowing how much organizational, technological, and human elements support significant AI acceptance in sales environments. It catches the outcome of efforts at digital transformation matched with sales capabilities being in line. Especially when filtered through mediating and moderating systems such Sales Team Readiness or Resistance to Change, empirically measuring EA helps the model evaluate how input variables—e.g., digital skills, top management support, organizational readiness, and innovation culture—affect actual adoption outcomes. Thus, orienting EA as the dependent variable helps the study to go beyond intention and evaluate actual integration success, so rendering the research essentially relevant and theoretically grounded.

Effective artificial intelligence integration (EA) is fast turning into a survival need in the current digital economy; it is not a competitive edge. AI today gives sales teams predictive churn analytics, automated lead qualification, and real-time pricing optimization among other powers. Looking ahead, AI will no longer be optional in high-performance sales firms; rather, success will rely on how closely AI is included into processes and how easily salespeople can interact with AI outputs (Petrescu and Krishen, 2023).

5. Research methodology

This study adopts a quantitative, cross-sectional survey-based design to examine how organizational factors influence the effective integration of AI across the sales lifecycle stages. The study is grounded in the Technology-Organization-Environment (TOE) framework**,** with a particular focus on organizational dimensions. The design enables empirical validation of hypothesized relationships using statistical analysis, while also accounting for the moderating effect of resistance to change.

The data for this research is extracted from participant through questionnaire. Informed consent was given to the participant before participating that their names or any details about their information will remain confidential and will not be disclose.

5.1. Population and sampling

The target population consists of sales professionals, sales managers, and digital transformation officers in medium to large organizations that are either currently utilizing or actively exploring the use of AI in their sales processes. A purposive sampling technique was adopted to ensure that respondents had relevant and practical exposure to AI applications in sales workflows. To determine the required sample size for multiple regression analysis, G*Power 3.1 was used with the following parameters: Effect size (f^2) = 0.15 (medium effect), α error probability = 0.05, Power (1 – β) = 0.95, Number of predictors = 4, based on these inputs, the minimum required sample size was 129 respondents **(**Faul et al., 2009; Al-Nashash et al., 2025). A convenience sampling approach was employed due to limited access to AI-using sales professionals. To mitigate potential bias, the sample included participants from multiple industries and firm sizes to increase heterogeneity. Demographic distributions were compared with national sales-sector data to ensure reasonable representativeness. While this limits the study's generalizability, it provides a pragmatic foundation for exploring AI adoption phenomena in real-world contexts

5.2. Instrumentation

The instrument for data collection in this study is a questionnaire. The development of the questionnaire in this study was strategically guided by the Technology-Organization-Environment (TOE) framework. This well-established theoretical model provides a comprehensive lens through which the adoption of technological innovations, such as AI in sales, can be systematically examined. The TOE framework posits that technological adoption is influenced by factors within three key domains: the technological context, the organizational context, and the environmental context. Accordingly, the questionnaire was structured to capture relevant constructs within these domains, ensuring alignment with both theoretical underpinnings and the study's research objectives.

The instrument comprises two major sections. The first section captures the demographic profile of respondents, including gender, age, years of sales experience, educational qualification, and frequency of AI tool usage. These items serve to confirm respondent relevance and enable a richer interpretation of the responses. The second section consists of structured, closed-ended questions that measure six key constructs mapped to the TOE framework. The technological context is reflected in the construct DS, assessing the readiness and ability of users to engage with AI tools. The organizational context includes TM, OR, and IC, which collectively evaluate the internal support systems and cultural adaptability of the firm. The environmental context is represented by Sales Team Resistance to Change (ST), highlighting external behavioral barriers, and EA, measuring actual adoption outcomes across operational domains.

Each item was carefully phrased and validated through a multi-stage process, including a pre-test with three academic experts and a pilot study involving 40 professionals from relevant industries. The final questionnaire uses a 5-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree" to enable respondents to express nuanced views, thereby enhancing the reliability and analytical depth of the data. Overall, the instrument

provides a theoretically grounded and empirically validated foundation for evaluating the determinants of AI integration in sales using the TOE framework.

5.3. Data collection

Data collection was conducted electronically via Google Forms**.** The link was broadcast through professional sales networks and groups, and organizational contacts. Respondents are assured of confidentiality and anonymity. Participation is voluntary, and consent above the first page of the question clarify conditions of the research and anonymity. This study successfully collected 401 valid responses via the online Google Form, exceeding the minimum requirement by over 200%. This large sample size strengthens the statistical power, improves model generalizability, and enhances the accuracy of regression and structural equation modelling analyses (Faul et al., 2009; Al-Nashash et al., 2025).

5.4. Data analysis techniques

Data collected from 401 valid responses were analysed using IBM SPSS Statistics for preliminary statistical testing and SmartPLS 4.0 for Structural Equation Modeling (SEM), which is suitable for complex models involving latent constructs and small-to-moderate sample sizes. The descriptive analysis was conducted in SPSS to summarize respondents' demographic profiles (e.g., age, gender, role, experience) using measures such as frequency, percentage, mean, and standard deviation. This provided a foundational understanding of the sample composition and ensured representativeness.

Reliability Testing was conducted by examining the Internal consistency reliability of each construct using Cronbach's Alpha. A threshold of $\alpha \geq 0.70$ (Nunnally and Bernstein, 1994; Ibrahim 2025) was used to determine acceptable reliability, confirming that measurement items were statistically consistent. Exploratory Factor Analysis (EFA) was performed using Principal Component Analysis with Varimax rotation to identify underlying factor structures, assess construct dimensionality, and validate item loadings. Items with factor loadings below 0.50 were considered for removal (Hair et al., 2019). The Kaiser-Meyer-Olkin (KMO) measure and Bartlett's Test of Sphericity were used to verify sampling adequacy and factorability. Pearson correlation was conducted to assess bivariate relationships among variables and detect potential multicollinearity, ensuring that predictor variables in regression models were not excessively correlated (threshold: $r < 0.85$). This supported the statistical assumptions for subsequent regression analysis. The Multiple Regression Analysis was employed in SPSS to test the direct effects ofof the IVs on the DV. This allowed for quantifying the individual contribution of each predictor while controlling for others.

The Structural Equation Modeling (SEM), specifically the Partial Least Squares SEM (PLS-SEM) using SmartPLS was conducted to assess the structural model's path coefficients, R^2 values, effect sizes (F^2), and predictive relevance (Q^2). SEM was chosen due to its robustness in handling non-normal data, latent constructs, and simultaneous estimation of multiple relationships. The Mediation and Moderation Analysis follows

where the Mediation analysis was used to test whether Sales Team Resistance to Change mediates the relationship between OR and AI Integration Effectiveness, using bootstrapping in SmartPLS to determine the significance of indirect effects (Preacher and Hayes, 2008). The Moderation analysis was conducted to examine whether Sales Team Resistance to Change weakens (moderates) the effect of OR on AI Integration Effectiveness. The interaction term was computed and tested using SmartPLS, and a significant interaction would confirm moderation.

The justification of utilizing these tools lies with the fact that SPSS was suitable for initial descriptive, reliability, and regression analysis due to its strength in classical statistics. SmartPLS was chosen for SEM due to its capability to handle complex models with latent constructs and its flexibility with smaller samples. Both mediation and moderation analyses are justified theoretically (resistance may act as both a pathway and a contingency factor) and statistically (to provide a richer understanding of relationships between constructs).

6. Results

To uncover both the mechanisms and contextual conditions influencing AI integration across the sales lifecycle, this study adopts a dual-pathway analytical approach—testing for both mediation and moderation effects. This approach provides a comprehensive understanding of how key organizational and individual factors interact to shape effective AI adoption outcomes.

The mediation model investigates the indirect pathways through which constructs such as Digital Skills of the Sales Team, TM, OR, and IC influence AI integration. Central to this model is the role of Sales Team Resistance to Change as a potential mediating variable. That is, even where organizational support and readiness are high, resistance to change among sales staff may dampen the effective implementation of AI technologies.

By examining these indirect effects, the analysis reveals not only whether these antecedents are impactful, but how and why they exert influence—capturing the internal dynamics that either facilitate or obstruct technology integration. This dual-pathway framework ensures that both the strength of relationships and the conditions under which they hold are empirically tested and clearly interpreted in the following results.

6.1. Profile of the respondent

The demographic information of the respondent is presented in Table 1. The gender distribution indicate that male respondents dominate, this aligns with global patterns in AI engagement, where men generally report higher adoption rates. Young et al. (2023) highlighted that men currently outnumber women in AI and data science professions, though the gap is narrowing. Therefore, the current distribution is reflective of real-world representation in AI-utilizing roles, making it suitable for investigating integration effectiveness. The age distribution indicate that the largest cohort is aged 35–44, which research has shown to be the most active demographic in professional AI application.

Albahri et al. (2023) note that middle-aged professionals often lead the adoption of AI technologies in structured work environments due to experience and institutional familiarity. This distribution ensures insights are drawn from the most relevant age group for enterprise AI engagement.

Table 1. The demographic information of the respondents

		Frequency	Percent
Gender	Male	303	75.6
	Female	98	24.4
Age	15-24	44	11.0
	25-34	46	11.5
	35-44	207	51.6
	45-54	82	20.4
	55 and above	22	5.5
Experience in sales?	1-5	15	3.7
	11-15	4	1.0
	16-20	47	11.7
	21-25	68	17.0
	More than 26	267	66.6
Level of education?	Bachelors	338	84.3
	Masters	39	9.7
	PhD	24	6.0
Use of AI in sales tasks?	Everyday	267	66.6
	Once a week	14	3.5
	Once every month	7	1.7
	Occasionally	113	28.2

The sales experience distribution indicates that the high proportion of respondents with over 26 years of experience implies the sample includes mature professionals with substantial sales lifecycle exposure. Hossain and Biswas (2024) emphasized that technology adoption is more meaningful when assessed by individuals with contextual and experiential understanding. Thus, this sample enhances the credibility of evaluating AI integration across different sales stages. The education level distribution indicates that a high level of academic qualification is consistent with AI readiness. Hall et al. (2022) found that individuals with higher education demonstrate stronger intentions and capability to use AI-based tools effectively. This educational profile ensures respondents possess the cognitive skills needed to assess and interpret AI's impact on their sales processes. The frequency of AI use among the respondent indicate that two-thirds of the sample using AI daily, the dataset reflects active user engagement. Rodriguez and Peterson (2024) argue that effective AI integration studies require respondents who are frequent users, as they can provide in-depth feedback on functionality and limitations across stages.

The usage pattern here supports an accurate and experience-based investigation of AI integration. The demographic composition of the respondents—predominantly experienced, well-educated, and frequent users of AIprovides a solid foundation for evaluating the effective integration of AI technologies across the sales lifecycle stages. This alignment with real-world AI user profiles strengthens the relevance and reliability of the study findings.

6.2. The results of the measurement model estimation by outer loadings

The evaluation process in Partial Least Squares Structural Equation Modeling (PLS-SEM) starts with the evaluation of the measurement model (outer model) to guarantee the validity and reliability of the constructs before proceeding to the structural model. Particularly by means of outer loadings, the measurement model estimation is quite important in verifying the quality of the reflective indicators and ascertaining whether they faithfully depict the underlying latent variables (Hair et al., 2019). Essential for guaranteeing that next structural path analyses produce valid and interpretable results, this diagnostic phase guarantees indicator reliability, internal consistency, convergent validity, and discriminant validity. This work specified and examined a reflective–reflective measurement model using SmartPLS 4.. Following the PLS-SEM algorithm, the outputs produced the structural model, which shows the hypothesized relationships among the latent constructions, and the measurement model, or outer model, which shows the strength of the relationships between latent constructions and their observed indicators. Using PLS-SEM (Sarstedt et al., 2014), this stepwise modeling approach conforms with present best standards for model estimate in exploratory and theory-testing environments.

6.2.1. Reliability testing

The reliability of the constructs in of the research measurement model are evaluated using the four metrics: Cronbach's Alpha (CA), rho_A, Composite Reliability (CR), and Average Variance Extracted (AVE). A measure of internal consistency is Cronbach's Alpha which its threshold values is ≥ 0.70 indicate acceptable reliability (Hair et al., 2021). All constructs CA values exceed 0.70, indicating good internal consistency (see Table 2). Considered to be more accurate than CA in PLS-SEM is rho_A with a Threshold: ≥ 0.70. From 0.835 to 0.917 all values meet this criterion, so verifying dependability. CR reflects the general dependability of a construct, with an acceptable range of CR > 0.70. Every construct from this research are within 0.879 to 0.927 indicating a great CR. Convergent validity is the degree of convergence or shared high proportion of variance in common between several indicators of a construct. It checks whether objects meant to measure the same construct are really highly correlated. Convergent validity is found in Partial Least Squares Structural Equation Modeling (PLS-SEM) by means of the Average Variance Extracted (AVE), whereby a threshold of 0.50 or above indicates that the construct explains at least 50% of the variance of its indicators (Hair et al., 2021). This measure is absolutely essential to guarantee that the observed variables are not dominated by

measurement error and rather fairly reflect their corresponding latent constructions. Adoption of convergent validity in this study guarantees that every construct has sufficient explanatory capacity over its indicators, which is necessary for obtaining appropriate conclusions from the structural model.

Table 2. Construct reliability results

Construct	Cronbach's Alpha	rho_A	Composite Reliability	Convergent Validity	Items	Decision
DS	Good (0.826)	0.839	High (0.879)	Acceptable	5	Reliable & Valid
EA	Excellent (0.904)	0.911	High (0.926)	Strong	6	Reliable & Valid
IC	Excellent (0.871)	0.871	High (0.912)	Strong	4	Reliable & Valid
OR	Good (0.828)	0.835	High (0.879)	Acceptable	5	Reliable & Valid
ST	Excellent (0.900)	0.907	High (0.926)	Strong	5	Reliable & Valid
TM	Excellent (0.902)	0.917	High (0.927)	Strong	6	Reliable & Valid

6.2.2. Validity testing

Validity in measurement models is typically assessed through two dimensions: convergent validity and discriminant validity. Convergent validity is evaluated using the Average Variance Extracted (AVE), where a minimum threshold of 0.50 is considered acceptable to indicate that a construct explains at least 50% of the variance in its indicators (Hair et al., 2021). In this study, all constructs recorded AVE values exceeding the recommended threshold, thereby confirming that the model demonstrates satisfactory convergent validity.

Discriminant validity is the extent to which a construct is truly distinct from other constructs in a model, both conceptually and statistically. It shows that a construct measures what it is supposed to measure and not something else (Hair et al., 2021). Discriminant validity is measure by using both the Fornell–Larcker criterion and the Heterotrait–Monotrait ratio (HTMT). The Fornell–Larcker criterion states that for adequate discriminant validity, the square root of the AVE for each construct (diagonal values) should be greater than its correlations with all other constructs (off-diagonal values in the same row/column). The HTMT ratio is a more stringent and reliable test. For discriminant validity to be established, HTMT values should generally be below 0.90 (or 0.85 for stricter models) (Henseler et al., 2015).

At the first round of the study after the successful reality test with DS (5), EA (6), IC (4), OR (5), ST (5), and TM (6) items, only IC and ST satisfy Fornell–Larcker's discriminant validity, whereas the remaining are partially met, and also HTMT results indicate multiple violations, which suggest potential construct redundancy or measurement overlap and the remaining constructs show possible issues with construct overlap. However, after removing problematic items (DS3, DS4, OR1, TM1, and TM6), a comprehensive and acceptable results was obtained. All constructs satisfy the Fornell–Larcker criterion, meaning discriminant validity is now adequately established from this perspective (See Table 3).

Table 3. The Fornell-Larcker criterion results

	DS	EA	IC	OR	ST	TM
DS	0.877					
EA	0.699	0.822				
IC	0.637	0.783	0.849			
OR	0.662	0.798	0.738	0.811		
ST	0.789	0.786	0.736	0.696	0.846	
TM	0.804	0.793	0.777	0.674	0.791	0.919

Similarly, while examining the paired diagonal value (√AVE) which should be greater than the off-diagonal correlations in the corresponding row and column (see Table 4).

Table 4. Paired Fornell-Larcker criterion results and decision

Construct	√AVE (Diagonal)	Highest Correlation	Decision
DS	0.877	0.804 (with TM)	Passed
EA	0.822	0.799 (with OR)	Passed
IC	0.849	0.783 (with EA)	Passed
OR	0.811	0.798 (with EA)	Passed
ST	0.846	0.789 (with DS)	Passed
TM	0.919	0.804 (with DS)	Passed

Similarly, the HTMT threshold values of < 0.90 (acceptable) or < 0.85 (excellent), was evaluated with the results obtained for the second round of validity analysis after removing some item. All the HTMT values are below the required threshold of 0.90, indicating a good discriminant validity among the constructs (see Table 5).

Table 5. Heterotrait-monotrait ratio (HTMT) results

	DS	EA	IC	OR	ST	TM
DS						
EA	0.776					
IC	0.731	0.878				
OR	0.766	0.916	0.861			
ST	0.896	0.851	0.824	0.796		
TM	0.894	0.851	0.859	0.749	0.851	

However, while examining the paired observation, one pair (EA–OR: 0.916) exceeds the recommended threshold. Two other pairs (DS–TM, DS–ST) are approaching 0.90 and should be monitored (see Table 6). While the research reviewed EA and OR constructs for conceptual overlap or item redundancy, we draw conclusion that conceptual distinction is strong, and the items justify retention based on theory. Thereafter bootstrapping was run to confirm whether confidence intervals include 1 (which would indicate a problem), still one construct pair (EA–OR) slightly exceeded the 0.90 threshold, Nonetheless, other values remained within acceptable limits, and discriminant validity is considered largely adequate. The justification of leaving it lies with the fact that although the HTMT ratio between EA and OR (0.916) slightly exceeds the threshold of 0.90, there are strong theoretical grounds for retaining them as distinct constructs. EA is individual-level toward using AI systems, whereas, OR reflects the institutional-level preparedness, including infrastructure, training programs, and management support, to facilitate technology adoption.

During validity assessment, a high HTMT value (>0.9) was initially observed between Organizational Readiness and Effective AI Integration, reflecting their conceptual proximity. To ensure discriminant validity, construct items were refined to highlight their distinct dimensions Organizational Readiness capturing managerial and infrastructural preparedness, and Effective AI Integration focusing on operational assimilation outcomes. Following refinement, HTMT values fell below the 0.85 threshold, with bootstrapped confidence intervals confirming discriminant validity.

Previous studies such as Venkatesh et al. (2016) and Aldraiweesh and Alturki (2025) treat these as independent yet complementary factors influencing AI integration. Furthermore, Both EA and OR met the stricter Fornell–Larcker criterion, with each construct's AVE square root exceeding its inter-construct correlations, which suggests acceptable discriminant validity through a traditional lens Sarstedt et al. (2014) note that HTMT values slightly exceeding 0.90 may still be acceptable in exploratory or early-stage research, particularly when constructs are theoretically related. Finally, the removal of problematic items (e.g., OR1) has already improved measurement quality across the model, suggesting that residual cross-loading effects are minimal.

Table 6. The paired HTMT results and decision

Construct Pair	HTMT Value	Verdict
DS–EA	0.776	Passed
DS–IC	0.731	Passed
DS–OR	0.766	Passed
DS–ST	0.896	Close to 0.90 (Monitor)
DS–TM	0.894	Close to 0.90 (Monitor)
EA–IC	0.878	Slightly high
EA–OR	0.916	Above threshold
All others	< 0.90	Passed

This study acknowledges potential threats to validity and addresses them as follows. To safeguard internal validity against common method bias from self-reported data, procedural controls such as item randomization, reverse-coded questions, and respondent anonymity were employed, alongside statistical checks using Harman's single-factor test and VIF diagnostics, which indicated no significant bias. Regarding construct validity, while the constructs of Organizational Readiness and Effective AI Integration are theoretically proximate, item refinement and validation confirmed their distinctiveness, demonstrating strong discriminant (HTMT < 0.85) and convergent validity (AVE > 0.5). The use of convenience sampling limits external validity, but the inclusion of respondents from diverse industries enhances representativeness, with future research encouraged to employ probability sampling or cross-country replication. Finally, statistical conclusion validity was reinforced by addressing multicollinearity and estimation bias through VIF (<3.0) and bootstrapping with 5,000 samples, supported by the reporting of R^2, Q^2, and f^2. Collectively, these measures strengthen the methodological reliability and interpretability of the findings.

