



Міністерство освіти і науки України  
Харківський національний університет радіоелектроніки

Ministry of Education and Science of Ukraine  
Kharkiv National University of Radio Electronics

## Автоматизовані системи управління та прилади автоматики

### Випуск 189

*Всеукраїнський міжвідомчий  
науково-технічний збірник*

**Заснований у 1965 р.**

Затверджений до друку  
Науково-технічною Радою  
Харківського національного  
університету радіоелектроніки  
(протокол № 6 від 21 травня 2026 р.)

#### **ЗАСНОВНИК**

Харківський національний  
університет радіоелектроніки,

#### **АДРЕСА РЕДАКЦІЇ:**

Україна, 61166, м. Харків, проспект Науки, 14  
Інформаційний сайт: <https://asu-pa.nure.ua>  
E-mail редколегії: [asu@nure.ua](mailto:asu@nure.ua)

## Management Information System and Devices

### Issue 189

*All-Ukrainian Interdepartmental  
Scientific and Technical Collection*

**Founded in 1965**

Approved for publication  
by the Scientific and Technical Council  
of the Kharkiv National  
University of Radio Electronics  
(Protocol No. 63 d.d. 21 May 2026)

#### **ESTABLISHER**

Kharkiv National  
University of Radio Electronics,

#### **EDITORIAL OFFICE ADDRESS:**

14 Nauky Ave., Kharkiv, Ukraine, 61166  
Information site: <https://asu-pa.nure.ua>  
E-mail of the editorial board: [asu@nure.ua](mailto:asu@nure.ua)

Ідентифікатор медіа R30-03874  
Витяг з реєстру суб'єктів у сфері медіа – реєстрантів від 25.04.2024 № 1410

*Журнал включено до Переліку наукових фахових видань України, в яких можуть публікуватися  
результати дисертаційних робіт на здобуття наукових ступенів доктора і кандидата наук  
наказом Міністерства освіти і науки України від 16.07.2018 №775 (додаток 7)*

Харків – 2026

## РЕДАКЦІЙНА КОЛЕГІЯ

### *Головний редактор*

**Рубан Ігор Вікторович,**  
доктор технічних наук, професор

### *Заступник головного редактора*

**Левикін Віктор Макарович,**  
доктор технічних наук, професор

### *Члени редколегії:*

**Волк Максим Олександрович,**  
доктор технічних наук, професор;

**Гребеннік Ігор Валерійович,**  
доктор технічних наук, професор;

**Євсєєв Владислав В'ячеславович,**  
доктор технічних наук, професор;

**Єрохін Андрій Леонідович,**  
доктор технічних наук, професор;

**Каргін Анатолій Олексійович,**  
доктор технічних наук, професор;

**Кошовий Микола Дмитрович,**  
доктор технічних наук, професор;

**Невлюдов Ігор Шакирович,**  
доктор технічних наук, професор;

**Новоселов Сергій Павлович,**  
доктор технічних наук, доцент;

**Осадчий Сергій Іванович,**  
доктор технічних наук, професор;

**Пітерська Варвара Михайлівна,**  
доктор технічних наук, професор;

**Удовенко Сергій Григорович,**  
доктор технічних наук, професор;

**Федорович Олег Євгенович,**  
доктор технічних наук, професор;

**Філатов Валентин Олександрович,**  
доктор технічних наук, професор;

**Филипенко Олександр Іванович,**  
доктор технічних наук, професор;

**Варламіс Іракліс,**  
PhD, Professor;

**Девадзе Давид,**  
Candidate of Sciences in Physics and Mathematics, Professor;  
**Сергієнко Олег,**

доктор технічних наук, професор;  
**Койфман Олексій Олександрович,**

кандидат технічних наук, доцент;  
**Семенов Сергій,**

доктор технічних наук, професор

### *Відповідальний секретар:*

**Цимбал Олександр Михайлович,**  
доктор технічних наук, професор;

## EDITORIAL BOARD

### *Editor in Chief*

**Ihor Ruban,**  
Doctor of Technical Sciences, Professor

### *Deputy Chief Editor*

**Viktor Levykin,**  
Doctor of Technical Sciences, Professor

### *Editorial Board Members:*

**Maksym Volk,**  
Doctor of Technical Sciences, Professor;

**Igor Grebennik,**  
Doctor of Technical Sciences, Professor;

**Vladyslav Yevsieiev,**  
Doctor of Technical Sciences, Professor;

**Andriy Yerokhin,**  
Doctor of Technical Sciences, Professor;

**Anatolii Kargin,**  
Doctor of Technical Sciences, Professor;

**Mykola Koshovyi,**  
Doctor of Technical Sciences, Professor;

**Igor Nevliudov,**  
Doctor of Technical Sciences, Professor;

**Sergiy Novoselov,**  
Doctor of Technical Sciences, Associate Professor;

**Sergei Osadchy,**  
Doctor of Technical Sciences, Professor;

**Varvara Pitera,**  
Doctor of Technical Sciences, Professor;

**Serhiy Udoenko,**  
Doctor of Technical Sciences, Professor;

**Oleg Fedorovich,**  
Doctor of Technical Sciences, Professor;

**Valentin Filatov,**  
Doctor of Technical Sciences, Professor;

**Oleksandr Filipenko,**  
Doctor of Technical Sciences, Professor;

**Iraklis Varlamis,**  
PhD, Professor;

**David Devadze,**  
Candidate of Sciences in Physics and Mathematics, Professor;  
**Oleg Sergiyenko,**

DDSc, Professor;

**Oleksiy Koyfman,**  
Candidate of Technical Sciences, Associate Professor;

**Serhii Semenov,**  
Doctor of Technical Sciences, Professor

### *Responsible secretary:*

**Oleksandr Tsymbal,**  
Doctor of Technical Sciences, Professor;

## ЗМІСТ

### Інтелектуальні інформаційні технології

- Білак Ю. Ю.** 5  
 Інформаційна технологія аналізу спектральної чутливості в задачах параметричного синтезу  
**Григоренко С. М., Савельєва О. С., Протасова Л. І.** 17  
 Модель нейромережевого планування кіберфізичної системи на основі аферентного синтезу  
**Нечипуренко С. І., Сорокін А. Р.** 34  
 Метод структурного прунінгу згорткових нейронних мереж з огляду на апаратні характеристики edge-пристроїв  
**Роговий А. І., Ахїєзер О. Б., Решетнікова С. М., Колбасін В. О., Пахомов Ю. В.** 47  
 Інтелектуальна модель формування інноваційно-інжинірингових кластерів  
**Склярів І. І.** 57  
 Інформаційна технологія інтелектуального управління сервісними процесами в автомобільній галузі на основі Big Data  
**Ткачов В. М.** 72  
 Метод локального відновлення процесів інформаційної системи на мобільній платформі

### Комп'ютерні системи, мережі та вбудовані технології

- Бобровнікова К. Ю., Гурман І. В., Олексюк Д. А.** 95  
 Порівняльний аналіз міжсервісної взаємодії з використанням REST і gRPC у середовищі .NET 8  
**Дяченко Д. О., Міхаль О. П.** 109  
 Метод оцінювання ризику дефектів програмних компонентів розподілених комп'ютерних систем для пріоритизації автоматизованих випробувань

### Програмні системи та безпека інформаційних технологій

- Віпшовський Ю. А., Грицюк Ю. І.** 128  
 Локалізація дефектних зон на поверхнях металів методом аналізу кумулятивних проєкцій інтенсивності пікселів  
**Вовк О. В., Григор'єв О. В., Глюза М. П., Чеботаров Р. І., Менделєва М. В.** 138  
 Тестування юзабіліті страхових вебсайтів за допомогою проксі-користувачів  
**Даценко С. С., Кучук Г. А.** 153  
 Міжвибіркова узагальненість для виявлення фейкових новин із використанням контрастивного адаптаційного навчання

## CONTENTS

### Intelligent information technology

- Bilak Y.** 5  
 Modification of the spectral sensitivity analysis method in parametric synthesis problems  
**Hryhorenko S., Savielieva O., Protasova L.** 17  
 A neural network model for planning the movement of a cyber-physical system based on afferent synthesis  
**Nechypurenko S., Sorokin A.** 34  
 Method of structural pruning of convolutional neural networks with regard to edge device hardware characteristics  
**Rogovyi A., Akhiezer O., Reshetnikova S., Kolbasin V., Pakhomov Y.** 47  
 Intelligent Model for the Formation of Innovation and Engineering Clusters  
**Skliarov I.** 57  
 Information technology for intelligent management of service processes in the automotive industry based on big data  
**Tkachov V.** 72  
 A method for local recovery of information system processes on a mobile platform

### Computer systems, networks and built-in technologies

- Bobrovnikova K., Hurman I., Oleksiuk D.** 95  
 Comparative analysis of inter-service communication using REST and gRPC in .NET 8 environment  
**Diachenko, D., Mikhal, O.** 109  
 Method for assessing the risk of defects in software components of distributed computer systems for automated test prioritization

### Software systems and information technology security

- Vipshovsky Y., Hrytsiuk Y.** 128  
 Localization of Defective Zones on Metal Surfaces via Cumulative Pixel Intensity Projection Analysis  
**Vovk O., Hryhoryev O., Hliuza M., Chebotarov R., Mendieliava M.** 138  
 Usability testing of insurance websites via proxy users  
**Datsenko S., Kuchuk H.** 153  
 Cross-Dataset Generalization for Fake News Detection Using Contrastive Fine-Tuning

|                                                                                                                                                                                                                  |           |                                                                                                                                                                                                                     |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Кириченко І. В., Демчук В. Г., Луценко В. В.</b><br>Розроблення й дослідження підходу до виявлення аномалій у потокових конвеєрах даних на основі операційної телеметрії                                      | 165<br>ua | <b>Kyrychenko I., Demchuk V., Lutsenko V.</b><br>Research of automated data quality testing and anomaly detection methods in data pipelines                                                                         |
| <b>Кучапін М. Ю., Смеляков К. С., Чуприна А. С., Лученко С. В.</b><br>Комплексний аналіз методів та інструментів гібридного анотування зображень у системах комп'ютерного зору                                   | 182<br>en | <b>Kuchapin M., Smelyakov K., Chupryna A., Luchenko S.</b><br>Comprehensive analysis of methods and tools for hybrid image annotation in computer vision systems                                                    |
| <b>Кушнір І. С., Купій В. В.</b><br>Модель оцінювання трудовитрат на розроблення програмного забезпечення на основі попередньо навчених zero-shot-моделей                                                        | 196<br>en | <b>Kushnir I., Kyrii V.</b><br>Software development effort estimation model based on pre-trained zero-shot models                                                                                                   |
| <b>Маковецький С. О., Каук В. І.</b><br>Метод проєктування програмного забезпечення крайових вузлів інтернет-речей mesh-мереж в умовах хибних спрацювань                                                         | 211<br>ua | <b>Makovetskyi S., Kauk V.</b><br>Architectural model and design method of software for edge nodes of IoT mesh networks with three-stage false-trigger suppression                                                  |
| <b>Просолов В. В., Халімов Г. З.</b><br>Формування простору ознак для ризик-скорингу Bitcoin на основі локальних і топологічних метрик                                                                           | 224<br>ua | <b>Prosolov V., Khalimov H.</b><br>Feature space construction for Bitcoin risk scoring based on local and topological metrics                                                                                       |
| <b>Сидоренко З. М., Северінов О. В.</b><br>Метод забезпечення цілісності даних із використанням модифікованого завадостійкого коду в системах промислового інтернету речей                                       | 235<br>ua | <b>Sydorenko Z., Sievierinov O.</b><br>Method for ensuring data integrity using a modified error-correcting code in industrial internet of things systems                                                           |
| <b>Ситник О. О.</b><br>Підхід до диспетчеризації бізнес-процесів в умовах функціонування виробничих об'єктів на основі антропоцентричності                                                                       | 247<br>ua | <b>Sytnik O.</b><br>Approach to business process dispatching in the operation of industrial facilities based on anthropocentricity                                                                                  |
| <b>Системне моделювання та обчислювальні методи</b>                                                                                                                                                              |           | <b>Sistem modeling and computational method</b>                                                                                                                                                                     |
| <b>Далека В. Х., Бомбандьоров Ю. П., Метешкін К. О., Андрусевич А. О., Чумаченко І. В.</b><br>Інформаційне забезпечення енергетичної ємності електротранспорту в проєктах і програмах сталого розвитку України   | 259<br>ua | <b>Daleka V., Bombandyorov Y., Meteshkin K., Andrushevich A., Chumachenko I.</b><br>Information support of the energy capacity of electric transport in projects and programs of sustainable development of Ukraine |
| <b>Димерцов Д. О.</b><br>Комп'ютерне моделювання складних процесів на основі одновимірного підходу                                                                                                               | 267<br>en | <b>Dymertsov D.</b><br>Computer simulation of gas-dynamic processes: a one-dimensional approach                                                                                                                     |
| <b>Євтушенко О. С., Заковоротний О. Ю.</b><br>Апроксимація критичної швидкості поїзда методами машинного навчання                                                                                                | 276<br>ua | <b>Yevtushenko O., Zakovorotnyi O.</b><br>Estimation of a train's critical speed using machine learning methods                                                                                                     |
| <b>Ніколайчук А. І., Кобилін О. А., Кобилін І. О., Пуятіна О. Є.</b><br>Комплексне оцінювання трансформерних моделей для задачі семантичної подібності речень для англійської та української мов                 | 284<br>en | <b>Nikolaichuk A., Kobylin O., Kobylin I., Putyatina O.</b><br>A comprehensive evaluation of transformer models for sentence-level semantic similarity in English and Ukrainian                                     |
| <b>Островецький С. В.</b><br>Застосування штучного інтелекту в транспортній логістиці: комплексний огляд моделей, методів і практичних рішень                                                                    | 297<br>ua | <b>Ostrovetskyi S.</b><br>Application of artificial intelligence in transport logistics: a comprehensive review of models, methods and practical solutions                                                          |
| <b>Погорєлов С. В., Балаба Я. А., Татар'янц М. С., Беркунський Є. Ю., Павленко А. Ю.</b><br>Систематичний аналіз чутливості фізично-інформованих нейронних мереж у задачах моделювання епідемічних часових рядів | 309<br>en | <b>Pohorielov S., Balaba Y., Tatariants M., Berkunskyi Y., Pavlenko A.</b><br>A systematic sensitivity study of physics-informed neural networks for epidemic time series                                           |
| <b>Алфавітний покажчик</b>                                                                                                                                                                                       | 325       | <b>Alphabetical index</b>                                                                                                                                                                                           |



Yurii Bilak

## MODIFICATION OF THE SPECTRAL SENSITIVITY ANALYSIS METHOD IN PARAMETRIC SYNTHESIS PROBLEMS

**The subject** of the study is methods for analyzing the spectral sensitivity of parameters in parametric synthesis problems related to the reconstruction of medium characteristics from spectral data. **The aim** of the work is to develop a modification and investigate the applicability of the spectral sensitivity analysis method to improve the stability and identifiability of parameters in parametric synthesis problems. **To achieve this goal**, the following tasks were addressed: analysis of the causes of instability in parametric synthesis problems; investigation of approaches for evaluating spectral sensitivity of parameters; conducting numerical experiments to study the influence of parameters on the spectral response; and assessment of parameter informativeness using appropriate quantitative metrics. **The research methods** are based on the use of an analytical model of the spectral response, differential analysis of sensitivity functions, and numerical methods for computing integral characteristics and eigenvalues of the spectral sensitivity matrix. **The obtained results** demonstrate that the parameters have significantly different impacts on the spectral shape, which is confirmed both by the visual analysis of sensitivity functions and by numerical evaluation of their integral measures. It is established that the dominance of certain parameters leads to substantial differences in influence scales and results in the presence of poorly identifiable parameters. The spectrum of eigenvalues of the sensitivity matrix is analyzed, revealing features indicating the proximity of the problem to a degenerate case. Additionally, the condition number is evaluated, confirming the potential instability of the solution. It is concluded that spectral sensitivity analysis is an effective tool for preliminary assessment of parameter informativeness and can be used to improve the stability of parametric synthesis problems. The obtained **results** can be applied in the development of more complex spectral analysis models and in improving parameter identification methods.

**Keywords:** spectral sensitivity; parametric synthesis; parameter identifiability; problem stability; spectral analysis; numerical experiment; analytical model; parameter informativeness; information technologies.

### 1. Introduction

Parametric synthesis in spectroscopy is a key component of analyzing complex physical systems, in which the parameters of the medium are determined based on experimentally measured spectral characteristics. Such problems arise in various fields, including optics, materials science, laser diagnostics, and biomedical research, where spectral data serve as the primary source of information about the properties of the objects under study. Traditionally, parametric synthesis is formulated as the problem of minimizing the functional difference between the experimental and model spectra. However, in practice, such problems are often characterized by solution instability, which is caused by the limited nature of spectral data, the presence of noise, and the complex interdependence of model parameters. As a result, even small changes in the input data can lead to significant deviations in the recovered parameters, complicating their physical interpretation.

One of the key causes of instability is the non-uniform spectral sensitivity of the parameters, i.e., the varying degree of influence that individual parameters have on the shape of the spectrum within a given spectral

range. In cases where a change in a parameter has practically no effect on the spectral response, or can be compensated for by other parameters, the problem of their weak identifiability arises. This, in turn, leads to the degeneracy of the parametric synthesis problem and the emergence of non-physical or unstable estimates. In this regard, it is important to apply methods that allow for the analysis of the informativeness of spectral data regarding model parameters even before the stage of their numerical reconstruction. One such method is the spectral sensitivity analysis method, which is based on the study of changes in the spectral response as parameters vary. This method allows one to assess the contribution of each parameter to the formation of the spectrum, identify weakly sensitive or interdependent parameters, and determine the conditions under which the synthesis problem is well-posed.

This scientific article investigates the possibilities of applying the spectral sensitivity analysis method to improve the stability of parametric synthesis problems. The main focus is on evaluating the identifiability of parameters based on their influence on the spectral signal and the analysis of corresponding numerical characteristics. To illustrate the advantages, model

spectral dependencies are used, which clearly demonstrate the peculiarities of parameter behavior in cases of full and partial identifiability. The results obtained can be used for preliminary analysis of parametric synthesis problems, improving their robustness, and making a well-founded selection of parameters to be estimated, which creates the prerequisites for a more reliable application of numerical and neural network methods for processing spectroscopic data in practical problems. From the perspective of information technology, the modified method can be viewed as a component of the spectral data processing and decision support process in parametric synthesis problems. In particular, the analysis of spectral sensitivity enables the formation of informative features, the evaluation of the parametric space, and the selection of an effective configuration of parameters, which are characteristic tasks of modern information-analytical systems.

## 2. Literature Review and Problem Statement

---

The problems of parametric synthesis and parameter recovery based on spectral characteristics belong to the class of inverse problems, for which solutions are typically unstable, the formulation is ill-posed, and results are highly dependent on errors in the input data. Fundamental works on the theory of inverse problems have shown that even minor noise in spectral measurements can significantly affect the accuracy of parameter recovery, especially when there is strong correlation between them or the experimental data are insufficiently informative [1–3]. Furthermore, for complex multiparameter models, the problem is complicated by the limited amount of available information and the possibility of multiple close solutions [4].

Parameter identifiability analysis plays a crucial role in ensuring the correctness of parametric synthesis. In [5, 6], methods were developed based on the analysis of the profile likelihood function and the study of parametric dependencies, which allow for the determination of parameters that cannot be unambiguously estimated from the available data. However, most of these methods were developed primarily for biological or dynamic models, so when applied to spectroscopic problems with a large number of parameters, they often prove computationally complex.

The study [7] examines the relationship between spectral sensitivity and parameter identifiability in inverse uncertainty estimation problems. The authors

demonstrated that parameters with a weak influence on the spectral signal are significantly more difficult to recover correctly. A similar approach is used in [8], where the Fisher information matrix is applied to assess the informativeness of parameters. At the same time, such methods are primarily focused on statistical analysis of the model and do not always provide a sufficiently clear explanation of exactly how individual parameters influence the shape of the spectral dependence.

The problem of instability is particularly noticeable in modern spectroscopic problems, where spectral models are characterized by complex structures and a significant number of interrelated parameters. For example, in [9], a paper on photoacoustic tomography, the high sensitivity of the results to noise and prior model assumptions is emphasized. Similar difficulties are described in [10] for optical coherence tomography problems, where even high-quality spectral data do not guarantee the uniqueness of parameter recovery due to the possibility of multiple equivalent solutions.

A separate area of current research involves the use of neural network and hybrid methods for solving spectral parameter synthesis problems. In [11], a neural network model for analyzing diffuse reflection spectra is presented, which enables rapid parameter estimation. However, the authors note that the accuracy of such models depends significantly on the quality and representativeness of the training sample, and in cases of weak parameter identifiability, the effectiveness decreases noticeably. In [12], machine learning is applied with consideration of physical laws for spectral reconstruction; however, the problem of instability and insufficient informativeness of the parameters remains relevant. Thus, even modern neural network methods do not completely eliminate the problem of degeneracy in the parametric space, but only partially compensate for it due to the approximation properties of the models.

Regularization methods are traditionally used to stabilize solutions to inverse problems. In [13], a modern interpretation of Tikhonov regularization is presented as a tool for reducing the influence of noise and stabilizing ill-posed problems. However, regularization alone does not eliminate the problem of weak parameter informativeness. If the parameters have similar or nearly linearly dependent sensitivity functions, even regularized solutions may remain unstable or physically ambiguous.

In recent years, considerable attention has been paid to the use of automatic differentiation and methods based on partial differential equations in parametric

reconstruction problems. In [14], the effectiveness of automatic differentiation for the reconstruction of spectral functions is demonstrated, while in [15], the problem of parameter identifiability in PDE models is investigated. However, these methods focus primarily on the numerical aspect of solving problems and take into account, to a lesser extent, the preliminary analysis of parameter informativeness based on their spectral sensitivity.

In [16–18], methods for adaptive parametric synthesis and the creation of hybrid systems for analyzing the spectral characteristics of multilayer structures were developed. In particular, [16] considers the use of an adaptive informative subset of parameters to improve the stability of spectral synthesis. In the studies [17, 18], hybrid models of inverse analysis and the concept of a digital spectral twin of thin-film structures are presented. The obtained results confirm the promise of adaptive parameter space restriction; however, the issues of analyzing the spectral sensitivity of parameters and its impact on the degeneracy of the problem remain insufficiently studied.

Thus, the above analysis of current scientific works indicates that most existing methods of parametric synthesis are primarily focused on the numerical solution of inverse problems or the stabilization of results using regularization or machine learning. In contrast, insufficient attention is paid to the preliminary analysis of spectral sensitivity of parameters, which allows for the assessment of their informativeness and identifiability even before the synthesis stage. This is precisely why it is relevant to study methods of spectral sensitivity analysis to identify poorly informative parameters, evaluate the structure of the parametric space, and improve the stability of parametric synthesis problems.

### 3. Research Objectives and Tasks

The object of the study is the process of parametric synthesis of multilayer thin films.

The subject of the study is the information technology for analyzing the spectral sensitivity of parameters in parametric synthesis problems based on spectral data and the corresponding methods for evaluating their informativeness and identifiability.

The aim of the work is to develop a modified method for analyzing spectral sensitivity and to formalize the information technology for its application to improve the stability of solutions to parametric synthesis problems.

To achieve this goal, the following tasks must be solved:

Analyze the characteristics and causes of instability in parametric synthesis problems in spectroscopy and examine methods for analyzing the spectral sensitivity of parameters based on the differential characteristics of the spectral model.

Investigate the influence of spectral sensitivity on the identifiability of parameters and the correctness of their recovery, and develop a modified method for spectral sensitivity analysis.

Develop and formalize the structure of the information technology for spectral sensitivity analysis of parameters in parametric synthesis problems.

To conduct a numerical experiment to evaluate the behavior of parameters in cases of full and partial identifiability.

To analyze the possibility of using the results of spectral sensitivity analysis to improve the stability of parametric synthesis.

### 4. Materials and Methods

Research hypothesis, accepted assumptions, and simplifications. The hypothesis of the research conducted in this scientific work is that the use of a modified method for analyzing the spectral sensitivity of parameters, which includes the normalization of sensitivity functions and spectral analysis of the sensitivity matrix, makes it possible to improve the identifiability of parameters and the stability of parametric synthesis problems as early as the stage of their preliminary analysis.

A number of assumptions and simplifications are made in this work. The spectral response is a continuous and differentiable function of the parameters, while the parameters affect the spectrum independently or are weakly correlated in a local region. The study is conducted in a fixed spectral range with sufficient discretization, and the influence of noise in the data is not taken into account, which allows us to focus on the internal properties of the model. Another simplification is the use of a generalized analytical model of the spectrum instead of a complete physical model. A local (linear) approximation via sensitivity functions is considered, and continuous integral relations are replaced with their discrete counterparts. Additionally, normalization of sensitivity functions is applied to eliminate the influence of different parameter scales, which allows for a correct

comparison of their contributions to the formation of the spectral response.

### **Presentation of the main material**

This paper presents a modified method for analyzing the spectral sensitivity of parameters in parametric synthesis problems. The method is based on classical methods of sensitivity analysis and parameter informativeness estimation, which are widely used in inverse modeling and spectroscopy problems [1, 19]. The developed modification involves the use of normalization of sensitivity functions, as well as spectral analysis of the sensitivity matrix to evaluate the stability of the synthesis problem.

The spectral response of the system is defined as a mapping

$$S = S(\lambda, \vec{p}), \quad (1)$$

where  $\lambda$  is a spectral variable, and  $\vec{p} = (p_1, p_2, \dots, p_N)$  is a parameter vector.

To investigate the influence of the parameters, the spectral sensitivity functions are defined as

$$\Phi_i(\lambda) = \frac{\partial S(\lambda, \vec{p})}{\partial p_i}. \quad (2)$$

To analyze sensitivity, the partial derivatives of the spectrum with respect to the parameters that determine the spectral sensitivity functions are considered. These functions characterize the change in the spectrum as the corresponding parameter varies and allow us to evaluate its contribution to the formation of the spectral signal. The integral estimate of the parameter's influence is determined as follows

$$I_i = \int_{\Lambda} \left| \frac{\partial S(\lambda, \vec{p})}{\partial p_i} \right|^2 d\lambda = \int_{\Lambda} \Phi_i^2(\lambda) d\lambda, \quad (3)$$

which characterizes the total contribution of the parameter to the formation of the spectrum.

To account for differences in scale, it is proposed to use normalized sensitivity functions

$$\Phi_i^n(\lambda) = \frac{\Phi_i(\lambda)}{\|\Phi_i(\lambda)\|}, \quad (4)$$

where  $\|\Phi_i(\lambda)\|$  is the norm of the function, defined as the square root of the integral of the square of the sensitivity function over the spectral range:

$$\|\Phi_i\| = \sqrt{\int_{\Lambda} \Phi_i^2(\lambda) d\lambda}. \quad (5)$$

In addition, to analyze the interdependence of parameters, a spectral sensitivity matrix is formed using the formula:

$$M = [M_{ij}], \quad M_{ij} = \int_{\Lambda} \frac{\partial S}{\partial p_i} \frac{\partial S}{\partial p_j} d\lambda = \int_{\Lambda} \Phi_i(\lambda) \Phi_j(\lambda) d\lambda. \quad (6)$$

This matrix allows us to assess the linear independence of the parameters and determine the effective dimension of the parameter space. In the case where the rank of the matrix decreases, i.e.,

$$\text{rank}(M) < n, \quad (7)$$

the parametric synthesis problem becomes degenerate, leading to unstable solutions ( $n$  is the number of spectrally independent model parameters).

In normalized form, the matrix takes the form

$$M_{ij}^n = \int_{\Lambda} \Phi_i^n(\lambda) \Phi_j^n(\lambda) d\lambda. \quad (8)$$

To analyze the identifiability of the parameters, spectral decomposition is used

$$Mv_k = \lambda_k v_k, \quad (9)$$

where  $\lambda_k$  – eigenvalues,  $v_k$  – eigenvectors.

The condition number is defined as

$$k(M) = \lambda_{\max} / \lambda_{\min}. \quad (10)$$

The effective rank of the matrix is estimated using the following condition

$$r_{\text{eff}} = |\{\lambda_k : \lambda_k > \varepsilon\}|, \quad (11)$$

where  $\varepsilon$  is the threshold value.

To demonstrate the effectiveness of the modified method, a test analytical model of the spectrum

$$S(\lambda) = p_1 \cdot \exp\left(-\frac{\lambda - \lambda_0}{2\sigma^2}\right) + p_2(\lambda - \lambda_0) + p_3, \quad (12)$$

which reproduces the typical features of spectral dependencies (amplitude component, slope, offset).

For this test model, the corresponding sensitivity functions take the form:

$$\Phi_1(\lambda) = \exp\left(-\frac{\lambda - \lambda_0}{2\sigma^2}\right), \quad (13)$$

$$\Phi_2(\lambda) = \lambda - \lambda_0, \quad (14)$$

$$\Phi_3(\lambda) = 1. \quad (15)$$

A numerical experiment was conducted on synthetic data in the range  $\lambda \in [400, 600]$ . Sensitivity functions, integral indices  $I_i$ , the matrix  $M$ , and its eigenvalues were computed. The results showed that without normalization of the sensitivity functions, certain parameters dominate, leading to a loss of information from others. The developed modification, which includes normalization and analysis of the spectral characteristics

of the matrix, allows this effect to be eliminated, ensures a correct interpretation of the influence of parameters, and increases the stability of the parametric synthesis problem. Thus, the theoretical framework described above can be regarded as a modification of the spectral sensitivity analysis method, oriented toward practical application in spectroscopy and related fields.

For the practical implementation of the modified method, a discrete representation of spectral data is used. Let the spectral variable be given as a discrete set of values  $\lambda_k$ ,  $k=1, \dots, N$ . In this case, the integral relations (3), (5)–(8) are replaced by the corresponding discrete sums, which enables the numerical implementation of the algorithm. In particular, the integral sensitivity index is calculated as the sum of the squares of the values of the sensitivity function on a discrete grid, and the norm of the function is defined as the square root of this sum. This method allows one to transition from a continuous formulation of the problem to a computationally efficient discrete model without loss of information. The spectral sensitivity matrix in the discrete case is formed as a matrix of scalar products of discrete sensitivity vectors, which ensures convenient implementation in the form of matrix operations and allows for the effective use of standard numerical methods of linear algebra to find eigenvalues and estimate rank.

The numerical implementation was carried out using the Python programming language with standard scientific computing libraries. Sensitivity functions were computed analytically and numerically. To evaluate the properties of the problem, the eigenvalues of the matrix, its rank, and the diagonal elements of the inverse matrix were analyzed as indicators of the stability of parameter estimation. It is precisely this implementation that allows us to investigate the influence of spectral sensitivity on the stability of parametric synthesis, as well as to justify the feasibility of restricting the parametric space to parameters with a sufficient level of informativeness.

The developed modification can also be interpreted from the perspective of inverse modeling problems. In particular, the spectral sensitivity matrix is analogous to the sensitivity matrix or Jacobian used in least-squares methods. In this context, the analysis of its spectral characteristics allows one to assess the conditionedness of the problem and predict the behavior of numerical optimization methods. The algorithm for applying the method includes the following main steps: forming a spectral response model, computing spectral sensitivity

functions, normalizing the functions, constructing the sensitivity matrix, performing spectral analysis of the matrix, and evaluating integral indices, the condition number, and the effective rank. It is precisely this sequence of algorithmic steps that allows for a preliminary diagnosis of the parametric synthesis problem prior to the stage of solving the inverse problem.

The method modified by the author of this article can be used as a tool for the preliminary analysis of parametric synthesis problems. Its practical application involves the following sequence of steps: spectral sensitivity functions of the parameters are computed based on a given spectral response model, after which they are normalized to eliminate the influence of scales. Next, a spectral sensitivity matrix is formed and its eigenvalues are computed. In the final stage, integral sensitivity indices and the condition number are determined. Based on the obtained characteristics, the informativeness of the parameters is evaluated. Parameters with low integral sensitivity values or those corresponding to small eigenvalues of the matrix can be excluded or fixed to improve the stability of the problem. Additionally, the method can be integrated into optimization procedures as a preliminary stage of parameter selection. Thus, the method allows for a justified reduction in the dimensionality of the parameter space even before the numerical synthesis stage.

As a component of information technology, the modified method can be implemented as a sequence of computational procedures, including the processing of input spectral data, the construction of sensitivity functions, the formation of a matrix of characteristics, and their spectral analysis. It is precisely this structure that corresponds to the typical stages of information technology for data processing and can be integrated into spectral analysis software packages as a module for the preliminary diagnosis of parametric models.

The formalized structure of the information technology takes the form of an ordered set of components:

$$IT = \langle D, M, A, S, U, R \rangle, \quad (16)$$

where  $D$  is the input spectral data,  $M$  is the spectral response model,  $A = \{A1, A2, A3\}$  is the set of procedures for calculating and analyzing spectral sensitivity functions,  $S$  is the parameter informativeness evaluation block,  $U$  – the decision-making block for their selection,  $R$  – the resulting parametric model. Figure 1 presents the developed scheme of the information technology for analyzing the spectral sensitivity of parameters.



This information technology implements a sequence of spectral data processing from the model-building stage to the evaluation of parameter informativeness and decision-making regarding their selection, which ensures an increase in the robustness of parametric synthesis tasks. The obtained characteristics are used to evaluate the informativeness of parameters and make decisions regarding their selection, which in turn

ensures an increase in the robustness of parametric synthesis tasks. It should be noted that the modified method can be implemented as a software module or framework that integrates into spectral analysis systems and provides automated assessment of parameter informativeness and support for decision-making regarding their selection.

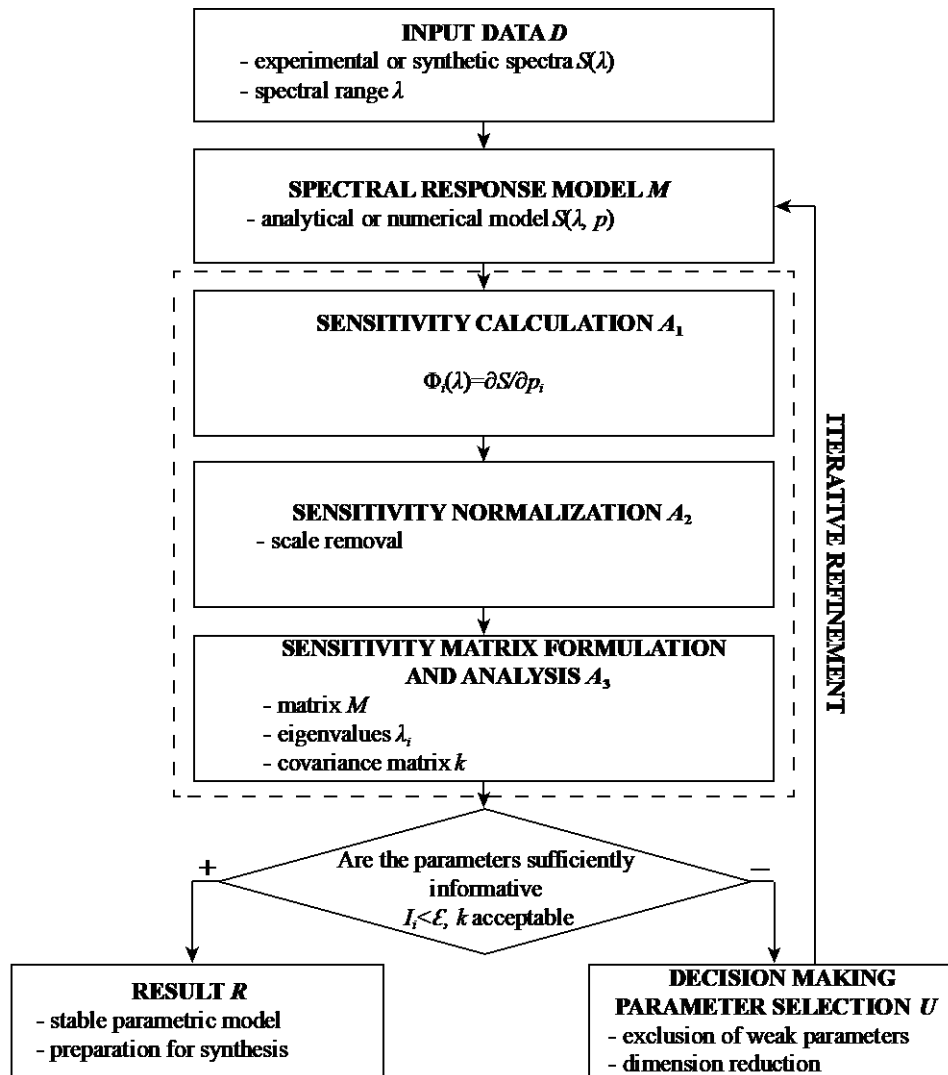


Fig. 1. Block diagram of the information technology for analyzing the spectral sensitivity of parameters

## 5. Research Results

To investigate the effect of spectral sensitivity on the stability of parametric synthesis, a numerical experiment was conducted based on the analytical spectral model described by formula (12). The use of synthetic data allows us to control the behavior of parameters, vary the conditions of identifiability, and clearly demonstrate the key properties of the problem

without the influence of experimental noise and errors. This model reproduces the generalized features of spectral dependencies characteristic of optical systems, in particular multilayer thin films, where the spectrum is formed as a result of interference and resonance effects. In particular, the localized spectral component models resonant transmission or reflection peaks, the linear trend accounts for background variations associated with the dispersion properties of the medium, and the



additive offset accounts for the baseline signal level. Furthermore, this choice of model is not intended to reproduce a specific physical structure but is used as a test bed for investigating general patterns of how parameters affect the spectral response, allowing us to focus on analyzing spectral sensitivity and parameter identifiability, which are universal characteristics of parametric synthesis problems.

The spectral dependence was defined as a superposition of three components: a localized resonant component (Gaussian peak), a linear trend, and an additive offset. Accordingly, these parameters have

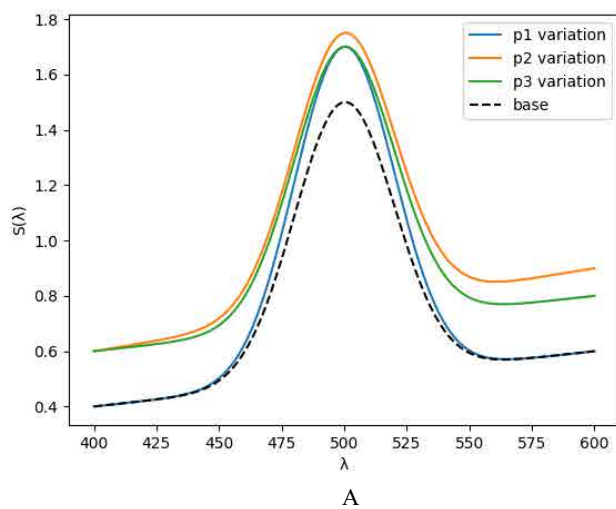


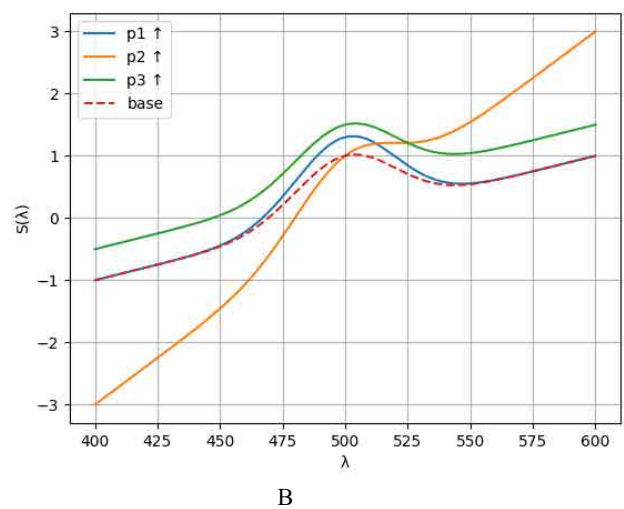
Fig. 2. The influence of parameters on the spectrum

Fig. 2A shows the effect of parameters on the spectral dependence in its original form. In this case, the slope parameter  $p_2$  and the shift parameter  $p_3$  are visually weak due to the use of an uncentered spectral variable and relatively small variations in the parameters. In Fig. 2B, the spectral variable is centered ( $\lambda - \lambda_0$ ), and the parameter variations are increased, which allowed the effects to be clearly distinguished: parameter  $p_2$  is responsible for the change in the spectral slope, while  $p_3$  is responsible for its vertical shift.

Fig. 3A shows the spectral sensitivity functions on the original scale. Due to a significant difference in orders of magnitude, the linear function  $\Phi_2(\lambda)$  dominates the others, with the result that  $\Phi_1(\lambda)$  and  $\Phi_3(\lambda)$  are virtually indistinguishable visually. The linear function  $\Phi_2(\lambda)$  increases in the range of  $\lambda$  values (400–600) and significantly exceeds the other functions in magnitude. As a result, the localized function  $\Phi_1(\lambda)$

different physical interpretations: parameter  $p_1$  determines the amplitude of the spectral line,  $p_2$  is responsible for the background slope, and  $p_3$  for the overall signal level. A model constructed in this way allows for the investigation of both local and global effects in the spectrum.

Figures 2–4 show the results of a numerical experiment in two visualization variants: (A) – the initial representation without additional data processing, (B) – an improved representation that accounts for scaling and normalization to enhance the clarity of the results.



has a significantly smaller amplitude and visually "merges" with the axis, while the constant function  $\Phi_3(\lambda)$  practically overlaps the lower part of the graph. Thus, all three functions are actually present, but their simultaneous display without normalization leads to a loss of visual distinctiveness.

Fig. 3B shows the normalized sensitivity functions, which allowed them to be brought to a comparable scale and clearly distinguished all three components: localized ( $\Phi_1$ ), linear ( $\Phi_2$ ), and constant ( $\Phi_3$ ).

Fig. 4A shows the eigenvalues of the spectral sensitivity matrix on a linear scale. In this case, a single dominant peak is observed, while the other values are practically imperceptible, creating the impression of degeneracy without a quantitative assessment. Fig. 4B uses a logarithmic scale, which allows us to visualize the entire spectrum of eigenvalues and assess their relative magnitudes. The rank of the matrix, which characterizes the effective dimension of the parametric space, is 3 in our case.

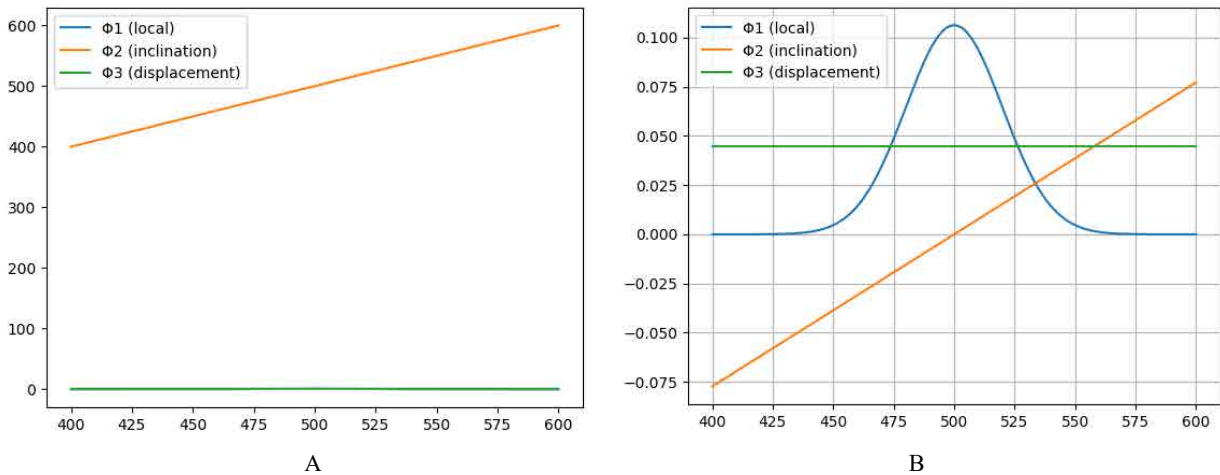


Fig. 3. Spectral sensitivity functions

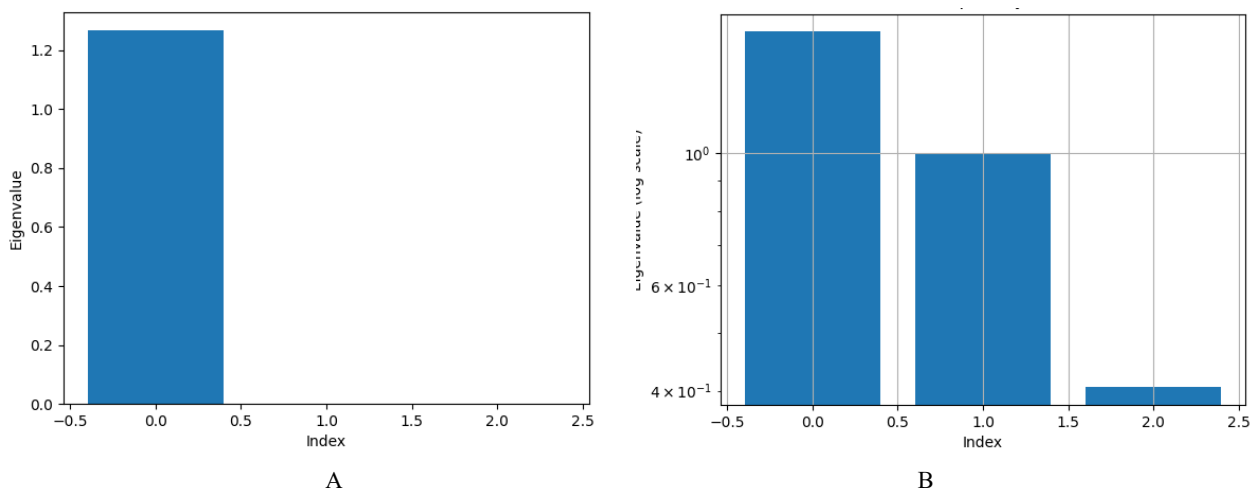


Fig. 4. Eigenvalues of the sensitivity matrix

The results shown in Figure 4 indicate a significant difference in the informativeness of the parameters: one direction in the parametric space is determined much better than the others. This situation is typical for problems with sensitivities of varying scales and can be interpreted as an approximation to a degenerate case. Smaller eigenvalues correspond to poorly identified parameters, which potentially leads to instability in their estimation.