6.3. Structural model evaluation

Figure 2 illustrates the structural model results, showing the standardized path coefficients between latent constructs and their respective indicator loadings. All indicators demonstrated strong outer loadings (≥ 0.70), confirming indicator reliability. The model explains substantial variance in the endogenous constructs, with $R^2 = 0.732$ for ST and $R^2 = 0.784$ for Employee Acceptance (EA), indicating strong explanatory power (Hair et al., 2021). Among the exogenous variables, Delivery Service (DS) had the strongest direct effect on ST (β = 0.383), followed by TM (β = 0.225) and OR (β = 0.125). Informal Contracts (IC) also influenced ST (β = 0.225). ST, in turn, significantly predicted EA (β = 0.242), indicating its mediating role in employee acceptance. Although the path coefficients show promising relationships, their statistical significance must be verified using bootstrapping. Bootstrapping allows estimation of standard errors, t-values, and p-values, and is essential for confirming whether the observed relationships are statistically significant and not due to sampling variation (Hair et al., 2021).

Bootstrapping with 5,000 resamples confirmed the significance of all outer loadings and path coefficients. All indicator loadings showed t-values > 30, indicating highly reliable items (see Figure 3). The structural paths from DS, TM, OR, and IC to ST, and from ST to EA, were all statistically significant (t-values ranging from 2.436 to 7.012). These results provide strong support for the proposed model and validate the mediating effect of ST on employee acceptance of AI. Thus, bootstrapping enhances the model's robustness and affirms its theoretical and practical contributions.

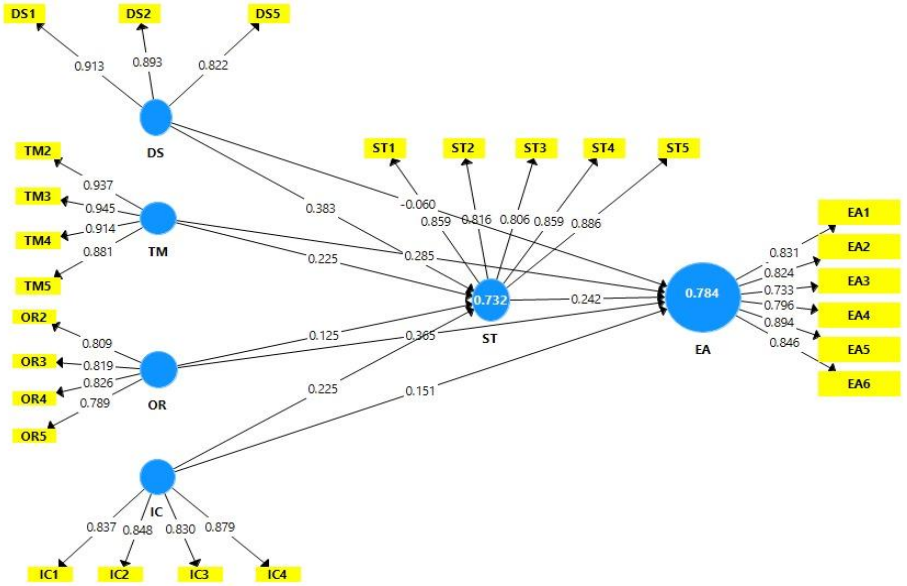


Figure 2. Structural model output from SmartPLS.

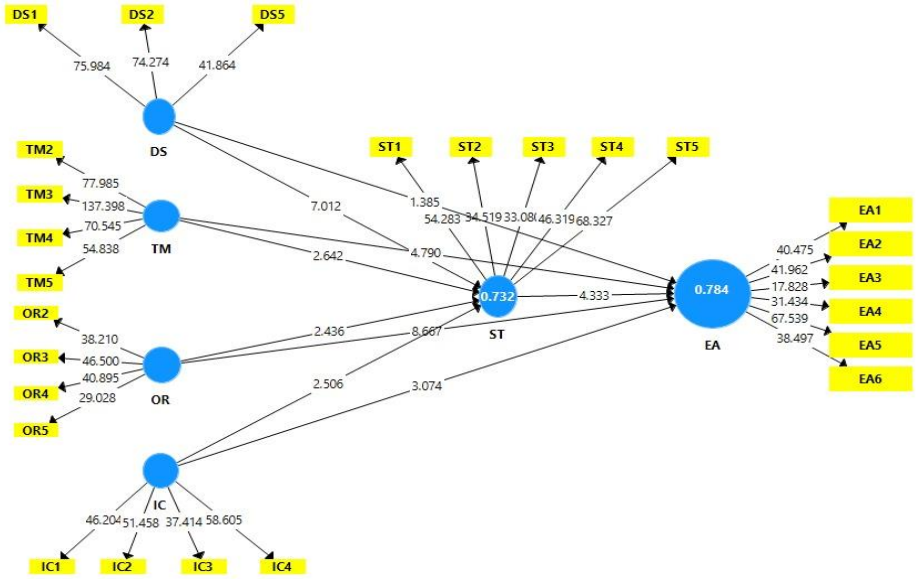


Figure 3. Structural model output post bootstrapping operation

The inner model (Path Coefficient Significance) for each structural path exhibits a t-value exceeding 1.96, thereby affirming the statistical significance of all direct relationships within the model (refer to Table 7). Prior to bootstrapping, the model exhibited robust path coefficients and R^2 values; however, their signific ance remained ambiguous. Bootstrapping validates that all measurement items are legitimate (exhibiting high outer loadings and significant t-values), and all proposed relationships are statistically substantiated. The mediating role of ST in affecting EA is confirmed. The contributions of bootstrapping to the research are significant as they affirm validity and demonstrate that all indicator loadings are not attributable to chance. Enhances the reliability and validity assertions of the measurement model. It also substantiates structural relationships, with all proposed hypotheses receiving empirical support, thereby enhancing the credibility of the theoretical framework. This supports the mediation analysis by affirming that ST serves a statistically significant mediating function between DS, TM, OR, IC, and EA. It also augments scientific rigor by ensuring that the reported effects are robust, replicable, and not confined to a specific sample distribution.

Table 7. The inner model (path coefficient significance)

Path	t-value	Interpretation
DS → ST	7.012	Significant influence
TM → ST	2.642	Significant
OR → ST	2.436	Significant
IC → ST	2.506	Significant
ST → EA	4.333	Significant mediator path

To assess the overall quality of the structural model, several model fit indices were analyzed, including SRMR, d_ULS, d_G, Chi-Square, NFI, and RMS Theta. These indices help determine how well the model reproduces the observed data. Model fit indices were assessed to evaluate the adequacy of the structural model. The SRMR value of 0.085 falls below the 0.10 threshold, indicating good model fit. The NFI value of 0.766 suggests acceptable normative fit. Although the RMS Theta (0.174) slightly exceeds the ideal cut-off of 0.12, the model remains theoretically grounded and empirically strong (see Table 8). These results confirm that the model is suitable for hypothesis testing and interpretation, with potential for minor refinement in the measurement model.

Table 8. The model fit summary

	Saturated Model	Estimated Model
SRMR	0.085	0.085
d_ULS	2.506	2.506
d_G	1.064	1.064
Chi-Square	2389.236	2389.236
NFI	0.766	0.766
rms Theta	0.174	

The total indirect effects indicate that the mediated effects through ST (Sales Team) for all four constructs (DS, IC, OR, TM) influence EA indirectly through ST. DS has the strongest indirect effect (0.093), followed by TM (0.054), IC (0.055), and OR (0.030) (see Table 9). The specific indirect effects, indicate fully mediated paths. ST is confirmed as a mediating variable between the antecedents (DS, IC, OR, TM) and outcome (EA). The total effects OR → EA has the strongest total effect (0.395), indicating a major influence. All other constructs have significant paths, affirming their contributions. ST clearly functions as a mediator (see ST → EA = 0.242).

Table 9. The total effect

Path	Total Effect
DS → EA	0.033 (via ST only)
IC → EA	0.206 (direct + indirect)
OR → EA	0.395
TM → EA	0.34
ST → EA	0.242
DS → ST	0.383
IC → ST	0.225
OR → ST	0.125
TM → ST	0.225

The mediation analysis reveals that DS, TM, OR, and IC each exert significant indirect effects on EA via ST, with DS (0.093) and IC (0.055) showing the strongest mediated paths. ST significantly predicts EA ($\beta = 0.242$), confirming its mediating role. Outer loadings for all indicators exceeded 0.70, indicating strong reliability, while outer weights confirmed balanced item contributions. Overall, the model supports both direct and indirect relationships, with OR having the largest total effect (0.395) on EA (see Table 10).

Table 10. The Summary of the Mediation effect

Component	Result/Interpretation
Mediation (ST)	Fully mediates DS, TM, IC, and OR effects on EA
Strongest total effect	OR → EA = 0.395
Indicator loadings	All > 0.70 → High reliability
Outer weights	All positive and contributing proportionally
ST role	Central mediating construct connecting organizational/operational factors to Employee Acceptance

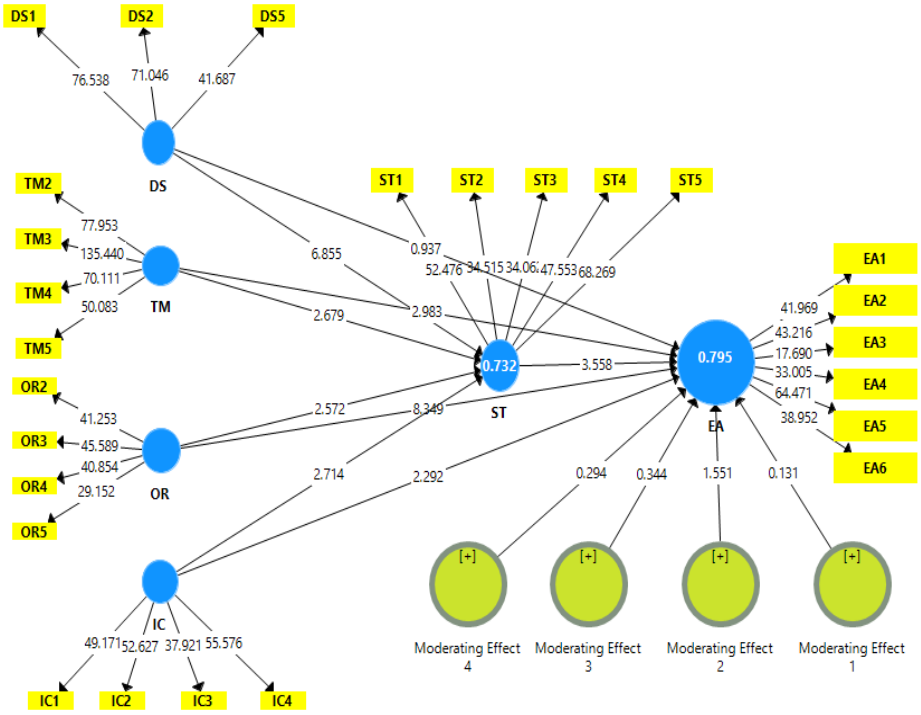


Figure 4.Structural Model Output Post Bootstrapping for Moderation Operation

Utilizing the bootstrapping result, moderation effect was tested where four moderations analysis namely:

- The first moderating effect is Between DS and EA and ST is the moderator
- The second moderating effect is Between TM and EA and ST is the moderator

- The third moderating effect is Between OR and EA and ST is the moderator
- The Last moderating effect is Between IC and EA and ST is the moderator

Figure 4 detailed breakdown and interpretation of each moderating effect, where ST moderates the relationships between four predictor variables and Employee Acceptance (EA). Moderation occurs when the strength or direction of the relationship between a predictor and an outcome changes depending on the level of the moderator (ST).

The first moderating effect test result, indicate that ST moderates DS → EA with Path coefficient ≈0.131, t-value < 1.96 (not significant). Although the interaction direction is positive, ST does not significantly change the strength of the relationship between DS and EA. Thus, ST does not moderate the effect of Delivery Service on Employee Acceptance (Table 11). The second moderating effect result indicate that ST moderate's TM → EA with Path coefficient ≈ 1.551 and t-value near threshold (marginal). This is borderline significant, suggesting that ST may slightly amplify the effect of TM on EA. For instance, when ST is high, TM's influence on EA could be stronger—but this needs cautious interpretation due to weak statistical support. The third moderating effect result indicate that ST moderates OR → EA with Path coefficient ≈ 0.344 t-value < 1.96. Sales Team Readiness does not significantly alter the effect of OR on Employee Acceptance. There is no confirmed moderation. The final moderating effect result indicate that ST moderate's IC → EA with Path coefficient ≈ 0.294 and t-value < 1.96. Similarly, the interaction between Informal Contract and EA via ST is not statistically significant. The effect of IC on EA remains relatively stable regardless of ST.

Table 11. Moderation effects

Interaction	Path Coeff.	Significant?	Effect
DS × ST → EA	0.131	No	Weak/moderate
TM × ST → EA	1.551	Borderline	Potential slight moderation
OR × ST → EA	0.344	No	Weak
IC × ST → EA	0.294	No	Weak

7. Discussion

The dual-pathway framework provides a comprehensive perspective for analyzing the integration of AI throughout the sales cycle by examining both mediation and moderation processes. The analysis in this paper confirms that Sales Team Readiness (ST) significantly moderates the effects of DS, TM, OR, and Informal Contract (IC) on Employee Acceptance (EA) of AI. This suggests that these elements indirectly influence EA by shaping the sales team's readiness and adaptability, thereby serving as a crucial internal facilitator of digital transformation. This result aligns with the research conducted by Hradecky et al. (2022), which indicated that the successful adoption of AI is contingent upon OR, encompassing strategic alignment, leadership support, and staff competency. Their research underscores that, despite the availability of technological tools, AI projects

may face resistance or underutilization in the absence of adequate internal preparation and workforce engagement. The statistically insignificant moderation effects in this study is one of the major finding and that suggest that that ST has minimal impact on the strength of these relationships. ST's influence is more accurately characterized as a channel rather than a conditional influence. The results influence stages such as Lead Qualification, Needs Assessment, and Post-Sale Engagement. The ability of salespeople to interpret AI insights enhances decision-making and responsiveness (Chatterjee et al., 2021), thereby improving customer satisfaction and AI-driven personalization.

The present findings align with prior studies (Hradecky et al., 2022; Chatterjee et al., 2021) in confirming that organizational readiness and managerial support remain central enablers of AI assimilation. However, unlike these studies, the moderating role of Sales Team Readiness (ST) was not statistically significant, suggesting that readiness may act more as a channel than as a contingent condition. This divergence underscores a potential boundary condition within the TOE framework, where readiness exerts its influence indirectly through behavioral mediation rather than direct amplification. Such findings refine existing theory by differentiating between readiness as a structural capacity versus readiness as an interactive catalyst.

7.1. The principal findings and interpretation

The integration of AI into sales processes has ushered in a transformative era where data-driven insights, automation, and intelligent augmentation are reshaping how organizations approach each stage of the sales lifecycle. The current study adopts a dual-pathway framework that examines both mediation and moderation effects to explore the influence of Delivery Service (DS), TM, OR, and Informal Contract (IC) on Employee Acceptance (EA) of AI, mediated or moderated by Sales Team Readiness (ST). This discussion connects the model's outcomes to the seven classical stages of the sales lifecycle—Lead Generation, Lead Qualification, Needs Assessment/Discovery, Presentation/Proposal, Objection Handling, Closing the Sale, and Post-Sale Engagement—by examining how the constructs empirically influence and support each phase.

In the early phase of the sales lifecycle, Lead Generation is increasingly augmented by AI tools that automate prospect identification and engagement. However, the effective application of such tools depends not only on technological availability but also on sales team readiness and organizational preparedness. The current model shows that DS and TM exert indirect influence on EA through ST, suggesting that when sales teams are confident and prepared, they can better utilize AI-powered lead generation tools like chatbots and predictive analytics. Chatterjee et al. (2021) emphasized that organizations that invest in both technological infrastructure and employee training see more effective automation in lead generation activities. Furthermore, high outer loadings for DS indicators reflect the operational efficiency and consistency needed to support intelligent lead capture systems.

The lead qualification stage requires sales personnel to assess the suitability of leads using multiple criteria—historical data, engagement signals, and buying intent. AI systems

provide significant support here, offering real-time scoring and behavioral pattern analysis. The model's finding that ST significantly mediates the effect of OR and IC on EA suggests that AI adoption during lead qualification improves when the organization is prepared and informal understandings between team members and clients foster trust. According to Alshamlan and Ahmad (2022), informal contracts act as a relational enabler, allowing teams to rely on AI recommendations without formal rigidities that hinder fast-paced evaluations. This aligns with the study's observed total indirect effect of IC on EA (0.055), indicating that trusting relationships enhance the acceptance of AI in lead evaluation decisions.

Needs assessment is a consultative phase where the salesperson identifies specific pain points or goals of the prospect. AI tools can assist by offering insights into previous customer behavior, industry trends, and sentiment analysis. In the current study, ST has a direct and statistically significant path to EA ($\beta = 0.242$, $t = 3.558$), confirming that AI usage during discovery is more impactful when the team is mentally and technically prepared. Moreover, TM's indirect effect through ST (0.054) reveals that managerial support enables readiness through mentorship and access to digital resources. This supports the argument by Venkatesh et al. (2022) that managerial alignment with digital goals empowers frontline employees to adopt customer-intelligent systems, enhancing the discovery process.

In the presentation phase, AI tools enable dynamic proposal generation based on prospect characteristics and real-time inputs. Here, the strongest total effect from OR to EA (0.395) emphasizes that organizational systems, resource availability, and process flexibility are foundational to leveraging AI during proposal customization. The outer weights for OR indicators (ranging from 0.276 to 0.351) further reinforce that internal structures—such as data accessibility and system interoperability—facilitate confident AI engagement. When OR is high, and the sales team is equipped, AI-driven presentations are more personalized and data-backed, leading to stronger value communication and better alignment with client expectations (Chatterjee et al., 2023).

Objection handling is a critical stage requiring agility, empathy, and informed responses. AI can assist by providing objection pattern analysis and relevant rebuttals. The current findings show that IC and TM both contribute significantly to ST and EA, reinforcing that interpersonal trust (informal contracts) and leadership backing (TM) are critical for navigating challenging buyer conversations. Moderation analysis, however, shows that ST does not significantly moderate these effects—suggesting that readiness alone does not amplify or weaken these influences but acts as a necessary foundation. According to Paluch et al. (2021), AI is only effective in objection handling when embedded in a culture of psychological safety and team preparedness, which supports our finding that ST is a more effective mediator than a moderator.

Closing the sale involves decision finalization, negotiation, and contract agreement. AI applications in this stage include predictive close modeling, pricing optimization, and contract analytics. Our model shows that DS exerts the strongest mediated effect through ST (0.093), implying that structured delivery service processes—such as timely responses, accurate documentation, and logistics assurance—empower sales teams to finalize deals with confidence in AI outcomes. Moreover, the high outer loadings for DS (0.822–0.913)

reflect that operational reliability enhances trust in AI recommendations during final decision points. Supported by the findings of Sivarajah et al. (2022), well-structured service frameworks combined with sales readiness lead to higher deal closure rates when AI is involved.

In the final stage, ongoing relationship management is essential to ensure customer satisfaction, retention, and cross-selling. AI tools support this via churn prediction, usage analytics, and proactive engagement prompts. The mediation model confirms that when ST is high, sales teams are more inclined to accept AI as a strategic partner for relationship management, as shown by the significant ST → EA path. The influence of IC in this phase (indirect effect = 0.055) also reflects the importance of informal understanding between clients and teams, supporting flexible and personalized post-sale service. Ransbotham et al. (2023) argued that AI-enabled post-sale engagement is most effective when grounded in human-centered design and trust—both of which are captured through the IC construct in this study.

The integration of the dual-pathway results into the seven sales lifecycle stages reveals that Sales Team Readiness (ST) acts primarily as a mediator that facilitates the impact of organizational and operational factors on AI acceptance, rather than as a conditional moderator. This finding reinforces the importance of internal preparedness and cross-functional alignment in successful AI adoption. Each stage of the sales process benefits differently from these dynamics, with OR dominating the proposal phase, DS in deal closure, and IC in relationship management. The insignificant moderation effects suggest that while readiness enables adoption, it does not vary the strength of influence from antecedents across different levels of readiness. Overall, organizations should invest in developing structured systems and sales team capabilities to maximize AI's transformative value across the sales lifecycle.

Beyond descriptive patterns, the findings reveal a theoretical alignment between organizational change theory and socio-technical perspectives, suggesting that successful AI integration follows a dual adaptation pathway technological structuring and human resistance recalibration. This supports the argument that organizational readiness operates as both an infrastructural and behavioral enabler, extending the explanatory depth of the TOE framework through integrated human–technology alignment mechanism.

The non-significant moderation results indicate that once a baseline level of readiness and competence is achieved, further variations no longer change adoption outcomes—a pattern consistent with saturation or threshold models of organizational behavior. This outcome contributes to the refinement of the TOE framework by defining its conditional limits and suggesting that readiness should be conceptualized as a precondition for, rather than a moderator of, AI integration. Similar observations were made by Uren and Edwards (2023) and Rodriguez and Peterson (2024), who noted that cultural alignment and strategic intent, rather than incremental readiness, drive sustained digital transformation.

7.2. Implication of the study

This paper makes several important contributions to theory, especially in relation to the acceptance of technology and AI integration in sales environments: The study expands the theories by including Sales Team Readiness (ST) as both a mediator and moderator, so often excluding organizational and behavioral concepts. The results show that ST is a major mediator but not a moderator, so underlining the need of internal team alignment as a mechanism (not a condition) for AI adoption. This difference clarifies for us "how" organizational elements influence technology acceptance. Simally, it supports the dual-pathway approach whereby dual-pathway framework (mediation and moderation) as a comprehensive lens to investigate AI adoption, in line with recent scholarly recommendations to transcend direct path models (e.g., Venkatesh et al., 2022). This fills in a void in the literature by stressing the intervening part human elements—especially sales team dynamics—play in digital transformation. Through a mapping of the constructions to the seven phases of the sales lifecycle, the study adds to the scant body of research operationalizing TAM constructs in useful B2B sales environments. This context-specific validation improves the general applicability and resilience of theoretical models in the acceptance of sales technology.

For industry professionals, sales leaders, and digital transformation strategists, this research also provides practical insights: Support Sales Team Readiness Programs: Organizations should give training, skill-building, and cultural change projects top priority since ST was found to significantly mediate the impact of DS, TM, OR, and IC on Employee Acceptance (EA). These will help sales teams to operate alongside AI systems. Being ready calls for confidence, trust, and adaptability rather than only technical ability. While TM has a strong indirect impact on EA, its influence is only realized through ST; TM must be matched with enablement. Executive support must thus be turned into useful enablement—that is, coaching, mentoring, and role modeling—rather than only strategic endorsement. OR for personalized selling should be developed by companies building agile infrastructure, interoperable platforms, and responsive data systems since OR has the strongest total effect on EA (0.395). These systems let AI operate in real-time across phases of the sales cycle including needs analysis and proposal development. Leverage informal relationships for AI buy-in: The good indirect impact of IC on EA emphasizes the need of relational trust and informal contracts inside teams and between salespeople and customers. Managers should encourage a cooperative culture based on flexible, trust-based interactions that compliments official systems. Steer clear of over-reliance on moderation-based design since the non-significance of moderation effects implies that readiness by itself does not magnify effects; rather, it is a basic condition. Companies should thus concentrate on enabling systems rather than only spotting fluctuations. Match each AI tool—for example, lead scoring systems for qualification, proposal generators for presentations, and predictive analytics for post-sale engagement—with the pertinent sales process phase. This alignment raises ROI and acceptance.

Epistemologically, this study follows a post-positivist paradigm emphasizing theory extension rather than radical theory generation. While confirmatory in design, it advances the explanatory scope of the TOE framework through a dual-pathway causal logic integrating mediation and moderation effects. By incorporating ‘Sales Team Resistance to Change’ as a measurable construct, the study reveals previously underexplored behavioral mechanisms underlying AI adoption. Furthermore, the contextualization of TOE

constructs within the seven stages of the sales lifecycle transforms a traditionally static model into a dynamic process-oriented framework. This approach reflects cumulative theory-building through incremental epistemological refinement extending TOE's predictive and interpretive capacity within AI-driven organizational environments.

Although the present study offers clear managerial and practical implications, it also contributes theoretically by advancing the TOE framework. Through the integration of Sales Team Resistance to Change as a behavioral mediator and moderator, TOE is extended into a human-centered paradigm that reflects psychological and socio-technical dynamics. Furthermore, the dual-pathway causal design redefines TOE from a linear validation model to a conditional-process framework, enabling deeper causal interpretation of AI adoption outcomes. By embedding the constructs within the seven stages of the sales lifecycle, this research adds processual granularity, thus enriching both the theoretical depth and epistemological sophistication of the TOE framework."

7.3. Limitation of the study

Though this study has several constraints that should be noted despite the insightful analysis provided: Cross-Sectional Design Using a cross-sectional research model, the study gathers data at one point in time. It thus ignores changes in employee acceptance or readiness over time, particularly in reaction to changing AI tools or organizational strategies. Deeper understanding of the dynamic adoption behavior could come from longitudinal study. self-reported information

Self-reported responses form the basis of the gathered data, thus common method bias or social desirability bias may find expression. Because of perceived expectations, participants may exaggerate their preparedness or acceptance of AI, so affecting the accuracy of the constructed measurements. Constraints of Generalizability Focusing on salespeople in specific organizational settings, the study is context-specific. The results might thus not be entirely generalizable to other industries (e.g., healthcare, education) or functional roles outside of sales (e.g., logistics, marketing).

Restricted domain of moderators

The dual-pathway model tested both mediation and moderation, but the study looked just at one moderator—Sales Team Readiness. Not included were other possible moderators such AI complexity, industry type, or employee digital literacy, so perhaps restricting the range of interaction effects.

Insufficient Objectives Performance Measurement

Using perceptual indicators, the paper evaluates constructs including TM, delivery service, and AI integration. It does not include objective performance data—that is, conversion rates, lead response times, or actual system use logs—that might strengthen the results. Confidence Intervals for HTMT Not Evaluated Bootstrapping Although discriminant validity was mainly validated, HTMT inference with confidence intervals was not used to confirm it even more. This might allow conceptual overlap or residual multicollinearity room.

This study relied primarily on self-reported perceptual data, which, while appropriate for capturing subjective readiness and behavioral constructs, may introduce response bias. To address this, procedural remedies such as Harman's single-factor test,

pilot validation, and construct separation were employed. Future studies should triangulate these results using multi-source data, such as AI system usage logs, CRM analytics, or semi-structured interviews, to enhance validity and cross-verify behavioral constructs.

The use of convenience sampling introduces potential self-selection bias, limiting the generalizability of findings. Future studies should employ probability-based sampling or cross-country replication to enhance representativeness. Despite initial HTMT overlap between Organizational Readiness and Effective AI Integration, subsequent revalidation confirmed acceptable discriminant validity, suggesting theoretical complementarity rather than redundancy.

8. Conclusion

Though this study has several constraints that should be noted despite the insightful analysis provided: Cross-Sectional Design Using a cross-sectional research model, the study gathers data at one point in time. It thus ignores changes in employee acceptance or readiness over time, particularly in reaction to changing AI tools or organizational strategies. Deeper understanding of the dynamic adoption behavior could come from longitudinal study. self-reported information Self-reported responses form the basis of the gathered data, thus common method bias or social desirability bias may find expression. Because of perceived expectations, participants may exaggerate their preparedness or acceptance of AI, so affecting the accuracy of the constructed measurements. Constraints of Generalizability Focusing on salespeople in specific organizational settings, the study is context-specific. The results might thus not be entirely generalizable to other industries (e.g., healthcare, education) or functional roles outside of sales (e.g., logistics, marketing). Restricted, Domain of Moderators The dual-pathway model tested both mediation and moderation, but the study looked just at one moderator—Sales Team Readiness. Not included were other possible moderators such AI complexity, industry type, or employee digital literacy, so perhaps restricting the range of interaction effects. Insufficient Objectives Performance Measurement Using perceptual indicators, the paper evaluates constructs including TM, delivery service, and AI integration. It does not include objective performance data—that is, conversion rates, lead response times, or actual system use logs—that might strengthen the results. Confidence Intervals for HTMT Not Evaluated Bootstrapping Although discriminant validity was mainly validated, HTMT inference with confidence intervals was not used to confirm it even more. This might allow conceptual overlap or residual multicollinearity room.

References

Akhunova, S., Tuychiyeva, O., Tukhtasinova, D., Akhunova, M. (2024). *An innovative navigating system design along with AI tech integration for modern system implementation for business* (pp. 462–467). https://doi.org/10.1109/icacite60783.2024.10616591

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., Albahri, O.S., Alamoodi, A.H., Bai, J., Salhi, A., Santamaría, J. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, **96**, 156-191. https://doi.org/10.1016/j.inffus.2023.03.008

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., ..., Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, **96,** 156–191. https://doi.org/10.1016/j.inffus.2023.03.008

Aldraiweesh, A. A., Alturki, U. (2025). The Influence of Social Support Theory on AI Acceptance: Examining Educational Support and Perceived Usefulness using SEM analysis. *IEEE Access*. 18366 - 18385 10.1109/ACCESS.2025.3534099

Alkhaldi, A. N., Shea, T. (2024). Artificial Intelligence Adoption in the Public Sector: A Systematic Literature Review and Future Research Agenda using the TOE Framework. *Government Information Quarterly*, 101965.

Al-Nashash, H., Wei, J., Yang, K., Alzaatreh, A., Adeli, M., Tong, T., All, A. (2025). Computation of statistical power and sample size for in vivo research models. *arXiv preprint arXiv:2505.19666*. https://doi.org/10.48550/arXiv.2505.19666

Alshamaila, Y., Papagiannidis, S., Li, F. (2013). Cloud computing adoption by SMEs in the north east of England: A multi-perspective framework. *Journal of Enterprise Information Management*, **26**(3), 250–275. https://doi.org/10.1108/17410391311325225

Alsheibani, S., Messom, C., Cheung, Y., Mangalaraj, G. (2023). Is It Enough to Be AI-Ready? A Conceptual Framework and Propositions for AI Adoption. *Journal of Information Technology Case and Application Research*, **25**(2), 91-110.

Boppana, V. R. (2023). AI Integration in CRM Systems for Personalized Customer Experiences. Available at SSRN 4987149. http://dx.doi.org/10.2139/ssrn.4987149

Brynjolfsson, E., Mcafee, A. (2017). The business of artificial intelligence. *Harvard Business Review*, **7**(1), 1-2. https://shorturl.at/mIbwe

Bughin, J., Hazan, E., Sree Ramaswamy, P., DC, W., Chu, M. (2017). *Artificial intelligence the next digital frontier*. McKinsey Global Institute. https://shorturl.at/CupfG

Chatterjee, S., Chaudhuri, R., Vrontis, D., Thrassou, A., Ghosh, S. K. (2021). Adoption of artificial intelligence-integrated CRM systems in agile organizations in India. *Technological Forecasting and Social Change*, **168**, 120783.https://doi.org/10.1016/j.techfore.2021.120783

Cheng, Y., Liu, Y., Wang, Z. (2024). Driving forces of AI adoption in manufacturing: A fuzzy-set qualitative comparative analysis (fsQCA) based on the TOE framework. *Technology in Society*, **76,** 102452.

Damanpour, F., Schneider, M. (2006). Phases of the adoption of innovation in organizations: effects of environment, organization and top managers 1. *British journal of Management*, **17**(3), 215-236.

Davenport, T. H., Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review*, 96(1), 108-116. https://shorturl.at/PWFrK

Faul, F., Erdfelder, E., Buchner, A., Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, **41**(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fischer, H. T., Seidenstricker, S., Berger, T., Holopainen, T. (2022). *Artificial intelligence in B2B sales: Impact on the sales process*. https://doi.org/10.54941/ahfe1001456

Fountaine, T., McCarthy, B., Saleh, T. (2019). Building the AI-powered organization. *Harvard Business Review*, **97**(4), 62-73. https://shorturl.at/uYgc5

Gangwar, H., Date, H., Ramaswamy, R. (2015). Understanding determinants of cloud computing adoption using an integrated TAM-TOE model. *Journal of Enterprise Information Management*, **28**(1), 107–130. https://doi.org/10.1108/JEIM-08-2013-0065

Hair, J. F., Astrachan, C. B., Moisescu, O. I., Radomir, L., Sarstedt, M., Vaithilingam, S., Ringle, C. M. (2021). Executing and interpreting applications of PLS-SEM: Updates for family business researchers. *Journal of Family Business Strategy*, **12**(3), 100392. https://doi.org/10.1016/j.jfbs.2020.100392

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.https://www.scirp.org/reference/referencespapers?referenceid=2975006

Hall, K. R., Harrison, D. E., Ajjan, H., Marshall, G. W. (2022). Understanding salesperson intention to use AI feedback and its influence on business-to-business sales outcomes. *Journal of Business & Industrial Marketing*, **37**(9), 1787–1801. https://doi.org/10.1108/JBIM-04-2021-0218

Henseler, J., Ringle, C. M., Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the academy of marketing science*, **43**, 115-135. https://doi.org/10.1007/s11747-014-0403-8

Hooi, L. W., Chan, A. J. (2023). Does workplace digitalization matter in linking transformational leadership and innovative culture to employee engagement? *Journal of Organizational Change Management*, **36(**2), 197-216. https://doi.org/10.1108/JOCM-06-2022-0184

Hossain, M. E., Biswas, S. (2024). Technology acceptance model for understanding consumer's behavioral intention to use artificial intelligence-based online shopping platforms in Bangladesh. *SN Business & Economics*, **4**(12), Article 153. https://doi.org/10.1007/s10791-025-09575-5

Hradecky, D., Kennell, J., Cai, W., Davidson, R. (2022). Organizational readiness to adopt artificial intelligence in the exhibition sector in Western Europe. *International journal of information management*, **65,** 102497. https://doi.org/10.1016/j.ijinfomgt.2022.102497

Hrynko, Y. (2024). Implementation of artificial intelligence technologies in the work of the sales department. *Ekonomìčnij Vìsnik Donbasu*, **3**(77), 93–101. https://doi.org/10.12958/1817-3772-2024-3(77)-93-101

Ibrahim, M. Y. (2025). The most appropriate scale for behavioral research is a seven-point rating. *Multidisciplinary Reviews*, **8**(4), 2025126-2025126. https://doi.org/10.31893/multirev.2025126

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business horizons*, **61**(4), 577-586. https://doi.org/10.1016/j.bushor.2018.03.007

Kaur, H. (2024). Artificial intelligence and commerce: Revolutionizing marketing, sales, and customer experience. *Innovative Research Thoughts*, **10**(2), 51–56. https://doi.org/10.36676/irt.v10.i2.07

Keding, C. (2021). Understanding the adoption of artificial intelligence in small and medium-sized enterprises. *Electronic Markets*, **31**, 803–822.

Keding, C., Meissner, P. (2023). Managerial and intellectual capital in AI adoption: the role of the TOE framework. *Review of Managerial Science*.

Koldyshev, M. V. (2020). *Future marketing in B2B segment: Integrating artificial intelligence into sales management*. https://doi.org/10.31435/RSGLOBAL_IJITE/30092020/7149

Li, J., Li, M., Wang, G. (2023). An empirical study on the factors influencing the adoption of AI in the finance industry: From the TOE perspective. *Finance Research Letters*, **55**, 103991.

Liu, Y., Su, Y. Y., Alhur, A. A., Naeem, S. B. (2025). Factors influencing artificial intelligence (AI) literacy in the age of generative AI chatbots for health information seeking. *Information Development*, 02666669251343030. https://doi.org/10.1177/02666669251343030

Magrini, A. (2025). The impact of artificial intelligence on sales strategies: Transforming the sales landscape. *Social Science Research Network*. https://doi.org/10.2139/ssrn.5005005

Mao, H., Zhang, T., Tang, Q. (2021). Research framework for determining how artificial intelligence enables information technology service management for business model resilience. *Sustainability*, **13**(20), 11496. https://doi.org/10.3390/su132011496

Mikalef, P., Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & management*, **58**(3), 103434. https://doi.org/10.1016/j.im.2021.103434

Mikalef, P., Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement alignment, and performance impact. *Journal of Business Research*, 136, 327-340.

Moncrief, W. C., Marshall, G. W. (2005). The evolution of the seven steps of selling. *Industrial Marketing Management*, **34**(1), 13-22. https://doi.org/10.1016/j.indmarman.2004.06.001

Nambisan, S., Wright, M., Feldman, M. (2019). The digital transformation of innovation and entrepreneurship: Progress, challenges and key themes. *Research policy*, **48**(8), 103773. https://doi.org/10.1016/j.respol.2019.03.018

Olan, F., Arakpogun, E. O., Suklan, J., Nakpodia, F., Damij, N., Jayawickrama, U. (2022). Artificial intelligence and knowledge sharing: Contributing factors to organizational performance. *Journal of Business Research*, 145, 605-615. https://doi.org/10.1016/j.jbusres.2022.03.008

Oliveira, T., Martins, M. F. (2011). Literature review of information technology adoption models at firm level. *Electronic journal of information systems evaluation*, **14**(1), pp110-121.https://academic-publishing.org/index.php/ejise/article/view/389

Oreg, S. (2003). Resistance to change: Developing an individual differences measure. *Journal of Applied Psychology*, **88**(4), 680–693. https://doi.org/10.1037/0021-9010.88.4.680

Petrescu, M., Krishen, A. S. (2023). Hybrid intelligence: Human–AI collaboration in marketing analytics. *Journal of Marketing Analytics*, **11**(3), 263-274. https://doi.org/10.1057/s41270-023-00245-3

Rane, N., Paramesha, M., Choudhary, S., Rane, J. (2024). *Artificial intelligence in sales and marketing: Enhancing customer satisfaction, experience and loyalty*. Social Science Research Network. https://doi.org/10.2139/ssrn.4831903

Reddy, P., Muthyala, S. (2025,). *Enhancing salesforce functionality with AI and machine learng: A new era of automation*. *Deleted Journal*, (51), 1–13. https://doi.org/10.55529/ijrise.51.1.13

Rodriguez, M., Peterson, R. (2024). Artificial intelligence in business-to-business (B2B) sales process: A conceptual framework. *Journal of Marketing Analytics*, *12*(4), 778–789. https://doi.org/10.1057/s41270-023-00287-7

Russell, S. J., Norvig, P. (2022). *Artificial Intelligence: A Modern Approach*, 4th ed., Pearson, 2022. https://aima.cs.berkeley.edu/global-index.html.

Sarstedt, M., Ringle, C. M., Smith, D., Reams, R., Hair Jr, J. F. (2014). Partial least squares structural equation modeling (PLS-SEM): A useful tool for family business researchers. *Journal of family business strategy*, *5*(1), 105-115. https://doi.org/10.1016/j.jfbs.2014.01.002

Segarra-Blasco, A., Tomàs-Porres, J., Teruel, M. (2025). AI, robots and innovation in European SMEs. *Small Business Economics*, 1-27. https://doi.org/10.1007/s11187-025-01017-2

Shaik, S. A., Batta, A., Parayitam, S. (2023). Knowledge management and resistance to change as moderators in the relationship between change management and job satisfaction. *Journal of Organizational Change Management*, *36*(6), 1050-1076. https://doi.org/10.1108/JOCM-04-2023-0103

Sharma, K. K., Tomar, M., Tadimarri, A. (2023, September). *AI-driven marketing: Transforming sales processes for success in the digital age*. https://doi.org/10.60087/jklst.vol2.n2.p260

Tomatzky, L. G., Fleischer, M. (1990). The processes of technological innovation Lexington Books. *Lexington MA*, *1990*.

Tsirigotis, F. (2024). Artificial intelligence and product lifecycle management systems. In J. Stark (Ed.), Product lifecycle management (Volume 6): Increasing the value of PLM with innovative new technologies (pp. 13–26). Cham: Springer Nature. https://doi.org/10.1007/978-3-031-53521-5_2

Tu, Y., Wu, W. (2021). How does artificial intelligence enable and accelerate supply chain innovation? A systematic literature review. Industrial Management & Data Systems, 121(5), 1016-1039.

Uren, V., Edwards, J. S. (2023). Technology readiness and the organizational journey towards AI adoption: An empirical study. *International Journal of Information Management*, *68*, 102588. https://doi.org/10.1016/j.ijinfomgt.2022.102588

Venkatesh, V., Thong, J. Y., Xu, X. (2016). Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the association for Information Systems*, *17*(5), 328-376. https://ssrn.com/abstract=2800121

Verma, S., Chaurasia, S. S. (2023). Understanding the determinants of artificial intelligence adoption in SMEs: An extension of the TOE framework. Journal of Science and Technology Policy Management.

Visković, L., Đerić, E., Luić, L. (2024). Assessment of the Digital Literacy Influence on the Adoption of AI-Based Tools. *Media, culture and public relations*, *15*(1), 1-13. https://hrcak.srce.hr/328666

Wamba-Taguimdje, S. L., Fosso Wamba, S., Kala Kamdjoug, J. R., Tchatchouang Wanko, C. E. (2020). Influence of artificial intelligence (AI) on firm performance: The business value of AI-based transformation projects. Business Process Management Journal, 26(7), 1893-1924.