Thus, the transition from option (A) to (B) makes it possible to eliminate the effects of dominance of individual components, increase the informativeness of the graphs, and ensure the correct interpretation of the results of spectral sensitivity analysis. Additionally, the obtained results confirm that spectral sensitivity analysis allows for the identification of parameters with varying levels of informativeness even before the synthesis problem is solved. Visualization of the sensitivity functions and the eigenvalue spectrum of the

matrix allows one to interpret the structure of the problem, explain the causes of instability, and justify the need to restrict the parameter space. Meanwhile, to improve visualization, it is advisable to use normalization of spectral sensitivity functions or plotting on a logarithmic scale, which will help avoid the dominance of individual components and allow for a more intuitive comparison of the parameters' contributions.

To quantitatively assess the contribution of parameters to the formation of the spectral response, integral indices of spectral sensitivity were calculated, defined as the integral of the square of the corresponding sensitivity functions, and which are  $I_1 \approx 35.4$ ,  $I_2 \approx 1.67 \times 10^5$ ,  $I_3 \approx 200$ . This confirms a significant difference in the magnitude of the parameters' influence on the spectral response, with the slope parameter ( $p_2$ ) being dominant. Additionally, the condition number of the spectral sensitivity matrix  $M$  was determined,

which for the case under consideration is of the order of  $10^3 - 10^4$ , indicating the potential instability of the parametric synthesis problem. Furthermore, the effective rank of the spectral sensitivity matrix is estimated, defined as the number of significant (non-zero, taking into account the threshold value) eigenvalues. The result obtained demonstrates that the effective rank is smaller than the formal number of parameters, which indicates the presence of weakly identified parameters and confirms the problem's approximation to a degenerate case. The obtained estimates are consistent with the shape of the eigenvalue spectrum (Fig. 4) and confirm the presence of significant non-uniformity in the informativeness of the parameters.

## 6. Discussion of Results

The obtained results confirm the general provisions of the theory of inverse problems regarding their instability and dependence on data informativeness [1–3]. In the present study, this manifests itself through significant non-uniformity in the spectral sensitivity of the parameters, as seen in the form of the functions  $\Phi_i(\lambda)$  (Fig. 3) and the integral indices  $I_i$  (Equation (3)). The dominance of certain parameters leads to a decrease in the identifiability of others, which is consistent with the conclusions regarding the possibility of the existence of close solutions [4]. Unlike identifiability analysis methods based on complex numerical procedures [5], the developed modification of the method allows for a visual assessment of parameter informativeness based on sensitivity functions (Equation (2)) and their normalization (Equations (4)–(5)), which in turn ensures better interpretability of results in spectral analysis problems. Analysis of the eigenvalue spectrum of the sensitivity matrix (Fig. 4) reveals the presence of dominant components, confirming the non-uniformity of the parameters' influence. A large value of the condition number  $k(M)$  (Equation (9)) indicates the potential instability of the problem, which is consistent with the results of studies in spectroscopy [9]. Compared to methods based on statistical estimates, in particular the information matrix [8], the developed modification allows for the direct analysis of the contribution of parameters to the formation of the spectrum. Furthermore, unlike regularization methods [12], it also allows for the identification of the causes of instability even before the problem is solved. Thus, the results

obtained demonstrate that the modified spectral sensitivity analysis method is an effective tool for the preliminary assessment of parameter identifiability and can be used to improve the stability of parametric synthesis problems.

The scientific novelty of this work lies in the modification of the spectral sensitivity analysis method for evaluating parameter identifiability and improving the stability of parametric synthesis problems in spectroscopy, as well as in the development of a method for its implementation as an information technology that includes the stages of spectral data processing, sensitivity analysis, and decision support for parameter selection.

The uniqueness of the method lies in the combination of the simplicity of the analytical model with the possibility of quantitative analysis of parameter informativeness through sensitivity functions and the eigenvalue spectrum of the corresponding matrix, which ensures a clear interpretation of the problem's properties.

An important advantage of the developed modification is its invariance to parameter scales, achieved through the normalization of sensitivity functions. It is this invariance that allows avoiding the dominance of parameters with large derivative values and ensures a more accurate comparison of their effects. Despite its advantages, the developed method has a number of limitations that should be taken into account when applying it. First, the use of an analytical model of spectral response simplifies real physical processes, which may limit the accuracy of parameter estimation in complex systems. Second, the results of the spectral sensitivity analysis depend on the choice of the spectral variable range and discretization, which can affect the values of integral indices and the spectral characteristics of the sensitivity matrix. Third, the method is based on local analysis, since sensitivity functions are defined as derivatives with respect to parameters, which assumes small variations in parameters in the vicinity of the selected point. In the case of highly nonlinear dependencies, this can lead to a loss of adequacy in the estimates. Furthermore, while normalizing the sensitivity functions eliminates scaling effects, it may partially reduce information about the absolute contribution of the parameters. It should also be noted that the analysis of the spectral sensitivity matrix does not account for noise in the data, which is characteristic of real experimental conditions. In such cases, it is necessary to additionally account for the statistical characteristics of the measurements and apply

regularization methods. Thus, the modified method is suitable for use as a tool for preliminary analysis and diagnosis of parametric synthesis problems, with subsequent refinement of the results based on more complex models and experimental data.

The results obtained can be used as a basis for developing information technology for spectral data analysis, in which the modified method performs the functions of processing, evaluating, and selecting parameters, opening up opportunities for its integration into spectral analysis software systems and intelligent decision support systems. The results of spectral sensitivity analysis obtained in this work are of practical significance for parametric identification problems in complex physical systems. The modified method allows for the identification of poorly identified parameters and the assessment of the stability of computational procedures at the problem formulation stage, which is critically important for improving the reliability of numerical modeling. In an industrial context, the method can be applied in spectroscopic material analysis, optoelectronics, non-destructive testing, and thin-film research technologies, where it is necessary to estimate medium parameters based on experimental spectral data. Furthermore, the method is relevant for automated process monitoring systems, where ensuring the stability of synthesis algorithms in the presence of noise and limited data is crucial. Thus, the spectral sensitivity method can be integrated into modern information technologies as a tool for preliminary diagnosis of the correctness of parametric synthesis and for improving its computational stability.

---

## 7. Conclusions

---

This research investigates the influence of the spectral sensitivity of parameters on the stability of parametric synthesis problems in spectroscopy. It is shown that the non-uniformity of sensitivities is one of the main causes of instability and deterioration of parameter identifiability.

A modification of the method for analyzing parameter informativeness based on spectral sensitivity functions and their correlation matrix has been developed, which allows for the estimation of the effective dimension of the parametric space. A structure of an information technology for spectral sensitivity analysis has been developed and formalized, which ensures the evaluation of parameter informativeness and

can be used to improve the stability of parametric synthesis problems. The developed modification of the method has been implemented as a component of an information technology for spectral data processing and decision support in parametric synthesis problems.

Based on a numerical experiment using synthetic spectral data, it has been demonstrated that parameters have varying degrees of influence on the spectral response, as confirmed by the form of the sensitivity functions and the eigenvalue spectrum of the corresponding matrix. It has been established that the presence of dominant eigenvalues of the spectral sensitivity matrix indicates the proximity of the problem to a degenerate case and points to the presence of weakly identified parameters. It is shown that the use of normalization and scaling allows for increasing the informativeness of the visualization and ensuring the correct interpretation of the results of spectral sensitivity analysis.

Prospects for further research include extending the method to more complex physical models, in particular multilayer thin-film structures, integration with numerical methods (TMM, RCWA), and use within hybrid and neural network architectures for solving parametric synthesis problems in spectroscopy. Furthermore, the obtained results can be used for preliminary analysis of parametric synthesis problems, the justified selection of parameters, and improving the stability of numerical methods for their recovery.

---

## Conflict of Interest

---

The authors declare that they have no conflict of interest, particularly of a financial, personal, authorial, or any other nature, that could influence the research or the results published in this article.

---

## Funding

---

The research was conducted without financial support.

---

## Data Availability

---

The manuscript has no associated data.

---

## Use of Artificial Intelligence

---

The ChatGPT artificial intelligence model (version GPT-5.3, OpenAI) was used in the preparation of the

---

manuscript. AI tools were used exclusively during the auxiliary stages of work on the manuscript, specifically to check the grammar, spelling, and punctuation of individual text fragments; to perform automated checks of linguistic formatting without altering the scientific content; and to conduct a preliminary search and compile a list of potentially relevant scientific sources from the past 5 years.

All selected literature sources were additionally and meticulously verified by the authors using scientific databases and, where necessary, edited. Linguistic edits were also manually checked for compliance with scientific writing style.

No substantive part of the manuscript, including the abstract, introduction, methods and models, results and discussion, as well as conclusions, was created or generated using AI tools. All main provisions, research results, and their interpretation were obtained by the authors independently.

The results obtained using artificial intelligence tools were thoroughly verified by the authors. The use of AI did not affect the scientific content of the work, the reliability of the results, their interpretation, or the conclusions drawn.

## References

1. Tarantola, A. (2005), *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia. DOI: <https://doi.org/10.1137/1.9780898717921>
2. Hansen, P. C. (2010), *Discrete Inverse Problems: Insight and Algorithms*, SIAM, Philadelphia. DOI: <https://doi.org/10.1137/1.9780898718836>
3. Kirsch, A. (2011), *An Introduction to the Mathematical Theory of Inverse Problems*, Springer, New York. DOI: <https://doi.org/10.1007/978-1-4419-8474-6>
4. Kaipio, J., Somersalo, E. (2005), *Statistical and Computational Inverse Problems*, Springer, New York.
5. Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., Timmer, J. (2009), "Structural and practical identifiability analysis via profile likelihood", *Bioinformatics*, Vol. 25, No. 15, pp. 1923–1929. DOI: <https://doi.org/10.1093/bioinformatics/btp358>
6. Kreutz, C. (2018), "Testing identifiability efficiently", *Bioinformatics*, Vol. 34, No. 11, pp. 1913–1921. DOI: <https://doi.org/10.1093/bioinformatics/bty035>
7. Wu, X., Shirvan, K., Kozłowski, T. (2019), "Sensitivity and identifiability in inverse uncertainty quantification", *Journal of Computational Physics*, Vol. 396, pp. 12–30. DOI: <https://doi.org/10.1016/j.jcp.2019.06.032>
8. Nordebo, S., Gustafsson, M., Nilsson, B., Sjöden, Th., Soldovieri, Fr. (2013), "Fisher information analysis in electrical impedance tomography", *Journal of Geophysics and Engineering*, Vol. 10, No. 6, Art. 064008. DOI: <https://doi.org/10.1088/1742-2132/10/6/064008>
9. Wang, Z., Tao, W., Zhao, H. (2023), "Optical inverse problem in quantitative photoacoustic tomography: A review", *Photonics*, Vol. 10, No. 5, Art. 487. DOI: <https://doi.org/10.3390/photonics10050487>
10. Macdonald, C. M., Arridge, S. R., Munro, P. R. T. (2023), "On the inverse problem in optical coherence tomography", *Scientific Reports*, Vol. 13, Art. 1507. DOI: <https://doi.org/10.1038/s41598-023-28366-w>
11. Lan, Q., McClarren, R. G., Vishwanath, K. (2023), "Neural network-based inverse model for diffuse reflectance spectroscopy", *Biomedical Optics Express*, Vol. 14, No. 9, pp. 4725–4738. DOI: <https://doi.org/10.1364/BOE.490164>
12. Park, H., Park, J. H., Hwang, J. (2024), "Physics-guided machine learning for optical spectrum inversion", *Scientific Reports*, Vol. 14, Art. 9042. DOI: <https://doi.org/10.1038/s41598-024-59594-3>
13. Gerth, D. (2021), "A new interpretation of Tikhonov regularization", *Inverse Problems*, Vol. 37, No. 6, Art. 064002. DOI: <https://doi.org/10.1088/1361-6420/abfb4d>
14. Wang, L., Shi, S., Zhou, K. (2022), "Reconstructing spectral functions via automatic differentiation", *Physical Review D*, Vol. 106, No. 5, Art. L051502. DOI: <https://doi.org/10.1103/PhysRevD.106.L051502>
15. Liu, Y., Suh, K., Maini, P. K., Cohen, D. J., Baker, R. E. (2024), "Parameter identifiability in PDE models of cell invasion", *Journal of the Royal Society Interface*, Vol. 21, No. 212, Art. 20230607. DOI: <https://doi.org/10.1098/rsif.2023.0607>
16. Bilak, Y. Y., Reblan, A. M., Mulesa, O. Y. (2026), "Controlled spectral parameter synthesis based on an adaptive informative subset", *Herald of Advanced Information Technology*, Vol. 9, No. 2, pp. 140–157. DOI: <https://doi.org/10.15276/hait.09.2026.10>
17. Bilak, Y. Y., Saibert, F. F., Reblan, A. M. (2025), "Development of a hybrid inverse analysis model for assessing spectral characteristics of multilayer structures", *Visnyk KNTU*, Vol. 1(92), Pt. 2, pp. 22–31. DOI: <https://doi.org/10.35546/kntu2078-4481.2025.1.2.3>



17. Bilak, Y. Y., Reblan, A. M., Mulesa, O. Y. (2026), "Architecture and implementation of a hybrid prototype system of a digital spectral twin of thin films", *Nauka i Tekhnika (AAIT)*, Vol. 9, No. 2, pp. 252–267. DOI: <https://doi.org/10.15276/aait.09.2026.18>
18. Press, W. H., Teukolsky, S. A. (2007), *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge.

Received (Надійшла) 06.04.2026

Accepted for publication (Прийнята до друку) 15.05.2026

Publication date (Дата публікації) 29.05.2026

#### *Відомості про авторів / About the Authors*

**Білак Юрій Юрійович** – кандидат фізико-математичних наук, доцент, Ужгородський національний університет, завідувач кафедри програмного забезпечення систем, Ужгород, Україна;

**Yurii Bilak** – PhD (Phys.-Math. Sciences), Associate Professor, Uzhhorod national university, Head of the department of software system, Uzhhorod, Ukraine;

e-mail: [yuriy.bilak@uzhnu.edu.ua](mailto:yuriy.bilak@uzhnu.edu.ua)

ORCID ID: <https://orcid.org/0000-0001-5989-1643>

## **ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ СПЕКТРАЛЬНОЇ ЧУТЛИВОСТІ В ЗАДАЧАХ ПАРАМЕТРИЧНОГО СИНТЕЗУ**

**Предметом вивчення** є інформаційна технологія аналізу спектральної чутливості параметрів у задачах параметричного синтезу за спектральними даними й відповідні методи оцінювання їх інформативності та ідентифікованості. **Мета дослідження** – розроблення модифікованого методу аналізу спектральної чутливості та формалізація інформаційної технології його застосування для підвищення стійкості розв’язків задач параметричного синтезу. Для досягнення поставленої мети розв’язано такі завдання: проаналізовано причини нестійкості задач параметричного синтезу; досліджено спосіб оцінювання спектральної чутливості параметрів; проведено чисельний експеримент для аналізу впливу параметрів на спектральний відгук; оцінено інформативність параметрів за допомогою відповідних кількісних показників. Крім того, сформовано й обґрунтовано структуру інформаційної технології аналізу спектральної чутливості, яка забезпечує послідовне опрацювання спектральних даних, оцінювання інформативності параметрів і підтримку прийняття рішень у задачах параметричного синтезу. **Методи дослідження** ґрунтуються на використанні аналітичної моделі спектрального відгуку, диференційного аналізу функцій чутливості та чисельних методів обчислення інтегральних характеристик і власних значень матриці спектральної чутливості. **Досягнуті результати** демонструють, що параметри мають суттєво різний вплив на форму спектра, що підтверджується як візуальним аналізом функцій чутливості, так і чисельними оцінками їх інтегральних показників. Встановлено, що домінування окремих параметрів призводить до значної різниці масштабів впливу й зумовлює появу слабо ідентифікованих параметрів. Проаналізовано спектр власних значень матриці спектральної чутливості, що дало змогу виявити ознаки наближення задачі до виродженого випадку. Додатково оцінено число обумовленості, яке підтверджує потенційну нестійкість розв’язків. **Висновки:** аналіз спектральної чутливості є ефективним інструментом попереднього оцінювання інформативності параметрів та може використовуватися для підвищення стійкості задач параметричного синтезу. Досягнуті результати можуть бути застосовані для побудови більш складних моделей спектрального аналізу та вдосконалення методів ідентифікації параметрів.

**Ключові слова:** спектральна чутливість; параметричний синтез; ідентифікованість параметрів; стійкість задач; спектральний аналіз; чисельний експеримент; аналітична модель; інформативність параметрів; інформаційні технології.

#### *Бібліографічні описи / Bibliographic descriptions*

Білак Ю. Ю. Інформаційна технологія аналізу спектральної чутливості в задачах параметричного синтезу. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 5–16. DOI: <https://doi.org/10.30837/0135-1710.2026.189.005>

Bilak, Y. (2026), "Modification of the spectral sensitivity analysis method in parametric synthesis problems", *Management Information System and Devices*, No. 2 (189), P. 5–16. DOI: <https://doi.org/10.30837/0135-1710.2026.189.005>



Григоренко С. М., Савельєва О. С., Протасова Л. І.

## МОДЕЛЬ НЕЙРОМЕРЕЖЕВОГО ПЛАНУВАННЯ КІБЕРФІЗИЧНОЇ СИСТЕМИ НА ОСНОВІ АФЕРЕНТНОГО СИНТЕЗУ

**Актуальність теми.** У статті розглянуто планування переміщення кіберфізичної системи в динамічному й невизначеному середовищі. Дослідження зумовлене потребою у створенні інтелектуальних систем управління, здатних забезпечувати автономне функціонування за умов неповної інформації, наявності перешкод і обмежених обчислювальних ресурсів. **Метою дослідження** є розроблення й формалізація моделі нейромережевого планування переміщення кіберфізичної системи на основі принципів аферентного синтезу, що забезпечує інтеграцію сенсорної інформації та формування керівних впливів у реальному масштабі часу. **Методологія вивчення** базується на використанні біоінспірованого підходу, теорії функціональних систем, методів штучних нейронних мереж, дискретної сіткової моделі, нейроподібної мережі регулярного типу, підсистеми аферентного синтезу, підсистеми прийняття рішень із механізмом конкурентного відбору типу "переможець отримує все", а також функціоналу складності траєкторій і методів дискретного моделювання. **Наукова новизна дослідження** полягає в розробленні біоінспірованої моделі нейромережевого планувальника, яка поєднує підсистему аферентного синтезу й підсистему прийняття рішень у єдиній архітектурі. **Результати.** Запропоновано підхід до подання зовнішнього середовища у вигляді дискретної сіткової моделі з подальшим відображенням у нейроподібну мережу, що відтворює топологію простору. Уперше обґрунтовано використання механізму хвильового поширення сигналів збурення в поєднанні з принципом конкурентного відбору ("переможець отримує все") для визначення оптимального напрямку руху кіберфізичної системи. Сформульовано математичну постановку задачі планування переміщення на основі функціоналу складності траєкторій і розроблено відповідну нейромережеву структуру для її реалізації. Продемонстровано, що запропонована модель забезпечує формування керівних впливів для позиційно-траєкторного регулятора й дає змогу реалізувати замкнений цикл управління від сенсорного сприйняття середовища до прийняття рішень і виконання руху. **Практична цінність роботи** полягає в можливості застосування запропонованого біоінспірованого нейромережевого планувальника переміщення кіберфізичної системи для створення інтелектуальних систем управління мобільними кіберфізичними об'єктами, що функціують у складних і динамічних умовах. **Висновки:** досягнуті результати підтверджують ефективність використання біоінспірованих нейромережевих підходів для задач планування переміщення та створюють передумови для подальшого розвитку інтелектуальних систем управління кіберфізичними системами.

**Ключові слова:** кіберфізична система; нейромережеве планування; аферентний синтез; біоінспірований підхід; планування переміщення; нейроподібна мережа; підсистема прийняття рішень; хвильове поширення сигналів; оптимальна траєкторія; адаптивне управління.

### Вступ

Сучасний розвиток кіберфізичних систем (КФС) супроводжується зростанням вимог до їх автономності, адаптивності та здатності ефективно функціонувати в умовах невизначеного й динамічного зовнішнього середовища. Особливої актуальності набувають завдання планування переміщення, що потребують оперативного оброблення сенсорної інформації, врахування наявності перешкод і забезпечення безпечного досягнення цільових станів.

Наявні підходи до планування руху, як правило, ґрунтуються на заздалегідь сформованих моделях середовища або вимагають значних обчислювальних ресурсів, що обмежує їх застосування в реальному масштабі часу. У зв'язку з цим зростає потреба в розробленні нових методів, здатних забезпечити

ефективне прийняття рішень на основі неповної та зашумленої інформації.

Перспективним напрямом розв'язання зазначеної проблеми є використання біоінспірованих підходів, зокрема концепції аферентного синтезу, що дає змогу інтегрувати різноманітні сенсорні дані та формувати цілеспрямовану поведінку системи. Поєднання таких підходів із нейромережевими моделями відкриває можливості для створення інтелектуальних планувальників, здатних працювати в умовах апріорної невизначеності середовища.

Розроблення моделі нейромережевого планування переміщення КФС на основі аферентного синтезу є актуальним науково-прикладним завданням, спрямованим на підвищення ефективності управління та забезпечення адаптивного функціонування сучасних інтелектуальних систем.

### Аналіз літературних джерел і постановка проблеми

Питання планування переміщення інтелектуальних систем, зокрема КФС, активно досліджується в сучасній науковій літературі у зв'язку з розвитком робототехніки, автономного транспорту та інтелектуальних виробничих систем [1].

У монографії [2] детально висвітлено ключові засади багаторівневої платформи кіберфізичних систем, окреслено напрями й особливості організації наукових досліджень у межах її функціонування. Розроблено основи автоматизованого синтезу комп'ютерних пристроїв для реконфігурованих апаратних платформ вузлів інтелектуальних сенсорів і актуаторів. Запропоновано підходи до побудови комп'ютерних засобів захисту інформації на основі алгоритмів у полях Галуа, а також визначено практичні аспекти імплементації положень багаторівневої платформи під час проєктування вендингових кіберфізичних систем.

У класичних підходах до планування руху використовуються алгоритми пошуку, зокрема RRT та їх модифікації, які забезпечують побудову оптимальних або квазіоптимальних траєкторій [1]. Однак такі методи, як правило, потребують повної або частково відомої моделі середовища й мають значні обчислювальні витрати, особливо в умовах високої розмірності простору або динамічних змін.

У зв'язку з цим останнім часом значного розвитку набули підходи, основані на використанні штучних нейронних мереж. Як показано в роботах, присвячених нейромережевому плануванню, такі методи дають змогу зменшити обчислювальну складність задачі та забезпечити адаптивність до змін середовища за допомогою навчання на даних [3]. Зокрема нейромережі застосовуються для апроксимації функцій вартості, побудови евристик пошуку й генерації траєкторій у реальному часі.

Окремим напрямом є використання біоінспірованих моделей, що імітують механізми функціонування біологічних нервових систем. Основи такого підходу закладені в класичних роботах, де нейрон розглядається як логічний елемент із пороговою функцією активації [4].

Наукові положення, висвітлені в праці [5], демонструють комплексний і міждисциплінарний підхід до формування сучасних біоінспірованих

систем. Особливо цінним є поєднання концепцій біомеханіки, штучного інтелекту, сенсорних технологій і адаптивного керування в єдиній інформаційно-фізичній архітектурі. Суттєвою перевагою дослідження є зосередження уваги на адаптивності та самоорганізації систем, що набуває особливого значення в умовах динамічного й невизначеного середовища функціонування. Використання біологічних принципів як основи для побудови технічних рішень дає змогу не лише підвищити ефективність управління та енергоефективність, але й забезпечити більш природний характер взаємодії між людиною та технічними системами.

Робота [6] присвячена дослідженню сучасних підходів до підвищення ефективності телекомунікаційних систем у разі впливу завад і групових помилок у каналах зв'язку. Автори розглянули особливості застосування алгебраїчних згорткових кодів перемешування й проаналізували методи їх декодування в умовах релієвського розсіювання сигналу. Автори обґрунтовують доцільність використання адаптивного підходу до декодування, який дає змогу підвищити якість передавання інформації та покращити енергетичні характеристики системи.

Крім того, увагу приділено побудові математичної моделі телекомунікаційної системи, яка зважає на специфіку виникнення групових помилок у каналі зв'язку. У роботі детально описано структуру моделі, функціональне призначення її елементів і особливості програмної реалізації. Це забезпечує можливість комплексного аналізу ефективності різних методів декодування й оцінювання їх впливу на характеристики системи зв'язку.

У праці [7] висвітлено концепцію функціональних систем і аферентного синтезу, яка дає змогу розглядати процес прийняття рішень як інтеграцію різномірної інформації про середовище, стан системи й цільові параметри. Використання цієї концепції в поєднанні з нейромережевими моделями створює передумови для побудови інтелектуальних систем керування з високим рівнем адаптивності. Сучасні дослідження демонструють ефективність поєднання нейромережових і біоінспірованих підходів. Зокрема розробляються гібридні моделі, які поєднують класичні алгоритми планування з методами машинного навчання, що допомагає досягти балансу між точністю, швидкістю та адаптивністю систем [8].

Автор роботи [9] обґрунтовує доцільність використання методів штучного інтелекту для аналізу складних логістичних і постачальницьких процесів, що сприяє підвищенню точності прогнозування та якості управлінських рішень. Цінність роботи полягає в поєднанні сучасних цифрових підходів із практичними завданнями управління постачанням, що є особливо важливим в умовах нестабільності ринкового середовища й цифрової трансформації бізнесу. Запропоновані нейромережеві підходи можуть стати основою для створення інтелектуальних систем підтримки прийняття рішень у сфері логістики, управління ресурсами й цифрового менеджменту підприємств.

У праці [10] проаналізовано сучасні методи, моделі та інформаційні технології класифікації й сегментації відеоданих із використанням нейронних мереж і апаратного прискорювача Google Coral. Дослідження охоплює моделі SSD/FPN MobileNet V1, MobileNet V2 і SSDLite MobileNet V2, для яких підібрано оптимальні параметри навчання. Крім того, активно вивчаються графові нейронні мережі та глибокі моделі навчання, що забезпечують оброблення складних просторових структур і дають змогу брати до уваги контекстні зв'язки між елементами середовища [11].

Незважаючи на значну кількість студій, залишається низка нерозв'язаних питань, пов'язаних із забезпеченням ефективного планування в умовах апріорної невизначеності середовища, обмежених обчислювальних ресурсів і необхідності функціонування в реальному масштабі часу. Це зумовлює доцільність подальших досліджень, присвячених розробленню нових моделей і методів, зокрема на основі біоінспірованих нейромережевих підходів.

#### Мета й завдання публікації

Метою дослідження є розроблення й формалізація моделі нейромережевого планування переміщення кіберфізичної системи на основі принципів аферентного синтезу, що забезпечує інтеграцію сенсорної інформації, формування оптимального напрямку руху й адаптивне функціонування в умовах невизначеного й динамічного зовнішнього середовища.

Для досягнення поставленої мети необхідно виконати такі завдання:

1) проаналізувати підходи до формування інформації про зовнішнє середовище КФС на основі даних системи технічного зору;

2) запропонувати спосіб подання зовнішнього середовища у вигляді дискретної моделі, придатної для нейромережевого оброблення;

3) сформулювати задачу планування переміщення КФС у формалізованому вигляді;

4) розробити нейромережеву структуру для інтеграції сенсорних показників і визначення напрямку руху;

5) описати принцип функціонування нейромережевого планувальника в процесі керування переміщенням кіберфізичної системи.

#### Виклад основного матеріалу

Залежно від властивостей зовнішнього середовища й типу використовуваної системи технічного зору (СТЗ) можливі різні варіанти формування сенсорної інформації [12]. У цьому контексті доцільно виокремити два основні класи СТЗ. До першого належать далекомірні системи, які формують на виході часткову інформацію про стан середовища в межах зони прямої видимості. Такі системи забезпечують визначення відстаней до перешкод, однак не дають змоги отримати повну уяву про структуру навколишнього простору. До другого класу належать СТЗ із просторовим оглядом, які формують більш повну інформацію про середовище в межах зони активності кіберфізичної системи. На основі даних, що містять просторово-часовий розподіл інтенсивності відбитого сигналу, стає можливим не лише виявлення перешкод і визначення відстаней до них, але й часткове відновлення структури простору поза межами безпосередньої видимості.

#### Мережеве подання моделі зовнішнього середовища

Мережеве подання моделі зовнішнього середовища є важливим етапом побудови інтелектуального планувальника переміщення кіберфізичної системи. На відміну від методів типу DVH і DVH-NN, орієнтованих на використання багатопроменевих далекомірних СТЗ [13], запропонований біонічно-орієнтований нейромережевий підхід передбачає застосування СТЗ із просторовим секторним оглядом у горизонтальній площині. Отримана від сенсорної підсистеми інформація в зовнішньому контурі попередньо обробляється й перетворюється у двовимірне дискретне подання із заданою точністю,





1.  $t_i = t_{i-1} + \Delta t_i$ , де  $\Delta t_i$  – інтервал часу між послідовними циклами аналізу й прийняття рішень.

2.  $X_{pi}$ ,  $X_{ii}$ ,  $X_{ci}$ ,  $X_{ni}$  – множини параметрів, що визначають стан зовнішнього середовища й положення КФС і перешкод у момент часу  $t_i$ .

3.  $L_{(X_i)}$  – функція, що визначає складність проходження кіберфізичною системою відповідної ділянки зовнішнього середовища. У найпростішому випадку ця функція може мати бінарний характер: якщо ділянка зайнята перешкодою, то  $L_{(X_i)} = \infty$ , якщо ж вона є вільною –  $L_{(X_i)} = 0$ . Загалом функція  $L_{(X_i)}$  може брати до уваги різні ступені складності переміщення, пов'язані з рельєфом, динамічністю середовища або іншими факторами.

4. Модель прохідності зовнішнього середовища може бути подана у вигляді  $W_i = W[X_{pi}, L(X_i), X_{ci}]$ , де  $W_i$  – узагальнена характеристика середовища, що відбиває його прохідність для КФС у момент часу  $t_i$ . Ця модель дає змогу абстрагуватися від другорядних властивостей середовища й зосередитися на ключовій властивості – можливості переміщення з певним рівнем складності або неможливості руху взагалі.

5.  $T_{j(W_i)}$  –  $j$ -та допустима траєкторія переміщення кіберфізичної системи, що забезпечує перехід із поточного положення  $X_{pi}$  до деякого наступного стану  $X_{ii}$  з огляду на функції складності  $L_{j(X_i)} : T_{j(W_i)} = T[X_{pi}, L_{j(X_i)}, X_{ii}]$ .

6. Для кожної траєкторії впроваджується її норма  $T_{j(W_i)}$ , яка визначає ступінь складності її проходження. З метою кількісного оцінювання цієї складності додається функціонал  $J_{(T_j)} = \int_{X_{pi}}^{X_{ii}} L_{j(X_i)} dT_{ji}$ , що відтворює інтегральну складність переміщення КФС уздовж траєкторії  $T_{ji}$ .

7. Множина всіх допустимих у момент часу  $t_i$  траєкторій позначається як  $E_{T_i}$ , до того ж кожна траєкторія  $T_{j(W_i)}$  належить цій множині:  $T_{j(W_i)} \in E_{T_i}$ . Відповідно, значення функціонала складності для всіх допустимих траєкторій формують множину  $E_{J_i} : J_{(T_j)} \in E_{J_i}$ .

8. Оскільки кожному значенню функціонала  $J_{(T_j)}$  відповідає певна траєкторія  $T_{ji}$ , відображення множини траєкторій  $E_{T_i}$  у множину їх оцінок складності  $E_{J_i}$  може бути подано у вигляді функціонального відтворення:  $J : E_{T_i} \rightarrow E_{J_i}$ .

З огляду на подані визначення задача планування переміщення КФС полягає у виборі оптимальної траєкторії з множини допустимих траєкторій  $T_{j(W_i)}$ , сформованих у момент часу  $t_i$ . Оптимальною вважається така траєкторія  $T_{(W_i)}^{opt}$ , для якої значення функціонала складності є мінімальним:  $J(T_{ji}^{opt}) = \min_{(T_j)} J_{(T_j)}$ .

Розв'язання цієї задачі можливе способом визначення напрямку найшвидшого зменшення функціонала  $J_{(T_i)}$  у поточній точці простору станів  $X_{pi}$ . Для цього використовується вектор антиградієнта функціонала, який задає напрямок оптимального руху КФС:

$$V_i = -\rho \nabla J_{(T_i)}, \quad (1)$$

де  $V_i$  – вектор швидкості кіберфізичної системи в точці  $X_{pi}$ ;  $\nabla J(T_i)$  – градієнт функціонала складності;  $\rho$  – коефіцієнт, що визначає модуль швидкості (у найпростішому випадку – одинична швидкість).

Важливо зауважити, що вектор градієнта  $\nabla J(T_i)$  визначає напрямок найбільшого зростання функціонала складності, тоді як вектор антиградієнта  $\nabla J(T_i)$  відповідає напрямку його найшвидшого зменшення. Саме цей напрямок є доцільним для формування керівного впливу, оскільки забезпечує мінімізацію інтегральної складності переміщення й, відповідно, вибір найбільш безпечної та ефективної траєкторії руху кіберфізичної системи.

На відміну від попередніх підходів, у запропонованому методі базовим обчислювальним елементом нейромережевої підсистеми застосовується формальний нейрон, що оперує бінарним поданням інформації [14]. Цей підхід ґрунтується на біонічній інтерпретації функціонування нервових клітин, які розглядаються як логічні елементи, що реалізують принцип "все або нічого", тобто перебувають або в активному стані, генеруючи вихідний сигнал, або в неактивному стані [14, 15].



Передбачається, що окремі елементи нейроподібної структури взаємодіють за аксондентритним принципом, а передача інформації моделюється у вигляді дискретних сигналів – імпульсів або потенціалів. У такій постановці вихідна функція нейрона  $y(t)$  розглядається як булева функція від множини вхідних змінних  $x_j(t)$ , а також синаптичних ваг  $\gamma_j$  порогового значення  $\theta$ , що приймають дискретні (цілочисельні) значення.

Формальний нейрон у межах запропонованої моделі інтерпретується як пороговий логічний елемент із такими властивостями:

а) нейрон має  $N$  синаптичних входів, кожен з яких може бути збудовальним ( $\gamma_j > 0$ ) або гальмівним ( $\gamma_j \leq 0$ ), де  $j = 1, 2, \dots, N$ ;

б) функціонування нейрона розглядається в дискретні моменти часу  $t_i = t_0 + i \cdot \Delta t$ , де  $\Delta t = t_i - t_{i-1}$  є кроком дискретизації;

в) незалежно від кількості активних вхідних сигналів у момент часу  $t_i$  вихідна реакція нейрона формується із затримкою на один крок дискретного часу, тобто в момент  $t_{i+1}$ ;

г) вхідні сигнали  $x_j(t_i)$  і вихід нейрона  $y(t_{i+1})$  можуть перебувати лише у двох станах – активному ( $x_{j(t_i)} = 1; y_{(t_{i+1})} = 1$ ) або неактивному ( $x_{j(t_i)} = 0; y_{(t_{i+1})} = 0$ );

д) нейрон визначається порогом збудження  $\theta$ , і якщо зважена сума вхідних сигналів у момент часу  $t_i$  задовольняє умову перевищення порога, формується вихідний сигнал  $y_{(t_{i+1})} = 1$ , проте в іншому разі нейрон залишається неактивним  $y_{(t_{i+1})} = 0$ .

Зважаючи на наведене вище, логіка функціонування формального нейрона описується пороговим правилом активації, яке визначає перехід системи в новий стан на основі інтеграції вхідних сигналів:

$$y_{(t_{i+1})} = \text{sign}(K \cdot V_{(t_i)}), \quad (2)$$

де  $V_{(t_i)} = \sum_{j=1}^N (\gamma_j \cdot X_{j(t_i)}) - \theta$ ;  $K$  – параметр, що визначає збудливість нейрона, а сама функція  $y_{(t_{i+1})}$  є знаковою

функцією, що приймає значення з множини  $[0, 1]$ .

У цьому разі, якщо прийняти  $\gamma_j \in 0, 1$ , блок-схема формального нейрона матиме такий вигляд (рис. 4):

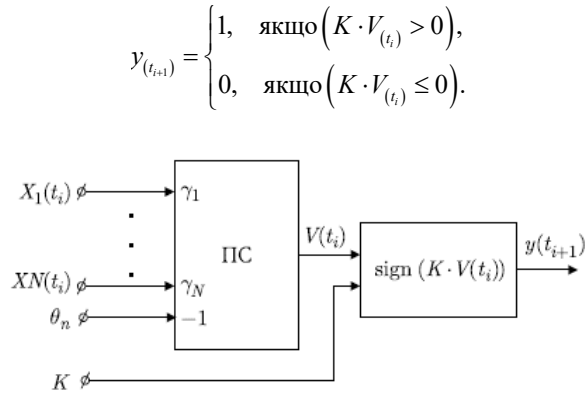


Рис. 4. Структурно-функціональна схема формального нейрона порогового типу

Структура формального нейрона є спрощеною та містить два основні функціональні блоки: блок просторової сумації (ПС) і блок формування вихідної функції  $\text{sign } K V_{(t_i)}$ , який може бути реалізований за імпульсним або статичним принципом. Зауважимо, що блок ПС забезпечує інтеграцію вхідних сигналів з огляду на їх вагові коефіцієнти, тоді як вихідний блок визначає характер реакції нейрона на перевищення порогового значення.

У разі імпульсної реалізації, виконання умови збудження ( $K \cdot V_{(t_i)} > 0$ ), на виході нейрона формується одиничний імпульсний сигнал фіксованої тривалості  $\tau$ . Після завершення імпульсу нейрон переходить у неактивний стан і залишається в ньому до наступного перевищення порогового рівня. Унаслідок статичної реалізації, виконання тієї самої умови, нейрон переходить у активний стан і зберігає його доти, доки умова збудження залишається виконаною.

#### Розроблення нейромережевої підсистеми аферентного синтезу (САС)

У найпростішому випадку зовнішнє середовище, що сприймається кіберфізичною системою, може бути подано як площина  $S$ , яка дискретизується на елементарні ділянки  $\Delta S_j$ . Подальше перетворення цих даних здійснюється способом побудови ортогональної сіткової моделі у вигляді плоского графа  $G(Q, F)$ , де  $Q$  – множина вершин  $q_j$ , що відповідають дискретним ділянкам середовища,

а  $F$  – множина ребер, які визначають топологічні зв'язки між ними.

Кожній вершині графа  $q_j$  відповідає нейроподібний процесорний елемент (ПЕ), а інформаційні зв'язки між елементами утворюються згідно із структурою графа  $G(Q, F)$ . Як наслідок, формується нейроподібна мережа, архітектура якої відтворює топологію середовища й забезпечує можливість поширення сигналів у межах доступних для переміщення ділянок (рис. 5).

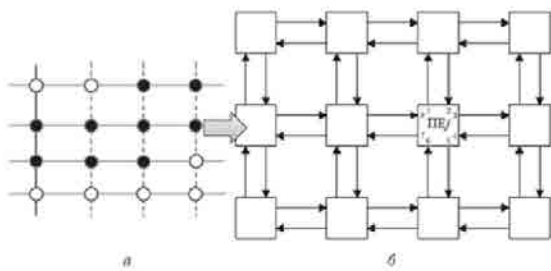


Рис. 5. Формування топологічної структури нейроподібної мережі регулярного типу:

- а – графова модель середовища  $G(Q, F)$ ;  
 б – відповідна структура мережі нейроподібних процесорних елементів

Функціонування такої мережі визначається правилами оброблення сигналів окремими процесорними елементами. Зокрема, якщо  $ПЕ_j$  відповідає вільній ділянці середовища, він забезпечує передачу сигналів збурення між сусідніми елементами, реалізуючи таким чином механізм поширення інформації в мережі. Якщо елемент відповідає ділянці, зайнятій перешкодою, він блокує передачу сигналів, що надходять від суміжних елементів, формуючи цим бар'єри для поширення збурення.

Описаний підхід дає змогу реалізувати просторово-розподілене оброблення інформації та формування шляхів переміщення КФС на основі топології середовища. У цьому разі нейромережа є обчислювальною структурою підсистеми аферентного синтезу, яка інтегрує сенсорні показники й формує основу для прийняття рішень щодо подальшого руху.

Якщо  $ПЕ_j$  відповідає ділянці середовища, в якій розташована цільова точка, він виконує функцію джерела збурення, генеруючи сигнали, що поширюються мережею через доступні (незаблоковані) елементи. У такий спосіб формується хвильовий

процес поширення збурення, який відтворює топологію прохідних шляхів у середовищі та слугує основою для визначення напрямку руху кіберфізичної системи.

Особливу роль у мережі відіграє  $ПЕ_0$ , який є виділеним, оскільки в кожному момент дискретного часу  $t_i$  відповідає поточному положенню КФС у зовнішньому середовищі. У зв'язку з цим  $ПЕ_0$  виконує функції підсистеми прийняття рішень, що передбачає визначення оптимального напрямку руху на основі аналізу сигналів збурення, які надходять від сусідніх елементів мережі. Крім того, цей елемент забезпечує фіксацію обраного напрямку протягом інтервалу часу, необхідного для реалізації елементарного переміщення КФС відповідною виконавчою підсистемою.

За умови, що вхідні та вихідні сигнали нейроподібних процесорних елементів мають бінарний характер, функціонування  $j$ -го елемента мережі може бути описано булевою функцією  $y_{j(t_i + \Delta t)}$ , що реалізується  $j$ -м елементом.

У більш конкретизованому вигляді ця функція може бути подана пороговим правилом активації:

$$y_{j(t_i + \Delta t)} = \bar{d} \wedge \left[ \bigvee_{\sigma \in l} V y_{\sigma(t_i)} \right] \vee P_{j(t_i)} \wedge \bar{P}_{j(t_i)}, \quad (3)$$

де  $y_{\sigma(t_i)}$  – значення функцій, що відтворюються в момент часу  $t_i$  нейроподібними процесорними елементами  $ПЕ_\sigma$ , які безпосередньо пов'язані з елементом  $ПЕ_j$ , а  $l$  – множина індексів таких суміжних елементів;  $P_{j(t_i)}$  – двійковий сигнал, що надходить із підсистеми формування моделі зовнішнього середовища (ПФМЗС) і визначає стан відповідної ділянки  $\Delta S_j$ : якщо  $P_{j(t_i)} = 0$ , то ділянка є вільною, якщо  $P_{j(t_i)} = 1$  – зайнятою перешкодою;  $P_{j(t_i)}$  – двійковий сигнал, що формується в керівній підсистемі КФС і відтворює наявність цільової точки в ділянці  $\Delta S_j$ ,  $d$  – двійковий сигнал, що надходить на всі елементи нейроподібної мережі з боку керівної підсистеми КФС.

Перед початком кожного нового циклу обчислення функцій  $y_{j(t_i + \Delta t)}$  сигнал  $d$  короткочасно приймає значення заборони ( $d = 1$ ), що призводить

до деактивації (гасіння) раніше збурених елементів мережі. У процесі подальших обчислень сигнал  $d$  встановлюється в стан дозволу ( $d = 0$ ), забезпечуючи можливість формування нової хвилі збурення відповідно до актуального стану середовища. Такий механізм відповідає циклічній природі функціонування САС і забезпечує узгоджене оновлення інформації.

### Розроблення підсистеми прийняття рішень

Відповідно до архітектури нейромережевого планувальника, для визначення напрямку руху КФС використовується спеціалізований  $ПЕ_0$ , який виконує функції підсистеми прийняття рішень. Цей елемент має  $n$  входів і  $n$  виходів та забезпечує вибір оптимального напрямку переміщення на основі аналізу сигналів збурення, що надходять від суміжних елементів мережі.

Функціонування  $ПЕ_0$  ґрунтується на логічному правилі відбору найбільш пріоритетного сигналу, що відповідає напрямку з найменшою складністю проходження або найвищим рівнем збурення. Як наслідок, формується керівний сигнал, що визначає напрямок подальшого переміщення КФС у поточному циклі:

$$Z_{j(t_{i+1})} = y_{j(t_i)} \begin{pmatrix} n \\ \wedge \\ \varphi = 1 Z_{\varphi(t_i)} \\ \varphi \neq j \\ \varphi \end{pmatrix}, \quad (4)$$

де  $Z_{j(t_{i+1})}$  ( $j=1,2,\dots,n$ ) – сигнал на  $j$ -му виході процесорного елемента  $ПЕ_0$ ;  $y_{i(t_i)}$  – сигнал на  $j$ -му вході  $ПЕ_0$ , що надходить від суміжних нейроподібних елементів мережі.