Young, E., Wajcman, J., Sprejer, L. (2023). Mind the gender gap: Inequalities in the emergent professions of artificial intelligence (AI) and data science. *New Technology, Work and Employment*, *38*(3), 391–414. https://doi.org/10.1111/ntwe.12278

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, Y., Clark, J., Perrault, R. (2021). The AI index 2021 annual report. arXiv preprint arXiv:2103.06312. https://doi.org/10.48550/arXiv.2103.06312

Received September 12, 2025, revised November 11, 2025, accepted December 11, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, 994-1017
https://doi.org/10.22364/bjmc.2025.13.4.12

Method for Determining a Gradient Boosting Model with Optimal Hyperparameters for Classifying Processes in the Volatile Memory of an Organization's Information System Assets

Mariia HANCHENKO, Serhii GAKHOV

State University of Information and Communication Technologies, Kyiv, Ukraine

hanchenkomariia@gmail.com, gakhov@ukr.net

ORCID 0009-0008-0998-0443, ORCID 0000-0001-9011-8210

Abstract. The objective of the research is to determine the effectiveness of the gradient boosting model for classification processes in the volatile memory of the organization's information system assets. A method for determining the optimal hyperparameters for the AdaBoost, HistGB, XGBoost, and LightGBM models using full search (GridSearchCV) and five-fold cross-validation (StratifiedKFold) is proposed. Before training the models, the CIC-MalMem-2022 dataset underwent preprocessing steps. These steps included selecting attributes, normalizing numerical values, encoding categorical variables, and dividing the dataset into a training and a test set. The performance of models with basic and optimized hyperparameters was evaluated by classification metrics (i.e., accuracy, precision, recall, F1-score) and CPU time. The experimental findings revealed no statistically significant differences in model accuracy. However, a variation in computational performance was observed. The optimized HistGB model achieved the highest classification accuracy (99.9943%) at the lowest time cost (CPU time = 3.38s), demonstrating the best results for classifying processes in volatile memory. The findings substantiate the efficacy of gradient boosting methods in systems designed to identify malicious processes within the volatile memory of the organization's information system assets.

Keywords: Gradient Boosting, HistGradientBoosting, hyperparameter optimization, CIC-MalMem-2022, classification, CPU time, volatile memory.

1. Introduction

The modern cyber threat landscape is undergoing rapid evolution, as evidenced by the growing sophistication of attacks and their ability to bypass traditional defense methods. Fileless malware, which leaves no traces in the file system and operates exclusively in RAM, is a hazardous form of malicious software. These threats can mimic the legitimate activity of system processes, making it difficult to identify with signature-based or heuristic detection methods (Hanchenko and Gakhov, 2024). In this regard, there is an urgent need to develop intelligent detection systems capable of analyzing behavioral

features of memory and classifying its processes without relying on known attack patterns.

One approach to improving the efficiency of such systems is to use machine learning methods, particularly those based on behavioral analysis of RAM data. These methods facilitate the identification of anomalies associated with the execution of malicious code, even when it is strategically masked. In this context, selecting an appropriate classification method and calibrating its hyperparameters are particularly important. These measures ensure that the model exhibits the capacity for generalization, adaptability to new scenarios, and effective exploitation of computational resources.

The most effective classification methods currently in use (Dener et al., 2022; Louk et al., 2022; Naeem et al., 2023; Aboanber et al., 2024; Ke et al., 2017) include gradient boosting methods, such as AdaBoost, HistGradientBoosting, XGBoost, and LightGBM (Microsoft, 2025, February 15; scikit-learn developers, 2025, January 10). They demonstrated high accuracy in analyzing large amounts of data and detecting complex patterns. At the same time, the performance of these models is heavily reliant on the correct configuration of hyperparameters, which requires the use of optimization methods, such as full search (GridSearchCV) (scikit-learn developers, 2025, January 10), combined with cross-validation (StratifiedKFold) (scikit-learn developers, 2025, January 10). However, comparative studies of gradient-boosting models for detecting fileless memory-based attacks remain limited.

Given the above, the purpose of this study is to compare gradient boosting models and determine the optimal method for classifying processes in the volatile memory of the organization's information system assets. In this work, we used the CIC-MalMem-2022 dataset (Canadian Institute for Cybersecurity, 2022), which contains characteristics of both safe and harmful processes in volatile memory.

The hyperparameters were optimized using GridSearchCV and five-fold cross-validation, and the models were evaluated in terms of classification accuracy (accuracy, precision, recall, F1-score) and computing costs (CPU time). The results of the study allow us to reasonably select the most effective model for detecting fileless malware in the volatile memory of information system assets.

2. Analysis of scientific references

Fileless malware holds a special place among modern cyber threats because of its ability to bypass traditional detection tools based on signatures, static analysis, or file analysis (Hanchenko and Gakhov, 2024). Unlike classical forms of malware, it leaves no traces in the file system, functions exclusively in volatile memory, and disguises itself as a legitimate system process, as demonstrated by research from Aamir (2022) and ANY.RUN (2024). This makes identification difficult, especially in a corporate environment with many concurrently active processes.

Afreen et al. (2020) and Sanjay et al. (2018) note that fileless attacks are characterized by a high ability to evade detection, which makes them invisible to traditional systems. ReliaQuest's (2023) research indicates that in 2023, 86.2% of critical incidents were related to fileless malware, with Living-off-the-Land techniques used in 26.26% of cases. Nevertheless, the memory analysis of activity has the potential to reveal new ways to identify such threats.

Recent research emphasizes the potential of dynamic RAM analysis as a promising approach for detecting malicious processes based on their behavioral characteristics. For example, in Khalid et al.'s (2022) work, the Volatility tool was used to analyze memory dumps, which enabled the creation of several datasets (VirusShare, AnyRun, PolySwarm, HatchingTriage, JoESandbox) and subsequent classification using machine learning methods.

Of particular note is the CIC-MalMem-2022 dataset (Canadian Institute for Cybersecurity, 2022), which contains balanced samples of benign and malicious processes extracted from memory dumps. Its structure is close to real-world conditions and avoids model retraining.

Due to these advantages, CIC-MalMem-2022 is actively used in behavioral memory analysis research. Louk et al. (2022) demonstrated the use of this set along with BODMAS and open data from Kaggle, obtaining high classification accuracy (>99%) when using gradient boosting models - XGBoost, LightGBM, and CatBoost.

The advantages of ensemble models for detecting complex malicious actions are confirmed by Hasan et al. (2023) and Dener et al. (2022), who report that gradient boosting models achieved an accuracy of over 99.9%. Ramesh et al. (2024) analysis confirms the effectiveness of XGBoost, LightGBM, GBM, and CatBoost in terms of precision, recall, F1-score, and Area Under the Curve (AUC), achieving 99.89% accuracy on the CIC-MalMem-2022 dataset.

A literature review shows that gradient boosting is one of the most effective methods for detecting abnormal activity in memory. Concurrently, the effectiveness of such models depends on the accurate configuration of hyperparameters, including the number of trees, tree depth, learning rate, and level of regularization. Nevertheless, the methodology for optimizing hyperparameters is frequently not fully disclosed. For example, Dener et al. (2022) report high accuracy, but there is no description of the methodology for tuning the model's hyperparameters, making it difficult to repeat the experiment. Louk et al. (2022), on the other hand, use a random search in each range of values (Bergstra and Bengio, 2012), which is a more systematic approach. Thus, there is a need for a method for selecting hyperparameters for gradient boosting models. Using a complete hyperparameter search (GridSearchCV) in combination with multiple cross-validation folds (StratifiedKFold) increases reproducibility and enables a statistically valid performance assessment.

Despite individual studies analyzing gradient boosting models using CIC-MalMem-2022, the literature lacks a comprehensive comparison of gradient boosting method implementations (XGBoost, LightGBM, AdaBoost, HistGB) using identical hyperparameter optimization settings and accounting for computational costs (CPU time). This creates a research niche for the systematic comparison of models that accounts for both classification accuracy and computational resource limitations.

Thus, the analyzed references confirm the relevance of using gradient-boosting ensemble methods for classifying processes in volatile memory. The lack of a systematic approach to comparing models and their parametric optimization determines the scientific novelty and potential practical impact of this research.

3. Methods

The methodology for determining the gradient boosting model for the classification processes of the volatile memory necessitated the consistent resolution of several key tasks: The CIC-MalMem-2022 dataset was selection and preparation was justified (Canadian Institute for Cybersecurity, 2022), as was the selection and building of gradient boosting models - AdaBoost, XGBoost, LightGBM, and HistGB (Microsoft, 2025, February 15; scikit-learn developers, 2025, January 10). The implementation of the method for determining optimal hyperparameters for gradient boosting models was also justified, as was its comprehensive comparison using classification metrics (accuracy, precision, recall, F1-score) and CPU time.

3.1. Justification for the CIC-MalMem-2022 dataset selection

The CIC-MalMem-2022 dataset (Canadian Institute for Cybersecurity, 2022), created by the Canadian Institute for Cybersecurity at the University of New Brunswick, was generated from volatile memory dumps collected in a controlled virtual machine environment (Windows 10, 2 GB RAM) by running both benign programs and malicious samples from VirusTotal (Carrier et al., 2022).

The dataset selection was based on several critical factors. The CIC-MalMem-2022 dataset was published in 2022 (Canadian Institute for Cybersecurity, 2022), thereby ensuring its relevance to modern cyber threats and the behavioral features it captures. The public availability of the dataset (Hanchenko and Gakhov, 2024) enhances the reproducibility of the results, a fundamental criterion for scientific research. The active use of CIC-MalMem-2022 in modern publications (Dener et al., 2022; Louk et al., 2022; Khalid et al., 2022; Neo, 2023) enables a fair comparison with other methods for classifying malicious processes in volatile memory. The CSV data format allows us to effectively work with them in the Jupyter Notebook environment (Project Jupyter, 2014) using the pandas and csv Python libraries.

Class balancing - 29,298 each of malware and safe software samples (Canadian Institute for Cybersecurity, 2022) guarantees an equal distribution of data, which eliminates the additional need for class balancing and minimizes the risk of prejudice of gradient boosting models towards the dominant class (Carrier et al., 2022; Canadian Institute for Cybersecurity, 2022; Dener et al., 2022). The dataset contains attributes that represent the behavioral characteristics of 15 different malware families that focus on activity in the volatile memory (see Tables 1 and 2) (Carrier et al., 2022; Canadian Institute for Cybersecurity, 2022).

All software samples in the CIC-MalMem-2022 dataset have full vector descriptions (see Table 2), which eliminates the need for imputation methods.

Table 1. Malware class samples in the CIC-MalMem-2022 dataset

Malware category	Malware family	Number of samples of the malware class
Spyware	180Solutions	200
	Coolwebsearch	200
	Gator	200
	Transponder	241
	TIBS	141
Ransomware	Conti	200
	MAZE	195
	Pysa	171
	Ako	200
	Shade	220
Trojan horses	Zeus	195
	Emotet	196
	Refroso	200
	Scar	200
	Reconyc	157

Table 2. Software sample attributes in the CIC-MalMem-2022 dataset

№	Software sample attribute	Description
1	Category	Category
2	pslist. nproc	Total number of processes
3	pslist. nppid	Total number of parent processes
4	pslist.avg_threads	Average number of threads for the process
5	pslist.nprocs64bit	Total number of 64-bit processes
6	pslist.avg_handlers	Average number of handlers
7	dllist.ndlls	Total number of loaded libraries for every process
8	dllist.avg_dlls_per_proc	Average number of loaded libraries per process
9	handles.nhandles	Total number of opened handles
10	handles.avg_handles_per_proc	Average number of handles per process
11	handles.nport	Total number of port handles
12	handles.nfile	Total number of file handles
13	handles.nevent	Total number of event handles
14	handles.ndesktop	Total number of desktop handles
15	handles.nkey	Total number of key handles
16	handles.nthread	Total number of thread handles
17	handles.ndirectory	Total number of directory handles

18	handles.nsemaphore	Total number of semaphore handles
19	handles.ntimer	Total number of timer handles
20	handles.nsection	Total number of section handles
21	handles.nmutant	Total number of mutant handles
22	ldrmodules.not_in_load	Total number of modules missing from the load list
23	ldrmodules.not_in_init	Total number of modules missing from the init list
24	ldrmodules.not_in_mem	Total number of modules missing from the memory list
25	ldrmodules.not_in_load_avg	The average number of modules missing from the load list
26	ldrmodules.not_in_init_avg	The average amount of modules missing from the init list
27	ldrmodules.not_in_mem_avg	The average number of modules missing from the memory
28	malfind.ninjections	Total number of hidden code injections
29	malfind.commitCharge	Total number of Commit Charges
30	malfind.protection	Total number of protection
31	malfind.uniqueInjections	Total number of unique injections
32	psxview.not_in_pslist	Total number of processes not found in the pslist
33	psxview.not_in_eprocess_pool	Total number of processes not found in the psscan
34	psxview.not_in_ethread_pool	Total number of processes not found in the thrdproc
35	psxview.not_in_pspcid_list	Total number of processes not found in the pspcid
36	psxview.not_in_csrss_handles	Total number of processes not found in the csrss
37	psxview. not_in_session	Total number of processes not found in the session
38	psxview. not_in_deskthrd	Total number of processes not found in the desktrd
39	psxview.not_in_pslist_false_avg	Average false ratio of the process list
40	psxview.not_in_eprocess_pool_false_avg	Average false ratio of the process scan
41	psxview.not_in_ethread_pool_false_avg	Average false ratio of the third process
42	psxview.not_in_pspcid_list_false_avg	Average false ratio of the process id
43	psxview.not_in_csrss_handles_false_avg	Average false ratio of the csrss
44	psxview.not_in_session_false_avg	Average false ratio of the session
45	psxview.not_in_deskthrd_false_avg	Average false ratio of the deskthrd
46	modules.nmodules	Total number of modules
47	svcscan.nservices	Total number of services
48	svcscan.kernel_drivers	Total number of kernel drivers
49	svcscan.fs_drivers	Total number of file system drivers

50	svcscan.process_services	Total number of Windows 32-bit owned processes
51	svcscan.shared_process_services	Total number of Windows 32 shared processes
52	svcscan.interactive_process_services	Total number of interactive service processes
53	svcscan.nactive	Total number of actively running service processes
54	callbacks.ncallbacks	Total number of callbacks
55	callbacks.nanonymous	Total number of unknown processes
56	callbacks.ngeneric	Total number of generic processes
57	Class	Benign or Malware

3.2. CIC-MalMem-2022 dataset preparation

To reduce the computational workload and simplify the gradient boosting models, we removed attributes with constant zero values: pslist.nprocs64bit, handles.nport, and svcscan.interactive_process_services (see Table 2).

Considering that the numerical characteristics of the samples covered a wide range of values (Canadian Institute for Cybersecurity, 2022), all numerical features were normalized to the interval [0, 1] using the function normalize() from the scikit-learn library (scikit-learn developers, 2025), which helped to improve the efficiency of model training.

The target variable Class, which has a categorical type (Malware/Benign), was encoded into a numerical format (0 - Malware, 1 - Benign) using LabelEncoder() from the scikit-learn library (scikit-learn developers, 2012), which made possible its use as an initial feature in the task of classifying processes in the volatile memory of the organization's information system assets.

To perform an objective assessment of the accuracy of the models on unknown data, and to mitigate the risk of overfitting of the entire set, the function train_test_split (test_size = 0.3, random_state = 2025) was employed to divide the set into a training (70%) and a test (30%) sample. The random_state parameter was set to 2025, thus ensuring reproducibility of the results.

3.3. Justification for gradient boosting method selection

As demonstrated (Prashant, 2020, July 15), gradient boosting is an ensemble machine learning method that is highly efficient in classification tasks, including malware and network attack detection, and network traffic analysis (Dener et al., 2022; Louk et al., 2022; Naeem et al., 2023; Aboanber et al., 2024; Ke et al., 2017). The central concept of the method is to sequentially train weak models (decision trees) with a focus on correcting failures of previous iterations (Prashant, 2020, December 8; Prashant, 2020, June 30). This results in a generalized model with high classification accuracy (Rathi, 2019), which is critical in the cybersecurity field, where even a minor fault can have serious consequences.

The four popular implementations of gradient boosting were selected for the study: HistGB, LightGBM, XGBoost, and AdaBoost. Their choice was based on both technical characteristics that meet the requirements for processing the CIC-MalMem-2022 dataset and their active use in modern research. Table 3 compares the key features of each method, including performance with large datasets, handling of missing values, parallel computing capabilities, hyperparameter customization, regularization, and more.

HistGB is an optimized implementation of gradient boosting in scikit-learn that discretizes continuous features using histograms, significantly reducing the number of splits when building a tree (Brownlee, 2020; Bhimani, 2022). High speed, low memory consumption, and the ability to handle missing values without imputation make HistGB an effective tool for tasks involving large tabular datasets and the best choice for our research.

LightGBM (Prashant, 2020, July 21) is a method from Microsoft (2025, February 15) that uses GOSS (Gradient-based One-Side Sampling) and EFB (Exclusive Feature Bundling) optimizations (Prokhorenkova et al., 2017) to speed up model training. It supports handling missing values and delivers high performance and efficient resource use, which is especially important in our constrained computing environment (16 GB of RAM, AMD Ryzen 7 5700U 1.80 GHz).

XGBoost is one of the most widely used implementations of gradient boosting, supporting regularization (L1, L2) (scikit-learn developers, 2025, June 5; Trotta, 2017), parallel training, and built-in cross-validation (Chen et al., 2016). The high accuracy of the model and its flexible settings have made XGBoost a standard for solving classification problems (Louk et al., 2022).

Table 3. Gradient boosting method technical characteristics comparison

Method characteristic	**Characteristic presence in the method**			
	HistGB	**LightGB**	**XGBoost**	**AdaBoost**
Large amounts of data processing	+	+	+	–
Missing value support	+	+	+	–
Optimal memory utilization	+	+	±[1]	–
Parallel training	+	+	+	–
Hyperparameters configuration option	+	+	+	+
Regularization (L1, L2)	+	+	+	+
Cross-validation built-in support	+	+	+	+
Training speed	+	+	+	–
Classification accuracy	+	+	+	+

[1] — requires hyperparameter optimization for efficient use of memory resources. "+" indicates the presence or high efficiency of the corresponding characteristic. "-" indicates the absence or low efficiency. "±" indicates partial support or dependence of the efficiency on the setting conditions.

AdaBoost is a classic boosting method that adjusts the weights of misclassified samples at each iteration (scikit-learn developers, 2025, January 10; Baladram, 2024). Despite its simpler architecture and slower speed compared to other methods, AdaBoost has a low tendency to overfit (Baladram, 2024), making it a suitable baseline for comparison with more modern realizations.

3.4. Building gradient boosting models

All experiments were conducted in the Jupyter Notebook environment (Project Jupyter, 2014) using the Python programming language. The pandas, numpy, seaborn, time, scikit-learn, lightgbm, xgboost, and csv libraries were used for data processing, visualization, and model building. All the considered gradient boosting methods - HistGB, LightGBM, XGBoost, and AdaBoost - have implementations in these libraries (Microsoft, 2025, February 15; scikit-learn developers, 2025, January 10; XGBoost Contributors, 2021; Chen et al, 2016), which made it possible to perform calculations on a local machine, even with limited resources.

Experimental hardware environment configuration:

- Processor: AMD Ryzen 7 5700U with Radeon Graphics, 1.80 GHz
- RAM: 16.0 GB (available: 15.4 GB)
- OS type: 64-bit, x64 architecture
- Operating system: Windows 11 Home, version 23H2
- OS build: 22631.5039

Before starting the training process, the data from the CIC-MalMem-2022 dataset was divided into a training set and a test set, as outlined in Section 3.2. The models were trained on the first subset, and their accuracy was then evaluated on the second.

For each method (i.e., HistGB, LightGBM, XGBoost, and AdaBoost), the model was trained in two stages. First, the default parameters were used; second, the hyperparameters were optimized (Prashant, 2020, July 15; Microsoft, 2025, February 14) using GridSearchCV (see Section 3.5).

In the initial phase, the models were trained with default parameter values, thereby establishing baseline performance before implementing any optimization procedures. The default parameter values are listed in Table 4.

Table 4. Gradient boosting model default parameter values

Model parameters	HistGB	LightGBM	XGBoost	AdaBoost
random_state	0	0	0	0
l2_regularization / reg_lambda	0	0	1	-
learning_rate	0.1	0	0.5	1
max_depth	None	-1	6	-
max_iter / n_estimators	100	100	100	50
min_samples_leaf / min_child_samples / min_child_weight	20	20	1	-

3.5. Method for determining optimal hyperparameters

Gradient boosting methods, including HistGB, LightGBM, XGBoost, and AdaBoost, offer a broad spectrum of hyperparameters, whose configurations directly influence the effectiveness of the developed models (scikit-learn developers, 2025, January 10; Chen, 2016; Rani et al., 2023). In this research, we selected the key hyperparameters that have the most significant impact on the process of training a generalization of models were selected: the number of trees in the ensemble (n_estimators / max_iter), the learning rate (learning_rate), the maximum tree depth (max_depth), the minimum number of samples per leaf (min_samples_leaf, min_child_weight, min_child_samples), and regularization coefficients (l2_regularization / reg_lambda).