Необхідно зазначити, що доцільно використовувати формальні нейрони логічного (порогового) типу (див. рис. 4). Такі нейрони, окрім збурювальних входів і порогового елемента, мають гальмівні входи, призначені для приймання сигналів  $П_j$  (стан ділянки середовища) та глобального сигналу керування  $d$ . Наявність гальмівних входів забезпечує можливість блокування активації нейрона в разі, якщо відповідна ділянка середовища є непрохідною або якщо виконується процедура скидання мережі перед новим циклом обчислень.

Умовне графічне зображення такого формального нейрона запропоновано на рис. 6, а, де подано поєднання збурювальних і гальмівних впливів у процесі формування вихідного сигналу. На рис. 6, б продемонстровано схему комутації входів і виходів формального нейрона за умови його використання як процесорного елемента  $ПЕ_j$  у складі нейроподібної мережі.

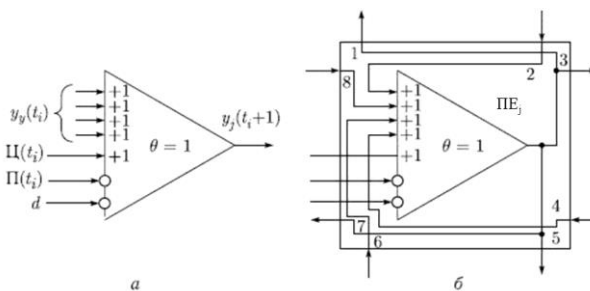


Рис. 6. Приклад структури європодібного процесорного елемента на основі формального нейрона

Поєднання процесорних елементів  $ПЕ_j$  відповідно до топології, наведеної на рис. 6, дає змогу сформувати нейроподібну мережу регулярного типу, яка моделює підсистему аферентного синтезу кіберфізичної системи. Така мережа забезпечує інтеграцію сенсорної інформації, формування хвиль збурення та визначення напрямків допустимого переміщення КФС у просторі.

Важливо зауважити, що для спрощення графічного подання на рис. 6 не зображено входи сигналів, що відповідають цільовим точкам, перешкодам і глобальному сигналу керування  $d$ , однак вони є невід'ємним складником функціонування мережі та беруть участь у формуванні керівних впливів.

Зазначимо, що функції спеціалізованого процесорного елемента  $ПЕ_0$  можуть бути реалізовані не лише як окремий вузол, а й у вигляді ансамблю з  $n$  формальних нейронів, об'єднаних у додатковий шар, структуру якого подано на рис. 7. Це дає змогу підвищити гнучкість реалізації підсистеми прийняття рішень (ППР) та узгодити її з нейромережевою природою всієї моделі.

З огляду на наведену структуру кожен формальний нейрон  $ФН_j$  ( $j=1, 2, \dots, n$ ) пов'язаний з іншими нейронами шару латеральними (бічними) інгібуючими зв'язками. Наявність таких зв'язків забезпечує реалізацію механізму конкуренції, що дає змогу шару виконувати функцію вибору одного активного стану з множини можливих, тобто

реалізовувати процедуру прийняття рішення. Зокрема мережа забезпечує виділення та фіксацію єдиного збудженого виходу, що відповідає обраному напрямку руху кіберфізичної системи.

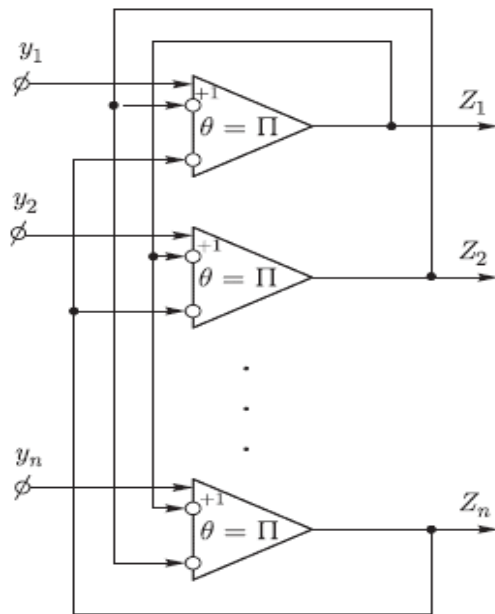


Рис. 7. Формальне подання структури євромережевого шару на основі формальних нейронів, що забезпечує прийняття рішення "переможець забирає все"

У початковому стані, за умови подання на мережу сигналу заборони  $d=1$ , усі входи  $y_{j(t_i)}$  і виходи  $Z_{j(t_i)}$  елементів ансамблю встановлюються в нульовий стан, що відповідає процедурі ініціалізації. У цьому разі інверсні сигнали з виходів надходять на гальмівні входи нейронів, забезпечуючи їх повну деактивацію. Після переходу сигналу керування в стан дозволу ( $d=0$ ) мережа стає чутливою до вхідних сигналів, що надходять від підсистеми аферентного синтезу.

У такому режимі перший активний вхідний сигнал  $y_{j(t_i)}=1$  викликає збудження відповідного нейрона  $\Phi N_j$ , що призводить до формування вихідного сигналу  $Z_j$ . Цей сигнал зі свого боку за допомогою латеральних інгібуючих зв'язків блокує активацію інших нейронів шару, запобігаючи появі альтернативних рішень у межах поточного циклу. У такий спосіб реалізується принцип "переможець отримує все", що забезпечує однозначний вибір напрямку руху КФС до моменту наступної ініціалізації системи за сигналом  $d$ .

### Розроблення нейромережевого планувальника

Об'єднання нейроподібних структур САС і ППР дає змогу сформувати біоінспірований нейромережевий планувальник переміщення кіберфізичної системи. Такий планувальник функціонує в дискретні моменти часу  $t_i$  і забезпечує визначення оптимального напрямку руху КФС в умовах апіорної невизначеності зовнішнього середовища.

Інформація про стан середовища надходить із СТЗ у вигляді бінарних сигналів  $C_j \in i P_j \in \{0,1\}$ , що відбивають, відповідно, наявність цілі та перешкод у дискретних ділянках простору. Ці сигнали подаються на вхід підсистеми аферентного синтезу, де кожен ПЕ виконує відповідну локальну функцію оброблення, забезпечуючи інтеграцію сенсорної інформації та формування просторового розподілу збудження.

Як наслідок, формується узгоджена картина доступних напрямків переміщення, яка передається до ППР для визначення оптимального керівного впливу. Отже, запропонований нейромережевий планувальник реалізує повний цикл оброблення інформації – від сенсорного сприйняття до формування сигналу керування – за принципами аферентного синтезу.

Процесорні елементи САС, реалізовані на основі формальних нейронів, з'єднані між собою відповідно до топології, поданої на рис. 5. У цьому разі виходи процесорних елементів крайнього рівня САС пов'язані не лише з іншими елементами мережі, але й під'єднані до збуджувальних входів формальних нейронів ППР. Така організація забезпечує передачу узагальненої інформації про стан середовища до блоку прийняття рішень.

Вихідні сигнали ППР надходять до позиційно-траєкторного регулятора кіберфізичної системи, формуючи керівні впливи, що забезпечують вчасне коригування параметрів руху з огляду на зміни в зовнішньому середовищі. Отже, реалізується замкнений контур керування, у якому нейромережева підсистема виконує функцію інтелектуального планування.

Після завершення чергового кроку переміщення КФС її виконавча (ефекторна) підсистема формує короткочасний сигнал керування  $d=1$ , який призводить до скидання станів усіх формальних нейронів у складі процесорних елементів підсистеми аферентного синтезу. Це забезпечує очищення попереднього стану мережі та підготовку до



оброблення нової порції сенсорної інформації, що відповідає актуальному положенню КФС у середовищі.

У процесорні елементи САС, що перебувають у початково неактивному стані, надходить сенсорна інформація від СТЗ про розташування цілі та перешкод у межах активної зони сприйняття. Для попереднього оброблення цих показників і формування дискретної мережевої моделі зовнішнього середовища на вході біоінспірованого нейромережевого планувальника передбачена спеціалізована ПФМЗС, яка перетворює та завантажує дані в процесорні елементи САС.

Процесорні елементи, що відповідають ділянкам середовища  $\Delta S_j$ , зайнятим перешкодами, переводяться в заблокований стан і не беруть участі в поширенні сигналів. Водночас елемент, що відповідає цільовій ділянці, активується сигналом  $C=1$  і є джерелом збурення для інших незаблокованих елементів підсистеми аферентного синтезу. Сформовані одиничні сигнали збурення поширюються мережею лише через доступні (прохідні) процесорні елементи, формуючи хвильовий процес, який одночасно відтворює всі можливі шляхи, що ведуть від цілі до ППР.

Оскільки сигнал поширюється з однаковою швидкістю у всіх напрямках, першим до ППР досягає сигнал, що відповідає найбільш короткому або найменш складному шляху. Підсистема прийняття рішень фіксує напрямок останньої ділянки цього шляху, який визначається як кутове відхилення від центральної осі секторної зони огляду, та формує відповідний сигнал керування. Отриманий сигнал передається до позиційно-траєкторного регулятора, який забезпечує переміщення КФС в обраному напрямку на один крок.

Після виконання елементарного переміщення положення КФС щодо цілі та перешкод змінюється, що приводить до оновлення сенсорних показників. Нова інформація про стан зовнішнього середовища знову надходить від СТЗ, проходить етап попереднього оброблення в ПФМЗС і завантажуються в підсистему аферентного синтезу. Подальше оброблення здійснюється після подання сигналу дозволу  $d$ , що ініціює новий цикл функціонування нейромережевого планувальника.

Для дослідження функціонування запропонованої моделі нейромережевого планування переміщення КФС виконано моделювання в середовищі Matlab.

Структуру системи планування переміщення КФС подано у вигляді функціональної схеми (рис. 8).

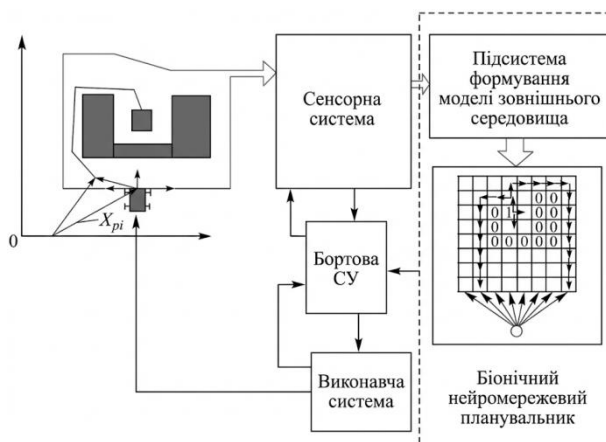


Рис. 8. Застосування біоінспірованого нейромережевого планувальника для керування переміщенням кіберфізичної системи

Під час моделювання запропонованого підходу кіберфізична система інтерпретується як бортова система управління безпілотного літального апарата, що виконує завдання автономного переміщення та уникнення перешкод у динамічному зовнішньому середовищі.

Програмна реалізація процедури визначення оптимального напрямку переміщення та формування локальної траєкторії ґрунтується на використанні хвильового алгоритму Лі, адаптованого до моделювання роботи біоінспірованого нейромережевого планувальника. Обраний підхід допомагає відтворити механізм поширення сигналів збурення в підсистемі аферентного синтезу й підсистемі прийняття рішень щодо вибору безпечного напрямку руху КФС.

Процедура моделювання роботи біоінспірованого нейромережевого планувальника в середовищі Matlab реалізується у вигляді послідовності кількох кроків.

1. У масиві цільових станів  $M_{ci}$  визначається адреса комірки, що відповідає поточному положенню цілі. Із цієї комірки ініціюється поширення сигналу збурення в суміжні комірки дискретного простору за чотирма ортогональними напрямками: вгору, вниз, ліворуч і праворуч.

2. Виконується логічна операція кон'юнкції масивів  $M_{ci}$  та  $M_{pi}$ , що дає змогу зважати на просторові обмеження середовища й вилучати ділянки, зайняті перешкодами.

3. Комірки масиву  $M_{ci}$ , що залишилися активними після операції кон'юнкції, повторно



поширюють сигнали збурення на сусідні елементи дискретного простору.

4. На черговому кроці знову здійснюється операція кон'юнкції масивів цільових станів і перешкод з метою актуалізації множини доступних напрямків поширення.

5. Ітераційний процес продовжується до моменту досягнення сигналом збурення комірки  $Y_0$  масиву  $M_{\Pi}$ , що відповідає поточному положенню КФС  $X_{P_i}$  у модельованому середовищі.

6. Положення активного розряду в комірці  $Y_0$  використовується для визначення оптимального напрямку локального переміщення та формування вектора руху  $u_i^{опт}$ .

7. Після виконання елементарного переміщення КФС на величину  $\Delta X_{P_i}$  у напрямку  $u_i^{опт}$  оновлюються параметри середовища й масивів станів  $(M_{\Pi(i+1)}, M_{\Pi(i+1)})$ , після чого запускається наступний цикл моделювання.

Для демонстрації функціонування алгоритму змодельовано переміщення КФС у дискретному середовищі (рис. 9).

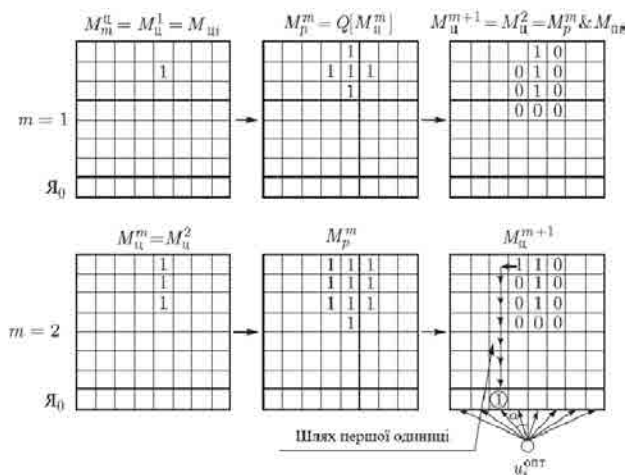


Рис. 9. Робота алгоритму моделювання переміщення кіберфізичної системи в дискретному середовищі

У масиві  $M_{\Pi}^{m+1}$  формується траєкторія поширення сигналу збурення, що першою досягає комірки  $Y_0$ , яка відповідає поточному положенню КФС у модельованому середовищі. Значення кута  $\alpha$  визначає напрямок оптимального локального переміщення та використовується для формування одиничного вектора руху  $u_i^{опт}$ .

Блок-схему реалізації розглянутого алгоритму зображено на рис. 10.

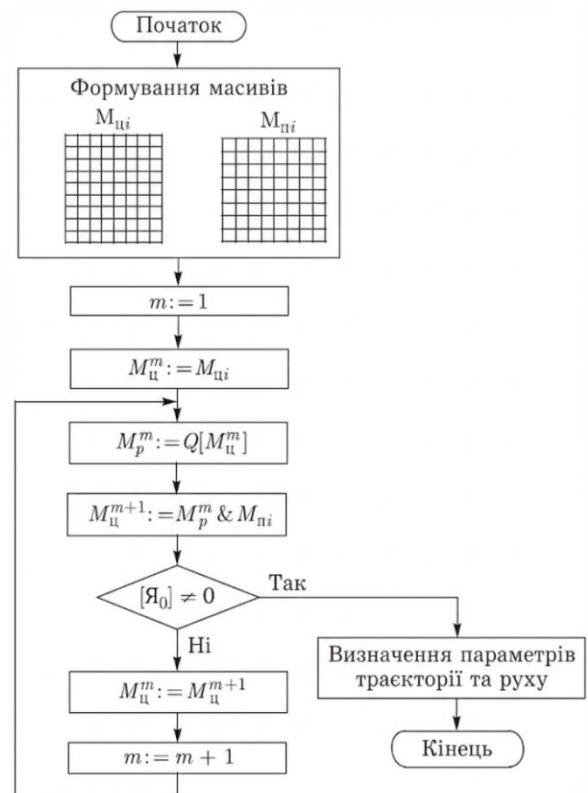


Рис. 10. Блок-схема алгоритму моделювання переміщення кіберфізичної системи в дискретному середовищі

Запропонований підхід уможливлює планування переміщення КФС у середовищі зі статичними й динамічними перешкодами, а також допускає зміну положення цільового об'єкта в процесі руху.

Для оцінювання працездатності та ефективності запропонованого підходу виконано серію модельних експериментів у середовищі Matlab. Дослідження проведено на шести тестових сценах (сцена 1, 2, 3, ..., 6), що розрізняються конфігурацією перешкод, умовами переміщення та взаємним розташуванням КФС і цілі.

Для кожної сцени задавалися координати початкового положення КФС  $(p_0^0, p_0^2)$ , координати цільової точки переміщення  $(p_k^1, p_k^2)$ . У базовому сценарії початкове положення системи відповідало точці  $(0,0; 0,0)$ , а цільова точка – координатам  $(20,0; 20,0)$ .

Для шостої тестової сцени було розширено множину початкових умов експерименту з метою оцінювання адаптивності алгоритму до зміни конфігурації середовища й варіації положення

стартових і цільових станів. Значення початкових і кінцевих координат формувалися з наборів альтернативних позицій, що дало змогу перевірити стійкість функціонування планувальника за різних умов переміщення.

У моделі встановлено обмеження на курсову лінійну швидкість КФС  $V_{\max} = 1,0$  м/с, що відповідає режиму руху за відсутності перешкод у зоні дії системи технічного зору. У разі зменшення відстані

```
if (IsSafetySituation(lidar_arr, 5, 45))
    if dzxy < deltaR
        % швидкість зменшити до 1/2 від відстані, що залишилася
        V = dzxy*(Vmax/deltaR)/2;
    else
        V = Vmax/2;
    end
else
    if dzxy < deltaR
        % швидкість зменшити до 1/2 від відстані, що залишилася
        V = dzxy*(Vmax/deltaR)/2;
    else
        % швидкість зменшити до 1/5 від максимуму
        V = Vmax/5;
    end
end
```

У моделі також встановлено обмеження на кутову швидкість приводів мобільної кіберфізичної системи. Граничне значення кутової швидкості прийнято рівним  $\omega^{\max} = 25,0$  град/с, що відповідає допустимому режиму маневрування платформи. У процесі моделювання розглядалася двоколісна мобільна конфігурація КФС. Радіус кожного колеса приймався рівним 0,2 м, половина довжини шасі – 0,5 м, а центр мас системи вважався розташованим у центральній частині платформи. За впроваджених обмежень на кутові швидкості приводів максимальний кут повороту системи протягом однієї секунди не перевищував  $\approx 10^\circ$ .

Параметри системи технічного зору задавалися відповідно до характеристик модельного лідара. Дальність дії сенсора прийнято рівною 5 м, а параметр дискретизації  $\delta = 0,02$  м. Якщо відстань до об'єкта перевищувала межі зони сприйняття, відповідним елементам сенсорного вектора надавалося значення 5,02 м. Додатково параметри лідара задавалися такими значеннями: кут огляду  $\Omega = 91^\circ$ , кутова роздільна здатність  $\lambda = 1^\circ$  і кількість променів  $N = 91^\circ$ . Для коректного врахування

до цілі швидкість руху коригувалася плавно, залежно від поточного положення КФС. За умови появи перешкод у зоні сприйняття швидкість додатково знижувалася з метою забезпечення безпечного маневрування. За результатами попереднього моделювання для стабільного проходження всіх тестових сцен значення швидкості було додатково скориговано способом її зменшення удвічі.

просторового впливу перешкод у модель додано параметр їх кутового розширення  $\eta = 3^\circ$ . Кроки дискретизації простору за координатами  $\Delta p_1$  та  $\Delta p_2$  прийнято рівними 0,02 м. Значення похибки відпрацювання регулятором траєкторії  $\xi_r$ , що визначає точність переходу КФС від поточного стану до прогнозованого положення, також встановлено на рівні 0,02 м.

Для програмної реалізації моделі в середовищі Matlab застосовано об'єктно-орієнтований підхід, що забезпечує модульне подання компонентів кіберфізичної системи, сенсорного контуру та нейромережевого планувальника згідно з алгоритмом, поданим на рис. 10.

Прогнозовані координати  $(p_1^{i+1}, p_2^{i+1})$  розраховувалися на основі класичних тригонометричних співвідношень із використанням значень синуса й косинуса сформованого нейромережевою моделлю курсового кута  $\varphi_{(i+1)}$ . На підставі отриманих координат визначалися параметри локальної траєкторії переміщення КФС, які, відповідно до рівнянь кінематики, застосовувалися для формування

керівних впливів і розрахунку швидкостей приводів мобільної платформи.

На рис. 11 подано результати візуалізації процесу моделювання в Matlab, що відтворюють різні етапи функціонування запропонованого підходу: загальний вигляд тестової сцени й траєкторію переміщення КФС (а); вміст активної зони системи технічного зору (б); модель середовища, сформовану за сенсорними даними (в); результати роботи біоінспірованого нейромережевого планувальника із визначенням рекомендованим напрямком переміщення КФС (г).

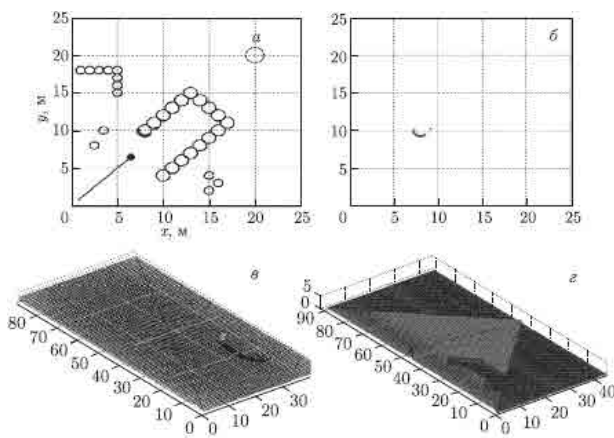


Рис. 11. Результати роботи алгоритму:

- а – загальний вигляд тестової сцени й траєкторія переміщення кіберфізичної системи;
- б – вміст активної зони системи технічного зору;
- в – модель середовища, сформована за сенсорними показниками;
- г – вміст нейромережі

Моделювання показало здатність запропонованої моделі оперативно формувати коригувальні керівні рішення відповідно до поточного стану середовища, забезпечуючи адаптивне перепланування руху й безпечне проходження тестових сцен.

Результати моделювання, подані на рис. 12, підтвердили ефективність застосування запропонованого підходу в плануванні переміщення КФС від початкового до цільового положення в межах тестових сцен 1–5 за однакових початкових умов.

Отримані результати свідчать про можливість оперативного формування маршруту, вчасного обходу перешкод і стабільного досягнення цільової точки без повного перепланування траєкторії. Результати проходження ускладненої тестової сцени за різних початкових умов подано на рис. 13.

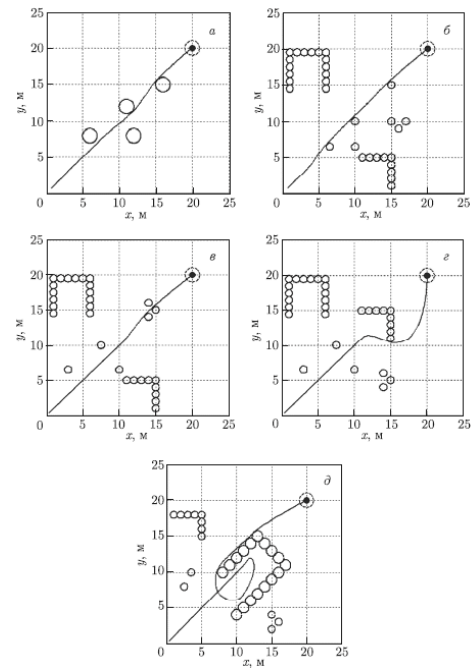


Рис. 12. Результати проходження сцен:

- а – приклад проходження сцени 1;
- б – приклад проходження сцени 2;
- в – приклад проходження сцени 3;
- г – приклад проходження сцени 4;
- д – приклад проходження сцени 5

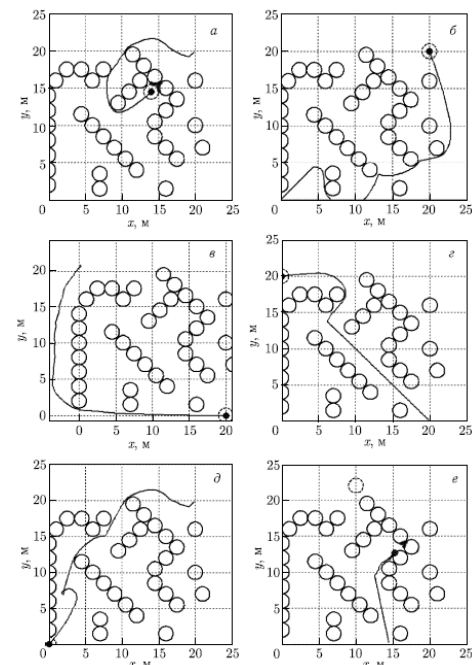


Рис. 13. Результати проходження методом сцени № 6:

- а – шлях від (20,0; 20,0) до (14,0; 14,5);
- б – шлях від (0,0; 0,0) до (20,0; 20,0);
- в – шлях від (0,0; 21,0) до (20,0; 0,0);
- г – шлях від (20,0; 0,0) до (0,0; 20,0);
- д – шлях від (20,0; 20,0) до (0,0; 0,0);
- е – шлях від (14,5; 0,0) до (10,0; 22,0)

Проведені експерименти в середовищі Matlab продемонстрували стійкість функціонування запропонованої моделі до зміни стартових і цільових параметрів, а також її здатність забезпечувати

коригування траєкторії в умовах підвищеної складності сцен і наявності перешкод. Результати серії експериментів, отримані в середовищі Matlab, зведено в табл. 1.

**Таблиця 1.** Результати роботи нейромережевого гібридного методу

| Показник                          | Номер сцени |        |        |        |        |        |        |
|-----------------------------------|-------------|--------|--------|--------|--------|--------|--------|
|                                   | порожня     | 1      | 2      | 3      | 4      | 5      | 6      |
| Безпека, м                        | –           | 0,09   | 0,10   | 0,10   | 0,04   | 0,05   | 0,11   |
| Довжина траєкторії, м             | 28,27       | 28,35  | 28,33  | 28,33  | 32,16  | 46,47  | 44,81  |
| Сумарний час регулятора $T_p$ , с | 12,67       | 23,84  | 22,40  | 21,46  | 25,90  | 53,41  | 74,88  |
| Сумарний час пошуку $T_s$ , с     | 12,77       | 23,94  | 22,50  | 21,56  | 26,00  | 53,51  | 74,98  |
| Кількість ітерацій                | 83,00       | 147,00 | 139,00 | 134,00 | 139,00 | 225,00 | 247,00 |
| Успіх місії                       | 1,00        | 1,00   | 1,00   | 1,00   | 1,00   | 1,00   | 1,00   |

Для забезпечення порівняності результатів у всіх тестових сценах спостерігалися однакові показники ефективності, зокрема значення безпеки переміщення, довжина сформованої траєкторії, сумарний час роботи регулятора й пошуку, кількість ітерацій і показник успішності виконання місії.

Аналіз досягнутих результатів демонструє, що запропонований біоінспірований нейромережевий підхід забезпечує успішне досягнення цільової точки у всіх досліджуваних сценаріях ( $Success = 1,0$ ), зокрема сцени з підвищеною складністю конфігурації середовища. У цьому разі зі зростанням складності сцени закономірно збільшуються довжина траєкторії, кількість ітерацій планування й сумарний час функціонування системи.

Для проходження лабіринтоподібних сцен із високою щільністю перешкод параметри експерименту було додатково скориговано: швидкість переміщення КФС зменшувалася, а кутове розширення перешкод збільшувалося. Незважаючи на це, у складних конфігураціях середовища залишалася можливість виникнення локальних мінімумів і повторного потрапляння системи в схожі просторові ситуації, що може призводити до циклічної поведінки траєкторії.

Досягнуті результати свідчать, що для підвищення ефективності функціонування запропонованої моделі в складних лабіринтоподібних середовищах доцільним є залучення додаткового інтелектуального рівня, призначеного для виявлення ситуацій локального мінімуму й формування альтернативної стратегії переміщення.

До переваг запропонованого підходу належать відносна простота реалізації, невисокі обчислювальні й апаратні витрати, можливість паралельного

оброблення інформації та потенціал інтеграції з іншими методами планування руху.

Крім того, результати моделювання дали змогу виявити низку обмежень запропонованої моделі. Ефективність планування залежить від точності визначення координат КФС і цільового положення, параметрів системи технічного зору, а також характеристик просторово-часової дискретизації середовища. Крім того, обмеження дальності прогнозування можуть призводити до збільшення ступінчастості траєкторії на значних відстанях, а показники безпеки руху залишаються чутливими до параметрів моделі перешкод.

Результати моделювання в середовищі Matlab підтвердили працездатність і ефективність запропонованої моделі біоінспірованого нейромережевого планування переміщення КФС в умовах динамічного й невизначеного зовнішнього середовища. Запропонований підхід забезпечує формування допустимої траєкторії руху, адаптивне врахування конфігурації середовища й успішне досягнення цільового положення в усіх досліджуваних тестових сценаріях.

Досягнуті результати демонструють здатність моделі оперативно реагувати на зміну просторової ситуації, коригувати напрямок переміщення та забезпечувати безпечне проходження сцен із різним рівнем складності. Використання механізму локального нейромережевого оброблення інформації дає змогу реалізувати планування з помірними обчислювальними витратами й забезпечує можливість функціонування системи в реальному масштабі часу.

Водночас результати експериментів продемонстрували наявність окремих обмежень моделі, зокрема чутливість до параметрів сенсорної

системи, характеристик дискретизації середовища й можливість виникнення локальних мінімумів у складних лабіринтоподібних сценах. Це визначає доцільність подальшого розвитку підходу шляхом інтеграції додаткових механізмів інтелектуального аналізу ситуацій та вдосконалення процедур адаптивного вибору траєкторії.

### Висновки

У межах дослідження розроблено модель нейромережевого планування переміщення кіберфізичної системи, що базується на використанні біоінспірованого підходу, зокрема концепції аферентного синтезу. Запропонований підхід дає змогу інтегрувати сенсорну інформацію про зовнішнє середовище, стан системи й цільові параметри в єдину структуру оброблення даних, що забезпечує формування керівних впливів у реальному масштабі часу.

Побудовано дискретну модель подання зовнішнього середовища й здійснено її відображення в нейроподібну мережу, структура якої відтворює топологічні особливості простору. Розроблена підсистема аферентного синтезу забезпечує поширення сигналів збурення в мережі, що дає змогу одночасно брати до уваги множину можливих траєкторій переміщення. Продемонстровано, що використання механізму хвильового поширення сигналів у поєднанні з підсистемою прийняття рішень на основі принципу "переможець отримує все" забезпечує вибір оптимального напрямку руху кіберфізичної системи.

Сформовано підхід до визначення керівних впливів, що передаються до позиційно-траєкторного регулятора, який забезпечує адаптивне коригування руху відповідно до змін у зовнішньому середовищі. Отже, запропонована модель дає змогу реалізувати замкнений цикл управління – від сприйняття середовища до формування керівних дій – у межах єдиної нейромережевої архітектури.

Досягнуті результати свідчать про доцільність використання біоінспірованих нейромережевих підходів

для розв'язання задач планування переміщення КФС у невизначених і динамічних умовах. Запропонованій моделі властиві простота реалізації, можливість роботи в реальному часі та адаптивність до змін середовища.

Перспективи подальших досліджень пов'язані з розширенням функціональних можливостей запропонованої моделі. Зокрема доцільним є врахування зашумленості та невизначеності сенсорних даних, розширення підходу на тривимірні середовища, а також дослідження можливостей інтеграції методів машинного й підкріплювального навчання для підвищення адаптивності планувальника. Важливим напрямом є також експериментальна перевірка моделі на реальних кіберфізичних платформах і встановлення її ефективності з огляду на обчислювальні та енергетичні обмеження.

### Конфлікт інтересів

Автори заявляють, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

### Фінансування

Це дослідження було проведено без фінансової підтримки.

### Доступність даних

Усі дані доступні в цифровій або графічній формі в основному тексті рукопису.

### Використання засобів штучного інтелекту

Автори підтверджують, що не застосовували інструментальні засоби ШІ для написання цієї роботи.

### References

1. Kliks, A., Member, S. (2022), "Neural Networks for Path Planning Salim Janji, Graduate Student Member", *IEEE*, v1. DOI: <https://doi.org/10.48550/arXiv.2207.00874>
2. Cyber-Physical Systems: Multilevel Organization and Design (2024): Monograph. A.O. Melnyk, V.A. Melnyk, V.S. Glukhov, A.M. Salo. Edited by Professor A.O. Melnyk. Lviv: "Magnolia 2006", 238 p. URL: [https://magnolia.lviv.ua/wp-content/uploads/2024/04/Kiberfizychni-systemy-1\\_zmist-1.pdf](https://magnolia.lviv.ua/wp-content/uploads/2024/04/Kiberfizychni-systemy-1_zmist-1.pdf)



3. Yang, S.X., Meng, M. (2000), "An efficient neural network approach to dynamic robot motion planning". *Neural Networks*, No.13(2), pp. 143–148. DOI: [https://doi.org/10.1016/S0893-6080\(99\)00103-3](https://doi.org/10.1016/S0893-6080(99)00103-3)
4. A Logical Calculus of the Ideas Immanent in Nervous Activity. URL: [https://en.wikipedia.org/wiki/A\\_Logical\\_Calculus\\_of\\_the\\_Ideas\\_Immanent\\_in\\_Nervous\\_Activity?utm\\_source=chatgpt.com](https://en.wikipedia.org/wiki/A_Logical_Calculus_of_the_Ideas_Immanent_in_Nervous_Activity?utm_source=chatgpt.com)
5. Kuznetsov, O. M. (2025), "Bio-inspired technologies: a synthesis of artificial intelligence and biomechanics". *Scientific Notes of Lviv University of Business and Law. Economic Series. Legal Series*, No. 46/246. DOI: <https://doi.org/10.5281/zenodo.17549924>
6. Prykhodko, S. I., Shtompel, M. A., Zhuchenko, O. S., Lisechko, V. P., Shuvalova, Y. S. (2019), "Study of the effectiveness of an adaptive decoding method for algebraic convolutional interleaved codes", *Information and Control Systems in Railway Transport*, No.4, pp.13–18. DOI: <https://doi.org/10.18664/ikszt.v0i4.178296>
7. Pyotr, K. (1975), *Anokhin Biology and Neurophysiology of the Conditioned Reflex*. DOI: <https://doi.org/10.1016/C2013-0-02871-X>
8. Qureshi, A. H., Miao, Y., Simeonov, A., Yip, M. C. (2019), "Motion Planning Networks: Bridging the Gap Between Learning-based and Classical Motion", *Computer Science*, Vol. 1. DOI: <https://doi.org/10.48550/arXiv.1907.06013>
9. Zemlyansky, D. (2025), "Neural Network Approaches to Evaluating Enterprise Supply Chain Activities in the Context of Digital Transformation", *Economy and Society*, No. 78. DOI: <https://doi.org/10.32782/2524-0072/2025-78-117>
10. Altukhov, O. V., Vasilyeva, L. V., Altukhov, V. O., Bogdanova, L. M., Tarasov, O. F. (2025), "Research on neural network approaches to the analysis of video objects using GOOGLE CORAL", *Tavriya Scientific Bulletin. Series: Technical Sciences*, No.1(5), pp. 3–12. DOI: <https://doi.org/10.32782/tnv-tech.2025.5.1.1>
11. Zhu, Y. (2023), "Using neural networks to explore path planning algorithms for robots". *Applied and Computational Engineering*, No. 5(1), pp. 566–572. DOI: <https://doi.org/10.54254/2755-2721/5/20230645>
12. Siciliano, B., Khatib, O. (2016), *Springer Handbook of Robotics*. Springer, Cham. DOI: <https://doi.org/10.1007/978-3-319-32552-1>
13. Borenstein J., Koren Y. (1991), "The Vector Field Histogram - Fast Obstacle Avoidance for Mobile Robots". *IEEE Transactions on Robotics and Automation*, No.7 (3), pp. 278–288. DOI: <https://doi.org/10.1109/70.8813>
14. McCulloch W., Pitts W.A (1943), "Logical Calculus of the Ideas Immanent in Nervous Activity", *Bulletin of Mathematical Biophysics*. No.5. pp. 115–133.
15. Haykin, S. (2009), *Neural Networks and Learning Machines*, Third Edition, Pearson Education, Inc., McMaster University, Hamilton. URL: <http://dai.fmph.uniba.sk/courses/NN/haykin.neural-networks.3ed.2009.pdf>

Received (Надійшла) 19.04.2026

Accepted for publication (Прийнята до друку) 17.05.2026

Publication date (Дата публікації) 29.05.2026

#### Відомості про авторів / About the Authors

**Григоренко Світлана Миколаївна** – кандидат технічних наук, доцент, Національний університет "Одеська політехніка", доцент кафедри комп'ютеризованих систем та програмних технологій, Одеса, Україна;

**Svitlana Hryhorenko** – Phd (Technical Sciences), Associate Professor, National University "Odesa Polytechnic", Associate Professor at the Department of Computerized Systems and Software Technologies, Odesa, Ukraine;

e-mail: [s.m.hryhorenko@op.edu.ua](mailto:s.m.hryhorenko@op.edu.ua)

ORCID ID: <https://orcid.org/0009-0006-4551-8243>

**Савельєва Оксана Степанівна** – доктор технічних наук, професор, Національний університет "Одеська політехніка", професор кафедри інтегрованих технологій управління, Одеса, Україна;

**Oksana Savielieva** – Doctor of Technical Sciences, National University "Odesa Polytechnic", Professor at the Department of Integrated Management Technologies, Odesa, Ukraine;

e-mail: [saveleva@op.edu.ua](mailto:saveleva@op.edu.ua)

ORCID ID: <https://orcid.org/0000-0002-0453-4777>

**Протасова Людмила Іванівна** – Національний університет "Одеська політехніка", старший викладач кафедри комп'ютеризованих систем та програмних технологій, Одеса, Україна;

**Liudmyla Protasova** – National University "Odesa Polytechnic", Senior Lecturer at the Department of Computerized Systems and Software Technologies, Odesa, Ukraine;

e-mail: [protasova.l.i@op.edu.ua](mailto:protasova.l.i@op.edu.ua)

ORCID ID: <https://orcid.org/0000-0002-8531-9357>

## A NEURAL NETWORK MODEL FOR PLANNING THE MOVEMENT OF A CYBER-PHYSICAL SYSTEM BASED ON AFFERENT SYNTHESIS

**Relevance of the topic.** The article addresses the problem of planning the motion of a cyber-physical system in an uncertain and dynamic external environment. The research is driven by the growing need to create intelligent control systems capable of ensuring the autonomous operation of cyber-physical systems in real time under conditions of incomplete information, the presence of static and dynamic disturbances, and limited computational resources. Modern approaches to motion planning are often characterized by high computational complexity or require a prior description of the environment, which limits their effectiveness in dynamic conditions. **Objective.** The objective of this work is to develop and formalize a neural network model for motion planning of a cyber-physical system based on the principles of afferent synthesis, which ensures the integration of sensory information and the generation of control actions in real time. **Methodology.** The research methodology is based on the use of a bio-inspired approach, the theory of functional systems, and artificial neural network methods. A discrete grid model is used to represent the external environment, which is mapped to a regular-type neural network. Control actions are generated by propagating excitation signals in the afferent synthesis subsystem and subsequently selecting the optimal direction of movement using a decision-making subsystem implemented on the basis of a "winner-takes-all" competitive selection mechanism. Trajectory complexity functional and discrete modeling methods are used to formalize the problem. **Originality.** The scientific novelty of the study lies in the proposed bio-inspired model of a neural network planner, which combines the afferent synthesis subsystem and the decision-making subsystem into a single architecture. **Results.** An approach is proposed for representing the external environment as a discrete grid model, which is then mapped onto a neural-like network that reproduces the topology of the space. For the first time, the use of a mechanism for the wave propagation of excitation signals in combination with the principle of competitive selection ("winner takes all") is justified to determine the optimal direction of motion for a cyber-physical system. The paper formulates a mathematical formulation of the motion planning problem based on the trajectory complexity functional and develops a corresponding neural network structure for its implementation. It is shown that the proposed model ensures the generation of control inputs for a position-trajectory controller and allows for the implementation of a closed-loop control cycle from environmental sensing to decision-making and motion execution. **Practical value.** The practical value of this work lies in the possibility of applying the proposed model of a bio-inspired neural network motion planner for cyber-physical systems to create intelligent control systems for mobile cyber-physical objects operating in complex and dynamic conditions. **Conclusions.** The obtained results confirm the effectiveness of using bio-inspired neural network approaches for motion planning tasks and lay the groundwork for the further development of intelligent control systems for cyber-physical systems.

**Keywords:** cyber-physical system; neural network planning; afferent synthesis; bio-inspired approach; motion planning; neural-like network; decision-making subsystem; wave propagation of signals; optimal trajectory; adaptive control.

### Бібліографічні описи / Bibliographic descriptions

Григоренко С. М., Савельєва О. С., Протасова Л. І. Модель неймережевого планування кіберфізичної системи на основі аферентного синтезу. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 17–33. DOI: <https://doi.org/10.30837/0135-1710.2026.189.017>

Hryhorenko, S., Savielieva, O., Protasova, L. (2026), "A neural network model for planning the movement of a cyber-physical system based on afferent synthesis", *Management Information System and Devices*, No. 2 (189), P. 17–33. DOI: <https://doi.org/10.30837/0135-1710.2026.189.017>

Stanislav Nechypurenko, Anton Sorokin

## METHOD OF STRUCTURAL PRUNING OF CONVOLUTIONAL NEURAL NETWORKS WITH REGARD TO EDGE DEVICE HARDWARE CHARACTERISTICS

**Research subject** – methods of structural pruning of convolutional neural networks applied to reduce the computational complexity of object detection models in video systems with local data processing. **Purpose of the study** – improvement of the structural pruning method through the introduction of a hardware-aware channel selection criterion that takes into account the actual operation latency on the target edge device. **Research objectives:** comparative analysis of existing structural pruning criteria, justification for selecting the baseline method, and development of its modification based on latency profiling of the target hardware platform. **Research methods** are based on the mathematical analysis of convolutional layer channel importance criteria, theoretical comparison of approaches to evaluating neural network parameter redundancy, and formalization of the hardware-aware optimization problem. **Research results** consist in the development of a modified channel importance criterion that combines the evaluation of a channel's impact on the loss function with a normalized latency coefficient of the corresponding layer, calculated by profiling the execution time of each convolutional layer on the target edge device. The proposed criterion enables prioritization of channel removal in layers that impose the greatest computational load on specific hardware while preserving channels with a high impact on model accuracy. A theoretical analysis of the criterion properties was performed, including proof of its reduction to the baseline Taylor criterion when the hyperparameter value is zero, as well as the monotonicity of pruning priority redistribution between layers. It was demonstrated that the proposed approach is modular, requires minimal additional computational overhead, and can be integrated into existing model optimization pipelines. Experimental validation on the YOLOv8n detector model confirmed the effectiveness of the proposed approach: with a pruning budget of 30% removed channels, the inference latency was reduced by 38.55% compared to 34.88% for the baseline Taylor criterion, corresponding to a relative improvement of +10.5%. **Conclusions.** The results confirm the theoretical validity of the proposed approach and its advantages over hardware-independent pruning criteria for deploying detection models on edge devices as part of real-time video systems.

**Keywords:** structural pruning; convolutional neural network; edge device; inference latency; channel importance criterion; model optimization; video analytics; object detection; local data processing; hardware-aware optimization.

### 1. Introduction

The rapid growth in the number of surveillance cameras, transportation sensors, and IoT devices is driving the need for video analytics systems capable of operating in real time directly at the data collection site. Transmitting raw video streams to cloud servers results in significant consumption of communication channel bandwidth, increased latency, and risks of data privacy breaches [1]. An alternative approach is to deploy deep learning models directly on edge devices – local gateways, embedded processors, or specialized accelerators – which enables rapid decision-making and system autonomy even under unstable network conditions.

However, modern convolutional neural networks (CNNs), which demonstrate high accuracy in object detection and classification tasks, are characterized by significant demands on RAM, computational power, and energy consumption. These requirements significantly exceed the resource capabilities of typical edge platforms, such as ARM processors, NVIDIA Jetson, or Google Coral [2]. Overcoming this trade-off

between model accuracy and computational complexity is one of the central challenges in the modern field of efficient deep learning.

Among the known approaches to reducing the computational complexity of CNNs are weight and activation quantization [3], knowledge distillation [4], neural architecture search [5], and structural pruning [6]. The latter approach, which involves removing entire channels or filters from convolutional layers, is attractive due to the direct reduction in the dimension of intermediate computation tensors without the need for specialized hardware for sparse arithmetic.

Existing criteria for structural pruning – based on the L1 norm [7], Taylor series expansion [8], and geometric median [9] – assess the importance of each channel solely in terms of its impact on model accuracy. This approach, however, does not account for the fact that reducing the number of floating-point operations (FLOPs) is not always proportional to the actual inference acceleration on specific hardware. Different network layers have varying execution efficiency depending on the processor architecture, cache size, memory access

characteristics, and the degree of computational parallelism. As a result, a model optimized using the " " criterion – which aims for the minimum accuracy loss given a fixed number of removed parameters – may exhibit suboptimal acceleration on the target edge device.