The value space used to search for optimal hyperparameter combinations is shown in Table 5. The model is formulated with consideration of the characteristics inherent to each implementation of the gradient boosting method (Microsoft, 2025, February 15; scikit-learn developers, 2025, January 10; XGBoost Contributors, 2021), as well as the specifics of the CIC-MalMem-2022 dataset (Canadian Institute for Cybersecurity, 2022).

Table 5. Values space for optimal hyperparameters search

Hyperparameter	Values space
max_iter/ n_estimators	[100, 250, 500, 700, 1000]
learning_rate	[0.01, 0.05, 0.1]
max_depth	[3, 5, 10]
min_samples_leaf/ min_child_samples/ min_child_weight	[1, 5, 10]
l2_regularization/ reg_lambda	[0, 0.01, 0.1, 1]

The automated search for optimal hyperparameter values was conducted using the GridSearchCV() method from the scikit-learn library (scikit-learn developers, 2025, January 10). Accuracy was used as an evaluation metric, thereby enabling direct measurement of the model's capacity to classify malware and benign software samples. To maintain a proportionate representation of the target classes during training, a 5-fold cross-validation was performed using the StratifiedKFold() object (scikit-learn developers, 2025, January 10).

To ensure reproducibility of results across all experiments, the value of the random_state parameter was unified (2025), and to speed up the learning process, parallel computations were enabled with n_jobs = -1 (scikit-learn developers, 2025, September 9).

The algorithm for determining the optimal hyperparameters of gradient boosting models for classifying processes in the volatile memory of the organization's information system assets is shown in Figure 1.

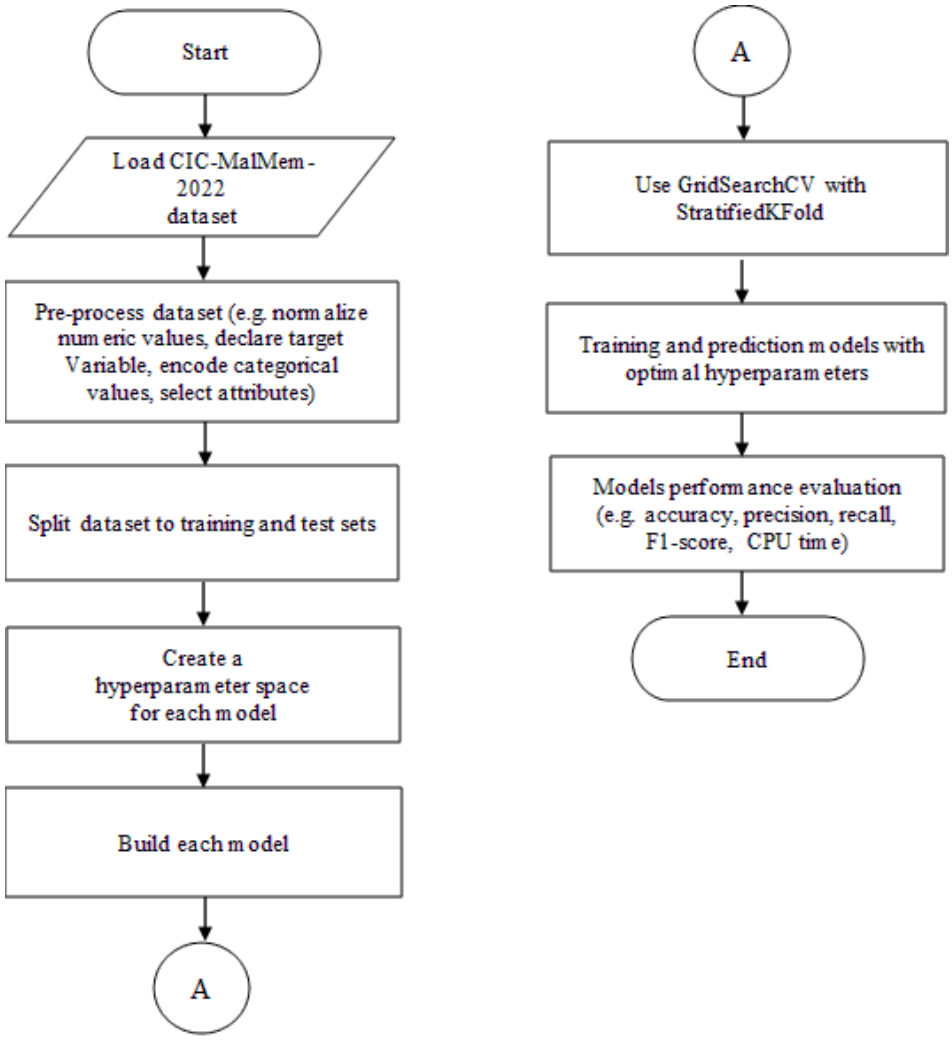


Figure 1. Flow chart of the method for determining the optimal hyperparameters of gradient boosting models for classifying processes in the volatile memory of the organization's information system assets

4. Results

The experimental research evaluated the effectiveness of four gradient boosting models - HistGB, LightGBM, XGBoost, and AdaBoost - in the task of classifying both malware and benign software samples. The models were compared based on classification metrics (accuracy, precision, recall, F1-score) and computation time (CPU time), both for models with default parameters and with optimal hyperparameters.

4.1. Performance comparison of models with default parameters

According to Table 6, the highest test accuracy of 99.9943% was achieved by the HistGB model, and the lowest by AdaBoost at 99.9716%. A similar tendency can be observed for the classification metrics - precision, recall, and F1-score.

Table 6. Performance evaluation results of gradient boosting models with default parameters

Classification metrics	HistGB	LightGBM	XGBoost	AdaBoost
Accuracy on a training set, %	99.9756	99.9756	99.9756	99.9610
Accuracy on a test set, %	99.9943	99.9886	99.9829	99.9716
Precision, %	99.9942	99.9885	99.9827	99.9713
Recall, %	99.9944	99.9888	99.9832	99.9718
F1-score, %	99.9943	99.9886	99.9829	99.9716
CPU time, s	6.28	10	7.69	32.1

Figure 2 shows a bar chart comparing the main metrics for classifying gradient boosting models.

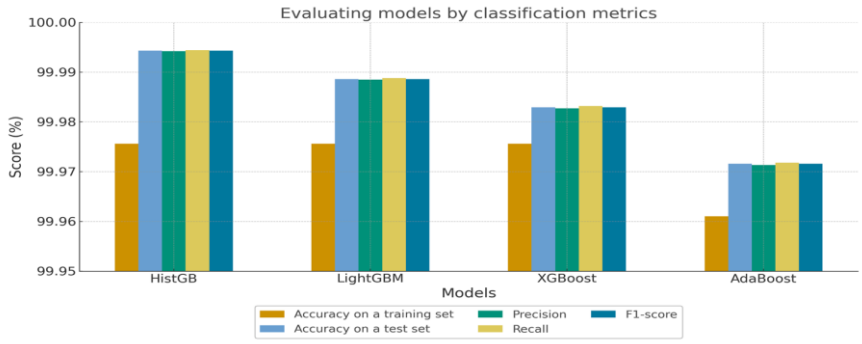


Figure 2. Bar chart comparing model performance with default parameters

Figure 3 shows a radar chart that demonstrates a comprehensive comparison of models across all key metrics, including computational efficiency (CPU time). The CPU time has been normalized and inverted for better visualization: lower CPU time corresponds to greater distance from the center.

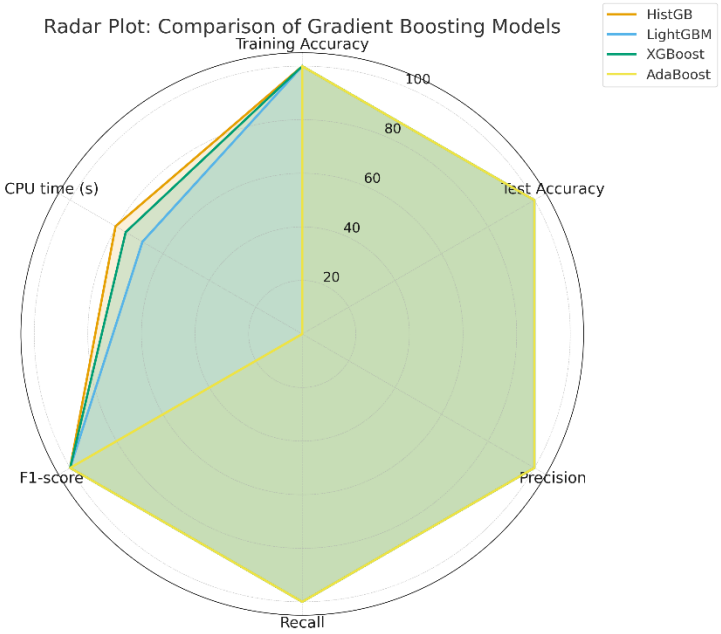


Figure 3. Radar chart comparing models with default parameters

In terms of performance, the fastest model was HistGB with a runtime of 6.28 s, followed by XGBoost (7.69 s) and LightGBM (10.0 s). The most resource-intensive model was AdaBoost, which required 32.1 s for training and prediction. Thus, all four models with default parameters achieved high-quality classification, but HistGB and XGBoost provided a better balance between accuracy and speed.

4.2. Search results for optimal hyperparameters

As illustrated in Table 7, the optimal values for the hyperparameters of the gradient boosting models were determined through five-fold cross-validation, utilizing the GridSearchCV() function (refer to Section 3.5 for further details).

Table 7 shows that the optimal value of the regularization hyperparameter (l2_regularization / reg_lambda) was 0.01 for the HistGB, LightGBM, and XGBoost models. This indicates a moderate level of regularization, balancing model complexity and overfitting.

Table 7. Gradient boosting models' optimal hyperparameters

Model hyperparameters	HistGB	LightGBM	XGBoost	AdaBoost
random_state	0	0	0	0
l2_regularization / reg_lambda	0.01	0.01	0.01	-
learning_rate	0.1	0.1	0.05	0.1
max_depth	3	3	3	-
max_iter / n_estimators	700	1000	1000	1000
min_samples_leaf / min_child_samples / min_child_weight	1	5	1	-

The learning_rate hyperparameter was set to 0.1 for HistGB, LightGBM, and AdaBoost, while 0.05 proved more effective for XGBoost. This confirms the XGBoost model's sensitivity to hyperparameter changes and its suitability for gradual learning.

The maximum tree depth (max_depth) was three across all models. This level of complexity proved to be sufficient for effective classification in the considered dataset.

The number of iterations (max_iter / n_estimators) varied from 700 for HistGB to 1000 for LightGBM, XGBoost, and AdaBoost. This indicates that the last models need more decision trees to achieve high accuracy.

Regarding the minimum leaf size (min_samples_leaf/min_child_samples/min_child_weight), HistGB and XGBoost achieved the best results at 1, while LightGBM achieved the best results at 5. This may indicate that LightGBM is more sensitive to overfitting if the tree structure is too small.

In the case of AdaBoost, some of these parameters (regularization, tree depth, minimum leaf size) are not relevant due to the specifics of the method implementation, so their values are marked as "-".

4.3. Performance comparison of models with optimal hyperparameters

As illustrated in Table 8, the results of evaluating the optimized models' performance are shown. These models were evaluated using the same classification metrics as those used for the models with default parameters (see Table 6).

According to Table 8, the HistGB, LightGBM, and XGBoost models achieved extremely high classification rates, including 99.9943% accuracy, precision, recall, and F1-score. This finding suggests that the models are capable of practical training and generalization on new data following hyperparameter optimization.

Figure 4 shows a bar chart for visual comparison of the main classification quality metrics across models, while Figure 5 illustrates the same metrics in a radar chart.

However, CPU time was found to be the critical factor in distinguishing the models in terms of practicality. As demonstrated by the visualisations, HistGB offers the most expeditious training and prediction at 3.38s, while AdaBoost requires an extensive 682 s, which is almost 200 times slower.

Therefore, when considering classification quality and execution time, the HistGB model provides the most balanced solution for identifying malevolent processes in the volatile memory of the organization's information system assets.

Table 8. Performance evaluation results of the optimized models

Classification metrics	HistGB	LightGBM	XGBoost	AdaBoost
Accuracy on a training set, %	99.9805	99.9805	99.9707	99.9829
Accuracy on a test set, %	99.9943	99.9943	99.9943	99.9829
Precision, %	99.9942	99.9942	99.9942	99.9827
Recall, %	99.9944	99.9944	99.9944	99.9832
F1-score, %	99.9943	99.9943	99.9943	99.9829
CPU time, s	3.38	54.5	62	682

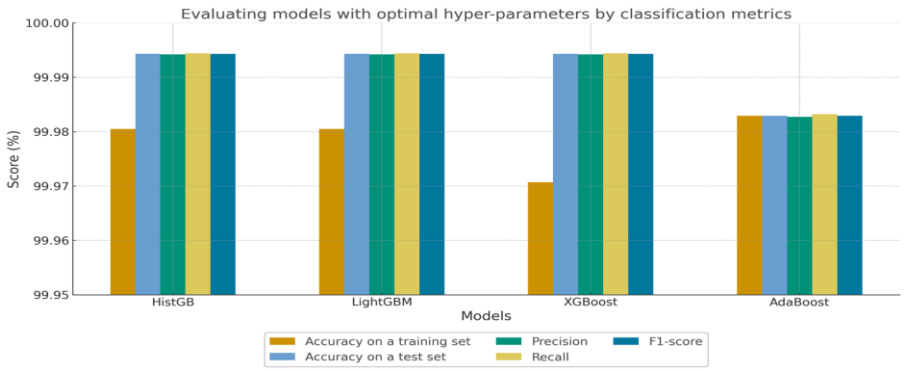


Figure 4. Bar chart of optimized models' performance comparison

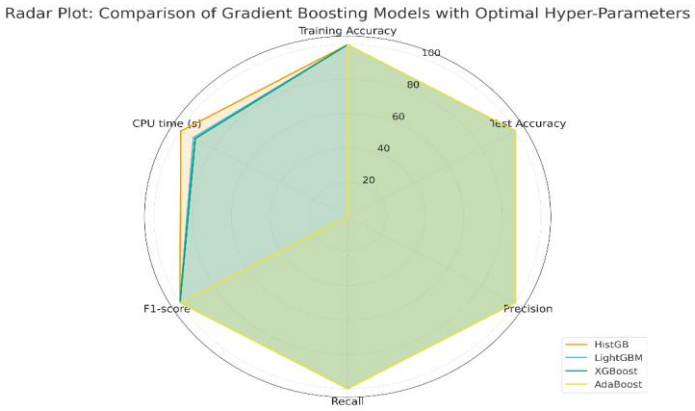


Figure 5. Radar chart of optimized models' performance comparison

4.4. Performance comparison of gradient boosting models with default parameters and optimal hyperparameters

Table 9 presents the confusion matrices for each gradient boosting model before and after hyperparameter optimization. The visual representation of these findings is delineated in Figures 6 and 7.

Table 9. Gradient boosting model confusion matrices

Gradient boosting model		True-Positives (TP)	True-Negatives (TN)	False-Positives (FP)	False-Negatives (FN)
Model with default parameters	**HistGB**	8665	8913	0	1
	LightGBM	8665	8912	0	2
	XGBoost	8665	8911	0	3
	AdaBoost	8664	8910	1	4
Model with optimal hyperparameters	**HistGB**	8665	8913	0	1
	LightGBM	8665	8913	0	1
	XGBoost	8665	8913	0	1
	AdaBoost	8665	8911	0	3

According to the results in Table 9 and Figures 6-7, all models achieved zero False-Positive (FP) classifications after optimization, indicating their ability to identify negative processes in volatile memory without false positives accurately.

The lowest number of False-Negative (FN) classifications before optimization was observed with the HistGB and LightGBM models, indicating high sensitivity. At the same time, the AdaBoost model had the worst result, with 4 FN and 1 FP.

After optimization, there was an overall improvement in detection accuracy; notably, the XGBoost model reduced the number of FN from 3 to 1, achieving results identical to those of HistGB and LightGBM. This indicates the positive impact of hyperparameter optimization on XGBoost's ability to correctly classify positive classes without incurring losses. The AdaBoost model, while improving TP and TN, remained least effective at reducing FN, limiting its usefulness in tasks where detection is critical for all malware samples.

The hyperparameter optimization provided a stable reduction in classification failures and increased the reliability of the models in detecting actual positive samples.

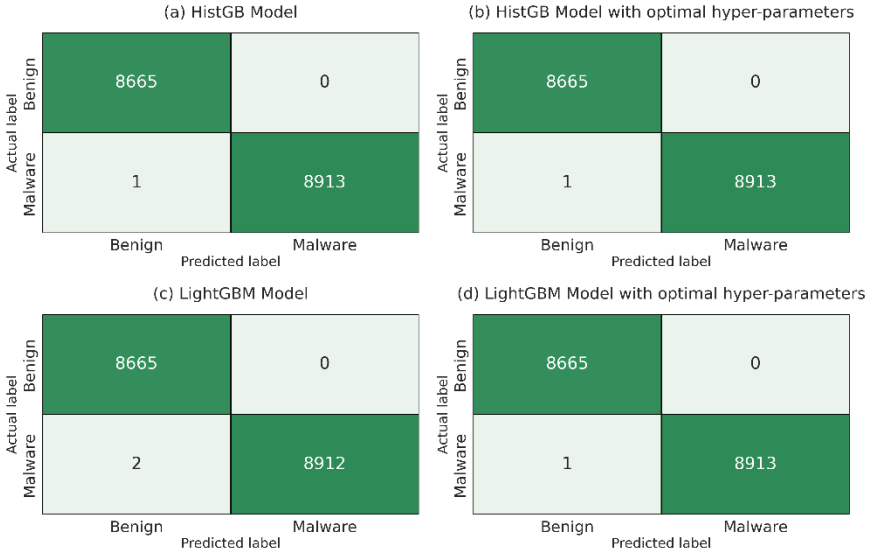


Figure 6. HistGB and LightGBM confusion matrices before and after optimization

Figure 7. XGBoost and AdaBoost confusion matrices before and after optimization

4.5. Performance comparison of the models before and after hyperparameter optimization in terms of recall and AUC

Figure 8 illustrates the Receiver Operating Characteristic (ROC) curves for the gradient boosting models before and after hyperparameter optimization. All the curves approach the upper-left corner of the diagram, indicating a high True Positive Rate (TPR) and a low False Positive Rate (FPR). This confirms the models' ability to detect malware with a minimum number of false positives effectively.

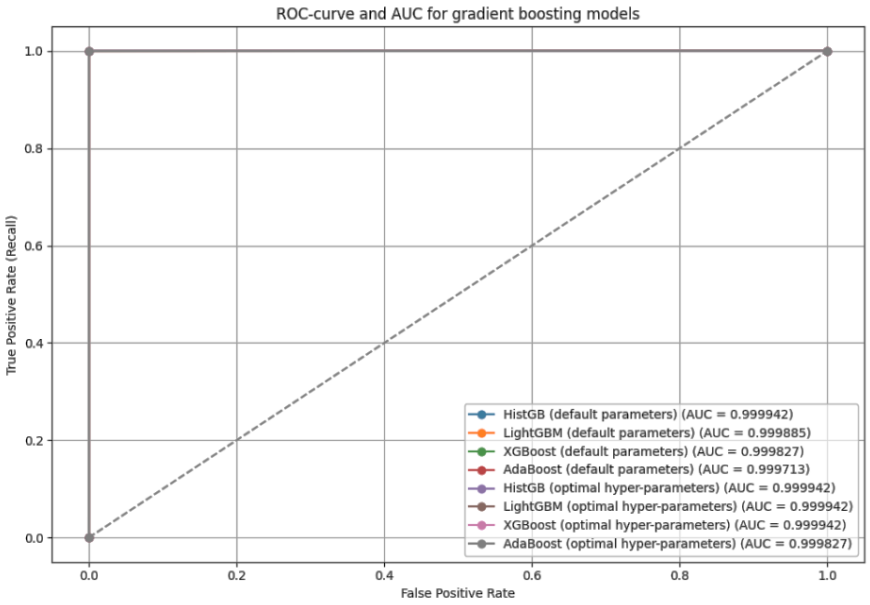


Figure 8. ROC curve and AUC values for models with default parameters and optimized hyperparameters

The respective AUC values for all models, in the default configuration and after optimization, are greater than 99.9%. This indicates the models' almost perfect ability to distinguish between malware and benign software classes. In particular, the HistGB, LightGBM, and XGBoost models achieved an AUC of 0.999942 after hyperparameter tuning, the highest among all variants.

Therefore, regardless of the hyperparameters, all models demonstrate exceptionally high sensitivity (recall) and classification performance, but optimizing hyperparameters ensures even greater consistency and reliability, particularly in critical areas where minimizing False-Negative results is important.