This paper proposes an improvement to the structural pruning method based on the Taylor series expansion by introducing a hardware-oriented coefficient that accounts for the actual latency of operations in each layer on the target platform. The rationale for selecting the base method is based on a comparative analysis of existing pruning criteria.

## 2. Analysis of the literature and definition of the research problem

Compression methods for convolutional neural networks encompass a wide range of approaches that differ in their mechanisms for reducing computational complexity and their hardware requirements. Quantization involves transitioning from 32-bit floating-point representation to reduced-precision integer arithmetic. This approach provides up to a fourfold speedup provided there is hardware support for INT8 operations [3]. Knowledge distillation involves training a compact student model under the guidance of a more powerful teacher model, allowing a significant portion of accuracy to be preserved while substantially reducing the number of parameters [4]. Neural Architecture Search (NAS) automates the design of efficient network topologies but requires significant computational resources during the search phase [5].

Structural pruning occupies a unique position among these methods due to a number of practical advantages. Unlike non-structural pruning, which removes individual weights and creates sparse matrices, the structural approach removes entire channels or filters, resulting in a network that retains a dense tensor structure and can run efficiently on standard hardware without specialized libraries for sparse arithmetic [6]. Furthermore, structural pruning can be combined with other compression methods, notably quantization, providing a cumulative acceleration effect. Recent studies, notably [10], address the issue of inter-layer dependencies in structural pruning, enabling the correct removal of channels in networks with complex topologies.

Among the criteria for determining the importance of channels for structural pruning, the following approaches have been most extensively studied.

The criterion based on the  $L1$  norm [7] evaluates the importance of the  $j$ -th filter in the  $l$ -th layer as the sum of the absolute values of its weights:

$$S_{L1}(f_j^l) = \sum_i w_i^{j,l}, \quad (1)$$

where  $w_i^{j,l}$  is the  $i$ -th weight of the filter  $f_j^l$  in layer  $l$ .

Filters with the smallest value of  $S_{L1}(f_j^l)$  are considered redundant and are removed. The advantage of the method is computational simplicity – calculating the criterion does not require a forward or backward pass of the network. However, the  $L1$  norm of the weights is only an indirect indicator of the channel's influence on the network's output signal and does not account for interactions between filters in different layers.

The criterion based on batch normalization scaling factors [11] uses the  $\gamma$  parameters of the BatchNorm layer as an indicator of the importance of the corresponding channel:

$$S_{BN}(c_j^l) = |\gamma_j^l|, \quad (2)$$

where  $\gamma_j^l$  is the batch normalization scaling coefficient for the  $j$ -th channel of the  $l$ -th layer. During training,  $L1$  regularization of the parameters  $\gamma$  is added to the loss function, which encourages the network to set to zero the coefficients of channels that have the least impact on the result. This approach integrates the process of identifying redundant channels directly into the training procedure, but requires retraining the network with a modified loss function.

A criterion based on the Taylor series [8] evaluates the impact of channel removal on the loss function value through the first term of the Taylor series:

$$S_T(c_j^l) = \left| \frac{\partial L}{\partial c_j^l} \cdot c_j^l \right|, \quad (3)$$

where  $L$  is the loss function, and  $c_j^l$  is the activation of the  $j$ -th channel in the  $l$ -th layer. Unlike the  $L1$  norm, this criterion takes into account not only the magnitude of the weights but also their actual impact on the loss function via gradient information. This provides a more accurate assessment of a channel's importance, especially in networks with complex architectures and residual connections. Computing the criterion requires one

forward and one backward pass on the training dataset, which is an acceptable computational cost.

The criterion based on the geometric median [9] is founded on a different idea: instead of estimating the absolute importance of each filter, its uniqueness relative to other filters in the same layer is determined:

$$S_{GM}(f_j^l) = \sum_{i \neq j} \|f_j^l - f_i^l\|_2, \quad (4)$$

where  $\|\cdot\|_2$  is the Euclidean norm. The filters located closest to the geometric median of the set of filters

in the layer are considered the most redundant, since their functionality can be replicated by other filters. The advantage of the method is its data independence – the criterion is calculated solely based on the network weights. However, the method does not account for the impact of filter removal on the network's overall accuracy in the context of a specific task. A comparative analysis of the considered criteria is presented in Table 1.

**Table 1.** Comparison of CNN structural pruning criteria

| Criterion        | What it evaluates             | Data requirements   | Advantage                      | Limitations                                          |
|------------------|-------------------------------|---------------------|--------------------------------|------------------------------------------------------|
| $L1$ norm        | Absolute value of weights     | Not required        | Ease of calculation            | Does not account for the impact on the loss function |
| BatchNorm        | Scaling coefficients $\gamma$ | Requires retraining | Integration with training      | Requires modification of the loss function           |
| Taylor           | Impact on the loss function   | Calibration sample  | Accurate importance estimation | Requires a backpass                                  |
| Geometric median | Filter uniqueness             | Does not require    | Data independence              | Does not account for impact on accuracy              |

An analysis of the approaches presented reveals a common limitation: all the criteria considered evaluate the importance of channels solely from the perspective of their impact on model accuracy or the structural redundancy of the network. None of them takes into account the actual computational cost of executing operations of a specific layer on the target hardware platform.

This limitation becomes critical for edge deployment, since theoretical computational complexity in FLOPs is not an adequate predictor of actual inference latency. As noted in [2], layers with the same number of FLOPs can have significantly different latencies depending on the tensor dimensions (which affects cache efficiency), the type of operations (standard convolution versus depth-wise convolution), and the degree of parallelism of the available computational units. The AMC approach [12] partially addresses this issue by using reinforcement learning to determine optimal pruning coefficients for each layer, subject to latency constraints. However, AMC requires a lengthy agent training process and does not directly modify the channel importance criterion, which complicates its integration with existing pruning pipelines.

Thus, the task of developing a structural pruning criterion that combines an accurate assessment of a channel's impact on the loss function with

consideration of the actual latency of the layer on the target edge device is of great relevance.

### 3. Research Objectives and Tasks

The objective of the research is to improve the method of structural pruning of convolutional neural networks for video systems with local data processing by introducing a hardware-oriented channel selection criterion that accounts for the actual latency of operations on the target edge device.

To achieve this goal, the following tasks have been defined:

- conduct a comparative analysis of existing structural pruning criteria and justify the selection of a baseline method for further improvement;
- develop a procedure for profiling the latency of CNN layers on the target hardware platform and formalize a hardware-oriented coefficient;
- formulate a modified channel importance criterion that combines the assessment of the impact on the loss function with the layer latency coefficient;
- perform a theoretical analysis of the properties of the proposed criterion and determine the conditions under which it provides advantages over the baseline method.



## 4. Materials and Methods

### 4.1 Justification for the choice of the baseline method

Based on the comparative analysis performed in Section 2, the criterion based on Taylor's expansion was selected as the baseline method for improvement. This choice is justified by the following arguments:

- the Taylor criterion provides the most accurate assessment of the impact of channel removal on the loss function among the approaches considered, as confirmed by the results of comparative studies on standard architectures [8];
- unlike the  $L1$  norm, the Taylor criterion takes gradient information into account, allowing channels with large but uninformative weights to be distinguished from channels with small but critically important weights;
- unlike the BatchNorm-based method [11], the Taylor criterion does not require retraining the network with a modified loss function, which simplifies its integration into the optimization pipeline;
- the computational cost of calculating the criterion (one forward and one backward pass on the calibration sample) is acceptable for practical application.

### 4.2 Latency Profiling Procedure

To obtain hardware-specific information about the computational cost of each network layer, we propose a profiling procedure consisting of the following steps:

- 1) specifying the target hardware platform and runtime environment (inference runtime);
- 2) sequential measurement of the execution time of each convolution layer  $l \in \{1, \dots, L\}$  on a test input tensor with a task-specific dimension;
- 3) repeating the measurements  $K$  times to reduce the impact of random fluctuations and calculating the average latency of each layer;
- 4) calculating the depthwise latency of each layer.

Let  $t_l$  be the measured average latency of the  $l$ -th layer, and  $N_l$  be the number of output channels of this layer. The depthwise latency is defined as:

$$\delta_l = t_l / N_l, \quad (5)$$

where  $\delta_l$  characterizes the contribution of a single channel of the  $l$ -th layer to the total inference latency, assuming a linear dependence of the layer's latency on the number of its channels.

This linear approximation is acceptable for most convolutional network architectures, since the computational complexity of a convolutional layer is proportional to the number of its output channels when other parameters are fixed (kernel size, number of input channels, spatial dimension). For depth-separated convolutions [13, 14], where the kernel is applied to each channel independently, the linear relationship holds directly.

To ensure comparability of depthwise latencies across different layers, normalization is performed:

$$\tilde{\delta}_l = \frac{\delta_l}{\max_{l'=1..L} (\delta_{l'})}, \quad (6)$$

where  $\tilde{\delta}_l \in (0,1]$  is the normalized depthwise latency, and the maximum is calculated across all convolutional layers of the network.

### 4.3 Modified channel importance criterion

The proposed hardware-oriented importance criterion for the  $j$ -th channel in the  $l$ -th layer is defined as:

$$S_{hw}(c_j^l) = \frac{S_T(c_j^l)}{(\tilde{\delta}_l)^\beta}, \quad (7)$$

where  $S_T(c_j^l)$  is the value of the Taylor's basic criterion according to equation (3);  $\tilde{\delta}_l$  is the normalized depthwise latency of the layer according to equation (6);  $\beta \in [0,1]$  is a hyperparameter that controls the degree of influence of hardware information on the criterion.

The mechanism of action of the proposed criterion is as follows. Channels with the smallest values of  $S_{hw}$  are subject to removal first. Dividing the value of the Taylor criterion by  $(\tilde{\delta}_l)^\beta$  leads to a reduction in the importance score of channels in layers with high depthwise latency. Accordingly, when ranking channels by  $S_{hw}$ , channels from computationally expensive layers receive lower scores and are more likely to be removed, which directly reduces inference latency.

The hyperparameter  $\beta$  allows adjusting the balance between model accuracy and actual acceleration. When  $\beta = 0$ , criterion (7) reduces to the standard Taylor criterion (3), i.e., pruning is performed without taking hardware information into account. When  $\beta = 1$ , the impact of latency is maximized. Intermediate values  $\beta$  ensure a smooth transition between these extreme cases.

#### 4.4 Algorithm for Hardware-Oriented Pruning

The general procedure for applying the proposed criterion consists of the following steps:

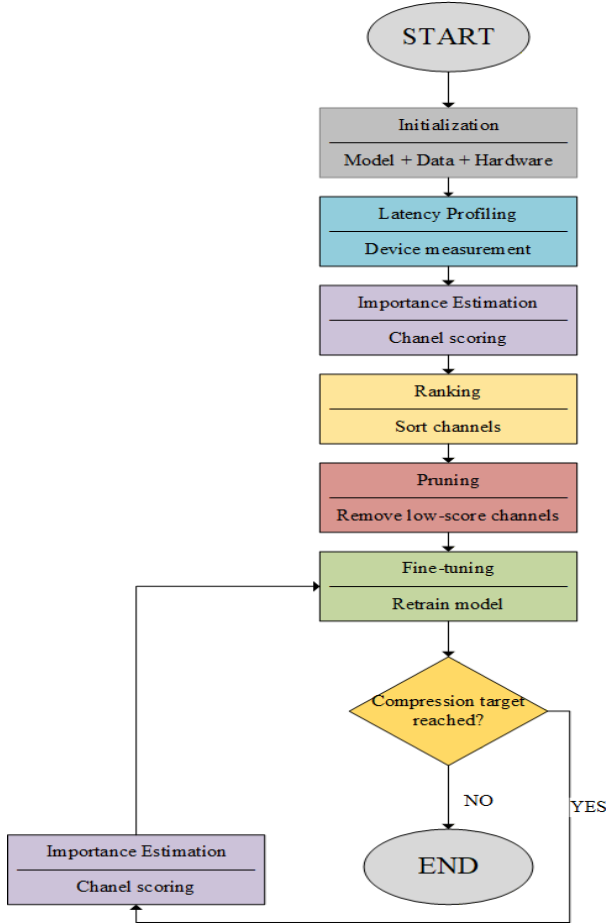


Fig. 1. Block diagram of the hardware-oriented structural pruning algorithm for CNNs

*Step 1. Preparation.* A pre-trained CNN model and a target hardware platform are selected. A calibration subsample is formed from the training dataset.

*Step 2. Latency profiling.* For each convolutional layer  $l$ , the average latency  $t_l$  is measured by repeating the inference  $K$  times on the target device. The depthwise latencies  $\delta_l$  are calculated using formula (5) and their normalized values  $\tilde{\delta}_l$  are calculated using formula (6).

*Step 3. Calculation of importance criteria.* A forward and backward pass of the network is performed on the calibration subsample. For each channel  $c_j^l$ , the value of the Taylor criterion  $S_T(c_j^l)$  is calculated using formula (3), after which the modified

criterion  $S_{hw}(c_j^l)$  is calculated using formula (7) with the selected value  $\beta$ .

*Step 4. Ranking and removal.* Channels from all network layers are ranked in ascending order by  $S_{hw}$ . A specified proportion of channels with the smallest values of  $S_{hw}$  is removed. The corresponding filters and BatchNorm parameters are removed from the model.

*Step 5. Fine-tuning.* The network obtained after pruning is fine-tuned on the full training set for a small number of epochs to restore accuracy.

*Step 6. Iteration.* If necessary to achieve a higher level of compression, steps 3–5 are repeated iteratively with a gradual increase in the proportion of channels removed.

## 5. Research Results

### 5.1 Analysis of the properties of the proposed criterion

The proposed criterion,  $S_{hw}$ , has several important properties that determine its behavior in various scenarios.

*Property 1* (reduction to the base method). Given  $\beta = 0$ , for any channel  $c_j^l$ ,  $S_{hw}(c_j^l) = S_T(c_j^l)$  holds; that is, the proposed criterion is a generalization of Taylor's criterion. This guarantees that, when the hardware-oriented component is disabled, the pruning results coincide with those of the base method.

*Property 2* (priority redistribution between layers). Let two channels  $c_a$  and  $c_b$  belong to layers  $l_a$  and  $l_b$  respectively, where  $S_T(c_a) = S_T(c_b)$ , but  $\tilde{\delta}_{l_a} > \tilde{\delta}_{l_b}$ . Then  $S_{hw}(c_a) < S_{hw}(c_b)$ , i.e., the channel from the computationally more expensive layer receives a lower importance score and will be removed earlier. Under Taylor's standard criterion, both channels would have the same score, and the choice between them would be arbitrary.

*Property 3* (monotonicity with respect to  $\beta$ ).

For a fixed channel  $c_j^l$ , the ratio  $\frac{S_{hw}(c_j^l)}{S_T(c_j^l)} = \frac{1}{(\tilde{\delta}_l)^\beta}$

monotonically depends on  $\beta$ : it increases as  $\tilde{\delta}_l < 1$  and equals one when  $\tilde{\delta}_l = 1$  (for a layer with maximum depthwise latency). This means that increasing  $\beta$  intensifies the pruning redistribution in favor of computationally expensive layers.

## 5.2 Theoretical Comparison with the Standard Approach

To illustrate the difference between standard and hardware-oriented pruning, let us consider a typical situation that arises when deploying models on edge devices.

Suppose the network contains two convolutional layers: layer  $A$  with standard  $3 \times 3$  convolutions and layer  $B$  with depth-separated  $3 \times 3$  convolutions [14]. With the same number of channels, layer  $A$  has significantly higher computational complexity in FLOPs. However, on real hardware, the situation may be different: standard convolutions make more efficient use of GPU/NPU parallelism, whereas depth-split convolutions, despite requiring fewer FLOPs, may execute more slowly due to low utilization of computational units [2].

In this case, the standard Taylor criterion, assuming equal importance of channels in both layers, will remove channels arbitrarily from both layers. The proposed criterion, on the other hand, will automatically identify the layer with higher actual depthwise latency and direct pruning primarily to it, providing greater actual acceleration at the same level of accuracy reduction.

This difference becomes particularly significant for modern detector architectures, such as YOLOv7 [15] or models based on MobileNetV2 [14], which contain layers with heterogeneous operations – standard convolutions, depthwise convolutions,  $1 \times 1$  pointwise convolutions, concatenations, and element-wise operations. For such networks, the discrepancy between FLOPs and actual latency is most pronounced, and a hardware-oriented approach potentially yields the greatest gains. Experimental validation of this mechanism on a specific detector model with real-world latency profiling is presented in Section 6.

## 5.3 Analysis of the Hyperparameter's Influence $\beta$

The hyperparameter  $\beta$  determines the trade-off between two optimization objectives. At small values of  $\beta$  (close to 0), pruning primarily optimizes model accuracy, as the criterion approaches the standard Taylor criterion. At large values of  $\beta$  (close to 1), pruning more aggressively removes channels from computationally expensive layers, which maximizes actual acceleration but may lead to a greater reduction in accuracy if these channels are critical to the task.

For each "model-hardware platform" pair, there is an optimal value for  $\beta^*$  that provides the best balance between accuracy and latency.  $\beta^*$  can be determined by iterating over a validation sample with a small number of candidate values, such as  $\beta \in \{0, 0.25, 0.5, 0.75, 1.0\}$ , which is computationally feasible since for each value of  $\beta$ , one only needs to re-rank the channels without re-profiling or computing gradients.

## 5.4 Computational Complexity of the Proposed Method

The proposed modification introduces minimal additional computational overhead compared to the baseline method. The latency profiling stage is performed once for each target platform and requires  $O(K \cdot L)$  direct passes of individual layers, where  $K$  is the number of iterations and  $L$  is the number of layers. Calculating the modified criterion (7) for all channels requires  $O\left(\sum_i N_i\right)$  additional division operations, which is negligibly small compared to the calculation of gradients for the Taylor criterion.

## 5.5 Comparison of the Proposed Criterion with Existing Approaches

To systematically compare the proposed criterion with existing approaches to structural pruning, six key characteristics have been identified that are critical for the deployment of detection models on edge devices within real-time video systems:

- accounting for the channel's impact on the loss function, which determines the accuracy of assessing the channel's importance for a specific task;
- accounting for the actual latency of the layer on the target hardware, which ensures that pruning is oriented toward actual acceleration rather than a theoretical reduction in FLOPs;
- adaptability to the target platform, i.e., the ability to automatically generate different pruning strategies for different edge devices;
- no need to retrain the model with a modified loss function, which simplifies integration into automated optimization pipelines;
- a mechanism for balancing model accuracy and actual inference acceleration;

• compatibility with existing structural pruning pipelines without changing the training procedure or network architecture.

The comparison results are presented in Table 2.

Analysis of Table 2 allows us to draw the following conclusions. None of the existing criteria takes into account the actual latency of operations on the target hardware platform and does not provide a mechanism for adapting the pruning strategy to a specific edge device. Criteria based on the  $L1$  norm [7] and the geometric median [9], despite their computational simplicity and data independence, do not utilize information about the channel's impact on the loss function, which may lead

to the removal of channels critical for detection accuracy. The criterion based on BatchNorm [11] only partially accounts for the impact on the loss function – through the  $L1$  regularization procedure during training, rather than through direct gradient computation – and requires retraining the network with a modified loss function, which complicates its integration into automated optimization pipelines. The basic Taylor criterion [8] provides the most accurate assessment of channel importance among the approaches considered, but does not distinguish between layers with different computational costs on specific hardware, which can lead to suboptimal acceleration on edge devices.

**Table 2.** Comparison of pruning criteria based on characteristics critical for edge deployment

| Characteristic                                     | $L1$ -norm | BatchNorm | Taylor | Geometric median | Proposed |
|----------------------------------------------------|------------|-----------|--------|------------------|----------|
| Takes into account the impact on the loss function | No         | Partially | Yes    | No               | Yes      |
| Takes into account the actual layer latency        | No         | No        | No     | No               | Yes      |
| Adapts to the target platform                      | No         | No        | No     | No               | Yes      |
| Does not require model retraining                  | Yes        | No        | Yes    | Yes              | Yes      |
| Adjustable balance between accuracy and speed      | No         | No        | No     | No               | Yes      |
| Compatibility with existing conveyors              | Yes        | No        | Yes    | Yes              | Yes      |

The proposed criterion  $S_{hw}$  (7) inherits from Taylor's basic criterion an accurate assessment of a channel's impact on the loss function and supplements it with a hardware-oriented component in the form of a normalized depthwise latency coefficient  $\tilde{\delta}_l$ . This provides three additional properties absent in all existing approaches: consideration of the actual latency of layers, automatic adaptation to the target platform through profiling, and an adjustable balance between accuracy and speed via the hyperparameter  $\beta$ . At the same time, the proposed criterion remains fully compatible with existing pruning pipelines and does not require model retraining.

## 6. Experimental verification of the proposed criterion

### 6.1 Experimental setup

To verify the proposed criterion  $S_{hw}$  (Equation 7), the YOLOv8n detector was used in its default configuration – 64 convolutional layers, with an input image size of  $640 \times 640$  pixels. Latency profiling was performed on a CPU as the most accessible edge computing configuration. For each layer,  $K = 50$  measurement repetitions were performed after five warm-up iterations, with the results averaged to minimize the impact of random fluctuations in system call durations.

Channel-wise latency  $\delta_l$  and its normalized value  $\tilde{\delta}_l$  were calculated using formulas (5) and (6), respectively. The value of the Taylor criterion  $S_r$  (Equation 3) was obtained by performing one forward and one backward pass of the network on the test tensor with a pseudo-functional in the form of the sum of the elements of all output feature maps of the detection heads – this choice of functional ensures the accumulation of gradients on all internal activations of the network. The hyperparameter  $\beta$  in formula (7) is set to 0.5, which corresponds to a balanced pruning redistribution mode.

### 6.2 Layer Latency Profile and Hardware Heterogeneity of the Model

Fig. 2(a) shows the absolute latency of each convolutional layer of the YOLOv8n model. The distribution is noticeably heterogeneous: values range from 0.009 ms to 0.135 ms, i.e., by more than one and a half orders of magnitude. The most computationally expensive operations are the convolutions  $3 \times 3$  in the middle part of the network (indices 20–24, 41–44) and the stream convolutions  $1 \times 1$  that operate on high-resolution maps (indices 22, 25–26, 41–42). In Fig. 2(b), the same profile is presented in the form of normalized depthwise latency  $\tilde{\delta}_l$ , which is directly included in formula (7). One of the detection head layers (index  $\approx 65$ )

exhibits  $\tilde{\delta}_l = 1,0$ , meaning its contribution to the depthwise inference time is tens of times greater than that of most other layers. The obtained profile experimentally confirms the initial hypothesis of

this work: the hardware cost of channels is significantly heterogeneous, and the uniform treatment of channels, characteristic of the classical Taylor criterion, ignores this heterogeneity.

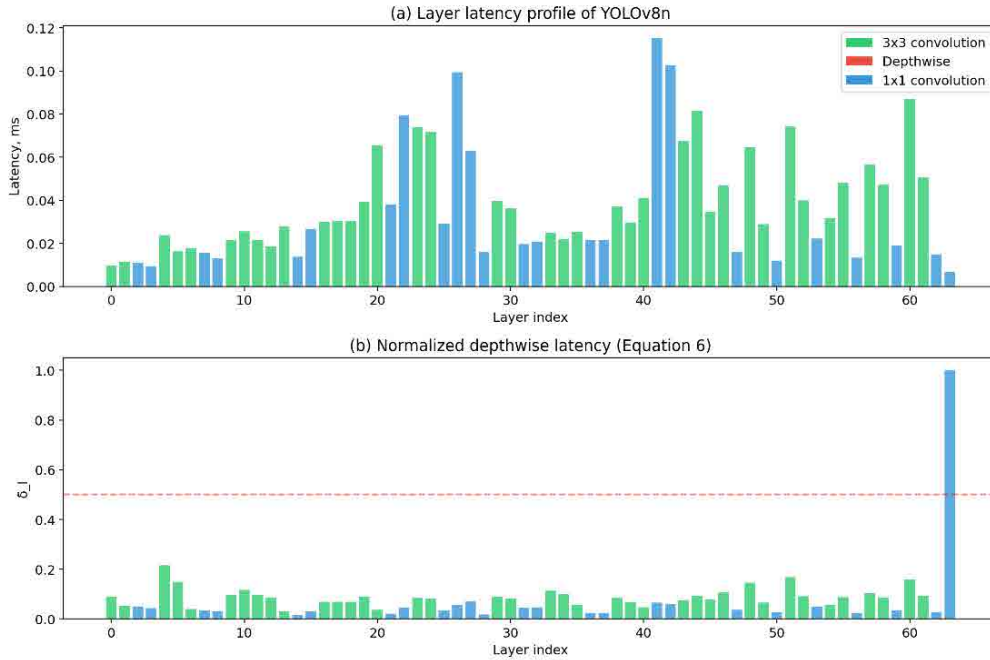


**Fig. 2.** Variation in the computational cost of the convolutional layers of the YOLOv8n model:  
 (a) absolute layer latency, ms; (b) normalized depthwise latency (Equation 6)

### 6.3 Pruning Redistribution and Quantitative Advantage of the Proposed Criterion

Fig. 3(a) presents normalized estimates of layer importance based on two criteria. The distribution shapes are similar, but the differences are concentrated precisely in the regions that have the greatest impact on inference time: for layers with low  $\tilde{\delta}_l$ , the estimates  $S_{hw}$  decrease relative to  $S_T$ , while for layers with high  $\tilde{\delta}_l$ , they increase. This constitutes the methodological core of the proposed approach: the importance of a channel is normalized by the hardware cost of the channel, as a result of which "cheap" channels are prioritized for retention, and "expensive" ones for removal.

Fig. 3(b) illustrates the result of applying this redistribution – the proportion of removed channels in each layer with a total pruning budget of 30%. As expected, both criteria do not affect the early layers of the backbone (indices 0–25), which have the highest importance score and are responsible for low-level features. However, in the middle and deep parts of the network (indexes 31–45), a systematic shift is observed:  $S_{hw}$  prunes slower convolutions more aggressively  $3 \times 3$

and preserves "cheap" ones  $1 \times 1$ , whereas the standard approach distributes removals without accounting for this difference.

Quantitative comparison results are presented in Table 3. With the same number of removed channels (1,718, corresponding to 30% of the total number of model channels), the proposed criterion achieved a projected latency reduction of 1.0727 ms versus 0.9705 ms for the standard Taylor criterion, or 38.55% and 34.88%, respectively, of the model's baseline latency. In relative terms, this represents a +10.5% additional performance gain achieved solely through hardware-oriented pruning redistribution, without changing the number of retained parameters and without additional overhead at the importance evaluation stage ( $\tilde{\delta}_l$  profiling is performed once). The number of layers in which pruning exceeds 5% increases slightly (33 versus 31), indicating a more balanced distribution of removals:  $S_{hw}$  avoids concentrating pruning in the few fastest layers and distributes it more broadly, but purposefully – toward the slower ones.



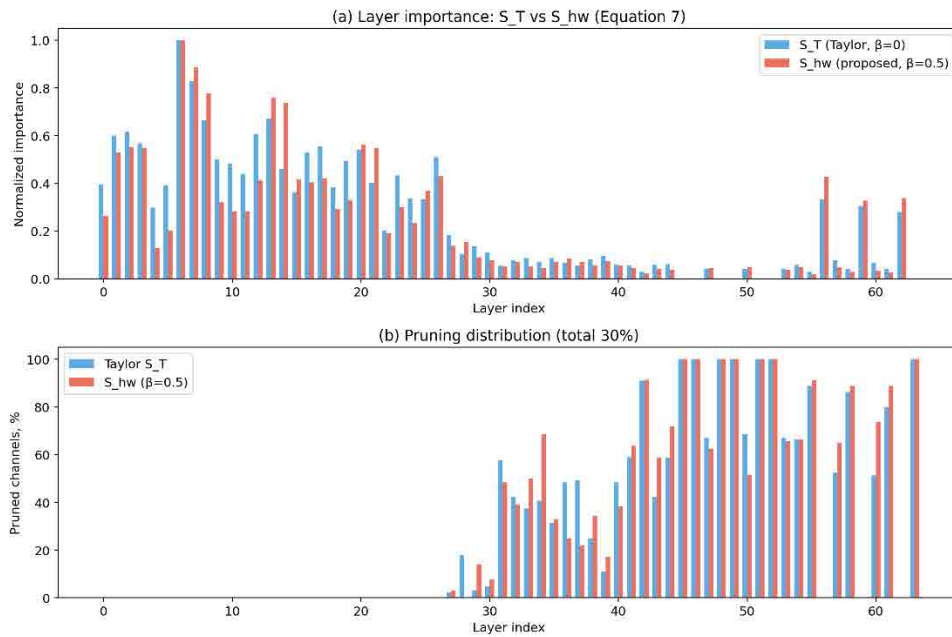


**Fig. 3.** Effect of the hardware-oriented component of the criterion (Formula 7) on the YOLOv8n pruning plan:

- a) redistribution of normalized layer importance between  $S_T$  and  $S_{hw}$ ;
- b) proportion of removed channels by layer with a 30% pruning budget.

**Table 3.** Comparison of pruning criteria for  $S_T$  and  $S_{hw}$  for the YOLOv8n model with a 30% pruning budget (CPU)

| Metric                               | Taylor $S_T$ ( $\beta = 0$ ) | $S_{hw}$ ( $\beta = 0.5$ ) |
|--------------------------------------|------------------------------|----------------------------|
| Channels removed                     | 1718                         | 1718                       |
| Layers with pruning > 5%             | 31                           | 33                         |
| Maximum pruning of a single layer, % | 100                          | 100                        |
| Latency reduction, ms                | 0.9705                       | 1.0727                     |
| Latency reduction, %                 | 34.88                        | 38.55                      |
| Advantage over the Taylor method     | —                            | +10.5%                     |

Note that the latency reduction shown in Table 3 is theoretical – based on the assumption of a linear relationship between a layer's inference time and the number of its channels, as given by Equation (5). The physical implementation of pruning, together with full model retraining, will provide final empirical verification of the preservation of detection quality and actual latency, which is a direct focus of future research.

## 7. Discussion of the research results

The proposed hardware-oriented criterion for structural pruning has several practical advantages for model deployment tasks in video systems with local data processing. Let us consider these advantages using specific examples.

### 7.1 Automatic detection of hardware "bottlenecks"

Criterion (7) automatically identifies layers with disproportionately high actual latency, even if their theoretical complexity in FLOPs is low. For example, when deploying MobileNetV2-SSD on an ARM Cortex-A53, depth-separated convolutions have significantly fewer FLOPs compared to standard ones, yet they execute inefficiently due to low arithmetic intensity. Profiling will reveal high depthwise latency  $\tilde{\delta}_l$  in these layers, and criterion (7) will lower the importance score of their channels, directing pruning specifically at them. The standard Taylor criterion lacks a mechanism for detecting such situations.

## 7.2 Adaptability to different hardware platforms

Since the coefficient  $\tilde{\delta}_l$  is calculated by profiling a specific device, the same algorithm automatically generates different pruning strategies for different platforms. For example, for YOLOv7-tiny, the convolutions  $3 \times 3$  on the NVIDIA Jetson Nano effectively utilize the parallelism of CUDA cores, whereas on the Google Coral Edge TPU, those same layers may be partially executed on the host device's CPU. Different latency profiles will result in different channel rankings based on criterion (7), which is optimal for each platform. The standard Taylor criterion will yield the same ranking regardless of the device.

## 7.3 Modularity and minimal additional costs

The modification does not require changes to the training procedure, network architecture, or pruning algorithm. Integration into any existing pipeline with depthwise ranking boils down to a one-time platform profiling and element-by-element decomposition of the Taylor criterion vector into a vector of normalized latencies according to formula (7), which requires

$O\left(\sum_l N_l\right)$  additional operations – negligibly few compared to gradient computation.

In the context of real-time video systems, the proposed method is particularly relevant, as inference latency requirements are stringent and determined by the video stream's frame rate. A typical video surveillance task requires processing 25–30 frames per second, corresponding to a latency budget of 33–40 ms per frame. Traditional pruning, focused on minimizing FLOPs, may fail to meet this budget even with a significant reduction in computational complexity if the "bottleneck" consists of layers with inefficient execution on specific hardware.

It is important to note the limitations of the proposed approach. The linear approximation of the relationship between layer latency and the number of channels (5) may be inaccurate for some configurations, particularly for layers with a very small number of channels, where the overhead of launching the computation kernel dominates the useful computations. In addition, the latency profile may vary depending on the driver version, the device's operating temperature, and other external factors.

The experimental validation performed in Section 6 using the YOLOv8n model confirmed the main hypothesis of this work: hardware-oriented channel

ranking provides a 10.5% reduction in latency for the same budget of removed channels, and this advantage is achieved without additional overhead during the importance evaluation stage. Note that the experiment was conducted on a CPU as the most accessible edge configuration; extending the verification to specialized accelerators – NVIDIA Jetson Nano (CUDA-core parallelism), Google Coral Edge TPU (quantized INT8 arithmetic), and the ARM Cortex-A series (no GPU acceleration) – constitutes the immediate focus of further research, since it is on such platforms that  $\tilde{\delta}_l$  profiles are expected to be the most heterogeneous. Separate research directions include the study of network architectures with a significant proportion of depthwise operations (MobileNetV2-SSD, EfficientDet-Lite), rigorous verification of the preservation of detection performance after the physical implementation of pruning according to both criteria, as well as combining the proposed pruning with subsequent quantization to achieve a cumulative acceleration effect.

## 8. Conclusion

This paper presents a comparative analysis of existing criteria for structural pruning of convolutional neural networks – based on the  $L1$  norm, batch normalization scaling factors, the Taylor series, and the geometric median. The choice of the Taylor criterion as the baseline method is justified due to its theoretical soundness and ability to account for the actual influence of the channel on the loss function.

An improvement to the Taylor criterion is proposed by introducing a normalized depthwise latency coefficient, calculated based on profiling the execution time of each layer on the target edge device. The modified criterion ensures a redistribution of pruning in favor of layers with high actual computational cost, which allows for greater inference acceleration compared to the hardware-independent approach with controlled loss of accuracy.

A theoretical analysis of the properties of the proposed criterion was performed; in particular, its reduction to the baseline method was shown for  $\beta = 0$ , the monotonicity of priority reallocation with respect to the hyperparameter  $\beta$ , and the minimal additional computational complexity.

Experimental verification on a YOLOv8n detector model with 64 convolutional layers confirmed the

theoretical predictions. With a pruning budget of 30% of removed channels, the modified criterion achieved a projected latency reduction of 38.55% compared to 34.88% for the baseline Taylor criterion, corresponding to a relative advantage of +10.5%. An analysis of the pruning distribution across layers showed that the achieved advantage is due precisely to the targeted removal of channels in layers with high depthwise latency ( $\tilde{\delta}_l$ ), which directly confirms the functioning of the mechanism described in Equation (7).

The proposed approach is modular and can be integrated into existing model optimization pipelines for video systems with local data processing. Further research is aimed at extending experimental verification to specialized edge accelerators and architectures with a higher proportion of depthwise operations, as well as investigating the combination of the proposed pruning with subsequent quantization to achieve a cumulative acceleration effect.

---

### Conflict of Interest

The authors declare that they have no conflicts of interest regarding this study, including financial, personal, authorship, or other conflicts that could influence the study and its results presented in this article.

---

### Funding

The study was conducted without financial support.

---

### Data Availability

The manuscript has no associated data.

---

### Use of Artificial Intelligence

The authors confirm that they did not use artificial intelligence technologies in the creation of this work.

## References

1. Wang, X., Han, Y., Leung, V. C. M., Niyato, D., Yan, X., Chen, X. (2023), "Edge computing with artificial intelligence: A machine learning perspective", *ACM Computing Surveys*, Vol. 55, No. 9, Article 184. DOI: <https://doi.org/10.1145/3555802>
  2. Yang, Z., Hu, J., Tang, X., Zhao, L., Li, K. (2022), "Optimization methods, challenges, and opportunities for edge inference", *Electronics*, Vol. 11, No. 14, Article 2189. DOI: <https://doi.org/10.3390/electronics11142189>
  3. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D. (2018), "Quantization and training of neural networks for efficient integer-arithmetic-only inference", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: <https://doi.org/10.1109/CVPR.2018.00286>
  4. Li, Z., Xu, P., Chang, X., Yang, L., Zhang, Y., Yao, L., Chen, X. (2023), "When object detection meets knowledge distillation: A survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI: <https://doi.org/10.1109/TPAMI.2023.3257546>
  5. Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S. (2022), "A ConvNet for the 2020s", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: <https://doi.org/10.1109/CVPR52688.2022.01167>
  6. Cheng, H., Zhang, M., Shi, J. Q. (2024), "A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations", *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI: <https://doi.org/10.1109/TPAMI.2024.3447085>
  7. Li, H., Kadav, A., Durdanovic, I., Samet, H., Graf, H. P. (2017), "Pruning filters for efficient convnets", *Proceedings of the International Conference on Learning Representations (ICLR)*. DOI: <https://doi.org/10.48550/arXiv.1608.08710>
  8. Molchanov, P., Mallya, A., Tyree, S., Frosio, I., Kautz, J. (2019), "Importance estimation for neural network pruning", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: <https://doi.org/10.1109/CVPR.2019.01161>
  9. He, Y., Liu, P., Wang, Z., Hu, Z., Yang, Y. (2019), "Filter pruning via geometric median for deep convolutional neural networks acceleration", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: <https://doi.org/10.1109/CVPR.2019.00447>
  10. Fang, G., Ma, X., Song, M., Mi, M. B., Wang, X. (2023), "DepGraph: Towards any structural pruning", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: <https://doi.org/10.1109/CVPR52729.2023.01544>
-

11. Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., Zhang, C. (2017), "Learning efficient convolutional networks through network slimming", *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. DOI: <https://doi.org/10.1109/ICCV.2017.298>
12. He, Y., Lin, J., Liu, Z., Wang, H., Li, L.-J., Han, S. (2018), "AMC: AutoML for model compression and acceleration on mobile devices", *Proceedings of the European Conference on Computer Vision (ECCV)*. DOI: [https://doi.org/10.1007/978-3-030-01234-2\\_48](https://doi.org/10.1007/978-3-030-01234-2_48)
13. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H. (2017), "MobileNets: Efficient convolutional neural networks for mobile vision applications", *arXiv preprint*. DOI: <https://doi.org/10.48550/arXiv.1704.04861>
14. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C. (2018), "MobileNetV2: Inverted residuals and linear bottlenecks", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: <https://doi.org/10.1109/CVPR.2018.00474>
15. Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y. M. (2023), "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: <https://doi.org/10.1109/CVPR52729.2023.00721>

Received (Надійшла) 18.03.2026

Accepted for publication (Прийнята до друку) 07.05.2026

Publication date (Дата публікації) 29.05.2026

#### Відомості про авторів / About the Authors

**Нечипуренко Станіслав Ігорович** – Харківський національний університет радіоелектроніки, аспірант кафедри електронних обчислювальних машин; Харків, Україна;

**Stanislav Nechypurenko** – Kharkiv National University of Radio Electronics, Postgraduate Student, Department of Electronic Computers; Kharkiv, Ukraine;

e-mail: stanislav.nechypurenko@nure.ua

ORCID ID: <https://orcid.org/0009-0007-2936-9935>

**Сорокін Антон Романович** – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, доцент кафедри електронних обчислювальних машин; Харків, Україна;

**Anton Sorokin** – Phd (Technical Sciences), Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor Department of Electronic Computers; Kharkiv, Ukraine;

e-mail: anton.sorokin@nure.ua

ORCID ID: <https://orcid.org/0000-0002-4383-2611>

## МЕТОД СТРУКТУРНОГО ПРУНІНГУ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ З ОГЛЯДУ НА АПАРАТНІ ХАРАКТЕРИСТИКИ EDGE-ПРИСТРОЇВ

**Предмет дослідження** – методи структурного прунінгу згорткових нейронних мереж, що застосовуються для зменшення обчислювальної складності моделей детекції об'єктів у відеосистемах з локальним обробленням даних.

**Мета роботи** – вдосконалення методу структурного прунінгу способом упровадження апаратно-орієнтованого критерію відбору каналів, який зважає на реальну латентність операцій на цільовому edge-пристрої.

**Завдання дослідження:** порівняльний аналіз наявних критеріїв структурного прунінгу, обґрунтування вибору базового методу й розроблення його модифікації на основі профілювання латентності цільової апаратної платформи.

**Методи дослідження** ґрунтовані на математичному аналізі критеріїв важливості каналів згорткових шарів, теоретичному порівнянні підходів до оцінювання надлишковості параметрів нейронної мережі та формалізації задачі апаратно-орієнтованої оптимізації. **Результати дослідження** полягають у розробленні модифікованого критерію важливості каналів, що поєднує оцінку впливу каналу на функцію втрат із нормалізованим коефіцієнтом латентності відповідного шару, який обчислюється способом профілювання часу виконання кожного згорткового шару на цільовому edge-пристрої. Запропонований критерій дає змогу пріоритизувати видалення каналів у шарах, що створюють найбільше навантаження на конкретному апаратному забезпеченні, водночас зберігаючи канали

з високим впливом на точність моделі. Виконано теоретичний аналіз властивостей критерію, зокрема доведено його зведення до базового критерію Тейлора за умови нульового значення гіперпараметра та монотонність перерозподілу пріоритетів прунінгу між шарами. Продемонстровано, що запропонований підхід є модульним, потребує мінімальних додаткових обчислювальних витрат і може бути інтегрований у наявні конвеєри оптимізації моделей. Експериментальна перевірка на моделі детектора YOLOv8n підтвердила запропонований підхід: за бюджету прунінгу 30% вилучених каналів отримано зниження латентності інференсу на 38,55% проти 34,88% у базового критерію Тейлора, що відповідає відносній перевазі +10,5%. **Висновки.** Результати підтверджують теоретичну обґрунтованість запропонованого підходу та його переваги порівняно з апаратно-незалежними критеріями прунінгу для задач розгортання моделей детекції на edge-пристроях у складі відеосистем реального часу.

**Ключові слова:** структурний прунінг; згорткова нейронна мережа; edge-пристрій; латентність інференсу; критерій важливості каналів; оптимізація моделі; відеоаналітика; детекція об'єктів; локальне оброблення даних; апаратно-орієнтована оптимізація.

#### Бібліографічні описи / Bibliographic descriptions

Нечипуренко С. І., Сорокін А. Р. Метод структурного прунінгу згорткових нейронних мереж з огляду на апаратні характеристики edge-пристроїв. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 34–46. DOI: <https://doi.org/10.30837/0135-1710.2026.189.034>

Nechypurenko, S., Sorokin, A. (2026), "Method of structural pruning of convolutional neural networks with regard to edge device hardware characteristics", *Management Information System and Devices*, No. 2 (189), P. 34–46. DOI: <https://doi.org/10.30837/0135-1710.2026.189.034>

---



Роговий А. І., Ахієзер О. Б., Решетнікова С. М., Колбасін В. О., Пахомов Ю. В.

## ІНТЕЛЕКТУАЛЬНА МОДЕЛЬ ФОРМУВАННЯ ІННОВАЦІЙНО-ІНЖИНІРИНГОВИХ КЛАСТЕРІВ

У сучасних умовах цифровізації та розвитку економіки знань інноваційно-інжинірингові кластери трансформуються в складні динамічні системи з високим рівнем невизначеності та багатофакторних взаємодій і значною залежністю від зовнішніх управлінських впливів. Традиційні економіко-математичні підходи не забезпечують достатньої гнучкості для моделювання таких систем, що зумовлює необхідність упровадження інтегрованих методів штучного інтелекту. **Мета дослідження** – розробити й апробувати гібридну модель формування інноваційно-інжинірингових кластерів, яка поєднує методи машинного навчання, графового аналізу, агентно орієнтованого моделювання й цифрових двійників, з подальшим оцінюванням синергетичного ефекту їх функціонування й аналізом впливу управлінських факторів на динаміку розвитку кластерів. У роботі використано багаторівневий підхід, що передбачає кластеризацію агентів (*k-means*, DBSCAN), побудову графової моделі взаємодій, застосування графових нейронних мереж для врахування топологічних залежностей, а також агентно орієнтованого моделювання для дослідження еволюції системи в часі. Додатково впроваджено функції сумісності агентів і ризику, що дають змогу формалізувати синергетичний ефект і стійкість кластера в межах оптимізаційної постановки. Усі компоненти інтегруються у вигляді цифрового двійника, що забезпечує можливість сценарного аналізу. **Результати дослідження.** Модель апробовано на змодельованому регіональному кластері, що охоплює підприємства, університети, інвесторів і державний орган. Результати симуляції продемонстрували, що державна підтримка суттєво підвищує інноваційну активність учасників кластера, збільшує щільність мережевих зв'язків і кількість коопераційних проєктів. Виявлено ключові вузли мережевої взаємодії, зокрема значення університетів як центрів трансферу знань та інвесторів, як фінансових хабів. **Висновки.** Результати дослідження підтверджують ефективність запропонованого підходу та його придатність для прогнозування розвитку інноваційних екосистем і підтримки прийняття стратегічних управлінських рішень. Розроблена модель забезпечує комплексне поєднання методів машинного навчання, графового аналізу, агентного моделювання й цифрових двійників, що дає змогу перейти від статичного аналізу кластерів до їх адаптивного й прогностичного управління. **Практична цінність** дослідження полягає в можливості використання запропонованого підходу для підтримки регіональної інноваційної політики, оцінювання ефективності управлінських рішень і формування стійких інноваційно-інжинірингових екосистем.

**Ключові слова:** інноваційно-інжинірингові кластери; штучний інтелект; машинне навчання; графовий аналіз; графові нейронні мережі; агентно орієнтоване моделювання; цифровий двійник.

### Вступ

Сучасна економіка знань розвивається в умовах глобалізації, цифровізації та швидкого поширення технологічних інновацій. Виробничі й економічні системи стають дедалі складнішими та вимагають нових підходів до аналізу, планування й управління. Якщо раніше наголошувалася на економіко-математичних моделях кластеризації, то нині дедалі більшої ваги набувають методи комп'ютерних наук – від алгоритмів машинного навчання до мультиагентних симуляцій і цифрових двійників. Це зумовлено стрімким зростанням обсягів інформації, необхідністю аналізувати багатофакторні взаємодії та потребою в інструментах, здатних підтримувати прийняття рішень у реальному часі.

Промислові кластери, що є сукупністю взаємопов'язаних підприємств, інституцій,

дослідницьких центрів і державних структур, у сучасних умовах трансформуються в складні кіберфізичні системи. Така трансформація супроводжується інтеграцією інтернету речей (IoT), великих даних (*Big Data*), штучного інтелекту (AI) і технологій хмарних обчислень. У цьому контексті кластери можуть розглядатися як динамічні мережі, що визначаються нелінійними взаємозв'язками, високим ступенем невизначеності та необхідністю швидкої адаптації до змін зовнішнього середовища.

Останнім часом зростає інтерес до використання графових моделей, які дають змогу формалізувати структуру кластерів через систему вузлів (підприємства, університети, органи влади, інвестори) та ребер (партнерські відносини, трансфер технологій, фінансові потоки). Водночас агентно орієнтовані моделі (ABM) створюють умови для моделювання поведінки окремих учасників кластера й дослідження того,

як локальні взаємодії призводять до макроскопічних ефектів. Ще одним перспективним напрямом є застосування цифрових двійників (*Digital Twins*), які допомагають відтворювати функціонування промислових кластерів у кіберпросторі, проводити симуляції *what-if* і розробляти сценарії стратегічного розвитку.

У сучасній економіці знань інноваційно-інжинірингові кластери трансформуються в складні кіберфізичні системи, що визначаються нелінійними взаємодіями, високою динамічністю та невизначеністю. Традиційні економіко-математичні підходи не дають змоги адекватно моделювати такі системи, особливо в умовах великих даних і багаторівневих взаємозв'язків.

У зв'язку з цим актуальним є застосування гібридних моделей, що інтегрують:

- методи машинного навчання;
- графовий аналіз;
- агентно орієнтоване моделювання (*Agent-Based Modeling*, ABM), яке дає змогу описувати автономну поведінку й адаптивну взаємодію учасників кластера;
- цифрові двійники (*Digital Twin*).

Однак наявні підходи здебільшого розглядають ці компоненти ізольовано, що обмежує їх здатність до комплексного аналізу й прогнозування.

#### Аналіз останніх досліджень і публікацій

Проблематика формування інноваційних кластерів набула значної уваги в сучасній науковій літературі. У роботі [1] запропоновано алгоритмічне визначення "кластерів споріднених індустрій" на основі ко-локалізації та економічних зв'язків, що продемонструвало прямий зв'язок між функціонуванням кластерів і зростанням зайнятості та рівнем патентування. Автори дослідження [2] проаналізували застосування графових методів для виявлення ієрархічних структур у міжнародній торгівлі та підтвердили ефективність спектральної кластеризації для виділення стійких промислових спільнот. Подібні висновки отримано, зокрема, у праці [3], де на основі бази WIOD автори продемонстрували стійкість виявлених індустріальних кластерів до шуму та зміни даних. Дослідження [4] присвячено визначенню "ключових секторів" у глобальних мережах торгівлі способом аналізу центральностей, що сприяло оцінюванню системної вразливості економічних структур. У роботі [5] доведено, що спільноти можуть бути відновлені навіть

без явних зв'язків, лише на основі часових слідів активності, що відкриває нові перспективи для аналізу транзакційних та інформаційних даних.

Окремий напрям досліджень становлять агентно орієнтовані моделі (ABM). Так, автори [6] розробили модель виникнення високотехнологічних кластерів, у якій локальні взаємодії фірм породжують макроекономічні ефекти концентрації. У дослідженні [7] описано вивчення процесу дифузії технологічних інновацій у китайських регіонах і виявлено порогові ефекти, пов'язані з масштабом мережі та швидкістю обміну знаннями. Автори праці [8] продемонстрували роль контрактів у розвитку індустріального симбіозу, що забезпечує стабільність виробничих обмінів. Аналогічно в роботі [9] застосовано ABM для аналізу формування міжфірмової довіри й знанневих потоків у кластерах. У наступних дослідженнях, зокрема [10], інтегровано агентні моделі з внутрішньофірмовим *input-output*-аналізом, доведено доцільність їх використання для оцінювання компромісів між економічними й екологічними цілями, тоді як автори роботи [11] наголошують на значенні поведінкових факторів у робастності симбіозних мереж.

Як продемонстрував аналіз сучасних досліджень, агентно орієнтоване моделювання є особливо ефективним для опису складних соціально-економічних систем, у яких учасники мають автономну поведінку, власні цілі та здатні адаптуватися до змін середовища. Саме тому використання агентного підходу є доцільним для дослідження інноваційно-інжинірингових кластерів.

Важливим доповненням до агентних моделей стають цифрові двійники (*Digital Twin*, DT). Так, наприклад, у роботі [12] запропоновано концепцію *Digital Twin Learning Ecosystem* для традиційних виробничих систем, що поєднує моніторинг, навчання й оптимізацію. Автори [13] реалізували мультиагентну систему з цифровим двійником для децентралізованого диспетчерування на виробничій фабриці, що сприяло зростанню ефективності прийняття рішень. У праці [14] запропоновано платформу DT для індустріальних парків із функціями моніторингу енергоспоживання й оцінювання сценаріїв зменшення CO<sub>2</sub>. На концептуальному рівні автори статті [15] підтвердили, що використання динамічних знанневих графів узгоджується з архітектурою універсального цифрового двійника, спрощуючи інтеграцію даних у промислових системах.

На окрему увагу заслуговує застосування методів графових нейронних мереж (*Graph Neural Networks*, GNN). Так, наприклад, у роботі [16] продемонстровано переваги GNN у прогнозуванні двосторонніх торгових потоків порівняно з класичними гравітаційними моделями. Автори ж статті [17] поєднали знання графів з нейросимвольним навчанням для виявлення ризиків у глобальних ланцюгах постачання. У дослідженні [18] розроблено ієрархічну GNN для оцінювання ризиків у ланцюгах постачання, яка перевершила традиційні алгоритми на реальних даних.

У роботі [19] розглянуто застосування інтелектуальних методів аналізу даних для дослідження складних соціально-економічних систем, зокрема із використанням підходів машинного навчання й мережевого аналізу. Автори приділяють увагу інтеграції структурних і поведінкових властивостей систем, що дає змогу підвищити точність моделювання взаємодій між елементами. Досягнуті результати підтверджують ефективність упровадження гібридних AI-підходів для аналізу динамічних систем, що узгоджується з напрямом дослідження, запропонованим у цій роботі.

Отже, сучасні розвідки демонструють перехід від класичних економіко-математичних моделей до комплексних підходів, що поєднують графову теорію, агентні моделі й цифрові двійники. Це свідчить про високу актуальність наступних робіт у галузі комп'ютерних наук, які дають змогу створювати інтегровані системи прогнозування, моніторингу й управління розвитком інноваційних кластерів.

#### Мета й завдання публікації

Метою дослідження є розроблення й формалізація моделі нейромережевого планування переміщення кіберфізичної системи на основі принципів аферентного синтезу, що забезпечує інтеграцію сенсорної інформації, формування оптимального напрямку руху й адаптивне функціонування в умовах невизначеного й динамічного зовнішнього середовища.

Для досягнення поставленої мети необхідно розв'язати такі завдання:

1) проаналізувати підходи до формування інформації про зовнішнє середовище КФС на основі даних системи технічного зору;

2) запропонувати спосіб подання зовнішнього середовища у вигляді дискретної моделі, придатної для нейромережевого оброблення;

3) сформулювати задачу планування переміщення КФС у формалізованому вигляді;

4) розробити нейромережеву структуру для інтеграції сенсорних показників і визначення напрямку руху;

5) описати принцип функціонування нейромережевого планувальника в процесі керування переміщенням кіберфізичної системи.

#### Мета дослідження

Метою є розроблення й апробування гібридної моделі формування інноваційно-інжинірингових кластерів, що поєднує методи машинного навчання, графового аналізу, агентно орієнтованого моделювання та цифрових двійників, з подальшим оцінюванням синергетичного ефекту їх функціонування й аналізом впливу управлінських і політичних факторів на динаміку розвитку кластерів.

#### Матеріали й методи дослідження

Для побудови гібридної моделі інноваційно-інжинірингового кластера впроваджено багаторівневий методологічний підхід, що поєднує методи машинного навчання, аналізу графів, агентного моделювання й цифрових двійників.

Розроблений підхід можна умовно подати у вигляді послідовності кількох етапів.

1) Збір і підготовка інформації: формування інтегрованої бази показників підприємств, університетів, інвесторів і державних інституцій; нормалізація та очищення даних.

2) Кластеризація та факторний аналіз: застосування алгоритмів машинного навчання (*k-means*, DBSCAN, ієрархічні методи, *spectral clustering*) для виділення стійких угруповань підприємств; використання PCA та *autoencoders* з метою скорочення розмірності.

3) Графове моделювання: побудова мережевої моделі кластера, де вузли позначають учасників, а ребра – зв'язки між ними; визначення центральних вузлів і підкластерів.

4) Агентне моделювання (ABM): створення симуляції взаємодії між підприємствами, університетами, інвесторами й державою для аналізу динаміки кластера в часі.

5) Інтеграція та сценарне моделювання кластера: об'єднання результатів кластеризації, групового аналізу й агентного моделювання в єдину інтегровану модель інноваційно-інжинірингового кластера.

6) Візуалізація та прийняття рішень: подання результатів у вигляді інтерактивних графів, таблиць і панелей для підтримки стратегічного управління.

Запропонований підхід відтворює послідовність кроків від збору даних до створення цифрового двійника, який слугує інтегрованим інструментом для оцінювання та прогнозування розвитку кластера.

У роботі використовується агентно орієнтований підхід до моделювання інноваційно-інжинірингових кластерів. На відміну від класичного об'єктного подання, де елементи системи переважно описуються як пасивні структури даних із фіксованими властивостями, агентний підхід дає змогу моделювати автономну поведінку учасників системи, їх адаптацію до змін середовища, прийняття рішень і динамічну взаємодію між собою.

У межах дослідження агентами є підприємства, університети, інвестори й державні органи, які здатні змінювати свою поведінку залежно від зовнішніх факторів, доступних ресурсів, рівня кооперації та управлінських впливів. Такий підхід дає змогу перейти від статичного опису структури кластера до моделювання процесів його розвитку в часі.

Використання агентного моделювання забезпечує вивчення впливів, коли глобальна поведінка кластера формується внаслідок локальних взаємодій окремих учасників. Це є особливо важливим для інноваційних екосистем, де результати їх функціонування значною мірою визначаються кооперацією, конкуренцією та адаптивністю учасників.

Інноваційно-інжиніринговий кластер у роботі розглядається як складна динамічна система, що містить множину автономних елементів моделі (агентів – підприємств, університетів, інвесторів, державних органів), взаємодія між якими визначає рівень інноваційної активності та синергетичного ефекту.

Структура кластера формалізується у вигляді зваженого графа:

$$G = (V, E, W),$$

де  $V = \{v_1, v_2, \dots, v_n\}$  – множина агентів (підприємства, університети, інвестори);  $E \subseteq V \times V$  – множина зв'язків;  $W = \{w_{ij}\}$  – ваги зв'язків (інтенсивність співпраці).

Такий граф задає структурну основу кластера, на якій будуються подальші аналітичні та імітаційні процедури.

Кожний агент  $v_i \in V$  описується вектором характеристик:

$$v_i = \{r_i, c_i, p_i, h_i\},$$

де  $r_i$  – інтенсивність інвестицій у R&D;  $c_i$  – рівень коопераційної активності;  $p_i$  – кількість спільних проєктів;  $h_i$  – кадровий потенціал.

Цей вектор формує простір станів системи, який використовується в задачах кластеризації, графовому аналізі та динамічному моделюванні.

Ключовою характеристикою ефективності кластера є синергетичний ефект, що виникає внаслідок взаємодії агентів. Його визначаємо таким чином:

$$S(G) = \sum_{i=1}^n \sum_{j=1}^m w_{ij} \cdot \varphi(x_i, x_j),$$

де  $w_{ij}$  – інтенсивність взаємодії між агентами  $i$  та  $j$ ;  $\varphi(x_i, x_j)$  – функція сумісності агентів.

Функцію  $\varphi(x_i, x_j)$  долучаємо для кількісної оцінки взаємодії між агентами:

$$\varphi(x_i, x_j) = \alpha \cdot \text{sim}(x_i, x_j) + \beta \cdot \text{comp}(x_i, x_j),$$

де  $\text{sim}(x_i, x_j)$  – міра подібності агентів;  $\text{comp}(x_i, x_j)$  – міра компліментарності;

$$\alpha, \beta \in [0, 1], \quad \alpha + \beta = 1.$$

Показник міри подібності відбиває схожість стратегій розвитку агентів. Для його розрахунку використовуємо косинусну подібність:

$$\text{sim}(x_i, x_j) = \frac{x_i \cdot x_j}{|x_i| \cdot |x_j|}.$$

Компліментарність визначається як різниця інноваційних характеристик і дає змогу зважати на ефект доповнюваності (наприклад, стартап + великий бізнес):

$$\text{comp}(x_i, x_j) = 1 - \frac{|r_i - r_j|}{\max(r)},$$

де  $r_i, r_j$  – інвестиції в R&D;  $\max(r)$  – максимальне значення у вибірці.

Отже,  $S(G)$  позначає інтегральний ефект функціонування кластера. Для оцінювання стійкості кластера додано функцію ризиків  $R(G)$ , яка має такий вигляд:

$$R(G) = \gamma \cdot (1 - D(G)) + \delta \cdot F(G),$$



де  $D(G) = \frac{2|E|}{|V|(|V|-1)}$  – щільність графа (низька щільність означає слабку взаємодію і, як наслідок, високий ризик);  $F(G) = 1 - |C_{\max}|/|V|$  – фрагментація мережі;  $C_{\max}$  – найбільша зв'язана компонента.

Щоб взяти до уваги складну топологію мережі та нелінійні залежності, впроваджується модель графових нейронних мереж (GNN):

$$h_i^{(k+1)} = \sigma \left( \sum_{j \in N(i)} w_{ij} \cdot W^k h_j^k \right),$$

де  $h_j^k$  – подання агента на  $k$ -му шарі;  $N(i)$  – множина сусідніх агентів;  $W^k$  – параметри моделі;  $\sigma$  – нелінійна функція активації.

GNN дає змогу зважати на мережеві ефекти, покращити ролі агентів у кластері.

Еволюція кластера в часі описується як дискретний динамічний процес:

$$x_i(t+1) = f(x_i(t), N_i(t), u(t)),$$

де  $f$  – функція, що описує формування нових зв'язків, зміну рівня інноваційності, адаптацію агентів до зовнішніх умов;  $x_i(t)$  – стан агента в момент часу  $t$ ;  $N_i(t)$  – сусідні агенти;  $u(t)$  – керівні впливи (державна політика, інвестиції).

Формування ефективного кластера розглядається як задача оптимізації:

$$\max_G \left( \sum_{i,j} w_{ij} \cdot \varphi(x_i, x_j) \right) - \lambda R(G),$$

де  $\lambda$  – коефіцієнт, що визначає компроміс між максимізацією синергетичного ефекту й мінімізацією ризиків. У разі нормалізації функції  $S(G)$  та  $R(G)$  до інтервалу  $[0,1]$  значення параметра  $\lambda$  доцільно обмежити таким само діапазоном.

У процесі аналізу встановлено, що параметр  $\lambda$  виконує роль регулятора кластера й сприяє моделюванню різних сценаріїв політики, наприклад, інноваційно-орієнтованому або стабілізаційному.

Отже, запропонований агентно орієнтований підхід дає змогу моделювати адаптивну поведінку окремих учасників кластера й досліджувати формування макроекономічних ефектів унаслідок їх локальних взаємодій.

Для переходу від статичного аналізу до динамічного й прогнозного управління складними

соціально-економічними системами в моделюванні інноваційно-інжинірингових кластерів використовують цифрові двійники. Вони допомагають інтегрувати структурну модель кластера (граф взаємодій)

із поведінковими характеристиками агентів і зовнішніми управлінськими впливами, забезпечуючи можливість проведення сценарного (*what-if*) аналізу в реальному або наближеному до реального часі. Це дає змогу оцінювати наслідки управлінських рішень, змін у політиці чи ринкових умовах без ризику для реальної системи, а також підвищує обґрунтованість стратегічного планування розвитку кластерів.

Тому всі попередні компоненти інтегруються в цифровий двійник:

$$DT(t) = (G(t), X(t), U(t)),$$

де  $G(t)$  – структура кластера;  $X(t) = \{x_i\}$  – стани агентів;  $U(t)$  – керівні впливи.

Розроблену модель запропоновано реалізовувати в три взаємопов'язані фази.

1. Аналітико-математична фаза. На основі агрегованих статистичних і експертних показників будується матриця ознак учасників кластера. Застосовується факторний аналіз для виявлення ключових груп змінних, які визначають розвиток кластера (підприємницький клімат, науковий потенціал, державна підтримка тощо).

2. Обчислювально-імітаційна фаза. Формується початкова структура кластера з використанням алгоритмів кластеризації та методів аналізу графів. Далі за допомогою агентно орієнтованого моделювання імітується розвиток системи:

- поява нових підприємств;
- формування коопераційних зв'язків;
- вплив державної політики;
- дифузія інновацій, унаслідок якої створюється

динамічна модель, здатна відтворювати поведінку реальних учасників.

3. Інтеграційно-прогнозна фаза. Побудований кластер віртуалізується у вигляді цифрового двійника, який відтворює як поточний стан системи, так і можливі сценарії розвитку. У цифровому двійнику інтегруються дані з різних джерел у реальному часі, що дає змогу проводити *what-if*-симуляції, прогнозувати результати управлінських рішень і оцінювати ризики.

Унаслідок використання запропонованої моделі маємо такі результати: кластерні карти регіонів,



графові структури взаємодії, імітаційні сценарії розвитку, цифровий двійник кластера з можливістю інтерактивного аналізу.

Для апробування розробленої моделі проведено імітаційне моделювання функціонування регіонального інноваційно-інжинірингового кластера. У межах симуляції сформовано модель кластерної екосистеми, яка містить:

- десять підприємств (виробничі та високотехнологічні компанії);
- два університети (освітньо-наукові центри);
- два інвестори (фонди венчурного капіталу);
- один державний орган (регіональна адміністрація).

Така структура дала змогу відтворити типову модель взаємодії ключових учасників інноваційного середовища й проаналізувати вплив коопераційних зв'язків і управлінських факторів на динаміку розвитку кластера.

Незважаючи на відносно незначну кількість агентів у змодельованому кластері, така конфігурація є репрезентативною та типовою для пілотного моделювання інноваційних екосистем. Вибір обмеженої кількості учасників зумовлений необхідністю забезпечення прозорості інтерпретації результатів, відстеження взаємодій між агентами та валідації поведінкових закономірностей моделі. Запропонована структура кластера відтворює ключові типи учасників реальних інноваційних систем (підприємства, університети, інвестори, державні інституції) та базові зв'язки між ними, що дає підстави вважати її узагальненою моделлю.

Важливо наголосити, що розроблена інтелектуальна модель не має обмежень щодо розміру системи: завдяки використанню графового подання, агентно орієнтованого підходу й алгоритмів машинного навчання вона може бути легко масштабована до значно більшої кількості агентів без зміни базових принципів функціонування. Отже, досягнуті результати можуть бути екстрапольовані на реальні регіональні або національні кластери значного масштабу.

Для кожного учасника бралися до уваги такі показники: обсяг інвестицій в інновації, рівень співпраці з іншими організаціями, кількість спільних проектів, кадровий потенціал.

#### Етап 1. Кластеризація підприємств

За допомогою алгоритму *k-means* підприємства було поділено на три підкластерні групи:

- 1) інноваційні виробники – підприємства з високими витратами на R&D;
- 2) традиційні виробники – підприємства з низьким рівнем інноваційності, але стійкими економічними показниками;
- 3) стартапи – малі компанії, орієнтовані на швидке зростання та співпрацю з університетами.

#### Етап 2. Графова модель кластера

Інноваційно-інжиніринговий кластер подано на рис. 1.

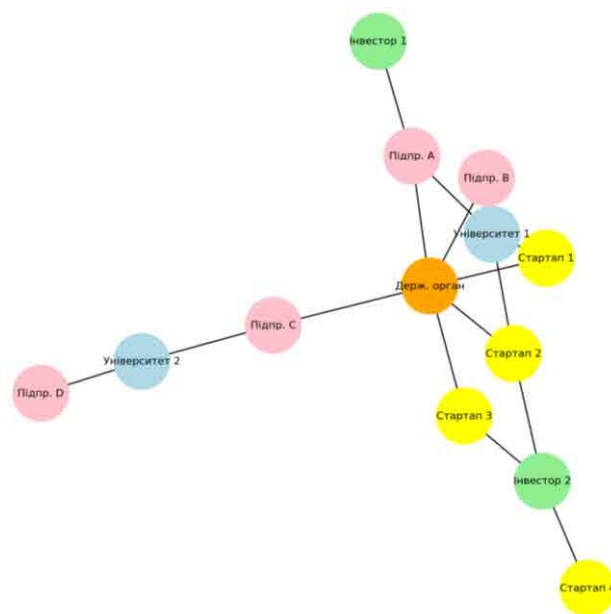


Рис. 1. Графова модель інноваційно-інжинірингового кластера

На основі рис. 1 можна зробити такі висновки:

- найбільш центральним вузлом виявився університет 1 (висока *betweenness centrality*), оскільки він пов'язує стартапи з великими виробниками;
- інвестор 2 утворює "хаб фінансування", підтримуючи відразу три стартапи;
- державний орган з'єднаний з усіма підприємствами через програми підтримки, але має низьку інтенсивність зв'язків.

#### Етап 3. Агентно орієнтована симуляція

Проведено симуляцію на 10 періодах (умовно – 10 років), унаслідок якої досягнуто таких результатів:

- за умови державних грантів частка підприємств, що впроваджують інновації, зросла з 40% до 70%;
- за відсутності такої підтримки рівень інноваційності залишався на рівні 45%;

– мережа знанневих зв'язків стала більш щільною: середній коефіцієнт кластеризації графа зріс із 0,32 до 0,55.

#### Етап 4. Цифровий двійник

На основі отриманих показників сформовано цифровий двійник кластера, який дає змогу:

– моделювати сценарії ("що буде, якщо інвестор піде з ринку", "що буде, якщо університет відкриє нову лабораторію");

Таблиця 1. Результати симуляції роботи моделі

| Показник                            | Початкове значення | Через 10 періодів<br>(з держпідтримкою) | Через 10 періодів<br>(без підтримки) |
|-------------------------------------|--------------------|-----------------------------------------|--------------------------------------|
| Частка підприємств із R&D           | 40%                | 70%                                     | 45%                                  |
| Середня кількість партнерств / фірм | 2,1                | 4,3                                     | 2,5                                  |
| Коефіцієнт кластеризації графа      | 0,32               | 0,55                                    | 0,36                                 |
| Кількість спільних проєктів         | 15                 | 34                                      | 18                                   |

– оцінювати зміни в продуктивності та кількості інноваційних проєктів;

– візуалізувати вплив політики на рівень співпраці. Результати симуляції роботи моделі подано в табл. 1.

Для більшої наочності результати симуляції подано на рис. 2.

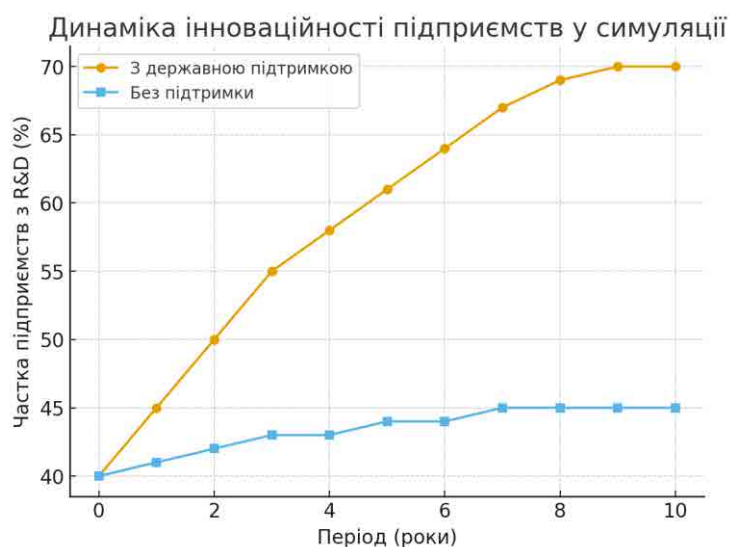


Рис. 2. Результати симуляції роботи запропонованої моделі

Отже, на основі проведеної симуляції можна зробити висновки.

1. Використання графових моделей дає змогу ідентифікувати ключових гравців (університети, інвестори).

2. Агентні симуляції показали значний вплив державної політики на інноваційну активність.

3. Цифровий двійник забезпечує інструментарій для прогнозування й візуалізації сценаріїв розвитку.

4. Досягнуті результати підтверджують ефективність інтеграції методів комп'ютерних наук у дослідженні промислових кластерів.

#### Результати дослідження

У процесі впровадження розробленої інтелектуальної гібридної моделі інноваційно-

інжинірингових кластерів досягнуто певних результатів.

1. Застосування алгоритмів кластеризації (*k-means*, DBSCAN) дало змогу виокремити три основні групи: інноваційні виробники, традиційні виробники й стартапи. Це дало змогу визначити їх особливі ролі у функціонуванні кластера.

2. Аналіз центральностей продемонстрував, що університети виконують роль "посередників знань", з'єднуючи стартапи й великі підприємства. Інвестори формують фінансові хаби, а державні органи забезпечують загальну інституційну підтримку.

3. Імітаційна симуляція на горизонті 10 періодів продемонструвала, що державна підтримка (гранти, податкові пільги) суттєво підвищує частку підприємств, які інвестують у R&D (з 40% до 70%), тоді як без підтримки цей показник зростає лише

до 45%. Також зафіксовано збільшення коефіцієнта кластеризації мережі з 0,32 до 0,55, що свідчить про посилення взаємодії між учасниками.

4. Створено прототип цифрового двійника кластера, який сприяє об'єднанню реальних показників (статистичних, експертних, операційних) з результатами симуляцій. Це забезпечує можливість моделювати сценарії розвитку (наприклад, вихід інвестора з ринку чи поява нових дослідницьких лабораторій) і оцінювати їх наслідки для інноваційної активності.

5. Застосування запропонованої методології підтвердило її ефективність: цифрова репрезентація кластера здатна відтворювати не лише статичний стан, а й динаміку взаємодії в реальному й прогнозованому часовому горизонті.

### Висновки

Розроблено гібридну модель інноваційно-інжинірингового кластера, яка поєднує методи машинного навчання, аналізу графів, агентного моделювання та цифрових двійників. Проведений експериментальний приклад продемонстрував, що впровадження запропонованого підходу дає змогу не лише виявити структуру й ключових учасників кластера, але й оцінити вплив управлінських рішень і політики на інноваційний розвиток. Модель підтвердила, що державна підтримка значно підсилює синергетичний ефект кластера: підвищується рівень інноваційності підприємств, зростає щільність мережових зв'язків і збільшується кількість спільних проєктів. Використання цифрового двійника забезпечує новий рівень інтеграції даних і симуляцій, створюючи інструмент для прогнозування сценаріїв розвитку регіональних інноваційних систем.

Досягнуті результати підтверджують високу ефективність застосування сучасних математичних методів для дослідження й розвитку інноваційно-

інжинірингових кластерів, що є особливо актуальним для України з огляду на відновлення та інтеграцію до європейського економічного простору.

Подальші дослідження необхідно спрямувати на розширення запропонованої моделі за допомогою інтеграції поточкових даних у реальному часі, використання більш складних архітектур графових нейронних мереж і методів глибокого навчання для прогнозування еволюції кластерів. Окрему увагу доцільно приділити практичній реалізації цифрового двійника регіонального інноваційно-інжинірингового кластера з інтеграцією відкритих державних, економічних і виробничих показників для підтримки стратегічного управління розвитком регіональних інноваційних екосистем.

### Конфлікт інтересів

Автор декларує, що не має конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

### Фінансування

Дослідження проводилося без фінансової підтримки.

### Доступність даних

Рукопис не має пов'язаних матеріалів у сховищі даних.

### Використання засобів штучного інтелекту

Автор підтверджує, що не застосовував технології ШІ для написання роботи.

### References

1. Delgado, M., Porter, M. E., Stern, S. (2016), "Defining clusters of related industries", *Journal of Economic Geography*, Vol. 16(1), pp. 1–38. DOI: <https://doi.org/10.1093/jeg/lbv017>
2. Bouet, A., Klimenko, M. (2019), "Graph clustering in industrial networks", *IMA Journal of Applied Mathematics*, Vol. 84(6), pp. 1177–1202. DOI: <https://doi.org/10.1093/imamat/hxz028>
3. Kireyev, A., et al. (2022), "Community detection in international production networks", *PLOS ONE*, Vol. 17(3), e0264623. DOI: <https://doi.org/10.1371/journal.pone.0264623>
4. De Paolis, L. T., et al. (2022), "Identifying key sectors in global trade networks", *Applied Network Science*, Vol. 7, p. 19. DOI: <https://doi.org/10.1007/s41109-022-00519-2>
5. Peel, L., Delvenne, J. C., Lambiotte, R. (2020), "Community detection without edges", *Science Advances*, Vol. 6(19), eaav1478. DOI: <https://doi.org/10.1126/sciadv.aav1478>

6. Zhang, J., Li, X. (2003), "Agent-based modeling of high-tech industrial clusters", *Journal of Evolutionary Economics*, Vol. 13(5), pp. 529–548. DOI: <https://doi.org/10.1007/s00191-003-0178-4>
7. Wang, J., et al. (2014), "Diffusion of technological innovations in Chinese regions: An agent-based model", *Papers in Regional Science*, Vol. 93(1), pp. 131–151. DOI: <https://doi.org/10.1111/pirs.12069>
8. Albino, V., Fraccascia, L., Giannoccaro, I. (2016), "Exploring the role of contracts in industrial symbiosis", *Journal of Cleaner Production*, Vol. 112, pp. 4353–4364. DOI: <https://doi.org/10.1016/j.jclepro.2015.06.070>
9. Zheng, L., et al. (2017), "Agent-based modeling of industrial symbiosis", *Sustainability*, Vol. 9(5), p. 765. DOI: <https://doi.org/10.3390/su9050765>
10. Yazan, D. M., et al. (2020), "Hybrid agent-based model of industrial symbiosis", *International Journal of Production Research*, Vol. 58(20), pp. 6188–6206. DOI: <https://doi.org/10.1080/00207543.2019.1590660>
11. Lange, S., Korevaar, G., Ohnishi, S. (2021), "Behavioral aspects in industrial symbiosis", *Journal of Artificial Societies and Social Simulation*, Vol. 24(3), p. 6.
12. García, F., Bregón, A., Martínez-Prieto, M. A. (2024), "A Digital Twin Learning Ecosystem for traditional manufacturing", *Internet of Things*, Vol. 23, p. 101094. DOI: <https://doi.org/10.1016/j.iot.2024.101094>
13. Siatras, V., et al. (2024), "Multi-agent digital twin for decentralized manufacturing scheduling", *Information*, Vol. 15(6), pp. 337. DOI: <https://doi.org/10.3390/info15060337>
14. Wang, J., et al. (2024), "Digital twin-based energy management in industrial parks", *Energies*, Vol. 17(23), p. 5972. DOI: <https://doi.org/10.3390/en17235972>
15. Akroyd, J., et al. (2021), "Dynamic knowledge graphs for universal digital twins", *Data-Centric Engineering*, Vol. 2, e6. DOI: <https://doi.org/10.1017/dce.2021.7>
16. Sellami, M., et al. (2024), "Graph neural networks for international trade prediction", *Big Data and Cognitive Computing*, Vol. 8(6), p. 65. DOI: <https://doi.org/10.3390/bdcc8060065>
17. Kosasih, F. D., et al. (2024), "Knowledge graphs and neurosymbolic AI for supply chain risk detection", *International Journal of Production Research*, Vol. 62(1), pp. 1–19. DOI: <https://doi.org/10.1080/00207543.2022.2100841>
18. Zhou, Z., et al. (2023), "Hierarchical graph neural networks for supply chain risk assessment", *arXiv preprint arXiv:2305.11291*. DOI: <https://doi.org/10.48550/arXiv.2311.04244>
19. Yanholenko, O., Grinchenko, M., Rohovyi, M., Yakovleva, O., Rogovyi, A. (2025), "Intelligent team work planning: a model for increasing sprint value", *Proceedings of the PhD Workshop on Artificial Intelligence in Computer Science at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025)*, pp. 134–149.

Received (Надійшло) 21.04.2026

Accepted for publication (Прийнята до друку) 14.05.2026

Publication date (Дата публікації) 29.05.2026

#### Відомості про авторів / About the Authors

**Роговий Антон Іванович** – кандидат технічних наук, Національний технічний університет "Харківський політехнічний інститут", доцент кафедри стратегічного управління, Харків, Україна;

**Anton Rogovyi** – PhD (Technical Sciences), National Technical University "Kharkiv Polytechnic Institute", Associate Professor of the Strategic Management Department, Kharkiv, Ukraine;

e-mail: [anton.rogovyi@khpi.edu.ua](mailto:anton.rogovyi@khpi.edu.ua)

ORCID ID: <https://orcid.org/0000-0002-8178-4585>

**Ахїєзер Олена Борисівна** – кандидат технічних наук, доцент, Національний технічний університет "Харківський політехнічний інститут", завідувачка кафедри комп'ютерної математики і аналізу даних, Харків, Україна;

**Olena Akhiezer** – Phd (Computer Sciences), Associate Professor, National Technical University "Kharkiv Polytechnic Institute", Head of the Department of Computer Mathematics and Data Analysis, Kharkiv, Ukraine;

e-mail: [Olena.Akhiezer@khpi.edu.ua](mailto:Olena.Akhiezer@khpi.edu.ua)

ORCID ID: <https://orcid.org/0000-0002-7087-9749>

**Решетнікова Світлана Миколаївна** – Національний технічний університет "Харківський політехнічний інститут", доцент кафедри комп'ютерної математики і аналізу даних, Харків, Україна;

**Svitlana Reshetnikova** – National Technical University "Kharkiv polytechnic institute", Associate Professor of the Department of Computer Mathematics and Data Analysis, Kharkiv, Ukraine;

e-mail: [Svetlana.Reshetnikova@khpi.edu.ua](mailto:Svetlana.Reshetnikova@khpi.edu.ua)

ORCID ID: <http://orcid.org/0009-0009-1435-8656>

**Колбасін Владислав Олександрович** – Національний технічний університет "Харківський політехнічний інститут", старший викладач кафедри комп'ютерної математики і аналізу даних, Харків, Україна;

**Vladyslav Kolbasin** – National Technical University "Kharkiv polytechnic institute", Senior Lecturer of the Department of Computer Mathematics and Data Analysis, Kharkiv, Ukraine;

e-mail: Vladyslav.Kolbasin@khpi.edu.ua

ORCID ID: <http://orcid.org/0009-0007-0835-2712>

**Пахомов Юрій Васильович** – кандидат технічних наук, доцент, Харківський національний університет міського господарства, доцент кафедри комп'ютерних наук та інформаційних технологій, Харків, Україна;

**Yuriy Pakhomov** – PhD (Technical Sciences), Associate Professor, O. M. Beketov National University of Urban Economy in Kharkiv, Associate Professor of the Department of Computer Science and Information Technologies, Kharkiv, Ukraine;

e-mail: abc050073@gmail.com

ORCID ID: <http://orcid.org/0000-0002-2267-8600>

## INTELLIGENT MODEL FOR THE FORMATION OF INNOVATION AND ENGINEERING CLUSTERS

In the context of digitalization and the development of the knowledge economy, innovation and engineering clusters are transforming into complex dynamic systems characterized by a high level of uncertainty, multifactor interactions, and significant dependence on external managerial influences. Traditional economic and mathematical approaches do not provide sufficient flexibility for modeling such systems, which necessitates the application of integrated artificial intelligence methods. The aim of this study is to develop and validate a hybrid model for the formation of innovation and engineering clusters that combines machine learning methods, graph analysis, agent-based modeling, and digital twins, followed by an assessment of the synergistic effect of their functioning and an analysis of the impact of managerial factors on cluster development dynamics. The study employs a multilevel approach that includes agent clustering (k-means, DBSCAN), construction of a graph-based interaction model, application of graph neural networks to account for topological dependencies, and agent-based modeling to investigate the evolution of the system over time. Additionally, agent compatibility and risk functions are introduced to formalize the synergistic effect and cluster resilience within an optimization framework. The integration of all components is implemented in the form of a digital twin, enabling scenario analysis capabilities.

**Research Results.** The model was validated on a simulated regional cluster consisting of enterprises, universities, investors, and a governmental body. Simulation results demonstrated that government support significantly increases the innovation activity of cluster participants, enhances the density of network connections, and raises the number of cooperative projects. Key nodes of network interaction were identified, particularly the role of universities as knowledge transfer centers and investors as financial hubs.

**Conclusions.** The obtained results confirm the effectiveness of the proposed approach and its suitability for forecasting the development of innovation ecosystems and supporting strategic managerial decision-making. The developed model provides a comprehensive integration of machine learning methods, graph analysis, agent-based modeling, and digital twins, enabling the transition from static cluster analysis to adaptive and predictive management. The practical significance of the study lies in the possibility of applying the proposed approach to support regional innovation policy, evaluate the effectiveness of managerial decisions, and form sustainable innovation and engineering ecosystems.

**Keywords:** innovation and engineering clusters, artificial intelligence, machine learning, graph analysis, graph neural networks, agent-based modeling, digital twin.

### Бібліографічні описи / Bibliographic descriptions

Роговий А. І., Ахїєзер О. Б., Решетнікова С. М., Колбасін В. О., Пахомов Ю. В. Інтелектуальна модель формування інноваційно-інжинірингових кластерів. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 47–56. DOI: <https://doi.org/10.30837/0135-1710.2026.189.047>

Rogovyi, A., Akhiezer, O., Reshetnikova, S., Kolbasin, V., Pakhomov, Y. (2026), "Intelligent Model for the Formation of Innovation and Engineering Clusters", *Management Information System and Devices*, No. 2 (189), P. 47–56. DOI: <https://doi.org/10.30837/0135-1710.2026.189.047>



Скляров І. І.

## ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ІНТЕЛЕКТУАЛЬНОГО УПРАВЛІННЯ СЕРВІСНИМИ ПРОЦЕСАМИ В АВТОМОБІЛЬНІЙ ГАЛУЗІ НА ОСНОВІ BIG DATA

**Предметом дослідження** є методи, моделі й засоби побудови інформаційної технології інтелектуального управління сервісними процесами на основі *Big Data*. **Мета роботи** – створення інформаційної технології інтелектуального управління сервісними процесами в автомобільній галузі на основі *Big Data* для персоналізації сервісних послуг, автоматизації прийняття рішень щодо обслуговування та підвищення ефективності взаємодії між сервісними центрами, виробниками й споживачами. З огляду на окреслену мету необхідно виконати такі **завдання**: дослідити сучасні підходи до управління сервісними процесами в автомобільній галузі та визначити можливості використання *Big Data* для їх інтелектуалізації; проаналізувати джерела й структури даних, які застосовуються в сервісних процесах автомобільної галузі; розробити структуру інформаційної технології інтелектуального управління сервісними процесами на основі інтеграції різномірних даних; реалізувати запропоновану інформаційну технологію та експериментально оцінити її працездатність. **Методи**. У роботі використано методи системного аналізу для дослідження сервісних процесів в автомобільній галузі та визначення основних інформаційних потоків між сервісними центрами, виробниками й споживачами. Для оброблення великих масивів різномірних даних впроваджено методи інтеграції, очищення й аналізу інформації. Інформаційна технологія розроблялася завдяки використанню методів інтелектуального аналізу даних, машинного навчання, класифікації та підтримки прийняття рішень. Для експериментальної перевірки працездатності запропонованого підходу застосовано методи обчислювального моделювання, порівняльного аналізу й оцінювання якості моделей за стандартними метриками класифікації. Програмну реалізацію виконано мовою *Python* у середовищі *Google Colab*. **Результати дослідження**. Розроблено інформаційну технологію управління сервісними процесами в автомобільній галузі на основі *Big Data*, що забезпечує інтеграцію технічних, експлуатаційних і сервісних даних, оцінювання ризику відмови й формування сервісних рішень. У середовищі *Google Colab* реалізовано програмний прототип цієї технології мовою *Python* на основі датасету *scania\_component\_x\_combined*, сформованого з відкритого набору *SCANIA Component X Dataset*. Експериментальна перевірка підтвердила працездатність запропонованого підходу й високу якість розподілу нормальних станів і станів ризику транспортних засобів. **Висновки**. Результати дослідження підтвердили доцільність використання *Big Data* для аналітичної підтримки сервісних процесів в автомобільній галузі. Запропонована інформаційна технологія може бути застосована як основа для підвищення обґрунтованості сервісних рішень, вчасного виявлення станів ризику й подальшого розвитку цифрових систем технічного обслуговування транспортних засобів.

**Ключові слова:** *Big Data*; автомобільна галузь; сервісні процеси; інформаційна технологія; технічне обслуговування; прогнозування ризику відмови; машинне навчання; аналіз даних; транспортні засоби; сервісні рішення.

### 1. Вступ

Сучасна автомобільна галузь визначається активною цифровізацією не лише процесів проєктування, виробництва та експлуатації транспортних засобів, а й систем обслуговування після продажу. Сервісні процеси дедалі більше залежать від швидкості оброблення інформації, повноти даних про транспортний засіб, історію його обслуговування, умов експлуатації та поведінкові якості споживача. У цих умовах традиційні підходи до організації сервісу, що ґрунтуються на регламентованих процедурах та ізольованому розгляді окремих звернень, втрачають ефективність, оскільки не забезпечують належного рівня гнучкості, персоналізації та оперативності прийняття

рішень. Розвиток технологій *Big Data* [1] відкриває нові можливості для побудови інтелектуальних сервісних систем, здатних інтегрувати дані телематики, діагностики, експлуатації разом із даними клієнтів в єдиний інформаційно-аналітичний простір. Використання такої інформації дає змогу не лише фіксувати стан транспортного засобу або історію сервісних операцій, а й формувати інтелектуальні механізми підтримки рішень щодо типу, часу, пріоритетності та персоналізації сервісного втручання. Особливої актуальності це набуває в умовах зростання вимог до якості обслуговування після продажу, скорочення часу реакції на сервісні події та підвищення ефективності взаємодії між усіма учасниками екосистеми сервісу.

Наявні інформаційні системи сервісного обслуговування в автомобільній галузі переважно орієнтовані на облік звернень, формування замовлень або збереження історії виконаних робіт. Водночас вони часто не забезпечують інтелектуального аналізу великих масивів даних для автоматизованого формування сервісних рекомендацій, адаптації обслуговування до особливостей конкретного клієнта чи транспортного засобу, а також для координації дій між сервісними центрами, виробниками й споживачами. Це зумовлює потребу у створенні інформаційної технології, яка поєднувала б інструменти збору, інтеграції, аналізу та інтерпретації *Big Data* з механізмами автоматизації сервісних рішень. Розроблення інформаційної технології інтелектуального управління сервісними процесами в автомобільній галузі на основі *Big Data* є актуальним науково-прикладним завданням, спрямованим на підвищення ефективності сервісної інфраструктури та якості обслуговування транспортних засобів.

## 2. Аналіз сучасних наукових публікацій. Визначення проблеми дослідження

Для виконання завдання підвищення ефективності управління сервісними процесами в автомобільній галузі необхідно брати до уваги сучасні підходи до використання *Big Data*, інтелектуального аналізу даних і цифрових сервісних платформ, визначити їх переваги й обмеження, а також обґрунтувати можливість розроблення інформаційної технології інтелектуального управління сервісом. З огляду на це доцільно проаналізувати сучасні дослідження, присвячені цифровізації сервісу після продажу, застосуванню аналітики *Big Data*, персоналізації сервісних послуг і автоматизації прийняття рішень у сервісних системах автомобільної галузі.

Роботу [2] присвячено аналізу використання машинного навчання для профілактичного обслуговування в автомобільній галузі. У дослідженні зазначено, що сучасні транспортні засоби генерують великі обсяги даних експлуатації, а тому машинне навчання стає одним із ключових інструментів забезпечення функціональної безпеки, надійності та економічно ефективного обслуговування протягом життєвого циклу автомобіля. Автори наголошують на необхідності поєднання кількох джерел інформації

та переходу до більш комплексних завдань, основаних на сучасних підходах у сфері обслуговування транспортних засобів. Висновки в цій роботі підтверджують доцільність розроблення більш комплексних рішень, де сервісне управління спирається не на окремі сигнали чи ізольовані моделі, а на інтеграцію інформації, аналітику *Big Data* й автоматизовану підтримку рішень.