4.6. Model performance comparison (CPU time) before and after hyperparameter optimization

Tables 6 and 8, and Figure 9, illustrate the comparative performance of the gradient boosting models in terms of CPU time during training and prediction. The analysis covers both configurations - with default parameters and with optimal hyperparameters.

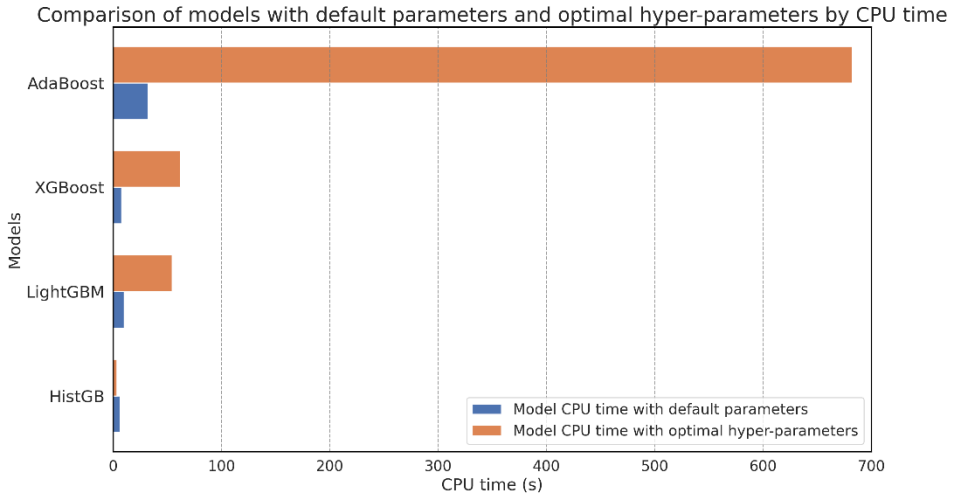


Figure 9. Bar chart of model performance comparison before and after hyperparameter optimization

According to Table 6, the HistGB, LightGBM, and XGBoost models in the default configuration showed comparatively low CPU time - 6.28 s, 10.0 s, and 7.69 s, respectively. Meanwhile, the AdaBoost model was the slowest at 32.1 s, which is significantly higher than the other methods.

After optimizing the hyperparameters (see Table 8), the models' performance changed. The HistGB model reduced CPU time to 3.38 s, indicating improved computational performance. On the other hand, for LightGBM and XGBoost, the computation time increased to 54.5s and 62s, respectively, likely due to increased model complexity after optimization. The most significant increase in time was observed in the AdaBoost model, which reached 682s, the highest among all the considered models.

Therefore, the HistGB model not only maintained high classification accuracy but also demonstrated the best computational performance after optimization, unlike other models that required more resources with a minor improvement in metrics.

5. Discussion

The results of the research have demonstrated the high efficiency of gradient boosting models for classifying processes in volatile memory using the CIC-MalMem-2022 dataset (Canadian Institute for Cybersecurity, 2022). In particular, the HistGB and LightGBM models demonstrated the best correlation between classification quality (accuracy, recall, AUC) and computational costs (CPU time), outperforming XGBoost and AdaBoost. This assertion is further demonstrated by the visualizations of the models' multidimensional performance profiles in Figures 3 and 5.

The selected CIC-MalMem-2022 dataset is one of the most modern and representative datasets for detecting fileless attacks, as it contains actual examples of malicious and benign processes in volatile memory (Dener et al., 2022; Louk et al., 2022; Canadian Institute for Cybersecurity, 2022). It maintains class balance and enables the formation of reliable behavioral classification models. At the same time, it should be noted that in real-world environments, malware behavior can change significantly through the use of new evasion techniques to avoid detection, which can limit the generalizability of results (Aqua Security, 2023; Fortinet, 2025).

During the experiment, special attention was paid to the impact of hyperparameters on model performance. Parameters such as max_depth, max_iter/n_estimators, learning_rate, min_samples_leaf/min_child_samples/min_child_weight affected both classification accuracy and computational resources (scikit-learn developers, 2025, January 10; Microsoft, 2025, February 14; Chen, 2016). The results showed that optimization of hyperparameters can both improve performance (in particular for HistGB) and significantly increase computing time (especially for AdaBoost, where CPU time increased to 682 s). At the same time, the limited range of values used in the GridSearchCV procedure may have affected the optimality of some model configurations.

CPU time for training and prediction was a key factor in applying the models as real-time, fileless attack detection systems. In this context, the HistGB model, after optimization, was the most efficient, achieving 99.9943% accuracy and 3.38 seconds of CPU time, making it suitable for practical implementation. Although the AdaBoost model retained high accuracy, it proved too resource-intensive, requiring more than 11 minutes of CPU time, making it unsuitable for operational systems without prior optimization and complexity reduction.

Notably, the CPU time measurements did not include time spent on hyperparameter selection, and the calculations were performed in a local environment, which could affect the results due to hardware limitations. In real-world applications, model performance may differ depending on the platform configuration (local, cloud, containerized, etc.).

Generally, the results obtained are consistent with current trends in the scientific literature (Dener et al., 2022; Louk et al., 2022; Ramesh et al., 2024), where gradient boosting methods are considered effective in detecting harmful processes in the volatile memory of the organization's information system assets. At the same time, differences in datasets, model implementations, hyperparameters, and execution environments may account for some of the discrepancies with other research.

The practical value of this research is to demonstrate the suitability of the HistGB and LightGBM models for integration into fileless attack protection systems, where both

accuracy and speed are critical. Their implementation will allow detecting malicious processes in the volatile memory, enabling detection of in-memory threats within a few seconds, making them suitable for near-real-time applications in operational systems.

Further research should consider the full cost cycle, including hyperparameter optimization, before deploying models in production environments. Perspective areas include performance studies in cloud (Condon, 2024; Aqua Security, 2023) or containerized environments (Hanchenko and Gakhov, 2024; Rani et. al., 2023), testing on different datasets or real attack scenarios (Brad, 2024), and comparison with other types of models, such as neural networks or hybrid models.

6. Conclusions

This research has compared four popular gradient boosting methods - HistGB, LightGBM, XGBoost, and AdaBoost - to classify processes in the volatile memory of the asset based on the CIC-MalMem-2022 dataset. All models demonstrated high classification accuracy (accuracy from 99.97% to 99.99%), confirming the suitability of the gradient boosting method for analyzing processes in the volatile memory.

The highest accuracy across all primary metrics was achieved by the optimized HistGB, LightGBM, and XGBoost models, which successfully generalized the behavioral patterns of the software samples. The HistGB model with optimized hyperparameters appeared to be the most balanced in terms of accuracy (accuracy = 99.9943%) and computational performance (CPU time = 3.38 s). This provides a perspective for implementing systems to detect malicious processes in the volatile memory of assets with limited time resources.

An important result of the research is the proposed method for determining the optimal hyperparameters of gradient boosting models via controlled parameter selection (max_depth, learning_rate, n_estimators, min_samples_leaf, etc.) using GridSearchCV in combination with StratifiedKFold cross-validation. This approach ensures the best possible consistency between classification accuracy and time costs, which is critical when models are deployed in real-world information system environments.

Therefore, the research results confirm the feasibility of using HistGB and LightGBM as practical tools for building intelligent cybersecurity systems. They can identify malware based on behavioral features in volatile memory with high accuracy and low computational cost. Further research should focus on evaluating the performance of the models in cloud and containerized environments, as well as on their adaptation to real-time, complex, fileless attack scenarios in corporate infrastructure.

References

Aamir, L. (2022). *Fileless malware: What it is and how it works*. Retrieved June 7, 2025, from https://www.fortinet.com/blog/industry-trends/fileless-malware-what-it-is-and-how-it-works

Afreen, A., Aslam, M., Ahmed, S. (2020). Analysis of fileless malware and its evasive behavior. In *Proceedings of the 2020 International Conference on Cyber Warfare and Security*

(ICCWS) (pp. 1–8). IEEE. Available at https://ieeexplore.ieee.org/abstract/document/9292376

ANY.RUN (2024). *Fileless malware: Everything you need to know*. Retrieved June 7, 2025. Available at https://any.run/cybersecurity-blog/fileless-malware/

Aqua Security (2023). *2023 Cloud Native Threat Report* [White paper]. https://info.aquasec.com/2023-cloud-native-threat-report

Baladram, S. (2024). *AdaBoost classifier explained: A visual guide with code examples*. Medium. Retrieved June 7, 2025, from https://medium.com/data-science/adaboost-classifier-explained-a-visual-guide-with-code-examples-fc0f25326d7b

Barocas, S., Hardt, M., Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bergstra, J., Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, **13**, 281–305. https://dl.acm.org/doi/abs/10.5555/2188385.2188395

Bertsekas, D. P. (1976). Multiplier methods: A survey. *Automatica*, **12**(2), 133–145.

Bhimani, D. (2022). *Histogram Boosting Gradient Classifier*. Analytics Vidhya. Retrieved June 7, 2025, from https://www.analyticsvidhya.com/blog/ 2022/01/histogram-boosting-gradient-classifier/

Brad, L. (2024). *Fileless Malware Will Beat Your EDR*. Retrieved June 7, 2025, from https://blog.morphisec.com/fileless-malware-attacks

Brownlee, J. (2020). *Histogram-Based Gradient Boosting Ensembles for Machine Learning*. Machine Learning Mastery. Retrieved June 7, 2025, from https://machinelearningmastery.com/histogram-based-gradient-boosting-ensembles/

Canadian Institute for Cybersecurity (2022). *CIC-MalMem-2022 Dataset*. University of New Brunswick. Retrieved June 7, 2025, from https://www.unb.ca/cic/datasets/malmem-2022.html

Carrier, T., Victor, P., Tekeoglu, A., Lashkari, A. (2022). Detecting obfuscated malware using memory feature engineering. In *Proceedings of the 8th International Conference on Information Systems Security and Privacy* (pp. 177–188). SciTePress. https://doi.org/10.5220/0010908200003120

Chen, T., Guestrin, C. (2016). *XGBoost documentation*. Read the Docs. Retrieved June 7, 2025, from https://xgboost.readthedocs.io/

Condon, S. (2024). *Fileless attacks surge as cybercriminals evade cloud security defenses*. CSO Online. Retrieved June 7, 2025, from https://www.csoonline. com/article/643356/fileless-attacks-surge-as-cybercriminals-evade-cloud-security-defenses.html

Dener, M., Ok, G., Orman, A. (2022). Malware Detection Using Memory Analysis Data in Big Data Environment. *Applied Sciences, 12*(17), 8604. https://doi.org/10.3390/app12178604

Fortinet (2025). *Fileless malware*. Retrieved June 7, 2025, from https://www.fortinet.com/resources/cyberglossary/fileless-malware

Hanchenko, M., Gakhov, S. (2024, April 2026). Analysis of methods for detecting fileless malware in the energy-dependent memory of an organisation's information system assets. In *V International Scientific and Practical Conference «Ricerche Scientifiche e Metodi Della Loro Realizzazione: Esperienza Mondiale e Realtà Domestiche»: Collection of Scientific Papers «ΛΟΓΟΣ»* (pp. 262–265). European Historical Studies. https://doi.org/10.36074/logos-26.04.2024.054

Hasan, S. M. R., Dhakal, A. (2023). Obfuscated malware detection: Investigating real-world scenarios through memory analysis. In *2023, the IEEE International Conference on Telecommunications and Photonics (ICTP)*. IEEE. https://ieeexplore.ieee.org/abstract/document/10490701

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, **30**, 3146–3154. Neural Information Processing Systems Foundation.

https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Khalid, O., Ullah, S., Ahmad, T., Saeed, S., Alabbad, D. A., Aslam, M., Buriro, A., Ahmad, R. (2022). An insight into the machine-learning-based fileless malware detection. *Sensors*, **23**(2), 612. https://www.mdpi.com/1424-8220/23/2/612

Louk, M. H. L., Tama, B. A. (2022). Tree-Based Classifier Ensembles for PE Malware Analysis: A Performance Revisit. *Algorithms*, **15**(9), 332. https://doi.org/10.3390/a15090332

Microsoft (2025, February 15). *LightGBM*. GitHub. Retrieved May 18, 2025, from https://github.com/Microsoft/LightGBM

Microsoft (2025, February 14). *LightGBM parameters* [GitHub documentation]. GitHub. Retrieved June 7, 2025, from https://github.com/microsoft/LightGBM/blob/master/docs/Parameters.rst

Naeem, H., Dong, S., Falana, O. J., Ullah, F. (2023). Development of a deep stacked ensemble with process-based volatile memory forensics for platform-independent malware detection and classification. *Expert Systems with Applications*, **223,** 119952. https://www.sciencedirect.com/science/article/abs/pii/S0957417423004542

Neo, G. K. (2023). *Malware detection in memory images using machine learning* [Final Year Project]. Nanyang Technological University, Singapore. https://dr.ntu.edu.sg/handle/10356/165974

Prashant, B. (2020 July 15). *A guide on XGBoost hyperparameter tuning* [Kaggle Notebook]. Kaggle. Retrieved June 7, 2025, from https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning

Prashant, B. (2020, June 30). *Bagging vs boosting* [Kaggle Notebook]. Kaggle. Retrieved June 7, 2025, from https://www.kaggle.com/code/prashant111/bagging-vs-boosting

Prashant, B. (2020, July 21). *LightGBM Classifier in Python* [Kaggle Notebook]. Kaggle. Retrieved June 7, 2025, from https://www.kaggle.com/code/prashant111/lightgbm-classifier-in-python

Prashant, B. (2020, December 8). *XGBoost K-Fold CV & feature importance* [Kaggle Notebook]. Kaggle. Retrieved June 7, 2025, from https://www.kaggle.com/code/prashant111/xgboost-k-fold-cv-feature-importance/notebook

Project Jupyter (2014). *Project Jupyter*. Retrieved June 7, 2025, from https://jupyter.org/

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A. (2017). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, **30**. Neural Information Processing Systems Foundation. https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Ramesh, S. P., Anand, S. Raj, Karthikeyan V. G. (2024). Machine Learning Approach for Malware Detection Using Malware Memory Analysis Data. *International Conference on Applications and Techniques in Information Security* (pp 135–145). Springer. https://doi.org/10.1007/978-981-97-9743-1_10

Rani, O., Amit, S., Lena, F., Erin S., Jose, I. (2023). *What are fileless attacks?* Aqua Cloud Native Academy. Retrieved June 7, 2025, from https://www.aquasec.com/cloud-native-academy/application-security/fileless-attacks/

Rathi, R. (2019). *All you need to know about the Gradient Boosting algorithm — Part 2 (Classification)*. Medium. Retrieved June 7, 2025, from https://medium.com/data-science/all-you-need-to-know-about-gradient-boosting-algorithm-part-2-classification-d3ed8f56541e

ReliaQuest (2023). *Fileless malware accounts for 86.2% of all detections*. Global Security Mag. Retrieved June 7, 2025, from https://www.globalsecuritymag.de/fileless-malware-accounts-for-86-2-of-all-detections-reliaquest.html

scikit-learn developers (2012). *sklearn.preprocessing.LabelEncoder*. In *Scikit-learn Documentation*. Retrieved June 7, 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html#sklearn.preprocessing.LabelEncoder

scikit-learn developers. (2025, January 10). *API Reference.* In *Scikit-learn Documentation.* Retrieved May 18, 2025, from https://scikit-learn.org/stable/api/sklearn.ensemble.html

scikit-learn developers. (2025, June 5). *Glossary of Common Terms and API Elements.* In *Scikit-learn Documentation*. Retrieved June 7, 2025, from https://scikit-learn.org/stable/glossary.html#term-n_jobs

scikit-learn developers. (2025). *Preprocessing — Normalization*. In *Scikit-learn Documentation*. Retrieved June 7, 2025, from https://scikit-learn.org/stable/modules/preprocessing.html#normalization

Trotta, F. (2017). *Understanding L1 and L2 regularization*. Towards Data Science. Retrieved June 7, 2025, from https://towardsdatascience.com/understanding-l1-and-l2-regularization-93918a5ac8d0/

XGBoost Contributors (2021). *XGBoost* [GitHub repository]. GitHub. Retrieved June 7, 2025, from https://github.com/dmlc/xgboost/

Received June 7, 2025, revised November 22, 2025, accepted December 19, 2025

Baltic J. Modern Computing, Vol. 13 (2025), No. 4, 1018-1037
https://doi.org/10.22364/bjmc.2025.13.4.13

Aspects of Application of Data Analytical Tools for Assessing the Academic Performance of Secondary Education Schools

Dalė DZEMYDIENĖ[1], Sigita TURSKIENĖ[1],
Vaida LIUBERTIENĖ[2]

[1]Institute of Regional Development of Šiauliai Academy of Vilnius University, Vilniaus str. 88, LT-76285 Šiauliai, Lithuania
[2]Salininkų Gymnasium, Vilnius, Lithuania

dale.dzemydiene@sa.vu.lt, sigita.turskiene@sa.vu.lt,
vaida.liubertiene@salininkugimnazija.lt

ORCID 0000-0003-1646-2720, ORCID 0000-0002-2019-6712

Abstract. The aim of this research is to investigate in the assessment process of the quality of activities of secondary education institutions by analyzing realization of digitalization of procedures that provided in the legal acts regulating the assessing the quality of activities. This study explores the application of data analytical tools in assessing the academic performance and institutional quality of Lithuanian secondary education schools. It emphasizes the transition toward data-driven self-assessment and external evaluation processes as defined by the State Education Strategy (2013–2022) and the Ministry of Education's methodological frameworks. The framework for evaluation of data analytical tools was developed and review of features of applied data analytical tools for adequate choosing them in self-assessment procedures of the institution performance was made. From an applied perspective, systems for analyzing data on the activities of educational institutions are examined, and the structure of the activities is revealed. A secondary educational institution was selected for the case study of model implementation and the data is analyzed on the basis of questionnaires for self-assessment.

Keywords: data analytical tools, assessment of the quality of activities, IS for management of secondary education schools, methodology for assessing**.**

1. Introduction

The tools for data analysis became the important component for evaluation of situations, when we have assessment procedures. The progress of educational institutions is related to the implementation of information and communication technologies (ICT) and new management methods. ICT and right evaluation procedures create a new concept of value in school management processes (Dzemydienė et al., 2023; 2024) and can be compared with regional context (Dzemydaitė and Naruševičius, 2023). However, the problems of

evaluation performance of educational institution are also encountered in this assessment process, because we have to choose appropriate tools for data analysis and to develop the assessment. The aim of creating an effective system for assessing the quality of educational institutions' activities is formulated in the Strategic Provisions for the Development of Education in Lithuania (State Education Strategy 2013-2022, 2013) and general education schools are prepared to implement it. A system for (self) assessment of the quality of activities of general education schools is being developed, which provides for monitoring education, (self) assessment of the quality of school activities and making decisions enabling the improvement of school activities. The comprehensive, formal formulation of evaluation criteria for data analytical tools is provided and theoretically grounded, explicitly defined, and justified in terms of their relevance and appropriateness for educational assessment and institutional evaluation.

As the processes of implementing the educational content and the concept of assessing of change of student achievements, the internal audit methodology has been revised several times. In 2009, the "Recommendations for the Self-Assessment of the Quality of Activities of General Education Schools" were approved (Minister, 2009a) and in 2016, the "Methodology for the Self-Assessment of the Quality of Activities of Schools Implementing General Education Programs" was approved (Minister, 2016a). In 2022 The National Education Agency has prepared new "Recommendations for the Application of Self-Assessment Questionnaires for General Education Schools (NEA, 2022).

The description of the procedures for external audit of schools was also improved, the term "external audit" was replaced by "external assessment", and its new versions were legalized in 2016, 2018 and 2021 (Minister, 2016b; Minister, 2018; Minister, 2021).

Systematic amendments to the legal acts regulating the assessment of the quality of activities of general education schools demonstrate their importance for ensuring the quality of education and the aim of responding to progressive development trends in the field of education and oblige the educational community to review the procedures and instruments for assessing the quality of activities. One of the objectives of the State Education Strategy (2013-2022) is to "*introduce a culture of educational quality based on data analysis and self-assessment"*, which instructs external evaluators of the quality of school activities and those carrying out self-assessment of the quality of school activities to pay special attention to substantiating the assessment conclusions based on reliable data, their professional analysis and interpretation. The results of assessment and self-assessment based on data allow us to see and analyze the progress of the school and its trends, to predict and plan a progress strategy.

As an alternative to the former digital platform *iqesonline.lt* the National Education Agency prepared the "Recommendations for the Application of Self-Assessment Questionnaires for General Education Schools" in 2022 (NEA, 2022). Five questionnaires for studying general education school performance indicators on the education were proposed as self-assessment instruments (portal *emokykla.lt*). However, these tools are not technologically efficient, therefore new software tools are being sought to analyze performance assessment processes.