У праці [3] описано гібридний підхід до профілактичного обслуговування, у якому поєднуються підходи на основі даних і знань та механізми підтримки рішень. Автори запропонували метод, у якому спочатку формується модель розпізнавання станів на основі LSTM, далі за допомогою методу найближчих сусідів аналізується зв'язок між ознаками робочого стану й станом несправності, а результати прогнозування передаються до графа знань для подальшого логічного виведення сервісних стратегій. Результатом є побудова системи, яка автоматично формує рішення щодо технічного обслуговування. Ця стаття підтверджує доцільність переходу від суто моделей прогнозування до систем підтримки рішень, у яких результати аналітики даних поєднуються з формалізованими знаннями предметної галузі.

Роботу [4] присвячено аналізу використання технологій машинного навчання, IoT та *Big Data* в завданнях трансформації технічного обслуговування. Запропоновано структуру, що охоплює збір і оброблення інформації, взаємодію людини й машини, обслуговування, зворотний зв'язок і забезпечення стійкості системи. Важливо зауважити, що автори розглядають профілактичне обслуговування не лише як задачу виявлення відмов, а як складник більш широкої цифрової екосистеми, у якій поєднуються дані, інтелектуальна аналітика, взаємодія з персоналом і підтримка прийняття рішень. Це дослідження підтверджує доцільність розроблення інформаційної технології, що буде орієнтована на персоналізацію сервісних послуг, автоматизацію обслуговування та підвищення ефективності взаємодії між сервісними центрами, виробниками й споживачами.

У статті [5] проаналізовано сучасний розвиток систем моніторингу стану автомобіля та діагностику несправностей для різних автономних транспортних засобів. Автори звертають увагу на те, що не існує єдиного узагальненого підходу, який би інтегрував різні методи й забезпечував масштабованість,

адаптивність і ефективність у різних умовах експлуатації. Також запропоновано таксономію підходів до моніторингу стану автомобіля, проаналізовано роль датчиків, IoT, машинне навчання, аналітику *Big Data* й мультисенсорну систему, а також сформовано концептуальну фреймворк-модель, орієнтовану на більш цілісне управління моніторингом і діагностикою транспортних засобів. Варто наголосити, що під час дослідження автори дійшли висновку, що більшість робіт зосереджена на окремих підсистемах, тоді як сучасний автомобіль потребує цілісного підходу, який поєднував би технічні, аналітичні та інформаційні засоби в межах єдиного фреймворк-рішення. Усе це робить доцільним перехід від локальних рішень для діагностики окремих вузлів до інтегрованих інтелектуальних систем управління сервісними процесами.

У контексті розвитку інтелектуальних сервісних технологій в автомобільній галузі окремий інтерес становлять дослідження, присвячені обробленню часових рядів бортової діагностики й використанню гібридних моделей глибокого навчання для підтримки профілактичної діагностики. Такі роботи є важливими, оскільки демонструють, яким чином показники, отримані через вбудовані системи діагностики автомобіля, можуть бути перетворені на основу для автоматизованого оцінювання стану транспортного засобу, раннього виявлення відхилень і формування сервісних рішень. Саме до цього напряму належить стаття [6], у якій розглянуто поєднання часового аналізу, кластеризації та глибокого навчання для прогнозування технічного обслуговування транспортних засобів. Основна ідея роботи полягає в тому, щоб використовувати кластеризацію для виявлення прихованих експлуатаційних шаблонів у даних, а LSTM – для прогнозування критичних параметрів стану двигуна. Авторі наголошують, що такий підхід є особливо корисним в умовах, коли набори автомобільних показників не мають явного маркування щодо відмов чи станів несправності. Застосування кластеризації дало змогу виділити режими експлуатації, а далі було використано LSTM-модель для прогнозування часової динаміки цих параметрів. Підтверджено, що показники вбудованих систем самодіагностики автомобіля можуть бути ефективною базою для побудови основаних на даних систем підтримки технічного

обслуговування транспортних засобів. Авторі не обмежуються класифікацією відмов, а пов'язують результати прогнозування з можливістю технічного обслуговування з огляду на стан транспортного засобу, зменшенням незапланованих простоїв, оптимізацією обслуговування та підтримкою прийняття рішень для фахівців сервісу й менеджерів з управління корпоративним автопарком. Хоча ця стаття й зосереджена переважно на прогнозуванні параметрів технічного стану й ранньому виявленні відхилень за результатами бортової діагностики, у ній не розкрито в повному обсязі питань персоналізації сервісних послуг, координації взаємодії між сервісними центрами, виробниками й споживачами, а також не запропоновано окремої інформаційної технології управління сервісною екосистемою.

У роботі [7] описано фреймворк для діагностики несправностей двигуна з використанням *MiniRocket*, багатокрокового відбору ознак і набору класифікаторів машинного навчання. На думку авторів статті, традиційні, основані на правилах, і статистичні підходи до оброблення сигналів не забезпечують належної якості діагностики в умовах високої динамічності та багатовимірності сучасних показників. У публікації фактично продемонстровано інтеграцію етапів попереднього оброблення даних, виокремлення ознак, відбору інформативних характеристик і машинного прийняття рішення. Водночас ця праця орієнтована насамперед на діагностику несправностей двигуна, тобто на локальну задачу класифікації, а не на повноцінне інтелектуальне управління сервісними процесами в автомобільній галузі. У ній не порушено питань персоналізації сервісних послуг, автоматизації взаємодії між сервісними центрами, виробниками й споживачами, а також не запропоновано окремої інформаційної технології керування обслуговуванням після продажу.

У роботі [8] проаналізовано сучасні підходи до профілактики в автомобільній сфері. Наголошено на тому, що більшість досліджень мають високі результати для прогнозування відмов окремих компонентів, зокрема підвісок, гальм, батарей та інших вузлів, однак такі рішення переважно функціонують ізольовано. Авторі зазначають, що основним нерозв'язаним завданням залишається перехід від прогнозування на рівні компонентів до цілісного управління станом транспортних засобів, тобто до комплексного оцінювання технічного стану автомобіля як взаємопов'язаної системи. Окремо

необхідно зауважити, що автори прямо формують одну з ключових проблем галузі: сучасні роботи часто реалізують однокомпонентний підхід, який не дає змоги повною мірою прогнозувати інтегровану потребу в технічному обслуговуванні всього транспортного засобу. Ця робота демонструє наявність реальної дослідницької прогалини між точковими профілактичними рішеннями й побудовою цілісної сервісно орієнтованої інформаційної технології. Отже, актуальною є інформаційна технологія, у межах якої *Big Data* використовується не лише для виявлення окремих технічних проблем, а й для персоналізації сервісних послуг, автоматизації рішень щодо обслуговування й координації взаємодії між сервісними центрами, виробниками та споживачами.

Деякі дослідження присвячені не лише прогнозуванню технічного стану транспортних засобів, а й прогнозуванню попиту на запасні частини й обслуговування після продажу. Такі роботи розширюють традиційне розуміння профілактичного обслуговування, оскільки демонструють, що ефективність сервісної інфраструктури залежить не тільки від вчасного виявлення технічних відхилень, а й від здатності передбачати потребу в запасних частинах, оптимізувати запаси й підтримувати належний рівень обслуговування клієнтів. У статті [9] автори досліджують можливості прогнозування технічного стану для підвищення точності передбачення попиту на запасні частини в автомобільній галузі, зокрема в країнах, що розвиваються. У роботі наголошено, що ринок автомобільних запасних частин визначається високою варіативністю попиту, значною кількістю одиниць складського обліку, нерегулярними й часовими рядами, а також суттєвою залежністю від обслуговування після продажу. Саме тому автори розглядають попит на запасні частини як критично важливий складник ефективного сервісного обслуговування, від якого залежать задоволеність клієнтів, рівень доступності деталей і загальна якість сервісу після продажу. Автори зробили висновок, що ефективність обслуговування після продажу значною мірою визначається здатністю правильно прогнозувати попит на запасні частини й обирати адекватні аналітичні інструменти для різних категорій номенклатури. Водночас стаття переважно зосереджена на прогнозуванні попиту на запасні частини, тобто на одному з важливих, але все ж локальних аспектів сервісного процесу. У ній не

порушуються питання персоналізації сервісних послуг, інтеграції сервісних показників і даними телематики, автоматизації взаємодії між сервісними центрами, виробниками й споживачами.

Іншим напрямом робіт з теми досліджень є ті, що зосереджуються на моделюванні процесів обслуговування після продажу в реальному середовищі автомобільного сервісу. У статті [10] автори розглядають проблему моделювання сервісних процесів у сегменті автомобільного ремонту як інструмент підвищення ефективності, прибутковості та задоволеності клієнтів. На відміну від багатьох праць, у яких порушено загальні питання, у цій роботі проаналізовано особливості обслуговування після продажу в автомобільній галузі. Автори наголошують, що більшість попередніх досліджень або залишаються на теоретичному рівні, або аналізують окремі процеси ізольовано, без огляду на їх взаємозалежності. Натомість у цій статті сервісний процес розглядається цілісно, як послідовність взаємопов'язаних етапів, що впливають на загальну результативність сервісної організації. Методологічно дослідження побудовано на поєднанні кількісного та якісного підходів. Автори використовували інформацію зі звітів сервісної діяльності, показники продуктивності, ефективності та задоволеності клієнтів, а також безпосереднє спостереження за участю в межах конкретного авторизованого сервісного центру. У роботі сформовано модель, що охоплює п'ять ключових етапів сервісного процесу: планування, прийом, діагностику, ремонт, передачу. Для кожного етапу визначено критичні фактори успіху й запропоновано практичні заходи вдосконалення, зокрема щодо розподілу ресурсів, організації запису, комунікації з клієнтами, дотримання технічних процедур, управління запасними частинами та збору зворотного зв'язку. Але водночас стаття має певні прогалини, зокрема вона ґрунтується на одному кейсі авторизованого сервісного центру, а тому не містить масштабної *Big Data*-аналітики, прогнозування сервісних подій у режимі реального часу чи побудови автоматизованої цифрової платформи на рівні всієї сервісної екосистеми.

Окремо важливо згадати наукові праці, присвячені комплексному прогнозуванню потреби в технічному обслуговуванні транспортного засобу, а не лише діагностиці окремих компонентів. У статті [11] автори виходять з того, що наявні підходи до прогнозування технічного обслуговування

переважно орієнтовані на окремі вузли чи компоненти автомобіля й не забезпечують повного прогнозування всіх потреб у сервісі. Основною проблемою, яку вони формують, є недостатність підходів типу CLP [12], оскільки транспортний засіб функціонує як взаємопов'язана система, а різні проєкти обслуговування мають як статичні, так і часові залежності між собою. У цьому полягає основна ідея роботи: перейти від ізольованого прогнозування окремих сервісних дій до комплексного прогнозування попиту на обслуговування на рівні всього автомобіля. У дослідженні запропоновано модель CoSDC, яка поєднує два рівні аналізу. Перший рівень – статичне кореляційно-усвідомлене навчання, а другий – динамічне навчання з огляду на кореляцію, де за допомогою рекурентної нейронної мережі та механізму подання моделюються часові залежності між записами технічного обслуговування на різних етапах життєвого циклу автомобіля. Після цього статичне й динамічне подання поєднується в кооперативний модуль прогнозування, що дає змогу передбачити наступний набір сервісних робіт. У дослідженні зазначено, що запропонована модель краще працює навіть за обмеженого обсягу навчальних даних і ефективно виявляє нові сервісні потреби, яких не було в попередній історії обслуговування. Крім того, автори зосереджені насамперед на прогнозуванні попиту на обслуговування як задачі одночасного визначення кількох можливих сервісних робіт. У роботі не порушуються питання персоналізації сервісних послуг, управління взаємодією між сервісними центрами, виробниками й споживачами, а також організації обслуговування після продажу.

Ще одним напрямом досліджень інтелектуального управління сервісними процесами в автомобільній галузі є студії, в яких узагальнено сучасні підходи до прогнозування технічного обслуговування на основі штучного інтелекту. У статті [13] продемонстровано, що сучасні підходи до прогнозування обслуговування спираються на різноманітні джерела інформації, зокрема показники бортової діагностики, сенсорні вимірювання, історію технічного обслуговування, експлуатаційні параметри та діагностичні коди несправностей, а основними прикладними завданнями є прогнозування відмов, оцінювання залишкового корисного ресурсу, планування технічного обслуговування, діагностика несправностей і моніторинг технічного стану. Особливістю цієї роботи є те, що

вона не обмежується оглядом машинного й глибокого навчання, а додатково розглядає пояснювальний штучний інтелект [14] як інструмент підвищення прозорості моделей, а також генеративний штучний інтелект [15] як перспективний напрям для автоматизованого формування звітів, генерації синтетичних показників, підтримки діагностики й оптимізації сервісних рішень. У статті також наведено публічно доступні набори даних, важливі для прогнозування технічного обслуговування. Отже, сучасна автомобільна галузь уже має потужну основу у вигляді методів штучного інтелекту для прогнозування технічного обслуговування, однак подальший розвиток цього напрямку потребує переходу до більш інтегрованих, інтерпретованих і сервісно орієнтованих рішень, які поєднували б прогнозування, аналітику *Big Data*, автоматизацію рішень і цифрову підтримку сервісної інфраструктури.

Аналіз сучасних наукових публікацій засвідчив, що наявні підходи до прогнозування технічного обслуговування й управління сервісними процесами в автомобільній галузі створюють вагоме теоретичне й прикладне підґрунтя для подальших досліджень, однак переважно орієнтуються або на прогнозування відмов окремих компонентів, або на розв'язання локальних завдань сервісної логістики й обслуговування після продажу. Водночас недостатньо опрацьованими залишаються питання побудови цілісної інформаційної технології, здатної інтегрувати різноманітні дані, забезпечувати інтелектуалізацію сервісних рішень, персоналізацію послуг і координацію взаємодії між основними учасниками сервісної інфраструктури. У зв'язку з цим доцільним є розроблення інформаційної технології інтелектуального управління сервісними процесами в автомобільній галузі на основі *Big Data*, опис її структури, принципів функціонування та особливостей практичної реалізації.

### 3. Мета й завдання дослідження

Метою роботи є створення інформаційної технології інтелектуального управління сервісними процесами в автомобільній галузі на основі *Big Data* для персоналізації сервісних послуг, автоматизації прийняття рішень щодо обслуговування та підвищення ефективності взаємодії між сервісними центрами, виробниками й споживачами. З огляду на окреслену мету необхідно виконати такі завдання: проаналізувати сучасні підходи до управління сервісними процесами



в автомобільній галузі та визначити можливості використання *Big Data* для їх інтелектуалізації; проаналізувати джерела й структури даних, які застосовуються в сервісних процесах автомобільної галузі; розробити структуру інформаційної технології інтелектуального управління сервісними процесами на основі інтеграції різнорідних даних; реалізувати запропоновану інформаційну технологію та експериментально оцінити її працездатність.

#### 4. Викладення основного матеріалу

Розроблення інформаційної технології інтелектуального управління сервісними процесами в автомобільній галузі зумовлено потребою переходу від фрагментарного використання окремих даних сервісу, експлуатації та діагностики до їх комплексного аналізу в межах єдиного цифрового середовища підтримки рішень. На відміну від традиційних підходів, за яких сервісні операції переважно виконуються на основі регламентів, локального досвіду персоналу або реакції на вже наявні несправності, запропонована інформаційна технологія орієнтована на раннє виявлення потреб у технічному обслуговуванні, координацію сервісних дій і підвищення ефективності взаємодії між учасниками сервісної інфраструктури. Її основою є використання технологій *Big Data* для інтеграції різнорідних масивів даних, їх аналітичного опрацювання й формування обґрунтованих сервісних рішень.

Сутність запропонованої інформаційної технології полягає в поєднанні засобів збору, накопичення, уніфікації, аналізу та інтерпретації показників, що надходять із різних джерел і визначають технічний стан транспортних засобів, умови їх експлуатації, історію обслуговування, інтенсивність сервісних звернень, доступність ресурсів та особливості взаємодії зі споживачами. У межах такої технології сервісний процес розглядається не як послідовність ізольованих операцій, а як цілісна система, в якій кожне рішення щодо обслуговування формується з огляду на сукупність технічних, часових, організаційних та інформаційних чинників. Це дає змогу перейти від реактивної моделі обслуговування до інтелектуального управління сервісними процесами, орієнтованого на вчасність, адаптивність і персоналізацію.

Функціонально інформаційна технологія охоплює кілька взаємопов'язаних рівнів. На початковому рівні

здійснюється збір та консолідація показників із бортових систем діагностики, сенсорних пристроїв, сервісних журналів, баз даних технічного обслуговування, клієнтських звернень та інших цифрових джерел. Далі виконується попереднє опрацювання інформації, що передбачає очищення, нормалізацію, узгодження форматів, усунення дублювання й формування єдиного інформаційного простору для подальшої аналітики. На наступному рівні реалізуються процедури інтелектуального аналізу даних, спрямовані на виявлення закономірностей у сервісній історії транспортних засобів, оцінювання технічних і сервісних ризиків, визначення пріоритетності обслуговування та формування рекомендацій щодо оптимального сервісного втручання. Завершальний рівень пов'язаний із підтримкою прийняття рішень, де результати аналітики трансформуються в конкретні сервісні дії, рекомендації персоналу, пропозиції для клієнтів і сигнали для координації роботи між сервісними центрами, виробниками й споживачами.

Ключовою особливістю запропонованої інформаційної технології є її орієнтація на використання не окремих показників, а саме сукупної інформації про поведінку транспортного засобу впродовж життєвого циклу. Це дає змогу зважати як на поточні параметри функціонування, так і накопичену історію сервісних подій, повторюваність певних несправностей, закономірності звернень, інтенсивність експлуатації та інші чинники, що можуть свідчити про майбутню потребу в обслуговуванні. Завдяки цьому інформаційна технологія забезпечує не лише фіксацію вже наявного стану, а й аналітичне передбачення подальших сервісних потреб, що є основою для підвищення ефективності управління.

Запропонований підхід також містить персоналізацію сервісних рішень. Це означає, що формування рекомендацій щодо обслуговування здійснюється з огляду на індивідуальні характеристики транспортного засобу, історію його використання, типові умови експлуатації, попередні звернення та доступність сервісних ресурсів. У такому разі одна й та сама сервісна дія не розглядається як універсальна для всіх автомобілів, а адаптується до конкретного об'єкта обслуговування. Саме це створює основу для більш гнучкої організації сервісу, підвищення задоволеності споживача й раціонального використання ресурсів сервісної інфраструктури.

Для формалізації інформаційної технології інтелектуального управління сервісними процесами в автомобільній галузі запропонуємо множину транспортних засобів, що перебувають у полі сервісного спостереження:

$$V = \{v_1, v_2, \dots, v_n\}, \quad (1)$$

де  $v_i$  –  $i$ -й транспортний засіб, для якого аналізуються технічний стан, експлуатаційні характеристики й сервісні історії.

Для кожного транспортного засобу формується сукупність різномірних показників, що надходять із кількох інформаційних джерел:

$$D_i = \{D_i^{sens}, D_i^{serv}, D_i^{oper}, D_i^{client}\}, \quad (2)$$

де  $D_i^{sens}$  – дані з сенсорів та діагностики;  $D_i^{serv}$  – історія технічного обслуговування;  $D_i^{oper}$  – параметри експлуатації;  $D_i^{client}$  – дані про сервісні звернення та взаємодію зі споживачем.

Оскільки зазначені показники є неоднорідними за структурою, обсягом і частотою оновлення, на першому етапі інформаційної технології виконується їх інтеграція, очищення, нормалізація та узгодження в єдиному просторі ознак. Унаслідок формується інтегрований вектор опису транспортного засобу:

$$X_i = \Phi(D_i^{sens}, D_i^{serv}, D_i^{oper}, D_i^{client}), \quad (3)$$

де  $\Phi$  – оператор попереднього оброблення й перетворення даних. Сформований вектор ознак  $X_i$  є основою для подальшого аналітичного оцінювання сервісної потреби транспортного засобу. Інтегральна оцінка потреби в сервісному втручанні визначається як

$$S_i = f(X_i), \quad (4)$$

де  $S_i$  – значення інтегральної сервісної оцінки для транспортного засобу  $v_i$ ;  $f$  – функція аналітичної моделі, реалізованої в межах інформаційної технології.

З метою підвищення інтерпретованості моделі інтегральну сервісну оцінку доцільно подати як зважену комбінацію основних факторів, що визначають технічний і сервісний стан транспортного засобу:

$$S_i = w_1 R_i + w_2 H_i + w_3 E_i + w_4 C_i, \quad (5)$$

де  $R_i$  – оцінка технічного ризику;  $H_i$  – показник сервісної історії;  $E_i$  – характеристика умов та інтенсивності експлуатації;  $C_i$  – показник клієнтської або сервісної активності;  $w_1, w_2, w_3, w_4$  – вагові коефіцієнти відповідних складників.

У цьому разі вагові коефіцієнти мають задовольняти умову нормування:

$$\sum_{j=1}^4 w_j = 1, \quad w_j \geq 0. \quad (6)$$

Одним із ключових складників запропонованої інформаційної технології є оцінювання ризику виникнення сервісної події або ймовірності появи стану, за якого транспортний засіб потребує технічного втручання. Така оцінка визначається як умовна ймовірність:

$$R_i = P(Y_i = 1 | X_i), \quad (7)$$

де  $Y_i = 1$  означає наявність потреби в сервісному втручанні в межах заданого горизонту прогнозування. Отримана інтегральна оцінка використовується для автоматизованого формування сервісного рішення. Залежно від її значення транспортний засіб належить до одного з класів сервісної реакції:

$$U_i = \begin{cases} u_1, & S_i < \theta_1 \\ u_2, & \theta_1 \leq S_i < \theta_2, \\ u_3, & S_i \geq \theta_2 \end{cases} \quad (8)$$

де  $u_1$  – стандартне планове обслуговування;  $u_2$  – додаткова діагностика;  $u_3$  – пріоритетне сервісне втручання;  $\theta_1, \theta_2$  – порогові значення для розподілу рішень за рівнем критичності. Оскільки в умовах реальної сервісної інфраструктури важливим є не лише факт потреби в обслуговуванні, а й черговість виконання відповідних робіт, у межах інформаційної технології додатково обчислюється пріоритет сервісного обслуговування:

$$Pr_i = \alpha S_i + \beta T_i + \gamma A_i, \quad (9)$$

де  $Pr_i$  – інтегральний пріоритет обслуговування транспортного засобу;  $T_i$  – часовий фактор критичності;  $A_i$  – показник доступності сервісних ресурсів;  $\alpha, \beta, \gamma$  – вагові коефіцієнти.

Для забезпечення персоналізації сервісних рішень остаточне рішення формується з огляду на індивідуальний профіль транспортного засобу, історію його використання та особливості взаємодії зі споживачем:

$$U_i^* = \Psi(U_i, Z_i), \quad (10)$$

де  $Z_i$  – профіль транспортного засобу та споживача, а  $\Psi$  – функція адаптації базового сервісного рішення до конкретних умов експлуатації та обслуговування. Узагальнено запропоновану інформаційну технологію

можна подати у вигляді композиції трьох основних функціональних компонентів:

$$IT = \langle \Phi, f, \Psi \rangle, \quad (11)$$

де  $\Phi$  реалізує інтеграцію та підготовку даних,  $f$  забезпечує аналітичне оцінювання сервісної потреби й ризику, а  $\Psi$  формує персоналізоване сервісне рішення.

## 5. Результати дослідження

Експериментальна реалізація запропонованої інформаційної технології в середовищі *Google Colab* з використанням *Python* була спрямована на відтворення її основних функціональних етапів у межах аналітичного моделювання. У програмній реалізації було зосереджено увагу не на повній цифровій інфраструктурі сервісного обслуговування, а на її ключовому функціональному ядрі, що забезпечує оброблення інтегрованих даних, оцінювання сервісної потреби транспортного засобу, визначення рівня ризику та формування сервісного рішення. Такий підхід дав змогу перевірити працездатність запропонованої інформаційної технології як засобу підтримки прийняття рішень у сервісних процесах автомобільної галузі.

В описаному експерименті використано попередньо сформований інтегрований датасет *scania\_component\_x\_combined*, отриманий на основі

відкритого набору *SCANIA Component X Dataset* [16]. На першому етапі реалізації виконано завантаження й первинний аналіз структури даних, після чого відібрано рядки з наявною цільовою міткою, оброблено пропущені значення, вилучено ознаки, що могли спричиняти витік цільової змінної, а також поділено параметри на числові й категоріальні. Це забезпечило формування узгодженого простору ознак, придатного для машинного аналізу. На наступному етапі було побудовано модель, яка на основі інтегрованих технічних, експлуатаційних і сервісних характеристик транспортних засобів формує оцінку ризику й дає змогу долучити об'єкт до одного з класів сервісного стану. Після цього оцінка ризику, яка була прогнозована, використовувалася для формування сервісного рішення, що інтерпретувалося як продовження експлуатації, додаткова діагностика або пріоритетне технічне втручання. Отже, у *Colab* реалізовано не просто задачу класифікації, а послідовність обчислювальних процедур, що відповідає за логіку функціонування запропонованої інформаційної технології.

На рис. 1 подано розподіл цільового класу у сформованій вибірці. Він свідчить про суттєву незбалансованість експериментальних показників, у яких переважають спостереження, що відповідають нормальному технічному стану транспортних засобів, тоді як частка записів із ризиком відмови є значно меншою.

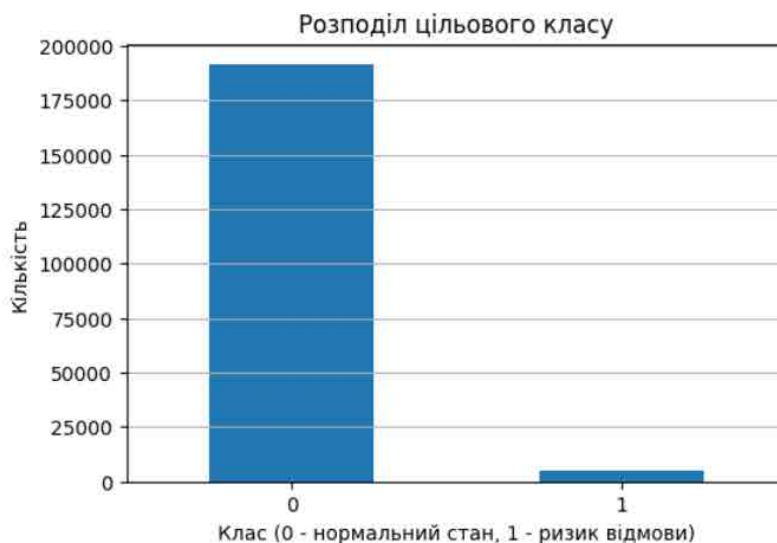


Рис. 1. Розподіл цільового класу

Така структура даних властива для реальних задач моніторингу технічного стану, оскільки в

умовах експлуатації більшість об'єктів у будь-який момент часу функціонує без критичних відхилень,

а потенційно небезпечні стани виникають відносно нечасто.

На рис. 2 запропоновано матрицю помилок, що відтворює результати класифікації нормальних і станів ризику транспортних засобів.

Аналіз отриманих значень демонструє, що модель правильно ідентифікує переважну більшість об'єктів, які належать до нормального технічного стану, а також забезпечує виявлення значної частини випадків ризику.

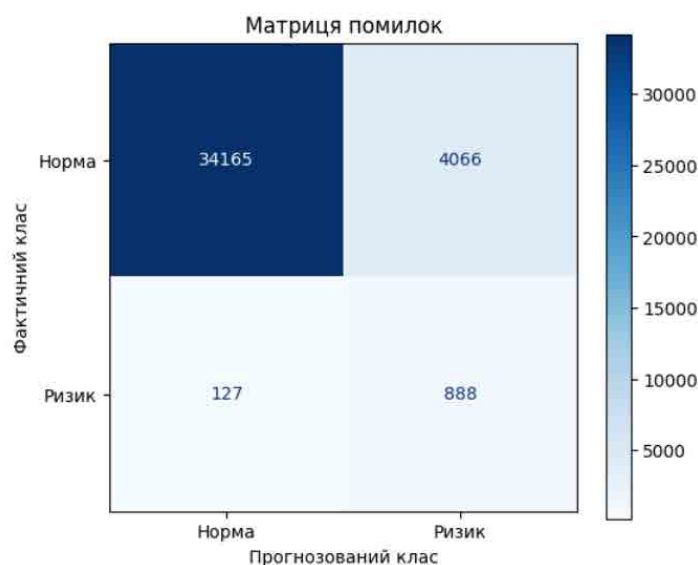


Рис. 2. Матриця помилок

Водночас спостерігається помітна кількість хибнопозитивних спрацювань, тобто ситуацій, коли нормальний стан помилково визначається як ризиковий. Така поведінка моделі є очікуваною в умовах незбалансованої вибірки та свідчить про її схильність до більш обережної стратегії прийняття рішень, за якої потенційно небезпечні стани виявляються з певним запасом чутливості. Кількість хибнонегативних результатів, навпаки, є відносно незначною, що має важливе практичне значення, оскільки знижує ризик пропуску дійсно небезпечних технічних станів. Отже, матриця помилок підтверджує, що реалізований прототип інформаційної технології загалом забезпечує ефективне розділення класів, хоча потребує подальшого налаштування для зменшення кількості помилкових спрацювань щодо класів ризику.

На рис. 3 подано ROC-криву, яка визначає здатність моделі розділяти нормальні стани й стани ризику транспортних засобів за всім діапазоном можливих порогів класифікації. Отримана крива суттєво відхиляється від діагоналі випадкового вгадування, що свідчить про високу якість

побудованої моделі з погляду відокремлення двох класів. Значення площі під ROC-кривою, яке становить  $AUC = 0.9608$ , підтверджує, що модель має високий рівень дискримінаційної здатності та здебільшого коректно визначає об'єкти за ступенем ризику. Це вказує на те, що ризикові стани здебільшого отримують вищі оцінки прогнозування, ніж нормальні, що є важливим для подальшого формування сервісних рішень. Водночас високе значення AUC необхідно інтерпретувати в поєднанні з іншими результатами, зокрема з матрицею помилок, оскільки загальна здатність до розділення класів не відкидає наявності хибнопозитивних спрацювань у виборі конкретного робочого порога. Загалом наведена ROC-крива підтверджує, що аналітичне ядро запропонованої інформаційної технології забезпечує якісну оцінку ризику й може бути використане як основа для підтримки прийняття сервісних рішень.

На рис. 4 подано розподіл прогнозованого ризику відмови, сформованого моделлю для об'єктів тестової вибірки.

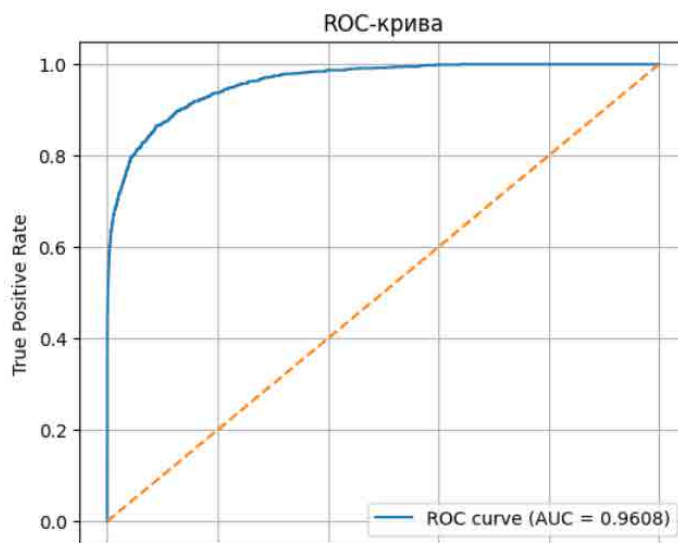


Рис. 3. ROC-крива

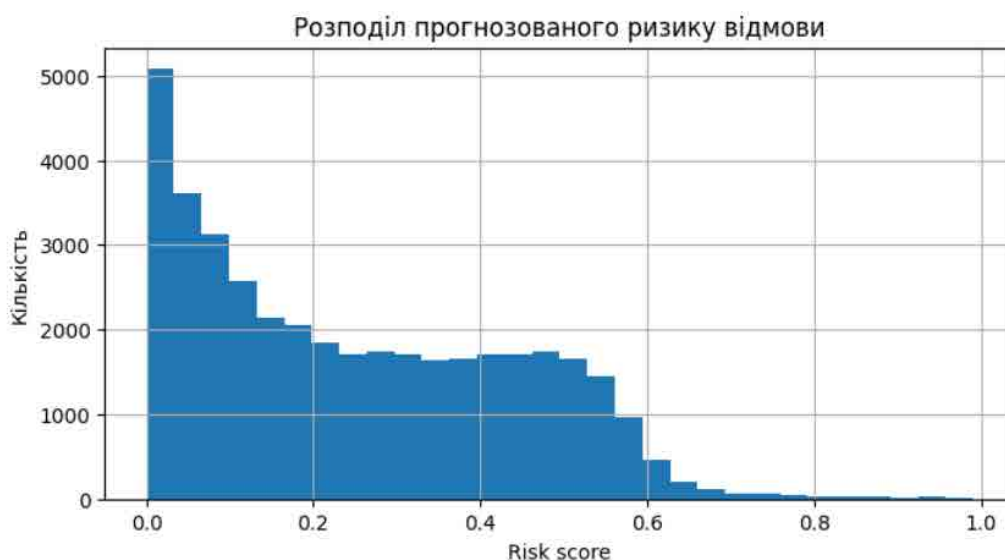


Рис. 4. Розподіл прогнозованого ризику відмови

Отриманий розподіл має виражену правосторонню асиметрію: найбільша частина спостережень зосереджена в ділянці низьких значень *risk score*, що відповідає перевазі транспортних засобів із нормальним технічним станом. Крім цього, у вибірці наявна й група об'єктів із середніми й високими значеннями прогнозованого ризику, для яких модель фіксує підвищену ймовірність виникнення небезпечного стану. Така форма розподілу є логічно узгодженою зі структурою даних і результатами класифікації, оскільки випадки ризику становлять меншу частку спостережень, але водночас мають бути достатньо чітко відокремлені від основної маси нормальних

станів. Практичне значення цього результату полягає в тому, що кількісна оцінка *risk score* може бути використана не лише для бінарної класифікації, а і як основа для подальшого ранжування транспортних засобів за рівнем критичності та формування диференційованих сервісних рішень. Отже, наведений розподіл підтверджує, що реалізоване аналітичне ядро інформаційної технології здатне генерувати безперервну оцінку ризику, яка є придатною для переходу від задачі класифікації до задачі інтелектуального управління сервісними процесами.

На рис. 5 зображено розподіл сервісних рішень, сформованих на основі прогнозованого ризику відмови.



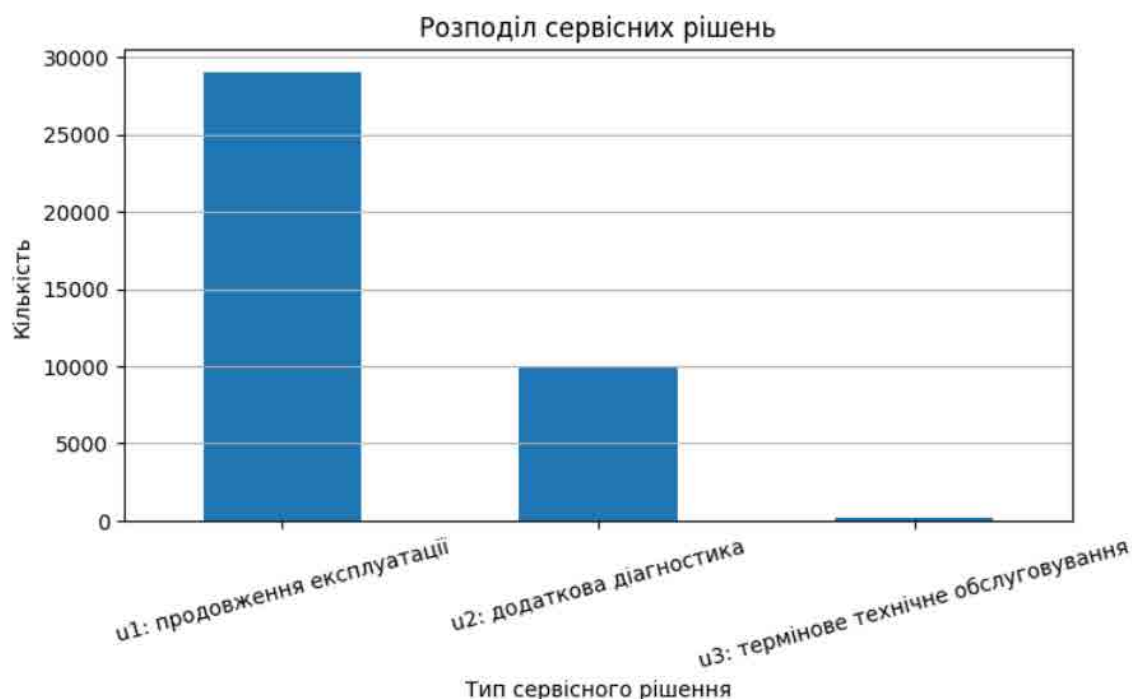


Рис. 5. Розподіл сервісних рішень

Отримані результати демонструють, що переважна більшість транспортних засобів була долучена до категорії продовження експлуатації, що узгоджується як із загальним розподілом цільового класу, так і з домінуванням низьких значень прогнозованого ризику. Помітно менша частка об'єктів була спрямована до категорії додаткової діагностики, що свідчить про наявність групи транспортних засобів із проміжним рівнем ризику, для яких модель не фіксує критичного стану, але виявляє підстави для поглибленої перевірки. Найменшу частку становлять рішення щодо термінового технічного обслуговування, що є логічним результатом, оскільки справді високий ризик відмови виявляється лише для обмеженої кількості об'єктів. Такий розподіл підтверджує, що запропонована інформаційна технологія не обмежується лише класифікацією стану, а забезпечує перехід до практично орієнтованого рівня підтримки сервісних рішень, у межах якого прогнозований ризик трансформується в конкретні дії щодо подальшої експлуатації, діагностики або обслуговування транспортного засобу.

На рис. 6 подано найважливіші ознаки, що мають найбільший вплив на формування прогнозу моделі. Отримані результати свідчать, що вирішальне значення в класифікації технічного стану транспортних засобів мають окремі числові

параметри, серед яких найбільш значущими виявилися ознаки 459\_17, 459\_15, 837\_0, 459\_14, 158\_9 та 158\_0. Водночас до факторів впливу увійшов також параметр *time\_step*, що вказує на важливість часового складника під час оцінювання ризику відмови. Такий результат підтверджує, що модель спирається не на одну окрему характеристику, а на сукупність технічних і часових параметрів, які спільно формують підґрунтя для розпізнавання станів ризику. Крім цього, необхідно брати до уваги, що наведені позначки ознак відтворюють внутрішню структуру набору даних і не мають безпосередньо інтерпретованого фізичного змісту без додаткового опису на рівні предметної галузі. Незважаючи на це, сам факт наявності чітко вираженої групи основних ознак є важливим, оскільки він засвідчує, що побудована модель виявляє стійкі закономірності в даних і використовує їх для формування прогнозної оцінки ризику й подальших сервісних рішень.

Отримані результати продемонстрували, що реалізований прототип здатний забезпечувати стійке розділення нормальних станів і станів ризику транспортних засобів, формувати кількісну оцінку сервісного ризику й автоматизовано генерувати рішення щодо подальших сервісних дій. Це підтверджує можливість використання запропонованої інформаційної технології як аналітичної основи для інтелектуального управління

сервісними процесами в автомобільній галузі. Водночас проведений експеримент необхідно розглядати саме як апробацію прототипу

в середовищі аналітичного моделювання, що демонструє принципову придатність розробленого підходу до практичної реалізації.

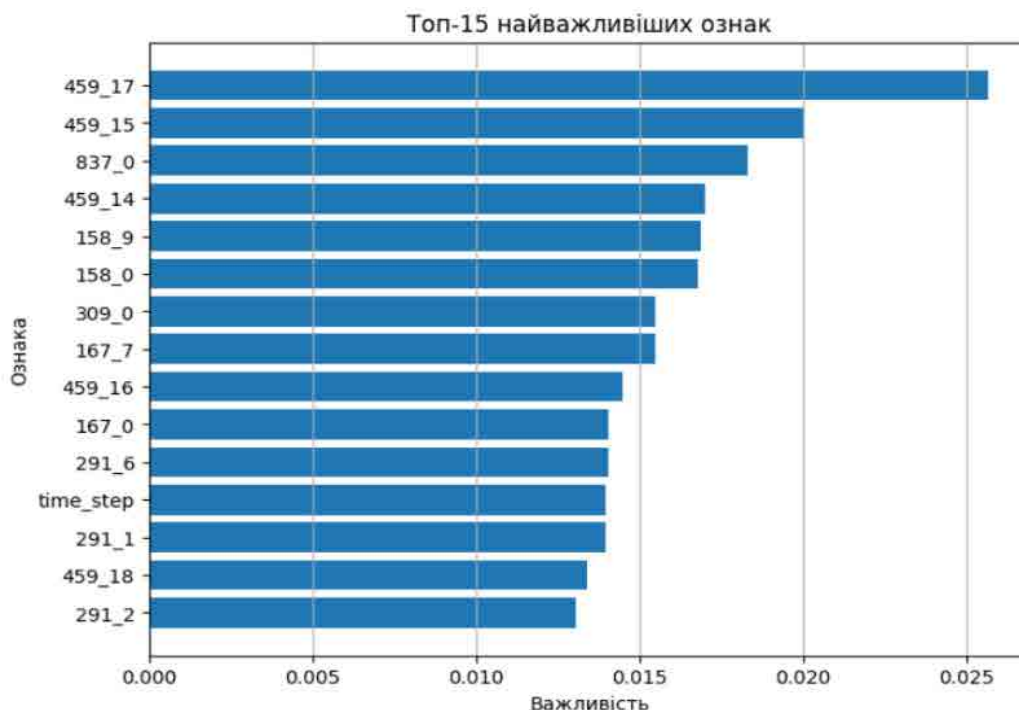


Рис. 6. Важливі ознаки

## 6. Обговорення результатів дослідження

Досягнуті результати підтвердили працездатність запропонованої інформаційної технології управління сервісними процесами в автомобільній галузі та доцільність використання підходу *Big Data* для підтримки сервісних рішень. Реалізований у середовищі *Google Colab* прототип відтворив основні етапи оброблення інтегрованих даних, оцінювання ризику відмови та автоматизованого формування сервісних рішень, що дає підстави розглядати його як придатну аналітичну основу для підтримки технічного обслуговування транспортних засобів.

Важливою особливістю експериментальних показників є суттєва незбалансованість цільового класу, оскільки кількість записів, що відповідають нормальному технічному стану, значно перевищує кількість ризикових випадків. З одного боку, це ускладнює побудову моделі, а з іншого – відтворює реальні умови експлуатації транспортних засобів, у яких потенційно небезпечні стани виникають відносно нечасто. Саме тому така структура вибірки може вважатися реалістичною для перевірки працездатності розробленого підходу.

Аналіз матриці помилок продемонстрував, що модель правильно розпізнає більшість нормальних станів і водночас виявляє значну частину ризикових випадків. У цьому разі спостерігається помітна кількість хибнопозитивних спрацювань, тобто ситуацій, коли нормальний стан визначається як ризиковий. У практичному сенсі це означає певну схильність моделі до перестраховування, однак така поведінка є менш небезпечною, ніж пропуск дійсно ризикових станів.

Високе значення площі під ROC-кривою підтвердило добру здатність моделі відокремлювати ризикові стани від нормальних. Це свідчить про те, що побудована модель не лише виконує класифікацію, а й формує коректну кількісну оцінку ризику, яка може бути використана для подальшого ранжування транспортних засобів за рівнем критичності.

Розподіл прогнозованого ризику та сформованих сервісних рішень показав, що більшість транспортних засобів долучено до категорії продовження експлуатації, меншу частку – до додаткової діагностики, і лише незначну кількість – до термінового технічного обслуговування. Такий результат є логічним

і підтверджує, що запропонована інформаційна технологія забезпечує перехід від прогнозування ризику до практичного формування сервісних дій.

Аналіз важливості ознак засвідчив, що модель спирається на групу технічних і часових параметрів, які мають найбільший вплив на результат прогнозування. Це підтверджує наявність стійких закономірностей у даних і придатність сформованого набору ознак для побудови аналітичної моделі.

Загалом результати дослідження свідчать, що запропонована інформаційна технологія є придатною для аналітичної підтримки сервісних процесів в автомобільній галузі. Водночас досягнуті результати необхідно розглядати як апробацію прототипу, що може бути надалі вдосконалений способом налаштування порогів прийняття рішень, зменшення кількості хибнопозитивних спрацювань та інтеграції з реальними сервісними системами.

## 7. Висновки

У роботі розв'язано актуальне науково-прикладне завдання, що полягає в розробленні інформаційної технології управління сервісними процесами в автомобільній галузі на основі *Big Data*. Аналіз сучасних наукових публікацій продемонстрував, що наявні підходи переважно зосереджені на прогнозуванні відмов окремих компонентів або на розв'язанні локальних задач сервісного обслуговування, тоді як питання побудови цілісної інформаційної технології, здатної інтегрувати різномірні дані, оцінювати ризик відмови й формувати сервісні рішення, залишаються недостатньо опрацьованими.

У межах дослідження розроблено математичний опис інформаційної технології, який охоплює етапи інтеграції даних, формування простору ознак, оцінку сервісної потреби, визначення рівня ризику й формування сервісного рішення. Запропонований підхід дає змогу перейти від аналізу окремих параметрів технічного стану до комплексної підтримки сервісних дій щодо транспортних засобів.

Для експериментальної перевірки працездатності розробленої інформаційної технології використано відкритий набір даних *SCANIA Component X Dataset*, на основі якого сформовано інтегрований датасет *scania\_component\_x\_combined*. Програмну реалізацію виконано мовою *Python* у середовищі *Google Colab* із використанням бібліотек *NumPy*, *Pandas*,

*Scikit-learn* та *Matplotlib*. У межах реалізації відтворено основні етапи функціонування інформаційної технології: завантаження й попереднє оброблення даних, формування ознак, побудова моделі прогнозування, оцінювання ризику відмови й автоматизоване формування сервісних рішень.

Результати експерименту підтвердили ефективність запропонованого підходу. Побудована модель продемонструвала високу здатність до розділення нормальних станів і станів ризику транспортних засобів, що підтверджується значенням площі під ROC-кривою на рівні 0.9608. Аналіз матриці помилок показав, що модель правильно розпізнає більшість нормальних станів і значну частину ризикових випадків, хоча зберігається помітна кількість хибнопозитивних спрацювань. Це свідчить про працездатність розробленої інформаційної технології та її придатність до використання як інструменту аналітичної підтримки сервісних процесів.

Додатково встановлено, що використання безперервної оцінки ризику дає змогу не лише класифікувати технічний стан транспортних засобів, а й формувати практично орієнтовані сервісні рішення, зокрема щодо продовження експлуатації, додаткової діагностики або термінового технічного обслуговування. Це підтверджує, що розроблена інформаційна технологія забезпечує перехід від прогнозування ризику до підтримки прийняття рішень у сервісній інфраструктурі автомобільної галузі.

Запропонована інформаційна технологія може бути використана як основа для подальшого розвитку цифрових сервісних систем в автомобільній галузі. Перспективою подальших досліджень є вдосконалення моделей оцінювання ризику, зменшення кількості хибнопозитивних спрацювань, адаптація порогів сервісних рішень до умов конкретної сервісної інфраструктури й інтеграція розробленого підходу з реальними інформаційними системами технічного обслуговування.

## Конфлікт інтересів

Автор статті декларує, що не має конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

**Фінансування**

Дослідження проводилося без фінансової підтримки.

**Доступність даних**

Рукопис не має пов'язаних даних.

**Використання засобів штучного інтелекту**

Автор підтверджує, що не використовував технології ШІ під час написання роботи.

**References**

- Han, X., Gstrein, O. J., Andrikopoulos, V. (2024), "When we talk about Big Data, What do we really mean? Toward a more precise definition of Big Data", *Front. Big Data*, Vol. 7:1441869. DOI: <https://doi.org/10.3389/fdata.2024.1441869>
- Theissler, A., Pérez-Velázquez, J., Kettelgerdes, M. Gordon Elger, (2021), "Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry", *Reliability Engineering & System Safety*, Vol. 215. 23 p. DOI: <https://doi.org/10.1016/j.ress.2021.107864>
- Wang, X., Liu M., Liu, C., Ling, L., Zhang, X., (2023) "Data-driven and Knowledge-based predictive maintenance method for industrial robots for the production stability of intelligent manufacturing", *Expert Systems with Applications*, Vol. 234. DOI: <https://doi.org/10.1016/j.eswa.2023.121136>
- Murtaza, A., Saher, A., Zafar, M., Moosavi, S., Aftab, M., Sanfilippo, F., (2024), "Paradigm shift for predictive maintenance and condition monitoring from Industry 4.0 to Industry 5.0: A systematic review, challenges and case study", *Results in Engineering*, Vol. 24. DOI: <https://doi.org/10.1016/j.rineng.2024.102935>
- Hossain, N., Rahman, M., Ramasamy, D., (2025), "Advances in intelligent vehicular health monitoring and fault diagnosis: Techniques, technologies, and future directions", *Measurement*, Vol. 253, Part B. DOI: <https://doi.org/10.1016/j.measurement.2025.117618>
- Errezgouny, A., Chater, Y., González, C., Cherkaoui, A., (2025), "An integrated deep learning approach for predictive vehicle maintenance", *Decision Analytics Journal*, Vol. 16. DOI: <https://doi.org/10.1016/j.dajour.2025.100597>
- Takara, L., Mariani, V., Coelho, L., (2025), "Novel engine fault diagnosis framework based on machine learning and MiniRocket feature extraction using multi-correlation feature selection and predictive power score", *Expert Systems with Applications*, Vol. 293. DOI: <https://doi.org/10.1016/j.eswa.2025.128662>
- Chen, F., Jia, H., Zhou, W., (2025), "Vehicle Maintenance Demand Prediction: A Survey", *Applied Sciences*, Vol. 15(20):11095. DOI: <https://doi.org/10.3390/app152011095>
- Tharinda, N., Kosgoda, D., Thibbotuwawa, A., Nielsen, I., (2025), "Predictive Analytics for Demand Forecasting in Automobile and Automotive Spare Parts Industry: Comparative Study", *IFAC-PapersOnLine*. Vol. 59, Issue 24, pp. 356–361. DOI: <https://doi.org/10.1016/j.ifacol.2025.11.891>
- Titu, A., Grecu, D., Pop, A., Şugar, I., (2025), "Service Process Modeling in Practice: A Case Study in an Automotive Repair Service Provider", *Applied Sciences*, No. 15(8):4171. DOI: <https://doi.org/10.3390/app15084171>
- Chen, F., Shang, D., Zhou, G., Xu, M., Ye, R., Ren, F., Wu, G., (2024), "Cooperative Static and Dynamic Correlation-Aware Learning for Vehicle Maintenance Demand Prediction", *IEEE Access*, pp. 1–13. DOI: <https://doi.org/10.1109/ACCESS.2024.3524433>
- Becker, S., Grunske, L., Mirandola, R., Overhage, S. (2006), "Performance Prediction of Component-Based Systems. In: Reussner, R.H., Stafford, J.A., Szyperski, C.A. (eds) Architecting Systems with Trustworthy Components", *Lecture Notes in Computer Science*, vol 3938, Springer, Berlin, Heidelberg. DOI: [https://doi.org/10.1007/11786160\\_10](https://doi.org/10.1007/11786160_10)
- Mahale, Y., Kolhar, S., More, A., (2025), "A comprehensive review on artificial intelligence driven predictive maintenance in vehicles: technologies, challenges and future research directions", *Discover Applied Sciences* 7(4), pp. 1–26. DOI: <https://doi.org/10.1007/s42452-025-06681-3>
- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F., (2023), "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence", *Information Fusion*, Vol. 99. pp. 1–52. DOI: <https://doi.org/10.1016/j.inffus.2023.101805>
- Andersen, J., Degn, L., Fishberg, R., Graversen, E., Horbach, S., Schmidt, E., Schneider, J., Sørensen, M., (2025), "Generative Artificial Intelligence (GenAI) in the research process – A survey of researchers' practices and perceptions", *Technology in Society*, Vol. 81, pp. 1–11. DOI: <https://doi.org/10.1016/j.techsoc.2025.102813>

16. SCANIA Component X Dataset: A Real-World Multivariate Time Series Dataset for Predictive Maintenance.  
 DOI: <https://doi.org/10.58141/1w9m-yz81>

*Received (Надійшла) 19.03.2026*

*Accepted for publication (Прийнята до друку) 12.04.2026*

*Publication date (Дата публікації) 29.05.2026*

#### *Відомості про авторів / About the Authors*

**Склярів Ілля Ігорович** – Харківський національний автомобільно-дорожній університет, аспірант кафедри комп'ютерних наук і інформаційних систем, Харків, Україна;

**Illia Skliarov** – Kharkiv National Automobile and Highway University, Postgraduate Student of the Department of Computer Science and Information Systems, Kharkiv, Ukraine;

e-mail: [is.ilyasklyarov@gmail.com](mailto:is.ilyasklyarov@gmail.com)

ORCID ID: <http://orcid.org/0009-0001-9116-5315>

### INFORMATION TECHNOLOGY FOR INTELLIGENT MANAGEMENT OF SERVICE PROCESSES IN THE AUTOMOTIVE INDUSTRY BASED ON BIG DATA

**The subject of the research** is the methods, models, and tools for building information technology of intelligent management of service processes based on Big Data. **The objective of the work** is to develop information technology for intelligent management of service processes in the automotive industry based on Big Data for the personalization of service offerings, automation of decision-making regarding maintenance, and improvement of interaction efficiency between service centers, manufacturers, and consumers. In accordance with the stated objective, **the following tasks** must be accomplished: to analyze modern approaches to the management of service processes in the automotive industry and determine the possibilities of using Big Data for their intellectualization; to analyze the sources and structure of data used in the service processes of the automotive industry; to develop the structure of the information technology for intelligent management of service processes based on the integration of heterogeneous data; to implement the proposed information technology and conduct an experimental evaluation of its performance. **Methods.** The work employs methods of system analysis to study service processes in the automotive industry and to determine the main information flows between service centers, manufacturers, and consumers. Methods of data integration, cleaning, and analysis are used for processing large volumes of heterogeneous data. The development of information technology is carried out using methods of data mining, machine learning, classification, and decision support. For the experimental validation of the proposed approach, methods of computational modeling, comparative analysis, and evaluation of model quality using standard classification metrics are applied. The software implementation is performed in Python within the Google Colab environment. **Research results.** Information technology for managing service processes in the automotive industry based on Big Data has been developed, which ensures the integration of technical, operational, and service data, failure risk assessment, and the formation of service decisions. A software prototype of this technology has been implemented in the Google Colab environment using Python, based on the `scania_component_x_combined` dataset, formed from the open SCANIA Component X Dataset. Experimental validation confirmed the operability of the proposed approach and the high quality of distinguishing between normal and risk states of vehicles. **Conclusions.** The results obtained confirmed the feasibility of using Big Data for analytical support of service processes in the automotive industry. The proposed information technology can be used as a basis for improving the validity of service decisions, timely identification of risk states, and further development of digital systems for vehicle maintenance.

**Keywords:** Big Data; automotive industry; service processes; information technology; maintenance; failure risk prediction; machine learning; data analysis; vehicles; service decisions.

#### *Бібліографічні описи / Bibliographic descriptions*

Склярів І. І. Інформаційна технологія інтелектуального управління сервісними процесами в автомобільній галузі на основі *Big Data*. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 57–71. DOI: <https://doi.org/10.30837/0135-1710.2026.189.057>

Skliarov, I. (2026), "Information technology for intelligent management of service processes in the automotive industry based on big data", *Management Information System and Devices*, No. 2 (189), P. 57–71. DOI: <https://doi.org/10.30837/0135-1710.2026.189.057>



Vitalii Tkachov

## A METHOD FOR LOCAL RECOVERY OF INFORMATION SYSTEM PROCESSES ON A MOBILE PLATFORM

**The subject of this study** is local disruptions in the internal state of individual processes within an information system on a mobile platform, which prevent the correct continuation of execution without a special recovery intervention. **The aim** of this work is to develop a method for the local recovery of information system processes on a mobile platform, which, in the event of a local disruption of a process's internal state, ensures the continuation of its execution through idempotent re-execution and the deterministic reconstruction of the minimally sufficient local state under conditions of limited resources and intermittent connectivity. The article aims to accomplish the following **tasks**: formalize the local state of the process and the signs of its disruption; establish classes of process steps based on the admissibility of re-execution; develop a rule for choosing between micro-recovery and micro-restart; propose a method for deterministic reproduction of the minimally sufficient local state of the process; conduct a model study of the proposed method's effectiveness compared to a full process restart and deferred service recovery. **Results Achieved.** A formalization of the local process state has been developed in the form of a vector containing the current step number, internal process data, step input data, service execution flags, and the frequency of result fixation outside the process. A local state violation is defined via a vector of inconsistency and a scalar measure of local process state damage. A local recovery rule is proposed that formalizes the choice between micro-recovery and micro-restart, taking into account the expected process state after the recovery action, the trade-off between action requirements and available time and resources, estimates of external result replication, and the compliance of the current state with restart conditions. A method has been developed for the deterministic reconstruction of the minimally sufficient local process state, which determines the selection of the restart reference point and the formation of the set of local state coordinates, input data, and service flags necessary for the correct continuation of execution. The results of the model study showed that the proposed method ensures a faster reduction in the local process state damage index, a lower cumulative recovery deviation index, a higher probability of successful return of the process to correct execution, a shorter average recovery time, a smaller amount of data required for restart, and a lower probability of duplicating the external result. **Conclusions:** The proposed method is appropriate as a separate process mechanism in a system for ensuring the resilience of an information system on a mobile platform, as it allows for reducing the duration of local process execution disruptions and increasing system resilience without resorting to a full restart or service recovery when local recovery is sufficient.

**Keywords:** information system on mobile platform; process recovery; idempotent re-execution; deterministic state reconstruction; survivability.

### 1. Introduction

In the classification of information systems, there is a class of systems designed to operate in distributed computing environments. One of the subclasses of this class consists of information systems on mobile platforms (ISMP). For ISMP, the key factors are the dynamic availability of mobile platform resources, the mobility of its nodes, the intermittent connectivity of platform components or nodes, the irregularity of telemetric monitoring of their status, and strict time constraints on decision-making regarding the control of nodes and components. Such operating conditions are often associated with the edge computing environment [1]. In such an environment, these properties are further complicated by the service-segmented architecture of the information system, the need to maintain the

continuity of components whose state changes during ISMP operation, and the variability of available resources during short intervals of the main function's execution. Under such conditions, the ISMP's ability to maintain or restore the execution of critical functions is most often associated with the provisioning, migration, or reconfiguration of service instances, although disruptions often occur on a smaller scale – during a single process that loses its correct local state before the ISMP proceeds to a full service or resource reconfiguration [2]–[4]. The availability of the core function in such a case is directly linked to the survivability of the ISMP as a coherent logical unit for executing the core function. The survivability of an information system is viewed as the property of maintaining the execution of target functions or restoring it under conditions of destructive influences, structural

disruptions, and resource constraints [5]–[7]. For ISMP, this interpretation is decisive, as it provides grounds for considering the local recovery of an individual process as one of the mechanisms for ensuring the survivability of the information system without resorting to a full service reconfiguration.

Current research in the area of system survivability and operational resilience in edge computing environments is primarily focused on service migration, reconfiguration of their deployment, redundancy, orchestration, restoration of processing sequences, and maintaining the continuity of service operation, the state of which changes during execution, under dynamic platform resource constraints [1, 3, 4]. Such approaches form the basis for adapting an information system to component connectivity failures or platform resource shortages. At the same time, the primary unit of recovery in these approaches is typically a service, a service chain, or a service segment. In cases where the disruption is local in nature and related to damage to the internal state of a single process, a full service reconfiguration proves excessive in terms of time, computational resources, and service data exchange [8]. This necessitates the identification of a separate task for rapid local process recovery without resorting to broader system-wide reconfiguration procedures.

The problem is that simply restarting a process after a failure does not guarantee correct recovery. For processes in the ISMP distributed environment, whose state changes during execution, restarting without accounting for the semantics of actions can cause duplication of external effects, repeated saving of intermediate results, disruption of local state consistency, and accumulation of inconsistent process artifacts [9]. This is why, in modern research in the area of fault tolerance for serverless systems with stateful processes – particularly their process control subsystems – idempotence is considered a fundamental prerequisite for permissible retries [10], while the uniqueness of external effect generation is ensured through special logging schemes, duplicate elimination, and action repeatability verification at the system level [11]. For mobile platforms, the execution of a restart function is considered permissible only for idempotent functional steps, whereas non-idempotent steps require a separate mechanism for coordinated recovery [12].

Another line of research links recovery to the replay of the execution flow, state saving, and recovery from a checkpoint, as well as to the deterministic replay of the

execution of a set of processes as a service [13, 14]. Such approaches have demonstrated their suitability for reproducible state reconstruction, fault analysis, and returning the system to coordinated continued operation. However, these approaches are focused on debugging and diagnosing distributed systems, on restoring execution variants, or on cluster and serverless environments. For ISMP, where available resources, time, and connectivity are variable and limited, the task of reconstructing the minimally sufficient local state of a process required for rapid resumption of execution within an acceptable recovery interval has not been formally defined separately.

For ISMP, there are approaches focused on restoring critical services within connectivity windows of variable duration and bandwidth, as well as on analyzing service survivability under resource-time variables. Their application has made it possible to formalize transitions between service profiles, assess the shortfall in meeting threshold requirements for QoS metrics, and perform analytical ranking of services for operational parametric resource management [15, 16]. However, these approaches do not separately address the task of restoring the locally damaged state of an ISMP component or node process, do not establish classes of process steps based on the admissibility of re-execution, and do not define rules for the deterministic reproduction of the minimally sufficient local state to continue execution without transitioning to full service recovery.

Thus, for ISMP, the local recovery of processes after a violation of their internal state under conditions of limited resources and intermittent connectivity remains a distinct scientific problem. Its solution requires the formalization of two interrelated principles. The first is that the re-execution of a process step must be either semantically valid or such that controlled compensation for the re-generated effect is ensured. The second is that the continuation of process execution after a failure must rely on the deterministic reconstruction of a minimally sufficient local state. Within the scope of this task, local process recovery should be implemented in two modes. The first mode consists of micro-recovery, i.e., the local correction of a damaged fragment of the internal state without returning to the re-execution of the preceding fragment of the process. The second mode consists of a micro-restart, in which the process returns to a previously valid state by re-executing only those steps necessary for correct continuation of operation. It is precisely the choice between these two modes

and the rules for their implementation that define the essence of the proposed method for local recovery of ISMP processes with idempotent re-execution and deterministic state reproduction.

## 2. Analysis of Recent Research and Publications

---

For multi-domain edge-to-cloud environments, special attention has been paid to the problem of idempotency in the service network. In [17], the situation is considered where a repeated HTTP POST request after a failure produces a different external effect than the initial execution. To eliminate this effect, a two-component solution is proposed that combines an idempotency control mechanism with a separate means of resolving functional failures. Within the scope of the subject area, the research aims to confirm the fundamental necessity of controlling the re-execution of actions associated with an external effect. At the same time, the proposed approach operates at the level of network interaction between microservices and the resending of external requests, rather than at the level of restoring the internal local state of an individual process. The approach does not formalize the composition of the minimally sufficient reproducible state of the process and does not implement a function for choosing between local state correction and local restart of the information system process.

The problem of duplicating already executed processes in the context of ensuring service operation after a failure (disruption) is considered in [18] for serverless computing environments. The authors proceed from the assumption that the usual re-execution of a process after a disruption can distort the operation of the entire service, therefore they ensure exactly-once semantics through asymmetric logging, where either read operations or write operations are logged for different load types, and introduce a selection criterion between the two protocols and a mechanism for switching between them without stopping process execution. Experiments have demonstrated a reduction in latency and a decrease in logging overhead compared to previous solutions.

In [19], the features of using a state-saving and restoration mechanism in user space for migrating a containerized application between nodes by saving its current state and subsequently restoring it in a different runtime environment were investigated. It is shown that the time required to create a checkpoint is primarily determined by the amount of memory used, whereas the

restoration time has a different dependency related to the procedure for reconstructing the saved state. The results obtained confirm the technical feasibility of this approach for resource-constrained infrastructures. This work focuses on the scenario of transferring process execution between nodes and includes local process recovery after a disruption of its internal state. The work does not formalize the division of process steps based on re-execution feasibility and does not define rules for constructing the minimally sufficient local state required to resume execution after a failure.

Maintaining delay-sensitive services in edge computing is the subject of [20], which examines the robust migration of edge microservices whose state changes during execution. The main focus is on two interrelated tasks: the migration of a containerized microservice and the preservation of its connection with a mobile user. To this end, a network solution based on OvS using Podman and CRIU is proposed for experimental evaluation of migration duration and service downtime. Additionally, an analytical model of the upper bounds of these metrics has been constructed. The significance of this work lies in the transition from a general concept of migrating a service whose state changes during execution to a concrete instrumental and analytical description. The subject of the study is the migration of a microservice between nodes. Consequently, local damage to the process state and the admissibility of re-executing a single step sufficient to continue process execution without full migration between information system components remain outside the scope of consideration.

For the tasks of migrating video analytics applications at edge nodes, an approach that accounts for the heterogeneity of their state is proposed in [21]. It is shown that existing solutions focused on cloud scenarios do not ensure proper migration of such applications in an edge environment, since different components of their state have distinct natures and require different processing methods. In view of this, the authors applied separate processing of three types of state using pre-processing, synchronization, and replay, and introduced a state store and an auxiliary control component, which allows for porting with minimal changes to the application code. Thus, the study confirms the feasibility of structured, rather than complete, state reproduction. At the same time, process replay is a component of the service migration procedure between nodes that does not account for local process

recovery following a disruption of its internal state. The aforementioned work does not address the choice between local state correction and micro-restart, nor does it formalize the verification of the validity of re-executing individual process steps.

Of particular interest is the work [22], which proposes a method for deterministic parallel execution of processes for modern data center applications (DORADD), in which the result depends on the current state of the processed data and the sequence of operations. The approach is based on ensuring the same result for the same input data by ordering data access and constructing a dynamic acyclic graph of dependencies between operations. It is shown that this approach allows combining high throughput with low latency and can be used for deterministic replay of a logging scheme during rapid recovery after failures. The authors confirm the feasibility of deterministic process replay in information systems where result reproducibility is of fundamental importance. At the same time, DORADD is focused on high-performance execution of ordered operations in data center systems and does not address the problem of local process recovery in subsystems. It does not formalize the locally damaged state of a process, does not establish classes of process steps based on the admissibility of re-execution, and does not define rules for constructing a minimally sufficient local state for rapid process recovery in the event of failures.

An analysis of works [17]–[22] shows that current research covers individual aspects related to the problem of this study, specifically, control of query replay at the network level, ensuring the uniqueness of result recording through external state logging, state preservation and recovery during process execution transfer, structured processing of various components of the process state, as well as deterministic process replay in distributed information systems. At the same time, these results do not solve, within a single formulation, the problem of local recovery of ISMP processes under resource constraints and intermittent connectivity of its components and nodes. In particular, the works under consideration lack a formalization of damage to the internal state of the process, do not establish classes of process steps based on the admissibility of re-execution, do not define the conditions for choosing between local state correction and micro-restart, and do not construct a method for deterministic reproduction of the minimally sufficient local state to continue process execution after a failure. It is precisely this unresolved part of the

problem that determines the feasibility of developing a method for local recovery of ISMP processes with idempotent retry and deterministic state reconstruction.

### 3. Research Objectives and Tasks

The objective of the research is to develop a method for the local recovery of ISMP processes that, in the event of a local violation of the process's internal state, ensures the continuation of its execution through idempotent re-execution and the deterministic reconstruction of a minimally sufficient local state under conditions of limited resources and intermittent connectivity of components or nodes.

To achieve this goal, the following tasks must be solved:

- formalize the local state of the ISMP process, identify signs of its violation, and establish classes of process steps based on the admissibility of re-execution, which will allow defining the process-level formulation of the local recovery problem;
- develop a local process recovery rule that determines the choice between micro-recovery of a local state fragment and micro-restart of the process with re-execution of a subset of steps, ensuring a well-founded selection of recovery actions within a single method;
- develop a procedure for deterministically reproducing the minimally sufficient local state of the process after a local failure, taking into account the idempotent semantics of retry and the resource- and time-based constraints of ISMP operation, which will ensure the continuation of process execution without transitioning to full service recovery;
- develop a set of metrics and conduct an experimental study of the proposed method's effectiveness on representative scenarios of local process state violations, which will enable the evaluation of its performance based on temporal, resource, and functional characteristics.

### 4. Materials and Methods

The subject of the study is local disruptions in the internal state of individual processes within the ISMP, which do not necessarily trigger the ISMP's transition to service reconfiguration but prevent the correct continuation of the target function's execution without special recovery intervention. Such disruptions include loss of consistency in intermediate data, disruption of the

internal sequence of execution steps, partial fixation of service data or other input information, and a mismatch between the current local state of the process and the state required for its further execution.

The generalized research framework encompasses event and data sources, computational nodes, the process execution environment, local logs of service data (markers), results of process execution steps, and local recovery mechanisms. Within this framework, each process receives input data, forms a sequence of execution steps, modifies its own local state, and, if necessary, generates a result for recording outside the process. Local recovery mechanisms use current service records, available local data, and process service markers, an assessment of the platform's available resources, and parameters of the current connectivity interval to determine whether a damaged state fragment can be corrected or a local process retry can be initiated. Thus, the proposed method is intended for ISMPs in which process recovery is an internal operation within the operational interval of functioning and does not amount to a complete migration of the process (or the service within which it may run), a transition between service profiles, or a general reconfiguration of the ISMP.

The research materials consist of time series of telemetric indicators of mobile platform resource usage, process execution logs, service records of completed and incomplete steps, and a set of situations involving local violations of the internal state of processes that ensure the functioning of the information system. Within these situations, the following are taken into account: partial execution of a step, failure after recording an intermediate result, loss of part of local data and service markers, re-arrival of an input event, and interruption of available communication between system components. For each situation, the following are recorded: available unoccupied resources of the mobile platform, permissible local recovery time, the volume of service data required to reproduce the state, and the presence or absence of results from the previous execution of the process recorded outside its boundaries.

Several interrelated approaches were used to solve the problem:

- the process behavior is described as a sequence of steps with local data and service completion flags;
- to analyze the admissibility of re-execution, the principles of idempotent semantics were used, according to which repeating the same step should not change the final result compared to a single execution. If an external

result is present, its controlled consideration or compensation must be ensured;

- for recovery after a local failure, a deterministic replay approach is used, whereby the continuation of process execution is based on the reconstruction of a minimally sufficient local state, ensuring the uniqueness of the subsequent process flow under identical input data, service markers, and the order of replay;

- The effectiveness of the proposed method was evaluated based on a simulation study comparing local recovery time, the proportion of successfully recovered processes, the scope of the reconstructed state, the number of permissible re-executions, and the frequency of re-executing actions whose results had already been recorded outside the process.

## 5. Research Results and Discussion

### 5.1. Formalization of the local process state and the semantics of re-execution

In previously proposed approaches to analyzing the process level of ISMP, the process state space was defined through invariants of the queue, time buffer, data validity, and causal-temporal order, which is sufficient for detecting violations at the level of the entire process environment [15]. In service-oriented approaches, the service profile served as the primary unit of evaluation, and recovery was described as a transition between profiles with feasibility checks within connectivity windows [16]. For the problem considered in this study, this is insufficient, since the object of recovery is a single process with a locally damaged internal state, for which it is necessary to determine the possibility of safe continuation of execution.

Let a component or node of the ISMP implement a set of processes:

$$P = \{p_i \mid i = 1, \dots, n\}. \quad (1)$$

For each process  $p_i$ , an ordered set of its steps is specified:

$$A_i = \{a_{i,r} \mid r = 1, \dots, m_i\}, \quad (2)$$

where:  $a_{i,r}$  – the  $r$ -th step of the process  $p_i$ ;  $m_i$  – the total number of steps in the local execution scenario of the process  $p_i$ .

The local state of the process  $p_i$  at discrete time  $t$  is given by the vector:

$$\vec{s}_i(t) = (r_i(t), \vec{d}_i(t), \vec{u}_i(t), \vec{v}_i(t), \mu_i(t))^T, \quad (3)$$



where  $r_i(t)$  – the step number of the process at time  $t$ ;  $\vec{d}_i(t)$  – a vector of internal process data generated in previous steps and required to execute the current step without recalculating the entire preceding portion of the process;  $\vec{u}_i(t)$  – a vector of input data for the current step, i.e., the values, events, and parameters with which this step is to be executed;  $\vec{v}_i(t)$  – a vector of status flags indicating which internal operations of the current step have already been completed, which intermediate results have already been recorded, and which service actions have already been performed;  $\mu_i(t)$  – the number of times the result of the current step has already been recorded outside the process (entry to an external log, message transmission, update of a shared repository, sending of a control command, etc.).

The vector  $\vec{\delta}_i(t)$  (3) specifies the minimum sufficient description of the process for the local recovery task. The component  $r_i(t)$  determines the process's position in the execution sequence;  $\vec{d}_i(t)$  – the presence of internal data without which it is impossible to proceed to the next action;  $\vec{u}_i(t)$  – the completeness of input data for the current step;  $\vec{v}_i(t)$  – the degree of completion of internal operations and service updates;  $\mu_i(t)$  – whether the result of the current step has been updated once or repeatedly outside the process. It is the combination of these components that makes it possible to determine whether the process can continue directly, whether a local correction of the state is needed, or whether some steps need to be re-executed.

To formally identify a local disturbance, the misalignment vector is introduced:

$$\vec{\delta}_i(t) = (\delta_{i,1}(t), \delta_{i,2}(t), \delta_{i,3}(t), \delta_{i,4}(t))^T, \quad (4)$$

where its components are defined by the following relations:

$$\delta_{i,1}(t) = \|\vec{u}_i(t) - \hat{\vec{u}}_{i,r_i(t)}\|,$$

$$\delta_{i,2}(t) = \|\vec{d}_i(t) - \hat{\vec{d}}_{i,r_i(t)}\|,$$

$$\delta_{i,3}(t) = |r_i(t) - v_i(\vec{v}_i(t))|,$$

$$\delta_{i,4}(t) = \max\{0, \mu_i(t) - \sigma_{i,r_i(t)}\}$$

where  $\hat{\vec{u}}_{i,r_i}$  – reference vector of input data for step  $a_{i,r}$ ;  $\hat{\vec{d}}_{i,r_i}$  – reference vector of internal data for which the execution of step  $a_{i,r}$  is valid;  $v_i(\vec{v}_i(t))$  – operator for

determining the number of the last step, the completion of which is confirmed by service flags;  $\sigma_{i,r_i}(t) \in \{0,1\}$  – indicator of the validity of recording the result of step  $a_{i,r}$  outside the process.

The components of  $\vec{\delta}_i(t)$  (4) have the following meanings:

–  $\delta_{i,1}(t)$  describes the deviation between the actual input data of the current step and the input data under which the execution of this step is correct. In other words, this component indicates whether the current set of input values matches the set with which the step  $a_{i,r}$  is intended to be executed. For example, if the current step is supposed to process a telemetry packet with a specific source ID, timestamp, and set of measured parameters, then  $\delta_{i,1}(t)$  increases if some of this data is lost, corrupted, or if an outdated data packet is received again;

– The  $\delta_{i,2}(t)$  characterizes the discrepancy between the actual internal data of the process and the internal state required to execute the current step. This component indicates whether the process has retained the intermediate data, results of previous calculations, and control values without which the current step cannot proceed correctly. For example, if a processing buffer, an intermediate feature vector, or an indicator of successful completion of the previous stage should already have been generated before executing the step, then corruption or loss of even a portion of this data leads to an increase in  $\delta_{i,2}(t)$ ;

–  $\delta_{i,3}(t)$  describes the discrepancy between the “step number at which the process is currently located and the number of the last step whose completion has been confirmed by system logs. This component indicates whether the process's current position in the execution sequence aligns with the entries actually recorded in the system log. For example, if the process has already moved to step  $r_i(t) = 5$ , but the status marks confirm the completion of only the third step, then  $\delta_{i,3}(t)$  indicates a discrepancy in the execution order and shows that the process has progressed further than is confirmed by its status;

–  $\delta_{i,4}(t)$  indicates that the permissible number of times the current step's result has been recorded outside the process has been exceeded. This component is zero if the current step's result has not yet been recorded or has been recorded within the permissible

number of times, and takes on a positive value if the re-execution has already led or may lead to excessive recording of that result. For example, if the current step has already generated a record in an external log, sent a message to another component, or saved a command to an external storage, then re-executing this same step without accounting for the value of  $\mu_i(t)$  may lead to duplication of the result, which is reflected by the  $\delta_{i,4}(t)$  component.

These four components represent the main types of local inconsistency that are critical to the process recovery problem: input data mismatch, corruption of internal process data, violation of the confirmed

$$\theta_i(t) = \bar{\omega}_i^T \bar{\delta}_i(t), \quad \bar{\omega}_i = (\omega_{i,1}, \omega_{i,2}, \omega_{i,3}, \omega_{i,4}), \quad \sum_{j=1}^4 \omega_{i,j} = 1, \quad \omega_{i,j} \geq 0. \quad (6)$$

The metric  $\theta_i(t)$  (6) is a scalar measure of the current degree of damage to the local state of the process. Its value is zero for a locally valid state and increases as the inconsistency grows across one or more components of the vector  $\bar{\delta}_i(t)$ . The coefficients  $\omega_{i,1}, \omega_{i,2}, \omega_{i,3}, \omega_{i,4}$  determine the contribution of, respectively, input data mismatch, damage to the process's internal data, violation of the confirmed step execution order, and re-fixation of the result outside the process to the final value  $\theta_i(t)$ . Subsequently,  $\theta_i(t)$  (6) is used for a quantitative comparison of possible recovery actions.

The change in the local state of the process at step  $a_{i,r}$  is represented as a transition operator:

$$\bar{s}_i(t+1) = \hat{f}_{i,r_i(t)}(\bar{s}_i(t), \bar{x}_i(t), \xi_i(t)), \quad (7)$$

where  $\bar{x}_i(t)$  is the vector of actual input data for the current step, i.e., the set of values, events, and parameters with which this step is actually executed at time  $t$ ;  $\xi_i(t)$  is the execution result code for the current step, distinguishing between successful completion, partial completion, and abnormal termination.

Equation (7) specifies the local state to which the process transitions after executing the current step based on the actual input data and the actual execution result.

To isolate the part of the process state that corresponds to the result already recorded outside the process, the following statement is introduced:

$$\bar{y}_i(t) = \pi_i(\bar{s}_i(t)). \quad (8)$$

The vector  $\bar{y}_i(t)$  (8) contains only those execution results that have an external fixation and are therefore

execution order, and excessive recording of the current step's result outside the process.

The set of locally valid process states  $p_i$  is defined as follows:

$$\Omega_i = \{\bar{s}_i(t) | \bar{\delta}_i(t) = \vec{0}\}. \quad (5)$$

Thus, a process state belongs to the set  $\Omega_i$  (5) if and only if the input data of the current step, the internal process data, the order of step execution, and the fact of recording the result outside the process are consistent.

To quantitatively assess the extent of local damage to the condition, a scalar index is introduced:

relevant for the analysis of re-execution. Such results may include input data, a record in an external log, a transmitted message, an update to a shared store, or a command sent to another ISMP component.

For the step  $a_{i,r}$ , single and double executions from the same specially fixed initial state  $\bar{s}_i(t)$  are considered:

$$\bar{s}_{i,r}^{(1)} = \hat{f}_{i,r_i}(\bar{s}_i, \bar{x}_i, 1), \quad \bar{s}_{i,r}^{(2)} = \hat{f}_{i,r_i}(\bar{s}_{i,r}^{(1)}, \bar{x}_i, 1) \quad (9)$$

In formula (9), the state  $\bar{s}_{i,r}^{(1)}$  corresponds to a single execution of the step  $a_{i,r}$ , whereas the state  $\bar{s}_{i,r}^{(2)}$  corresponds to a repeated execution of the same step with the same input data  $\bar{x}_i$ . Comparing these two states is necessary to determine whether the repeated execution changes the result already recorded outside the process.

Process steps with idempotent semantics of re-execution are defined by the set:

$$\Pi_i^{(1)} = \left\{ a_{i,r} \in A_i \mid \pi_i(\bar{s}_{i,r}^{(2)}) = \pi_i(\bar{s}_{i,r}^{(1)}) \right\}. \quad (10)$$

Thus, a step  $a_{i,r}$  belongs to the set  $\Pi_i^{(1)}$  (10) if its repeated execution does not change the result recorded outside the process compared to a single execution. For example, this class may include steps involving the recalculation of an intermediate value, the re-evaluation of a logical condition, or the re-generation of internal data, provided that such re-execution does not produce a new record, message, or other additional result outside the process.

Steps that can be repeated to produce a one-time result through a specific compensatory action are defined as follows:

$$\Pi_i^{(2)} = \left\{ a_{i,r} \in A_i \mid \exists \bar{g}_{i,r} : \pi_i \left( \bar{g}_{i,r} \left( \bar{s}_{i,r}^{(2)} \right) \right) = \pi_i \left( \bar{s}_{i,r}^{(1)} \right), a_{i,r} \notin \Pi_i^{(1)} \right\}. \quad (11)$$

The set  $\Pi_i^{(2)}$  (11) includes steps for which re-execution in itself changes the result recorded outside the process; however, there is an operator  $\bar{g}_{i,r}$  that eliminates redundant recording and reduces the result of two executions to that of a single execution. For example, if a repeat execution creates a duplicate entry in an external log, a duplicate service message, or a second instance of an already recorded result, and this duplicate can be detected and eliminated by a special procedure, then the corresponding step belongs to the set  $\Pi_i^{(2)}$ .

Finally, steps that cannot be repeated without a separate change in the process state are defined as follows:

$$\Pi_i^{(3)} = A_i \setminus \left( \Pi_i^{(1)} \Pi_i^{(2)} \right). \quad (12)$$

The set  $\Pi_i^{(3)}$  (12) encompasses steps for which re-execution results in a change to the result recorded outside the process, and for which there is no  $\bar{g}_{i,r}$  operator capable of mapping the result of two executions to the result of a single execution. In other words, this class includes steps whose re-execution is not permitted without first changing the state of the process. For example, this could be the formation of a unique external command, an irreversible update to a shared repository, or the re-transmission of an action that cannot be duplicated.

Thus, the relations  $A_i$  (2),  $\bar{s}_i(t)$  (3),  $\bar{\delta}_i(t)$  (4),  $\Omega_i$  (5),  $\theta_i(t)$  (6), and the classes  $\Pi_i^{(1)}$ ,  $\Pi_i^{(2)}$ ,  $\Pi_i^{(3)}$  (10)–(12) form a formal description of the local state of the process, the conditions for its validity, and the semantics of step repetition. This makes it possible to distinguish a locally valid state of the process from states in which the consistency of input data, internal process data, the confirmed execution order, or the recording of results outside the process is violated. Furthermore, the introduced classification of process steps based on the admissibility of re-execution establishes a formal basis for distinguishing between steps for which direct re-execution is admissible, admissible with a compensation operator, or inadmissible without a prior change in the process state.

## 5.2. The rule of local process recovery

Modern approaches to the local recovery of processes whose state changes during execution are based

on the premise that the decision to restart cannot be made solely upon the occurrence of a failure. Works on local component restart and controlled recovery of microservice systems [8, 10] emphasize the need to verify the boundaries of recovery and the ability to compensate for results already recorded outside the process. In studies focused on permissible request retries and the singleness of result recording [17, 10], it is shown that a retry is permissible only under controlled step semantics. In works devoted to deterministic replay of execution [14, 22], it is specifically emphasized that fast recovery must rely on reproducible input data, service markers, and execution order. That is why, in this study, the rule for local process recovery is formulated as a formal decision-making problem between micro-recovery and micro-restart, while simultaneously accounting for local state corruption, resource-time constraints, the semantics of re-execution, and the reproducibility of the data and service markers required to resume process execution.

For the process  $p_i$  at the time step  $t$ , the following set of local actions is considered:

$$U_i(t) = \{\rho_i, \kappa_i\} \quad (13)$$

where  $\rho_i$  is the micro-restoration of a local state fragment without restarting the step;  $\kappa_i$  is the micro-restart of the process with a return to a previously valid state and the re-execution of a portion of the steps.

The expected process inconsistency vector after applying the action  $u \in U_i(t)$  is given as:

$$\bar{\delta}_i''(t+1) = \bar{T}_i(u, t) \bar{\delta}_i(t) + \bar{\varepsilon}_i(u, t), \quad (14)$$

where  $\bar{T}_i(u, t)$  is the transformation matrix of the current misalignment vector under the action of  $u$ ;  $\bar{\varepsilon}_i(u, t)$  is an additional vector that accounts for the portion of misalignment that remains after the action is performed or arises due to the incomplete accuracy of the process state reconstruction.

Equation (14) can be interpreted as follows. First, the current inconsistency vector  $\bar{\delta}_i(t)$  is transformed by the matrix  $\bar{T}_i(u, t)$ , which shows exactly how the selected action changes each component of the inconsistency. After that, the vector  $\bar{\varepsilon}_i(u, t)$  is added, which represents the residual inconsistency not

eliminated by this action. Thus,  $\tilde{\delta}_i^u(t+1)$  is an estimate of what each component of the inconsistency will become after applying action  $u$ .

$$\bar{T}_i(\rho_i, t) = \text{diag}(\alpha_{i,1}(t), \alpha_{i,2}(t), \alpha_{i,3}(t), \alpha_{i,4}(t)) , \quad 0 \leq \alpha_{i,j}(t) < 1 . \quad (15)$$

The coefficient  $\alpha_{i,j}$  indicates what portion of the  $j$ -th component of misalignment remains after micro-restoration. If  $\alpha_{i,j} = 0$ , then the corresponding component is completely eliminated. If  $\alpha_{i,j}$  is close to one, then micro-recovery hardly reduces this

$$\bar{T}_i(\kappa_i, t) = \text{diag}(\beta_{i,1}(t), \beta_{i,2}(t), \beta_{i,3}(t), \beta_{i,4}(t)) , \quad 0 \leq \beta_{i,j}(t) < 1 . \quad (16)$$

The coefficient  $\beta_{i,j}(t)$  (16) indicates what portion of the  $j$ th inconsistency component remains after the process returns to a previously valid state and the relevant portion of the steps is re-executed. Unlike micro-recovery, a micro-restart does not change the process state locally at the current point, but rather

$$\beta_{i,1}(t) \leq \alpha_{i,1}(t), \quad \beta_{i,2}(t) \leq \alpha_{i,2}(t), \quad \beta_{i,3}(t) \leq \alpha_{i,3}(t), \quad \beta_{i,4}(t) \leq \alpha_{i,4}(t). \quad (17)$$

Equations (17) imply that, for the first three components, a micro-restart typically leaves no greater inconsistency than a micro-recovery. This is because returning to a previously valid state eliminates the discrepancy between the actual and confirmed step numbers and allows the input data and internal process data for the current step to be regenerated. For the fourth component, which involves re-recording the result outside the process, this advantage does not hold, so its evaluation is performed separately.

The expected scalar estimate of the process state after applying action  $u$  is determined as follows:

$$\Theta_i(u, t) = \bar{\omega}_i^T \tilde{\delta}_i^u(t+1) , \quad (18)$$

The quantity  $\Theta_i(u, t)$  (18) is the weighted sum of the components of the vector  $\tilde{\delta}_i^u(t+1)$  and indicates how consistent the state of the process is expected to be after applying action  $u$ . The smaller the value of  $\Theta_i(u, t)$ , the smaller the expected damage to the local state of the process after executing this action. That is why the quantity  $\Theta_i(u, t)$  is used to compare microrestore and microreset based on the result expected after their application.

To account for the limitations of the mobile platform, a vector of action requirements for time and resources is introduced:

For the micro-restoration  $\rho_i$ , the matrix  $\bar{T}_i(\rho_i, t)$  takes the form:

component. Thus, the diagonal matrix (15) provides a separate estimate of the micro-recovery result for each component of the vector  $\tilde{\delta}_i(t)$ .

For micro-restart  $\kappa_i$ , similarly:

by returning to a previous state in which the process execution was still valid.

For components related to the input data of the current step, the internal data of the process, and the step number, the following inequality typically holds:

$$\bar{c}_i(u, t) = (c_{i,1}(u, t), c_{i,2}(u, t), c_{i,3}(u, t), c_{i,4}(u, t))^T , \quad (19)$$

where  $c_{i,1}(u, t)$  – the time required to perform a local recovery operation;  $c_{i,2}(u, t)$  – the computational resource requirement (type 1, e.g., processor power);  $c_{i,3}(u, t)$  – the computational resource requirement (type 2, e.g., amount of RAM);  $c_{i,4}(u, t)$  – the amount of service data that must be read, transmitted, or written during state correction or reproduction.

Let the vector of resources available at time  $t$ , whose components are consistent with  $\bar{c}_i(u, t)$  (19), be:

$$\bar{o}_i(t) = (o_{i,1}(t), o_{i,2}(t), o_{i,3}(t), o_{i,4}(t))^T . \quad (20)$$

Then the relative load indicator of the operation is determined as follows:

$$\psi_i(u, t) = \sum_{j=1}^4 \varphi_j \frac{c_{i,j}(u, t)}{o_{i,j}(t)} , \quad \sum_{j=1}^4 \varphi_j = 1, \quad \varphi_j \geq 0 . \quad (21)$$

The value of  $\psi_i(u, t)$  (21) allows us to express the ratio between the requirements of action  $u$  and the time and resources available at time  $t$  as a single number. For each component separately, we compare exactly how much is required to perform the action and how much is actually available, after which these ratios are summed with weights  $\varphi_j$ . The smaller the value

of  $\psi_i(u, t)$ , the easier it is to perform the corresponding action in the current state of the mobile platform. Therefore,  $\psi_i(u, t)$  is used to compare micro-restarts and micro-reboots based on their resource requirements.