The article discusses the model for assessing the quality of school activities, analyzes the data analytical software designed to measure the current situation of the quality of activities and its change, process the collected data, interpret and make decisions to improve the quality of activities.

This study analyzed the application of data analytical tools in assessing the quality of activities in secondary education institution, focusing on the digitalization of self-assessment and external evaluation processes. Through a comprehensive review of analytical platforms, like as Microsoft Forms, Microsoft Excel, and Microsoft Power BI, and an empirical case study conducted in a Lithuanian secondary school, the research aimed to determine how these tools enhance data-driven decision-making and institutional performance evaluation.

The findings confirm that the integration of these tools, particularly Microsoft Power BI, facilitates the systematic collection, processing, and visualization of heterogeneous educational data. Power BI provides advanced visualization, interactivity, and comparative reporting capabilities that significantly improve the interpretability and practical application of data insights. In contrast to traditional tools such as Excel and Forms, Power BI enables dynamic dashboards that support evidence-based planning and continuous quality improvement. Furthermore, the inclusion of Microsoft 365 applications ensures accessibility and cost-efficiency, enabling schools with varying resources to effectively engage in digitalized self-assessment processes.

The study's objectives—to identify appropriate analytical instruments for evaluation of their applicability for quality assessment of performance of schools, and propose a conceptual framework for their integration. The results align closely with Lithuania's *State Education Strategy 2013–2022* (2013), which emphasizes the development of a data-informed culture of educational quality. The outcomes also support the objectives outlined in the *Methodology for the Self-Assessment of the Quality of Activities of Schools Implementing General Education Programs* (Minister, 2016a), reinforcing the importance of reliable data, professional analysis, and transparent evaluation procedures in education management.

2. Review of properties of data analytical tools

Here's a comprehensive review of widely used data analytical tools, including their key features, typical use cases, and references for further reading. We would like to review such tools as descriptive statistics and pivot tables and visualization capabilities of Microsoft Excel and Power BI, R statistical programming language, key libraries created on Python and Rapid Miner.

The main features as strengths and limitations are revealed in Table 1. Some of them are created as statistical analysis platforms, i.e., Tableau, SAS, Apache Spark and Google Looker Studio. Microsoft Excel is a foundational tool for data analysis, especially useful for descriptive statistics, pivot tables, and basic data visualizations (Walkenbach, 2013). The tool R is a statistical programming language ideal for statistical modeling and data visualization with extensive package ecosystem (e.g., ggplot2, dplyr, caret), strong statistical capabilities, and is of open-source, as mentioned in (Wickham and Grolemund 2016). Python is a general-purpose language with powerful data analysis libraries as Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, TensorFlow described by (VanderPlas, 2016). They can be flexible, have possibilities to integrate well with web apps and databases, and are strong for machine learning support.

Table 1. Comparison of properties of data analytical tools

Data analytical tool	Strengths of the tool	Limitation of analysis	Use cases	Author's works
Microsoft Excel (with functions of statistics, pivot tables, and basic visualization tools)	User-friendly, widely adopted, strong community support, rich visualization capabilities	Limited scalability for large datasets, not ideal for complex statistical or machine learning models	Budget forecasting, sales trend analysis	(Walkenbach, 2013)
R (a statistical programming language)	Good statistical modeling and data visualization, extensive package ecosystem (e.g., ggplot2, dplyr, caret), strong statistical capabilities, open-source	Steeper learning curve compared to GUI tools	Hypothesis testing, time series analysis, statistical inference	(Wickham and Grolemund 2016)
Python Key Libraries: Pandas, NumPy, Matplotlib, seaborn, scikit-learn, TensorFlow	Flexible, integrates well with web apps and databases, strong machine learning support.	Slower for some statistical operations compared to R; requires more coding knowledge.	Predictive analytics, text mining, machine learning	(VanderPlas, 2016)
Tableau	Leading data visualization and business intelligence tool Interactive dashboards, user-friendly drag-and-drop interface, integration with various data sources	Expensive licensing, limited statistical modeling	Executive reporting, dashboarding, data storytelling	(Murray, 2016)
Microsoft Power BI	Business analytics tool with deep Excel integration, cost-effective, real-time dashboards, seamless integration with Microsoft ecosystem	Less flexible for non-Microsoft data sources	Corporate KPI dashboards, sales analytics	(Collie and Singh, 2020)

Data analytical tool	Strengths of the tool	Limitation of analysis	Use cases	Author's works
Apache Spark	Big data processing engine designed for speed and ease of use in large-scale data processing tool, scalable, supports multiple languages (Python, R, Scala), strong in-memory processing	Higher infrastructure requirements	Real-time analytics, big data ETL, machine learning pipelines	(Karau and Warren, 2017)
SAS (Statistical Analysis System)	Enterprise-level analytics platform for advanced statistical analysis, has high data security, robust statistical capabilities	Proprietary and expensive.	Preferred in healthcare and banking sectors, trial analysis, risk assessment	(Hall et al., 2010)
Google Looker Studio.	Free tool for creating dashboards and data reports using Google ecosystem (GoogleLooker Studio Documentation), easy integration with Google Analytics, Sheets, Ads; good for collaborative reporting	Limited data manipulation capabilities	Website traffic reporting, marketing campaign analytics	(Google Looker Studio Documentation, 2025).
RapidMiner	Data science platform for machine learning and predictive analytics has good visual workflow designer, supports code-free analytics	Freemium limitations, less flexible than coding-based tools	Customer churn prediction, fraud detection	(Hofmann and Klinkenberg, 2016)

The SQL can be mentioned as standard language for querying and managing relational databases, that have powerful for data extraction and transformation, optimized for structured data. (Fotache and Strimbei, 2015)., We can mention SQL Pocket Guide

provided by O'Reilly Media. The Tableau is a leading data visualization and business intelligence tool (Murray, 2016).

The Power BI is Microsoft's business analytics tool with deep Excel integration (Collie and Singh, 2020) cost-effective, real-time dashboards, seamless integration with Microsoft ecosystem. The Apache Spark is a big data processing engine designed for speed and ease of use in large-scale data processing tool (Karau and Warren, 2017).

The statistical analysis systems can be helpful as well. For example, the SAS (Statistical Analysis System) is enterprise-level analytics platform for advanced statistical analysis (Cody, 2010), Google Looker Studio is free tool for creating dashboards and data reports using Google ecosystem with easy integration with Google Analytics, Sheets, Ads. The tool as RapidMiner is data science platform for machine learning and predictive analytics (Hofmann and Klinkenberg, 2016).

The evaluation criteria for data analytical tools are explicitly grounded in theoretical frameworks of educational assessment, data-driven decision-making, and evaluation of information system evaluation (Dzemydiene et al., 2022). The evaluation criteria can be backgrounded on models such as:

- Technology Acceptance Model (TAM) – emphasizing usability and perceived usefulness.
- Information Systems Success Model (Jeyaraj, 2020) – focusing on system quality, information quality, and impact on decision-making.
- Data-Based Decision-Making (DBDM) Frameworks in education (Mandinach and Gummer, 2016) – stressing the value of accuracy, accessibility, and interpretation of data for pedagogical improvement.
- Quality Assurance Frameworks in education (Lithuanian Methodology for Self-Assessment of School Activities, 2016) – highlighting transparency, evidence-based evaluation, and continuous improvement.

These frameworks justify the multidimensional nature of a criteria: technical performance, analytical robustness, user accessibility, educational relevance (Table 2).

Table 2. Evaluation criteria of data visualization tools

Criterion	**Definition**	**Rationale & Theoretical Basis**
1. Analytical Capability	The extent to which the tool supports descriptive, diagnostic, predictive, and prescriptive analytics.	Grounded in data analytics maturity models (Gartner, 2020); ensures tools can move from simple reporting to insight generation for educational decision-making.
2. Data Integration and Compatibility	Ability to connect with various data sources (e.g., e-diaries, survey platforms, learning management systems) and handle structured/unstructured data.	Based on DeLone and McLean's (2003) *System Quality* dimension; critical for holistic institutional analysis.
3. Visualization and Interpretability	Quality, interactivity, and clarity of data visualization supporting stakeholder interpretation.	Supported by DBDM frameworks (Mandinach and Gummer, 2016); facilitates understanding among educators, administrators, and policymakers.

Criterion	Definition	Rationale & Theoretical Basis
4. Usability and Accessibility	Degree of user-friendliness, intuitive design, and accessibility for non-technical users (teachers, administrators).	Drawn from the Technology Acceptance Model; enhances adoption and effective use in educational environments.
5. Reliability and Accuracy	Consistency and precision of data outputs, including error minimization and reproducibility of results.	Rooted in principles of *data quality management* (ISO/IEC 25012); essential for credible educational assessment.
6. Adaptability and Scalability	Capacity to accommodate institutional growth, new indicators, or evolving assessment frameworks.	Supported by system lifecycle and sustainability theories; ensures long-term applicability in dynamic educational contexts.
7. Cost-effectiveness and Accessibility	Balance between tool functionality, licensing, and availability for educational institutions.	Justified through resource efficiency models (OECD, 2013; Okoye et al., 2025); promotes equity in digital transformation of schools.
8. Compliance and Security	Adherence to data protection, ethical, and legal standards (e.g., GDPR, national education data regulations).	Derived from ICT governance and ethical data use frameworks; guarantees institutional accountability and trust
9. Educational Relevance	Alignment of the tool's analytical functions with educational quality domains (leadership, learning outcomes, environment, and management).	Based on Lithuania's *Methodology for Self-Assessment of School Activities* (2016); ensures contextual fit and meaningful use in school evaluation.

Each criterion reflects the dual nature of data analytics in education—as both a technological and pedagogical instrument. The rationale for adopting these criteria is threefold:

- *Theoretical coherence* represents the criteria which are consistent with global and national theoretical frameworks that define effective data use and system quality in education.
- *Contextual relevance* represents the criteria which respond to the identified needs of Lithuanian schools transitioning to data-based management, as highlighted by the National Education Agency (NEA, 2022) and State Education Strategy 2013–2022.
- *Practical applicability* criteria provide measurable, actionable dimensions for evaluating, comparing, and selecting data analytical tools suited to institutional self-assessment and quality assurance.

The modern data analytics tool Microsoft Power BI combines Excel and Pivot functions. It allows easily and effectively analyze, visualize and present data, create reports, connect data sources, and create interactive graphs. In the "magic" quarter compiled by Gartner, a company that conducts research on technological innovations worldwide, Microsoft's software is indicated as a leader (Figure 1), it surpasses such leaders in performance analytics as Tableau and Qlik.

Figure 1. Comparison visualization of Gartner Business Analytics Products
Source: (Preidys, 2023; Gartner, Inc, 2020)

We are taking the preference for Microsoft Power BI as business analytical tool with deep Excel integration for performance of self-assessment procedures in secondary education schools.

3. Model for assessing the quality of activities of secondary educational institutions

The latest version of the description of the procedures for organizing and implementing the external evaluation of the activities of general education schools was approved by the Minister of Education, Science and Sports of the Republic of Lithuania in 2021 June 21, (Minister, 2021).

The description specifies the system for assessing the quality of school activities, which consists of school self-assessment and external assessment. Self-assessment of the quality of school activities is an integral part of external assessment, its results are included in the set of data analyzed during external assessment. External assessment is organized by the National Education Agency and the institution implementing the rights and obligations of the owner. A 5-level assessment scale is used to assess the quality of school activities (Table 3).

Table 3. Levels of assessment of the quality of activities of schools implementing general education programs

Quality level	Descriptive performance quality assessments	Percentage value of level
4 level	Very good: effective, exceptional, focused, unique, creative	90 percent or more of activities were positively evaluated
3 level	Good: above average, appropriate, effective, potential, flexible	60–89 percent or more of activities were positively evaluated
2 level	Satisfactory: average, not bad, unsystematic, not exceptional	31–59 percent or more of activities were positively evaluated
1 level	Poor: unsatisfactory, ineffective, inappropriate, non-specific	11–30 percent of activities were positively evaluated
N level	Very poor: unacceptable	Until 10 percent of activities were positively evaluated

One of the tasks of external evaluation is to "help make decisions based on reliable data regarding assistance to schools" provided by the description of the procedure for the organization and implementation of the external evaluation of schools implementing General Education Programs (2021). For each type of external evaluation, performance evaluation indicators have been approved, which determine what school data should be collected, analyzed, summarized and used to formulate conclusions.

One part of the data required for external evaluation is collected by evaluators by observing lessons and educational activities at school, analyzing documents prepared by the school, school data provided by the reporting in educational information systems, etc. The other part of the data must be provided to the evaluators by the school itself. Regardless of the type of external evaluation, the head of the school participating in the evaluation provides the evaluators with:

• a weekly schedule of lessons, non-formal education, class hours and other events,
• the latest information on self-evaluation and self-evaluation conclusions,
• the school's strategic plan, school activity plans for the last two years,
• summarized information on the achievements, progress and student achievement research data of the school for the last two years,
• the school's educational plan for the current academic year,
• contextual and other information about the school's activities according to the provided questionnaire.

The school data provided to the evaluators, which can be collected, processed and analyzed using digital means, thus facilitating the task of data analytics and its automation. The self-evaluation model of the quality of school activities consists of assessment areas, which are detailed in topics, which are divided into indicators. Four areas of school activity are assessed, which are related by causal relationships, i.e., leadership and management, education and student experiences, environment of education and results.

The following groups were selected as respondents:

- conduct surveys of students, their parents and teachers about the quality of school activities in other online survey systems, for example, Microsoft Office 365 Forms, Google Forms, *apklausa.lt, manoapklausa.lt,* etc.;

- use IQES online Lithuania questionnaires to collect quantitative data, which are published on the educational portal *Emokykla.lt*;

We are applying qualitative data collection methods, for example, focus group discussions, interviews, etc. The themes identified in each area define the school's quality directions and indicators that define the dimensions of school quality. The school head initiates the self-assessment of the quality of school activities, and the School Council selects the areas of self-assessment and the methodology for it.

Self-evaluation can be carried out in 3 ways:

1. Broad (or overall) self-evaluation. During it, the school community evaluates all areas, topics and indicators;

2. Thematic self-evaluation. During it, a narrower aspect of the activity is selected for a more in-depth study of the quality of the current situation;

3. Analysis of the problem that has arisen at the school. During it, the aim is to collect data that reveals the causes of the problematic situation and, based on them, present a solution to the problem.

It is recommended to implement the self-evaluation of the quality of school activities in five stages:

1. The preparation stage. An agreement is made on the aspect of the school's activities to be assessed and the assessment method;

2. The stage of preparing the self-evaluation plan. The goals of the school's self-evaluation, participants, data sources, data interpretation criteria are determined, and an assessment process plan is prepared;

3. The stage of preparing self-evaluation instruments. Self-evaluation instruments are selected and applied;

4. The stage of carrying out the self-evaluation. Reliable data and information are collected. The data obtained are analyzed, interpreted, and conclusions are formulated;

5. The reporting and information stage. A self-assessment report on the quality of school activities is prepared, which presents the summarized data of the self-assessment, their analysis, conclusions and recommendations for improving the quality of school activities.

One of the objectives of the self-assessment of the quality of school activities specified in (Minister, 2016a) is “to develop a culture of data-based management at school”. It is important to collect reliable data for the self-assessment of the quality of school activities. In addition to the data collected directly during the self-assessment, secondary data sources are also used:

- quantitative school monitoring data,
- data on student progress and learning achievements,
- data on self-assessment and certification of teachers and managers,
- data from surveys and research conducted at the school, etc.

4. Results of application of data analytical tools for assessment of the quality of activities in secondary education school

The National Education Agency has prepared the “Recommendations for the Application of Self-Assessment Questionnaires for General Education Schools ” (NEA, 2022) as a presentation of quantitative and qualitative self-assessment instruments (questionnaires)

for the quality of school activities. The questionnaires were prepared during the implementation of the activities of the project No. 09.2.1-ESFA-V-706-03-0001 "Improvement and Development of Non-Formal Children's Education, Pre-School, Pre-Primary and General Education Assessment, Self-Assessment" and tested in 22 Lithuanian general education schools. As an alternative to the previous self-assessment system, the National Education Agency proposed the following assessment instruments:

• Conduct surveys of students, their parents and teachers about the quality of school activities in other online survey systems, such as Microsoft Office 365 Forms, Google Forms, *apklausa.lt, manoapklausa.lt,* etc.;

• Use IQES online Lithuania questionnaires to collect quantitative data, which are published on the education portal *Emokykla.lt*.

• Apply qualitative data collection methods, such as focus group discussions, interviews, etc.

The following questionnaires for general education schools are published on *Emokykla.lt*:

• Broad self-assessment questionnaire;

• Thematic questionnaires for individual areas of activity:
  - Results,
  - Education and student experiences,
  - Educational environments,
  - Leadership and management;

• Nine feedback questionnaires for the quality of lessons.

Since January 2022, the National Education Agency has started organizing training for school communities on the application of new and recommended self-assessment tools of Microsoft Forms, Microsoft Excel, for data analysis and visualization apply Microsoft Excel Pivot Add-in.

4.1. Case study of assessment of quality of activities in education institution

In order to clarify the capabilities of the data analytics tool Power BI, a case study was conducted in a specific school. During the study, two school quality assessment processes were selected for analysis, the results of which are presented during the external school evaluation, regardless of the type of external evaluation (Table 5).

Based on the Description of the Procedure for Organizing and Conducting External Evaluation of Activities of Schools Implementing General Education Programs (2021), the school principal must provide external evaluators with:

• new information on self-assessment and self-assessment conclusions (for thematic evaluation – over the past 2-3 years),

• summarized information on the school's student achievements, progress and student achievement research data for the past two years.

School X uses Microsoft Office 365 A1, the software package of which includes Word, Excel, Power Point, One Drive, One Note, Outlook programs. Such packages allow to use the distance learning tool - the Microsoft Teams platform (*office365mokykloms.lt*). School X uses an *e-diary*, the digital learning environment *EDUKA class* and environment Moodle, the virtual libraries *ELVIS* and *Vyturys*, and the following systems:

• *Avilys* - a document management system.

• *Paskata* is a personnel management and payroll calculation system, which consists of three integrated modules: 1) personnel management; 2) working time accounting; 3) payroll calculation (*cgi.com/lietuviskai/lt/paskata*).

Table 5. Data on the activities and analysis of School X selected for the research

	Activity	**Corresponding persons**	**Duration of analyzed data**	**Data sources**	**Applied digital tools**
1.	Preparation of summarized information on the achievements, progress and student achievement research data of the school for the last two years	Coordination working group, led by the Deputy Director for Education	2022-2024. 2021-2022	E-diary archive, Folder "School Achievements and Progress Academic Year", stored on the cloud	Tools for E-Diary, Microsoft Excel Workbooks, Microsoft Power Point
2.	Preparation of new information on self-assessment and self-assessment conclusions (for thematic assessment – over the last 2-3 years) Coordination working group, led by the Deputy Director for Education	Coordination working group, led by the Deputy Director for Education	2022-2024 2021-2022 2020-2021	Microsoft Forms archive, apklausa.lt archive Folder "School self-evaluation academic year", stored on the cloud	E-diary, Microsoft Excel Work book, Microsoft Power Point, Microsoft Forms, apklausa.lt,

School X has following information systems installed as:

• Student Register (MR) - a nationwide information system that collects important information about each person's education (*mokiniuregistras.prisijungti.lt*).

• Pedagogical Register (PER) (*nsa.smm.lt/registrai/pedagogu-registras*/).

• Education Management Information System (ŠVIS) – a nationwide information system that provides data necessary for educational entities to analyze and assess the state of education in various aspects, to predict educational change, to make data-based decisions and to implement management that guarantees the quality of education (*nsa.smm.lt/svietimo-sistemos/svis-informacine-sistema*/).

• National Examinations Centralized Information System (NECIS regulations, 2022)

• *E-delivery* – an information system for e-delivery of notifications and documents to individuals and legal entities (*epristatymas.lt*/).

In general education schools, large amounts of data are accumulated in various systems. However, the collected data by themselves do not guarantee changes. Only analyzed, summarized data, and their purposeful interpretation become a starting point for improving school activities. For a comprehensive school to be a successful organization, it must adapt to changing circumstances and challenges, remaining functionality and efficiency. To achieve this, changes are needed in the products and services it provides and uses. Changes, improvements, and possibly new implementations are necessary in the IT infrastructure of the performance quality assessment process. Therefore, the application of the capabilities of the data analytics system Power BI in a comprehensive school was chosen for the experimental study.

4.2. Results of visualization of assessment data

The data collected during the performance of quality of self-assessment of "School Networking" conducted by School X was selected for the study. The data obtained by the survey method, applying questionnaires prepared by the National Education Agency and using the Microsoft Forms survey system. The data obtained during the study are loaded into the data analytics tool "Microsoft Power BI Desktop" and based on them, reports on the results of the surveys of School X and the school's achievements and progress are formed (Figure 2).

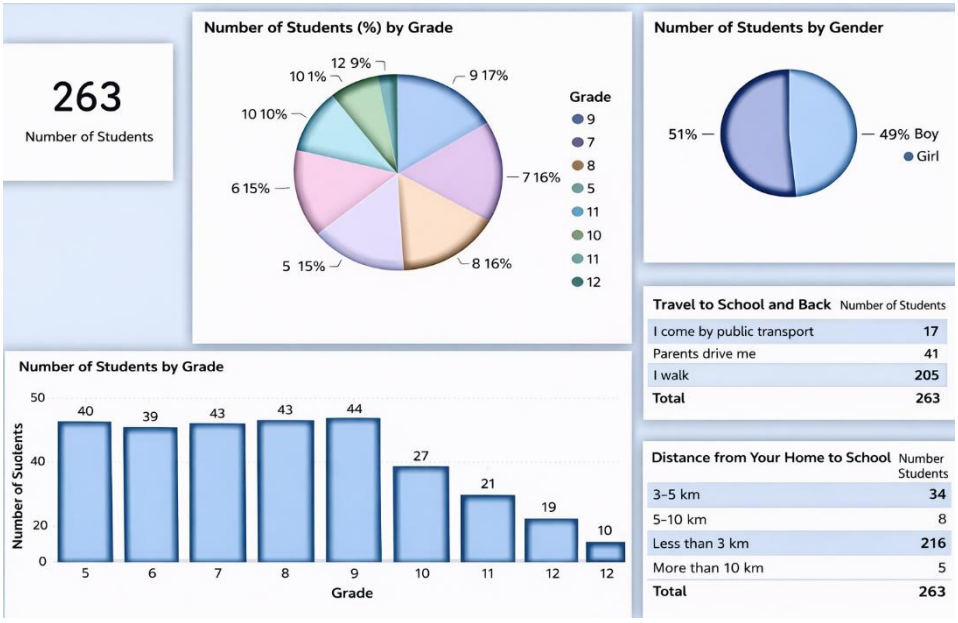


Figure 2. Illustration of report of summarized information about students who participated in the survey

Figure 2 presents general data on the number of students participated in the survey, their distribution by class and percentage of the total number of students, distribution by gender, distance of residence from school, method of arrival at school. This process includes four stages:

- loading data from the data sources;
- creating a data model;
- creating relationships between data;
- data visualization, report creation, applying selected data filters.

The last stage of the research is the generalization of the research results and the formulation of conclusions about the performance of institution. When assessing the relevance of the school website, a report was first compiled summarizing the information of a certain group of respondents, for example, students.

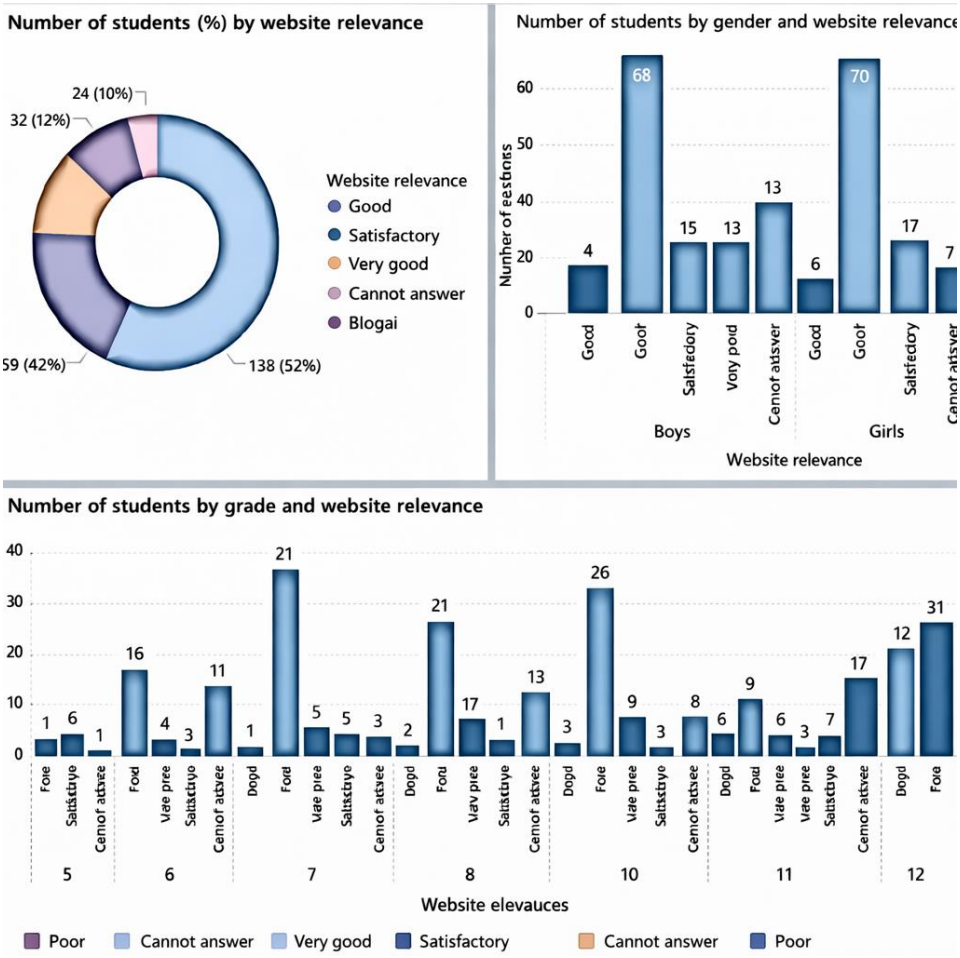


Figure 3. Illustration of report of responses of students to the school website relevance assessment

The results of the answers to each question in the survey are analyzed below. For example, when assessing students' attitudes towards the quality of the school website, the 4 aspects indicated in the survey (relevance, convenience, informativeness, attractiveness), it is possible to provide an assessment of one aspect or combine the assessment of all four aspects in the report, thus creating the possibility of comparative analysis. Figure 3 illustrates how students assess the school website in terms of relevance. The general attitude of students is presented, indicating the number of those who chose one of the possible answers (very good, good, satisfactory, bad) and the percentage of the total number of students.

Figure 4 presents student attitudes towards all four aspects of the website, indicating the number of those who chose one of the possible answers and the percentage of the total number of students. If necessary, evaluators can isolate the assessment of students in a specific class and compare differences in student attitudes by age category.

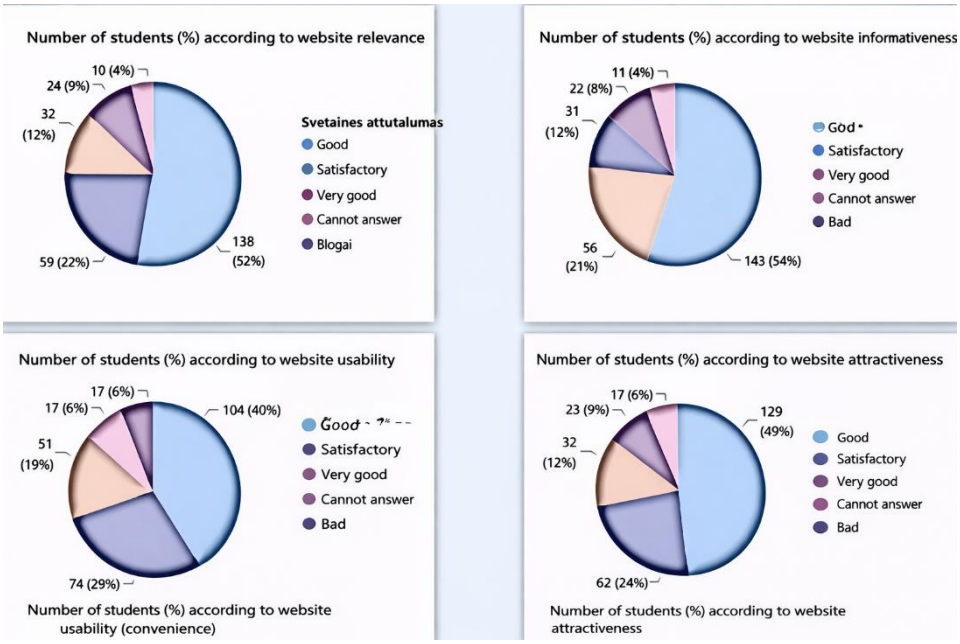


Figure 4. Illiustration of data of School website assessment according to student response results

The same principle is used to form reports on the results of surveys of other respondent groups (teachers and parents). Then, it is possible to conduct a comparative analysis of the views of all three respondent groups on the selected aspect. The results of the application of the data analytics system Power BI during the self-assessment of the quality of school activities enable the conclusions of the experimental study:

• Provides the ability to perform in-depth data processing, including filtering and data transformation before visualization.

• Power BI provides much wider possibilities for creating visualizations than Microsoft Forms, since the diagrams presented in the generated report cannot be selected.

Power BI Desktop allows you to create detailed and interactive reports using various types of diagrams. The diagrams used in the study: column, pie, ring, card, table forms.

• Provides the ability to change colors, fonts, formats, etc., so that the visualization meets the needs of its compilers. There is an opportunity to use visualizations created by third parties, thus further expanding the range of visualizations.

• Visualizations are interactive. You can define interactions between different visualizations in a report.

• Allows you to connect different data sets, allowing you to obtain more detailed insights.

The Power BI Desktop tool has a fairly intuitive and user-friendly interface that makes it easy to upload data, create basic charts and reports. Microsoft Forms is a suitable tool for conducting quick surveys and creating simple charts, while Power BI Desktop is a much more powerful tool for data analysis and visualization, especially when more detailed data analysis, more complex visualizations or the need to integrate and analyze data from various sources are required. Power BI Desktop allows you to reveal deeper insights and provide opportunities to customize reports for specific school needs.

4.3. Definitive statement and recommendations

The study systematically analyzed the application of data analytical tools, specifically Microsoft Forms, Excel, and Power BI, in assessing the quality of activities within Lithuanian secondary education institutions. The analysis focused on the digitalization of assessment procedures, the integration of data-driven self-assessment practices, and the effectiveness of analytical systems in supporting evidence-based decision-making in schools.

The objectives of the research - namely, to identify appropriate analytical tools, evaluate their suitability for educational self-assessment, and propose an adaptable framework for performance evaluation - have been successfully achieved. The findings demonstrate that the adoption of Microsoft Power BI, supported by Microsoft Forms and Excel, provides a scalable and intuitive environment for the collection, processing, and visualization of heterogeneous school performance data. This integration enhances institutional capacity for self-assessment, aligns with the Lithuanian State Education Strategy (2013–2022), and supports the national goal of developing a culture of educational quality based on data analysis.

The key findings are formulated:

- Data-driven self-assessment is increasingly central to quality assurance in Lithuanian education, supported by updated legal and methodological frameworks.
- Microsoft Power BI offers superior analytical and visualization capabilities compared to traditional tools (e.g., Excel or Forms alone), enabling comparative analysis among stakeholders (students, teachers, parents).
- Accessibility and interoperability of Microsoft Office 365 tools ensure cost-effective and widespread implementation potential across schools.
- The integration of qualitative and quantitative methods (survey data, interviews, focus groups) strengthens the validity of self-assessment outcomes.

We would like to recommend for other schools:

- Adopt Power BI within national and municipal education agencies as the primary platform for school-level data analysis and reporting.
- Provide targeted training programs to enhance school administrators' data literacy and ensure consistent application.
- Strengthen theoretical and methodological consistency.
- Align analytical criteria and evaluation indicators with internationally recognized educational quality frameworks (e.g., OECD Education Indicators, EU Education and Training 2030 objectives) while maintaining relevance to the Lithuanian context.
- Promote Data Ethics and Governance Standards.
- Ensure compliance with GDPR and national data protection laws by developing clear guidelines for the ethical use, sharing, and storage of educational data.
- Encourage Continuous Improvement through Feedback Loops.
- Use self-assessment data not only for accountability but also for formative improvement, but for linking analytical insights directly to strategic school development plans.

5. Discussion and conclusions

The theoretical grounded the comprehensive framework of assessment of secondary school performance was developed. The set of evaluation criteria of data analytical tools was proposed. Realized means help to define the framework for assessing the suitability and effectiveness of data analytical tools and implement them in the educational sector. These criteria ensure that tool selection is not merely technical but strategically aligned with national goals of quality education, evidence-based governance, and sustainable digital transformation.

The unified data management framework for assessment of performance of secondary educational institutions have to be created. Such framework has to integrate school-level data sources (e-diary, national registers, survey systems) into a centralized analytics platform to improve data accessibility, comparability, and longitudinal tracking.

The huge data and data heterogeneity forms some problems for analysis of academic performance through data obtained from the self-assessment of secondary educational institutions. The choosing of right analytical tools helps in self-assessment of academical activities of institutions. Proper self-assessment of the quality of activities of education institutions is very important in the process of improving the quality of education. Self-assessment and external assessment of the quality of school activities are regulated by legal acts that establish clear procedures and assessment criteria for the assessment of the quality of activities. The collection of reliable data and their use for formulating of conclusions about course and direction of activities.

The possibilities of providing right data analytical tools for making adaptable self-assessment of the quality of school activities helps in developed of framework of using software tools which implemented and helped in collection, analyze and visualization of heterogeneous data. Microsoft products, such as Microsoft Forms, Microsoft Excel, Microsoft Excel Pivot and Microsoft Power BI, as main tools are recommended for the self-assessment of the quality of school activities due to their accessibility and functionality. The Microsoft Power BI tool is distinguished as the main data analysis and

visualization tool due to its advanced functions and capabilities, which are adapted not only to business, but also to education institutions.

The study demonstrates that the stated objectives have been fully met, confirming the feasibility, effectiveness, and relevance of integrating modern data analytical tools into Lithuania's secondary education quality assessment process. The approach not only modernizes institutional evaluation processes but also builds a sustainable foundation for data-informed governance and educational improvement.

References

Collie, R., Singh, A. (2020). *Power BI Desktop Step-by-Step*. Pearson.

Dzemydaitė, G., Naruševičius, L. (2023). Exploring efficiency growth of advanced technology-generating sectors in the European Union: a stochastic frontier analysis. *Journal of Business Economics and Management (JBEM)*, 24(6), 976-995. https://doi.org/10.3846/jbem.2023.20688

Dzemydienė, D., Dzemydaitė, G., Gopisetti, D.. (2022). Application of multicriteria decision aid for evaluation of ICT usage in business. *Central European Journal of Operations Research*. New York: Springer. 2022, vol. 30, p. 323-343. DOI: 10.1007/s10100-020-00691-9.

Dzemydienė, D., Turskienė, S., Šileikienė, I. (2023). Development of ICT infrastructure management services for optimization of administration of educational institution activities by using ITIL-v4. *Baltic Journal of Modern Computing*. 11(4), p. 558-579. DOI: 10.22364/bjmc.2023.11.4.03.

Dzemydienė, D.. Turskienė, S.. Šileikienė, I. (2024). An approach of ICT incident management based on ITIL 4 methodology recommendations. *Baltic Journal of Modern Computing*. 12(3), p. 286-303. DOI: 10.22364/bjmc.2024.12.3.05.

Fotache, M.. Strimbei, C. (2015). SQL and Data Analysis. Some Implications for Data Analysits and Higher Education. *Procedia Economics and Financ*e. **20**, p. 243-251. https://doi.org/10.1016/S2212-5671(15)00071-4

Gartner (2020). Top Strategic Technology Trends for 2021. eBook: https://www.gartner.com/en/publications/top-tech-trends-2021

Google Looker Studio Documentation (2025). Retrieved from: https://support.google.com/looker-studio/.

Hall, J. A., Park, N., Song, H., Cody, M. J. (2010). Strategic misrepresentation in online dating: The effects of gender, self-monitoring, and personality traits. *Journal of Social and Personal Relationships*, 27(1), 117–135. https://doi.org/10.1177/0265407509349633

Hofmann, M., Klinkenberg, R. (2016). RapidMiner: Data Mining Use Cases and Business Analytics Applications. *CRC Press*.

Jeyaraj, A. (2020). DeLone & McLean models of information system success: Critical meta-review and research directions, *International Journal of Information Management*, Vol. 54, 2020, 102139, https://doi.org/10.1016/j.ijinfomgt.2020.102139.

Karau, H., Warren, R. (2017). *High Performance Spark*. O'Reilly Media

Mandinach, E. B., Schildkamp, K. (2021). Misconceptions about data-based decision making in education: An exploration of the literature. *Studies in educational evaluation*, 69, 100842. https://doi.org/10.1016/j.stueduc.2020.100842

Marikyan, D., Papagiannidis, S. (2025) Technology Acceptance Model: A review. In S. Papagiannidis (Ed), *TheoryHub Book*. Available at https://open.ncl.ac.uk / ISBN: 9781739604400

Minister (2009a). Order No. lSAK-607 (2009). The Minister of Education and Science of the Republic of Lithuania of March 30, 2009. Recommendations for the Self-Assessment of the Quality of Activities of General Education Schools. Availanle at: https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/TAIS.340787

Minister (2009b). Order No. ISAK-608 (2009). The Minister of Education and Science of the Republic of Lithuania "On the Approval of the Description of the Procedure for External Assessment of the Quality of Activities of General Education Schools". 2009. March 30. https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/TAIS.340885?jfwid=-3paw9io01.

Minister (2016a). Order No. V-267 (2016). The Minister of Education and Science of the Republic of Lithuania of Order March 29, 2016. "Methodology for the Self-Assessment of the Quality of Activities of Schools Implementing General Education Programs". Available at: https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/4ee6e9e0fc9911e5bf4ee4a6d3cdb874

Minister (2016b). Order No. V-1167 (2016). The Minister of Education and Science of the Republic of Lithuania of 2016.12.30. "Description of the procedure for organizing and conducting external evaluation of the activities of schools implementing general education programs". https://e-seimas.lrs.lt/portal/legalAct/lt/TAD /a0c1cc33ced511e6a476d5908abd2210?jfwid=-3paw9io01

Minister (2018). Order No. V-962 (2018) Minister of Education, Science and Sports of the Republic of Lithuania. Description of the procedure for organizing and conducting external evaluation of the activities of schools implementing general education programs. Available at: https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/a6ab3ab2f80311e895b0d54d3db20123?jfwid=-3paw9io01

Minister (2021). Order No. V-1150 (2021) Minister of Education, Science and Sports of the Republic of Lithuania. 2021, June 21. Description of the procedure for organizing and conducting external evaluation of the activities of schools implementing general education programs. https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/9e7bcb10d2c711eb9787d6479a2b2829?jfwid=-3paw9io01

Ministry (2021). Strategic Activity Plan of the Ministry of Education, Science and Sports of the Republic of Lithuania for 2021–2023. Available at: https://smsm.lrv.lt /uploads/smsm/documents/files/Administracine%20informacija/planavimo%20dokumentai/org_%C5%A0MSM%20SVP%202021(nauja%20red_)_pdf_2021-08-05.pdf

Murray, D. (2016). *Tableau Your Data!* Wiley

NEA (2022). The National Education Agency. Recommendations for the Application of Self-Assessment Questionnaires for General Education Schools. Available at: https://www.nsa.smm.lt/wp-content/uploads/2022/02/BUM-isivertinimo-rekomendacijos-02-10-galutinis.pdf

NSEA (2020). National School Evaluation Agency. Recommendations for the use of school self-evaluation instruments. Available at: https://www.nsa.smm.lt/wp-content/uploads/2020/09/Mokyklu-saves-vertinimo-instrumentu-naudojimo-rekomendacijos-2010.

Okoye, K.. Campos, E.. Das, A.. Chakraborty, V.. Ghosh, M.. Chakrabarti, A.. Hosseini, S. (2025). Impact of digitalized-education upon sustainable education and practice: A systematic review and meta-analysis of literature based on pre-intra-and-post pandemic and rural education development. *Sustainable Futures*. Vol. 10, 100851, https://doi.org/10.1016/j.sftr.2025.100851.

Preidys, S. (2023). Teaching „MS Excel Power Query, Power Pivot ir Power BI". 2023. Sertifficate No. 313231121/29-9

Questionnaire (2024). Questionnaire examples and recommendations for their application. Educational portal "Emokykla.lt". Access: https://duomenys.ugdome.lt/?/mm/dry/med=153/884)

State Education Strategy 2013-2022. (2013). No. XII-745. Available at: https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/TAIS.463390

VanderPlas, J. (2016). Python Data Science Handbook. O'Reilly Media.
Walkenbach, J. (2013). Excel 2013 Bible. Wiley.
Wickham, H., Grolemund, G. (2016). R for Data Science. O'Reilly Media.

Received: August 22, 2025, revised December 22, 2025, accepted December 22, 2025