A separate check is performed to determine whether repeating the current step would result in a duplicate of a result already recorded outside the process. For micro-restoration  $\rho_i$ , the current step is not repeated, therefore:

$$\lambda_i(\rho_i, t) = 0. \quad (22)$$

For micro-restart  $\kappa_i$ , the value of  $\lambda_i(\rho_i, t)$  is determined by the membership of the current step  $a_{i,r_i(t)}$  in the sets  $\Pi_i^{(1)}$ ,  $\Pi_i^{(2)}$ ,  $\Pi_i^{(3)}$  (10)–(12):

$$\lambda_i(\kappa_i, t) = \begin{cases} 0, & a_{i,r_i}(t) \in \Pi_i^{(2)}, \\ 1 - \zeta_i(t), & a_{i,r_i}(t) \in \Pi_i^{(1)}, \\ 1, & a_{i,r_i}(t) \in \Pi_i^{(3)}, \end{cases} \quad (23)$$

where  $\zeta_i(t) \in [0, 1]$  is a coefficient indicating what proportion of the duplicated result can be eliminated locally " " (if  $\zeta_i(t) = 1$ , the duplicated result can be eliminated entirely; if  $\zeta_i(t) = 0$ , local elimination is not possible).

Thus, if the current step belongs to the set  $\Pi_i^{(1)}$ , a micro-restart does not create a duplicate of an already recorded result. If the step belongs to  $\Pi_i^{(2)}$ , the consequences of re-execution depend on the possibility of locally eliminating the duplicate result. If the step belongs to  $\Pi_i^{(3)}$ , a micro-restart leads to an invalid repetition of an already recorded action.

To determine the extent to which the current step number, input data, and service flags match the set required for a deterministic restart, the following vector is entered

$$\bar{z}_i(t) = (r_i(t), \bar{u}_i(t), \bar{v}_i(t))^T, \quad (24)$$

and the reference vector

$$\hat{z}_{i,r_i(t)} = (r_i(t), \hat{u}_{i,r_i(t)}, \hat{v}_{i,r_i(t)})^T, \quad (25)$$

where  $\hat{v}_{i,r_i(t)}$  is the reference vector of service marks for step  $a_{i,r_i(t)}$ .

The vector  $\bar{z}_i(t)$  (24) contains those components of the current process state that are relevant specifically to the restart of the current step: the

step number, the input data for this step, and service flags that record the status of its execution. The vector  $\hat{z}_{i,r_i(t)}$  (25) specifies the reference set of these same values under which the restart of step  $a_{i,r_i(t)}$  is permissible.

Then, the indicator of the current state's compliance with the conditions for restarting is determined as follows:

$$\zeta_i(t) = \exp\left(-\|\bar{z}_i(t) - \hat{z}_{i,r_i(t)}\|_1\right). \quad (26)$$

The value of  $\zeta_i(t)$  (26) lies in the interval  $(0, 1]$ , equals one when the current and reference values required for restarting are exactly the same, and decreases as the deviation between them increases. Therefore, the closer the current state of the process is to the reference startup state, the larger the value of  $\zeta_i(t)$  is.

To use  $\zeta_i(t)$  in the selection rule, the following indicator is introduced:

$$t_i(u, t) = \begin{cases} 0, & u = \rho_i, \\ 1 - \zeta_i(t), & u = \kappa_i. \end{cases} \quad (27)$$

The value  $t_i(u, t)$  (27) is taken into account only for those actions that require the step to be restarted. For micro-reset  $\rho_i$ , a restart is not performed, so  $t_i(u, t) = 0$ . For micro-restart  $\kappa_i$ , the value of  $t_i(u, t)$  increases as the deviation between the current state and the reference state for the restart increases. This means that as the current state becomes less consistent with the restart conditions, the micro-restart becomes less suitable for use.

To verify whether a local action can be applied to the process in its current state, a barrier function is used, consistent with the approaches of admissible state sets and recovery estimation based on barrier functions [8, 22]. After the action is applied, the process must remain within a locally admissible state or return to it. In this regard, the following quantity is introduced:

$$\Gamma(u, t) = 1 - \frac{\Theta_i(u, t)}{\bar{\theta}_i}, \quad \bar{\theta}_i > 0, \quad (28)$$

where  $\bar{\theta}_i$  is the maximum permissible value of the scalar process state estimate after the action is executed.

The value  $\Gamma(u, t)$  (28) indicates the ratio between the expected state of the process after applying the action  $u$  and the maximum permissible state. If  $\Gamma(u, t) \geq 0$ , then after executing the action, the process



does not exceed the bounds of the locally permissible state. If  $\Gamma(u, t) < 0$ , then the action does not ensure that the process returns to a state from which its execution can be continued. Thus,  $\Gamma(u, t)$  characterizes whether the expected state of the process after the

$$\chi_i(u, t) = 1, \quad \psi_i(u, t) \leq 1, \quad \Gamma(u, t) \geq 1, \quad \lambda_i(u, t) \leq \bar{\lambda}, \quad \iota_i(u, t) = \bar{\iota}. \quad (29)$$

If at least one of the conditions in (29) is not satisfied, we assume that  $\chi_i(u, t) = 0$ . Therefore, action  $u$  is considered admissible for application only if it can be executed within the available time and resources, does not take the process beyond the bounds of a locally admissible state, does not

$$\wp_i(u, t) = \varpi_1 \Theta_i(u, t) + \varpi_2 \psi_i(u, t) + \varpi_3 \lambda_i(u, t) + \varpi_4 \nu_i(u, t) - \varpi_5 \Gamma_i(u, t), \quad (30)$$

where  $\varpi_1, \varpi_2, \varpi_3, \varpi_4, \varpi_5 \geq 0$  – importance coefficients of the objective function components, which are specified during method configuration in accordance with the requirements of a specific ISMP, in particular:  $\varpi_1$  corresponds to a scalar assessment of the process state after an action is performed;  $\varpi_2$  – the action's requirements regarding time and resources;  $\varpi_3$  – the re-recording of the result outside the process;  $\varpi_4$  – the deviation of the current state from the reference state of a restart;  $\varpi_5$  – maintaining the process within an acceptable state.

In the function  $\wp_i(u, t)$  (30), the first four terms increase its value as the expected damage to the state after an action increases, along with time and resource requirements, the re-recording of the result outside the process, and the deviation of the current state from the reference state at restart. The last term with a minus sign reduces the function's value for actions after which the process remains within the permissible state with a greater distance from the maximum permissible damage level.

The optimal local action is determined by the rule

$$u_i^*(t) = \arg \min_{\substack{u \in U_i(t), \\ \chi_i(u, t) = 1}} \wp_i(u, t) \quad (31)$$

Given the two-element set  $U_i(t)$  (13), rule (31) can be expressed in terms of the difference in the values of the objective function:

$$\Delta_i(t) = \wp_i(\rho_i, t) - \wp_i(\kappa_i, t). \quad (32)$$

Then, if both actions are locally feasible, the selection rule takes the form:

action satisfies the constraint on the permissible level of damage.

Then the admissibility criterion for action  $u$  for process  $p_i$  is defined as follows:

exceed the maximum permissible level of repeated fixation of the result outside the process, and does not require a restart in the event of an excessive deviation of the current state from the reference startup state.

Among the admissible actions, the one for which the value of the objective function is the smallest is selected:

$$u_i^*(t) = \begin{cases} \rho_i, & \Delta_i(t) \leq 0, \\ \kappa_i, & \Delta_i(t) > 0. \end{cases} \quad (33)$$

Thus, micro-restoration  $\rho_i$  is selected when the local correction of the state fragment based on the set of considered indicators yields a value of the objective function no greater than that of a micro-restart. A micro-restart  $\kappa_i$  is selected when returning to a previously feasible state with the re-execution of some steps yields a smaller value of the objective function.

If neither of the actions  $\rho_i$  nor  $\kappa_i$  satisfies condition (29), a set of actions admissible for the process is introduced:

$$U_i^*(t) = \{u \in U_i(t) \mid \chi_i(u, t) = 1\}. \quad (34)$$

If

$$U_i^*(t) = \emptyset, \quad (35)$$

then local recovery of the process at time  $t$  is considered impossible, and the process  $p_i$  is transferred to higher-level recovery mechanisms:

$$p_i \rightarrow \sum(t), \quad (36)$$

where  $\sum(t)$  is the set of processes requiring service-level recovery, profile switching, or other ISMP system reconfiguration [15].

Thus, relations (13)–(36) define a rule for selecting a local recovery action as the sole means of distinguishing between micro-recovery and micro-restart. Unlike selection based on a single damage criterion, the proposed rule simultaneously takes into account the scalar estimate of the process state after the action is performed  $\Theta_i(u, t)$  (18), the ratio between

the action's requirements and the available time and resources  $\psi_i(u, t)$  (21), the level of result re-fixation outside the process  $\lambda_i(u, t)$  (22), (23) and the deviation of the current state from the reference state of the restart  $t_i(u, t)$  (27). It is precisely this structure that makes it possible to formally determine which of the two local recovery actions is appropriate for the current state of the process.

### 5.3. Method for Deterministically Reconstructing the Minimally Sufficient State of a Process

Since rules (31)–(33) determine whether it is appropriate to apply micro-restoration  $\rho_i$  or micro-restart  $\kappa_i$  to the process  $p_i$ , then if for the moment  $t$  the following holds

$$u_i^*(t) = \kappa_i, \quad (37)$$

then, to continue execution, it is necessary to roll back the process to one of the previously confirmed checkpoints and restore a local state that is sufficient

$$B_i^*(t) = \{b_{i,\zeta} \in B_i(t) \mid \lambda_i(b_{i,\zeta}, t) \leq \bar{\lambda}, \nu_i(b_{i,\zeta}, t) \leq \bar{\nu}, \chi_i(\kappa_i, t) = 1\}. \quad (39)$$

In formula (39),  $\lambda_i(b_{i,\zeta}, t)$  is a scalar measure of the re-fixation of the result outside the process during the re-execution of the fragment from the reference point  $b_{i,\zeta}$ ;  $\nu_i(b_{i,\zeta}, t)$  is a scalar measure of the ratio between the resource requirements for such re-execution and the values of these quantities available at time  $t$ ;  $\bar{\lambda}$  and  $\bar{\nu}$  are the maximum permissible values of

$$\Xi_i(b, t) = \tau_1(r_i(t) - b) + \tau_2\nu_i(b, t) + \tau_3\lambda_i(b, t) + \tau_4\vartheta_i(b, t), \quad (40)$$

where  $(r_i(t) - b)$  – the number of steps that must be repeated after returning to point  $b$ ;  $\nu_i(b, t)$  – a scalar measure of the ratio between the time and resource requirements of this repetition and their available values;  $\lambda_i(b, t)$  – a scalar estimate of the repeated saving of the result outside the process during the re-execution from point  $b$ ;  $\vartheta_i(b, t)$  – the volume of internal data and service markers that need to be recreated to start from point  $b$ ;  $\tau_1, \tau_2, \tau_3, \tau_4 \geq 0$  – coefficients that specify the contribution according to the number of re-execution steps, time and resource constraints, re-capturing of results outside the process, and the volume of data that needs to be restored (their values are set during the configuration of a specific ISMP).

for an unambiguous restart and, at the same time, minimal in terms of service data. This is precisely the essence of the method for deterministically restoring the minimally sufficient state of the process.

Let the log of service marks and the stored internal data for the process  $p_i$  at time  $t$  contain a set of restart reference points:

$$B_i(t) = \{b_{i,\zeta} \mid \zeta = 1, \dots, M_i(t), b_{i,\zeta} < r_i(t)\}, \quad (38)$$

where  $b_{i,\zeta}$  is the step number after which a confirmed local state exists for the process, to which it can return for re-execution;  $M_i(t)$  is the number of such points, determined by the service marks  $\bar{v}_i(t)$  and the confirmed order of step execution.

Not every point  $b_{i,\zeta} \in B_i(t)$  is valid for replay.

For a point to be valid, replaying the segment starting from that point must be semantically valid and must not exceed the time and resource limits. Therefore, a set of valid reference points is introduced:

these indicators;  $\chi_i(\kappa_i, t)$  – a micro-restart feasibility indicator, defined in formula (29).

From the set  $B_i^*(t)$  (39), it is necessary to select the point for which reconstruction requires the lowest total cost. To do this, the objective function (criterion) for selecting the reference point is defined as follows:

In formula (40), the first term accounts for how far back the process needs to be rolled back in terms of the number of steps. The second term accounts for whether re-execution from the selected point is excessive in terms of resources and time for the current ISMP state. The third term accounts for whether re-execution from this point could lead to the result being recorded outside the process again. The fourth term accounts for the amount of data and service markers that need to be restored to start the process from the selected point. Thus, the  $\Xi_i(b, t)$  function allows selecting not the closest point automatically, but the one that is most appropriate based on the combination of the number of steps to be re-executed, resource and time requirements, re-capturing the

result, and the amount of data required to reproduce the process.

The optimal support point is determined as follows:

$$b_i^*(t) = \arg \min_{b \in B_i^*(t)} \Xi_i(b, t) \quad (41)$$

Equation (41) allows one to move to the nearest previous point for which the total cost of restarting is the lowest, taking into account the number of re-execution steps, resource and time requirements, the need to re-capture the result outside the process, and the volume of data that needs to be reproduced.

After selecting the point  $b_i^*(t)$  (41), a subset of the coordinates of the process's complete local state is formed, which are necessary for restarting from this point. Let the complete local state of the process at the point  $b_i^*(t)$  be given by  $\vec{s}_i(b_i^*(t)) \in \mathbb{R}^{d_i}$ . Then, the binary coordinate selection matrix is defined as:

$$\bar{G}_i(t) \in \{0, 1\}^{d_i^{\min}(t) \times d_i}, \quad d_i^{\min}(t) \leq d_i \quad (42)$$

where  $d_i^{\min}(t)$  is the number of coordinates of the complete local state stored for restart.

Each row of the matrix  $\bar{G}_i(t)$  contains exactly one element and determines which specific coordinate of the complete local state is included in the vector of coordinates to be reproduced.

The coordinate vector of the local state to be reconstructed is defined as follows:

$$\vec{h}_i(t) = \bar{G}_i(t) \vec{s}_i(b_i^*(t)). \quad \vec{h}_i(t) = \bar{G}_i(t) \vec{s}_i(b_i^*(t)) \quad (43)$$

The quantity  $\vec{h}_i(t)$  (43) contains only those coordinates of the complete local state without which restarting the process from the point  $b_i^*(t)$  is impossible. Therefore,  $\vec{h}_i(t)$  includes only those state data that are directly used when starting the process from the selected point.

To generate the complete set of data required for reproduction, the value of  $\vec{h}_i(t)$  alone is insufficient. It is also necessary to record the input data and service marks throughout the entire replay interval. For this reason, a sequence of input vectors is introduced:

$$\vec{X}_i(t) = (\vec{x}_i(b_i^*(t)), \vec{x}_i(b_i^*(t)+1), \dots, \vec{x}_i(r_i(t))) \quad (44)$$

where  $\vec{\sigma}_i(t)$  is a vector of restart service rules that specifies the order of input data presentation during the restart interval, the order of service mark

and the sequence of service marks:

$$\vec{V}_i(t) = (\vec{v}_i(b_i^*(t)), \vec{v}_i(b_i^*(t)+1), \dots, \vec{v}_i(r_i(t))). \quad (45)$$

Then, the minimally sufficient state of the process for reproduction is formulated as follows:

$$\vec{s}_i^{\min}(t) = (\vec{h}_i(t), \vec{X}_i(t), \vec{V}_i(t))^T. \quad (46)$$

The vector  $\vec{s}_i^{\min}(t)$  (46) is used as the input dataset for the reproduction procedure. Its sufficiency means that, based on  $\vec{s}_i^{\min}(t)$ , it is possible to restart the process from the point  $b_i^*(t)$  and bring the process to the same local state that corresponds to the correct initial execution. Its minimality means that removing at least one essential component from (46) results in the loss of the ability to ensure such a restart and such a reproduction of the process state.

The conditions under which the vector  $\vec{s}_i^{\min}(t)$  is sufficient and at the same time minimal for restarting the process are defined as follows:

$$\pi_i(\mathfrak{R}_i(\vec{s}_i^{\min}(t))) = \pi_i(\vec{s}_i^*(t)), \quad (47)$$

$$\forall \vec{s}_i^{\min}(t) \prec \vec{s}_i^{\min}(t) : \pi_i(\mathfrak{R}_i(\vec{s}_i^{\min}(t))) \neq \pi_i(\vec{s}_i^*(t)), \quad (48)$$

where  $\mathfrak{R}_i$  is the operator that generates the local state of the process after a restart;  $\vec{s}_i^*$  is the reference local state of the process after the restart is complete; the relation  $\prec$  denotes the removal of at least one component from the vector  $\vec{s}_i^{\min}(t)$ . Equation (47) implies that the vector  $\vec{s}_i^{\min}(t)$  contains all the data without which it is impossible to reach the reference local state  $\vec{s}_i^*$ . Equation (48) means that removing any component from  $\vec{s}_i^{\min}(t)$  violates this property. Thus, equation (47) means that state (46) ensures the attainment of the reference local state of the process, while equation (48) shows that removing any of its components violates this possibility.

Determinism in recovery means that for identical vectors  $\vec{s}_i^{\min}(t)$  and identical restart rules, the result of process recovery is the same. This requirement is stated as follows:

$$(\vec{s}_i^{\min}(t), \vec{\sigma}_i^{\min}(t)) = (\vec{s}_i^{\min}(t), \vec{\sigma}_i^{\min}(t)) \Rightarrow \mathfrak{R}_i(\vec{s}_i^{\min}(t), \vec{\sigma}_i^{\min}(t)) = \mathfrak{R}_i(\vec{s}_i^{\min}(t), \vec{\sigma}_i^{\min}(t)), \quad (49)$$

usage, restrictions on re-storing the result outside the process, and the order of eliminating redundant result storage.

The result of applying the method is the restored local state:

$$\vec{s}_i(t+1) = \mathfrak{R}_i(\vec{s}_i^{\min}(t), \vec{\sigma}_i(t)) \quad (50)$$

for which the condition for the process to transition to a locally stable state must be satisfied:

$$\vec{\sigma}_i(\vec{s}_i(t+1)) = \vec{0}, \quad \mu_i(\vec{s}_i(t+1)) \leq 1. \quad (51)$$

$$\mathfrak{I}_i(t) = \ell_1(r_i(t) - b_i^*(t)) + \ell_2 \|\vec{h}_i(t)\|_0 + \ell_3 \sum_{\zeta=b_i^*(t)}^{r_i(t)} \|\vec{x}_i(\zeta)\|_1 + \ell_4 \sum_{\zeta=b_i^*(t)}^{r_i(t)} \|\vec{v}_i(\zeta)\|_1, \quad (52)$$

where  $\ell_1, \ell_2, \ell_3, \ell_4 \geq 0$  are coefficients that determine the contribution based on the number of steps to be repeated, the size of the local state elements that must be saved for a subsequent run, the total size of the input data, and the total size of the service markers.

In formula (52), the term  $(r_i(t) - b_i^*(t))$  determines the number of steps that must be repeated after returning to the reference point  $b_i^*(t)$ . The value  $\|\vec{h}_i(t)\|_0$  determines the number of local state elements that must be saved and passed for the restart. The sum  $\sum_{\zeta=b_i^*(t)}^{r_i(t)} \|\vec{x}_i(\zeta)\|_1$  characterizes the total volume of input data over the restart interval, and the sum  $\sum_{\zeta=b_i^*(t)}^{r_i(t)} \|\vec{v}_i(\zeta)\|_1$  characterizes the total volume of service marks over the same interval. The smaller the value of  $\mathfrak{I}_i(t)$ , the lower the cost of restarting the process at the selected checkpoint. Therefore, the  $\mathfrak{I}_i(t)$  function is used as a quantitative measure of recovery costs to compare different options for returning the process to its functional (previous) state.

Thus, the sequence of steps for implementing the method is as follows.

**Step 1.** For the process  $p_i(t)$ , for which the condition  $u_i^*(t) = \kappa_i$  (37) holds, a set of reference points  $B_i(t)$  (38) is formed based on the service marks  $\vec{v}_i(t)$ .

**Step 2.** From the set  $B_i(t)$ , points are selected that satisfy the constraints regarding re-execution feasibility and resource-time characteristics. As a result, a set of admissible reference points  $B_i^\dagger(t)$  (39) is formed.

Equations (50) and (51) imply that after a restart and recovery, the process must transition to a state without local inconsistency (5), and the result of the current step cannot be recorded outside the process more than once.

The total cost of restarting the process can be estimated by the following functional:

**Step 3.** Based on criterion (40), the reference point  $b_i^*(t)$  is selected (41).

**Step 4.** For the point  $b_i^*(t)$ , a local state coordinate vector is generated, which must be saved for the subsequent execution of  $\vec{h}_i(t)$  (43); thereafter, the input data sequence  $\vec{X}_i(t)$  (44) and the control tag sequence  $\vec{V}_i(t)$  (45) are appended to it. As a result, a minimally sufficient state vector  $\vec{s}_i^{\min}(t)$  (46) is generated.

**Step 5.** It is checked whether the vector  $\vec{s}_i^{\min}(t)$  ensures the achievement of the reference local state of the process (47) and whether it contains redundant components (48). If condition (47) is not satisfied, the set of coordinates in  $\vec{h}_i(t)$  is expanded by re-tuning the matrix  $\vec{G}_i(t)$  (42), after which an updated vector  $\vec{s}_i^{\min}(t)$  is formed.

**Step 6.** For the constructed vector  $\vec{s}_i^{\min}(t)$ , the local state of the process  $\mathfrak{R}_i$  (50) is reconstructed while satisfying the determinism condition (49).

**Step 7.** Once the process recovery is complete, the condition for the process to transition to a locally valid state is checked (51). If this condition is met, the process proceeds to further execution. If not, the results of the restart are discarded, and the process is handed off to service-level recovery mechanisms or ISMP reconfiguration (36).

Thus, relations (38)–(52) and the given sequence of steps define a method for local recovery of ISMP processes with idempotent restart and deterministic state reproduction. Its difference from a full process restart lies in the fact that only that set of data and service flags is used for the restart, without which it is impossible to ensure that the reference local state of the process is achieved. And the difference from service-level recovery lies in the fact that recovery is performed for a single

process and does not require switching between service profiles or a general reconfiguration of the ISMP.

#### 5.4. Evaluation of the Proposed Method's Effectiveness

The effectiveness of the proposed method was evaluated using a series of model scenarios of local disruptions to the internal state of ISMP processes. A comparison was conducted with two alternative recovery modes: a full process restart and deferred service recovery. The simulation code used to generate the experimental results is presented in [23], and the FAIR data set with the original numerical arrays of the experiment is presented in [24].

Within the experiment, the same scenario set was used for all three modes, including identical sequences of input events, identical times and types of local process state violations, identical time series of available unoccupied mobile platform resources, and identical parameters of connectivity windows with variable duration and bandwidth.

The proposed local recovery method implements a process-level response to internal state violations and, for each failure episode, selects between micro-recovery and micro-restart according to rules (31)–(36), followed by deterministic reconstruction of the minimally sufficient local state according to relations (38)–(52). The full process restart mode is considered the basic recovery method, in which the process, upon detecting a violation, returns to its initial execution state without selecting a reference point, without selecting a minimally sufficient set of local data, and without considering the division of steps based on the admissibility of re-execution. The deferred service recovery mode is

considered a higher-level recovery method, in which no local action is performed at the process level, and recovery is deferred until it becomes possible to apply service procedures such as reconfiguration, profile switching, or other recovery actions at the service or ISMP node level.

The comparison is performed based on six metrics: the dynamics of local process state corruption, cumulative recovery deficit, probability of successful local recovery, average time for the process to return to correct execution, the amount of state required for recovery, and the risk of re-generating a side effect. This set of metrics allows us to evaluate the proposed method simultaneously across three groups of properties: the speed of the process's return to correct execution, the data and resource costs of recovery, and the ability to avoid undesirable consequences of re-executing process steps. First, it is advisable to examine the dynamics of the local process state damage metric. Fig. 1 shows that the proposed method ensures the fastest reduction of this metric. The maximum permissible level of 0.18 for the proposed method is reached at the 7th recovery step, whereas for deferred service recovery it is reached at the 19th step, and for a full process restart at the 25th step. Thus, based on this indicator, the proposed method reduces the time required to return the process to an acceptable state by approximately 2.7 times compared to deferred service recovery and by 3.6 times compared to a full process restart. Furthermore, after just 10–12 recovery steps, the value of the indicator for the proposed method decreases to a level close to zero and remains low thereafter. This indicates the rapid elimination of local inconsistencies in the process state without their recurrence.

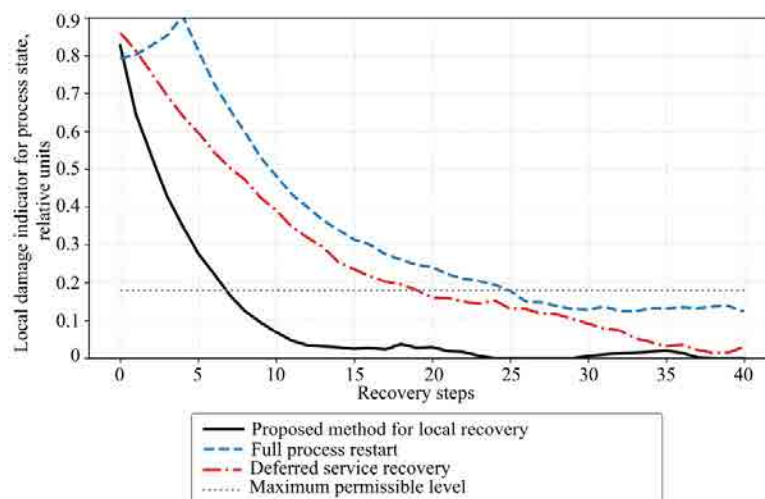


Fig. 1. Dynamics of the local process damage indicator



The cumulative effect of this reduction is confirmed by the values of the cumulative recovery deviation index (Fig. 2). The final value of this indicator for the proposed method is 2.03 relative units, whereas for a full process restart it is 7.64 relative units, and for a deferred service recovery it is 5.52 relative units. Accordingly, the reduction in the cumulative recovery deviation index is 73.5% compared to a full process restart and

63.3% compared to a deferred service recovery. Based on the average value across the entire simulation interval, the advantage also remains: 1.95 relative units for the proposed method versus 6.46 and 4.86 relative units for the two comparison modes. This means that the proposed method ensures a smaller cumulative recovery deviation across the entire interval of the process's return to correct execution.

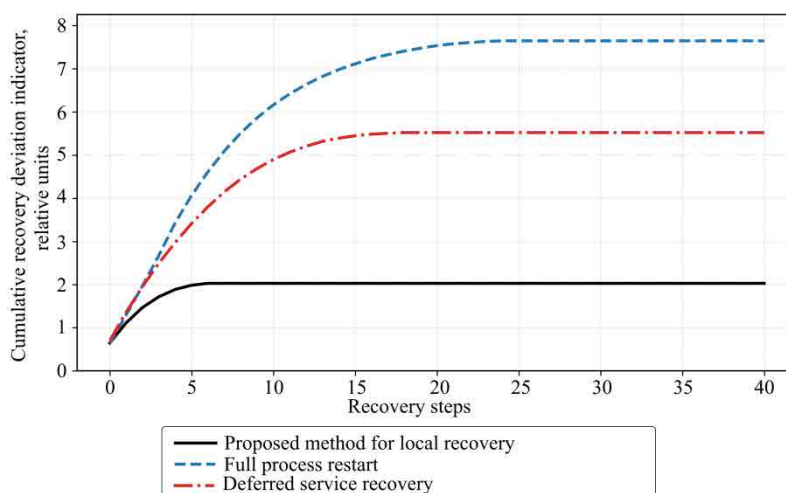


Fig. 2. Cumulative recovery deviation index

The next indicator is the probability of successful process recovery depending on the severity of local damage to its state. Figure 3 shows that, across the entire range of severities, the proposed method yields the highest probability of successful recovery completion. At the maximum intensity value, corresponding to the right boundary of the range, the probability of success for the proposed method is 0.706, for a full process restart – 0.542, and for a deferred service recovery – 0.435. Thus, even in the

most challenging scenario, the advantage is 30.3% compared to a full process restart and 62.3% compared to a deferred service recovery. The advantage also holds true for the average value across the entire interval: 0.863 versus 0.775 and 0.699. This confirms that the combination of local state correction, the elimination of invalid re-executions, and the deterministic reproduction of a minimally sufficient local state ensures higher recovery performance as the intensity of failures increases.

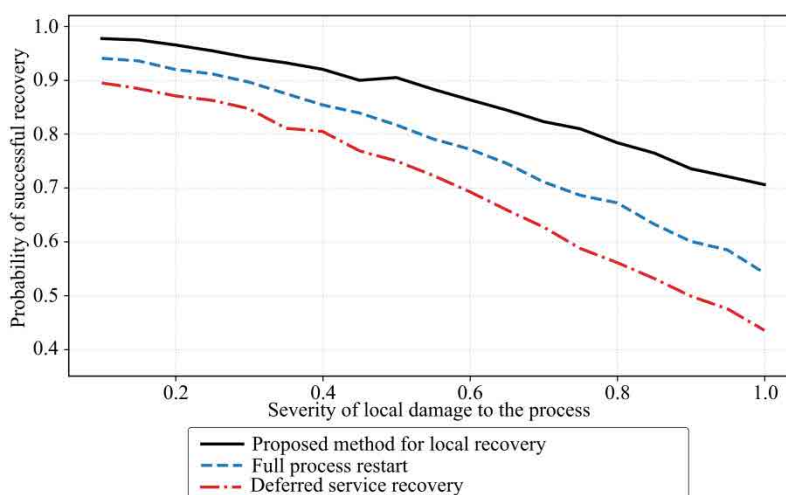


Fig. 3. Probability of successful process recovery depending on the severity of local damage to the state

A key performance indicator of the proposed method is the average time required for the process to return to correct execution. Figure 4 shows that, for the proposed method, this time increases at the slowest rate as the intensity of local state corruption increases. At the maximum intensity value, the average time for the process to return to correct execution is 6.13 for the proposed method, 9.96 for a full process restart, and 12.39 for deferred service recovery. Thus, the reduction in this metric is 38.4% compared to a full process restart and 50.5% compared to deferred

service recovery. Over the entire range of intensity variation, its average value is 4.05 for the proposed method, versus 6.55 and 8.22 for the two comparison modes. This is consistent with the formalism of the subdivisions of the introduced rule and the method based on it within the framework of this study, since the choice between micro-recovery and micro-restart reduces the length of the re-executed fragment, and the reconstruction of a minimally sufficient local state eliminates the need for a full process restart or a transition to service-level recovery.

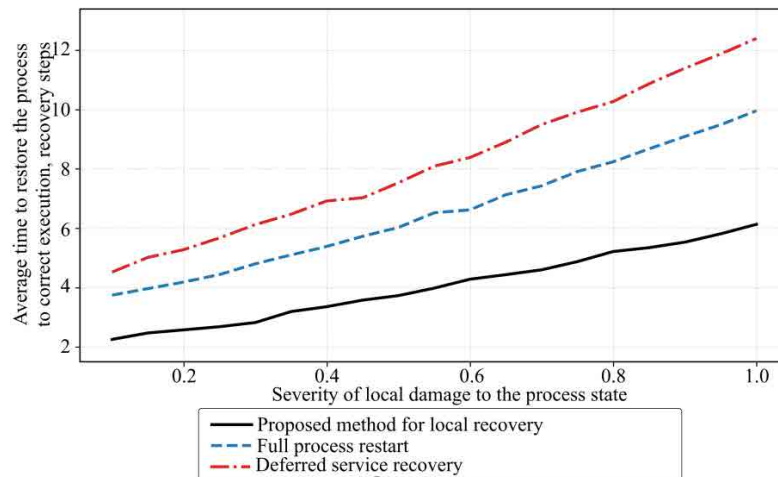


Fig. 4. Average time for the process to resume correct execution

Special attention should be paid to the amount of data required for process reconstruction. This metric directly characterizes the method's requirements regarding the volume of data used during reconstruction. Based on the results of modeling the method's operation [24], it was found that the proposed method requires the smallest amount of data needed for process recovery across the entire range of local damage intensity values of the process state (Fig. 5). At the maximum intensity value, this volume is 37.19 MB for the proposed method, whereas for a full process restart it is 96.60 MB, and for deferred service recovery it is 74.33 MB. Thus, the amount of data required to restore the process is reduced by 61.5% compared to a full process restart and by 50% compared to deferred service recovery. On average, across the entire range of changes in the intensity of local damage to the process state, these values are 25.02, 69.26, and 51.04 MB, respectively. This result is due to the fact that in the proposed method, only the set of process coordinates, input data, and service markers necessary to achieve the reference local state

of the process is generated for process restart, in accordance with equations (42)–(48).

Next, we analyze the probability of duplicating an external result depending on the proportion of steps involving non-idempotent re-execution in the process. Duplication of the external result is understood here as a repeated write to an external log, repeated transmission of a message, repeated update of a shared store, or repeated sending of a command to another node component or ISMP. Fig. 6 shows that as the proportion of such steps increases, the value of this metric rises for all three comparison modes; however, for the proposed method, it remains the lowest across the entire range of the argument. In the most complex scenario, where the proportion of steps with non-idempotent retries approaches one, the final value of the metric is 0.0449 for the proposed method, 0.2034 for a full process restart, and 0.1566 for deferred service recovery. Accordingly, the reduction in this metric compared to a full process restart is 77.9%, and compared to deferred service recovery, it is 71.3%. On average, across the entire range of changes in the proportion of steps with non-idempotent re-execution, the advantage

also persists: 0.0228 for the proposed method versus 0.1140 and 0.0868 for the two comparison modes. This is consistent with formulas (22) and (23), since for micro-recovery, the current step is not re-executed,

and for micro-restart, the consequences of re-execution are determined by the step's membership in sets (10)–(12) and the possibility of locally eliminating the duplicated result.

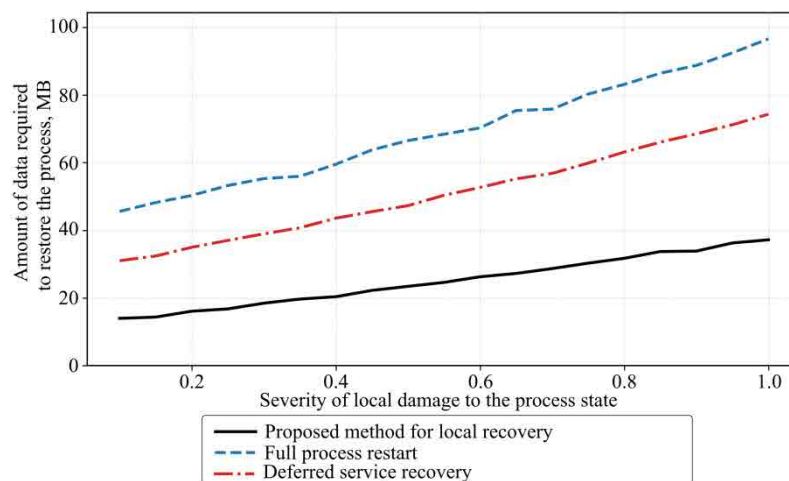


Fig. 5. Amount of data required to restore the process, depending on the severity of local damage to the process state

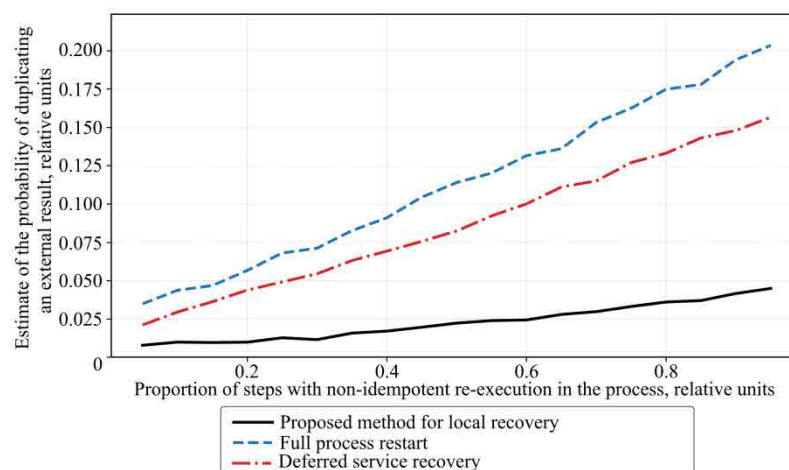


Fig. 6. Dependence of the estimated probability of duplicating an external result on the proportion of steps involving non-idempotent repetition in the process

The results of the evaluation of the proposed method's effectiveness show that it offers a consistent advantage over the two comparison methods across all studied metrics. This advantage is manifested in the fact that the local damage indicator of the process state decreases more rapidly, the cumulative recovery deviation indicator is significantly lower, the probability of successfully returning the process to correct execution is higher, the average recovery time is shorter, the amount of data required for process recovery, and the probability of duplicating the external result is reduced. This provides grounds for asserting that local recovery with idempotent re-execution and deterministic reproduction of the minimally sufficient local state is

appropriate for ISMP in those operating modes where a full process restart is excessive in terms of time and the volume of re-executed actions, and deferred service recovery does not ensure the timely return of the process to correct execution.

### 5.5. Discussion of Research Results

The results of the model study indicate that the proposed method provides a consistent advantage across all investigated metrics [24]. Within the framework of the adopted formulation, this means that local process recovery should be considered as an independent level of recovery intervention in ISMP. Its place is defined between a full process restart

and service-level recovery. This conclusion follows from the fact that the method simultaneously reduces the time required for the process to return to correct execution, decreases the cumulative recovery deviation metric, increases the probability of successful recovery, reduces the amount of data required for restart, and limits the probability of duplicating the external result. Thus, the method's advantage is systemic in nature and reflects the properties of the local recovery procedure itself, rather than a random improvement in a single metric.

The content of the obtained result is determined by a combination of two formalized decisions. The first decision is related to the selection of a recovery action based on relations (30)–(33). In this procedure, a local disruption of the process state is not equated with an automatic return to the start of execution. For each disruption episode, the following are evaluated: the local admissibility of the action, the relationship between its requirements and the available time and resources, the assessment of the replication of the external result, and the compliance of the current state with the conditions for restarting. The second solution involves constructing a minimally sufficient local state for restarting the process according to relations (41)–(49). As a result, not the entire volume of local data is involved in the recovery, but only that part of it which is necessary to achieve the reference local state of the process.

It is precisely this construction that explains why the achieved improvement manifests itself simultaneously in temporal, functional, and volumetric metrics.

Modern approaches related to this subject area confirm individual components of this conclusion but do not encompass them within a single process mechanism. The direction of checkpoint orchestration and accelerated restart of virtual functions [25] confirms the importance of reducing recovery latency. Here, the state of the execution environment serves as the control object. Consequently, the following remain outside the scope of consideration: classes of process steps based on re-execution feasibility, the procedure for choosing between micro-resumption and micro-restart, and the minimally sufficient local process state for continuing execution. This is fundamental for this article, since the time savings achieved are linked to the localization of the recovery effect itself at the level of an individual process.

A technologically related area is local recovery and partial state preservation in stream processing systems [26]. Its significance for this study lies in confirming that reducing the scope of recovery improves

the effectiveness of recovery procedures. At the same time, the object of recovery is a subset of stream graph operators. For ISMP, this formulation is insufficient, since the process level requires consideration of the semantics of individual steps, the recording of the external result, and the choice between two modes of local recovery. In this sense, the proposed method extends the mechanism of recovery localization to the process-step level, where the volume of data to be recovered and the admissibility of process re-execution are essential.

To justify deterministic reproduction, the results obtained for transactional systems with mutable state [27] and the results of behavioral consistency verification for applications with mutable state [28] are essential. These lines of research confirm that correct recovery is impossible without reproducibility of execution and without verifying that the obtained result matches the expected state. However, in [27], determinism is achieved through the repeated presentation of the input stream and the architecture of transactional execution, while in [28], the subject of study is the verification of the identity of results for different execution variants. In this work, the problem is formulated in a different context: it is necessary to construct a local data set sufficient for the correct resumption of a single process after a violation of its internal state. That is why the selection of a reference point, the construction of a recovery vector, and the verification of reaching the reference local state of the process occupy a central place in the formalism of the method.

Related areas include approaches to the porting of functions and data in edge environments [29] and state management systems for serverless cloud applications [30]. Their significance for this study lies in the fact that they emphasize the importance of explicit state management in environments with variable resource availability.

At the same time, these approaches focus on process migration between nodes, external state management, or infrastructure-level fault tolerance. For ISMP, this is insufficient in cases where the recovery action must be completed within an available local time window and with an available idle resource, without waiting for migration, profile switching, or other higher-level reconfiguration. It is here that the proposed method occupies a distinct application niche.

Thus, the results of the discussion provide grounds to believe that the task of local process recovery in ISMP constitutes a distinct subclass of tasks within the system for ensuring the survivability of information

systems. Its specificity is determined by the simultaneous consideration of four groups of factors: local damage to the internal state of the process, the admissibility of re-executing process steps, resource- and time-constraints for recovery, and the need for deterministic reconstruction of a minimally sufficient local state of the process. The comprehensive consideration of precisely these factors is absent in the contemporary approaches discussed [25–30]. Therefore, the obtained results define an independent process-based recovery mechanism for ISMP.

The results obtained should be interpreted taking into account the conditions adopted in the model study, which covers scenarios of local disruption of the state of a single process and a comparison of three recovery modes based on a fixed set of indicators [24]. Under the conditions adopted in this model study, the effectiveness of the local process recovery method and its advantage over the two comparison modes have been confirmed. Further study is required for scenarios involving simultaneous disruption of multiple interdependent processes, competition modes for an unoccupied mobile platform resource under higher load, and the impact of architectural heterogeneity of nodes on the selection of a recovery reference point. Such an expansion will enable the transition to studying the multiprocessing mode of ISMP operation.

The scientific novelty of the obtained results lies in the fact that, for the first time, a method for local recovery of information system processes on a mobile platform has been developed which, unlike known approaches, formalizes the choice between micro-recovery and micro-restart based on an assessment of local damage to the internal state of the process, the feasibility of re-executing steps, and the trade-off between recovery requirements and available time and resources, and defines a procedure for deterministically reproducing the minimally sufficient local state of the process, which allows for the correct resumption of process execution without resorting to a full restart or service recovery, reduces the duration of local operational disruptions, and thereby enhances the resilience of the information system on a mobile platform.

## **6. Conclusions and Prospects or Further Research**

This article solves the problem of developing a method for the local recovery of information system

processes on a mobile platform under conditions of local disruption of their internal state, limited resources, and intermittent connectivity. To this end, the local state of the process was formalized, a vector of inconsistency and a scalar measure of local damage to the process state were introduced, which made it possible to proceed to the process-level analysis of local recovery.

On this basis, a rule was developed for choosing between micro-recovery and micro-restart, taking into account the expected state of the process after the recovery action, the relationship between the requirements of the action and the available time and resources, the assessment of the replication of the external result, and the compliance of the current state with the conditions for restart. A method has been developed for the deterministic reproduction of the minimally sufficient local state of the process, which determines the selection of the restart reference point and the formation of the set of local state coordinates, input data, and service markers required for correct continuation of execution.

The results of the model study confirmed the effectiveness of the proposed method compared to a full process restart and deferred service recovery. It has been established that its application ensures a faster reduction in the local process state damage metric, a lower cumulative recovery deviation metric, a higher probability of successful process return to correct execution, a shorter average recovery time, a smaller volume of data required for restart, and a lower estimate of the probability of duplicating the external result.

The results obtained confirm that local process recovery with idempotent re-execution and deterministic reproduction of the minimally sufficient local state is appropriate as a separate process mechanism in the system for ensuring the survivability of an information system on a mobile platform, as it allows reducing the duration of local process execution failures without resorting to a full restart or service-level recovery in cases where local recovery is sufficient.

Prospects for further research should focus on extending the proposed method to cases of simultaneous state violations in several interdependent processes, accounting for inter-process dependencies when constructing a minimally sufficient local state for recovery, and integrating local process recovery with service recovery methods, dynamic resource reservation, and prediction of cascading transitions within a unified survivability assurance system.



**Conflict of Interest**

The author declares that there is no conflict of interest, particularly of a financial, personal, authorial, or any other nature, that could influence the research or the results published in this article.

**Data Availability**

The data supporting the results of this study are available in open access in the Zenodo repository [24]. The modeling code used to obtain the presented results is provided in [23].

**Funding**

The research was conducted without financial support.

**Use of Artificial Intelligence**

The author confirms that he did not use artificial intelligence technologies to write this paper.

**References**

1. Ergenç, D., Memedi, A., Fischer, M., Dressler, F. (2025), "Resilience in Edge Computing: Challenges and Concepts", *Foundations and Trends® in Networking*, Vol. 14, No. 4, pp. 254–340. DOI: <https://doi.org/10.1561/13000000074>
2. Shaikh, S., Jammal, M. (2024), "Survey of fault management techniques for edge-enabled distributed metaverse applications", *Computer Networks*, Vol. 254, Article No. 110803. DOI: <https://doi.org/10.1016/j.comnet.2024.110803>
3. Zhang, Y., Zhao, K., Yang, Y., Zhou, Z. (2025), "Real-Time Service Migration in Edge Networks: A Survey", *Journal of Sensor and Actuator Networks*, Vol. 14, No. 4, 79. DOI: <https://doi.org/10.3390/jsan14040079>
4. Trung, K.N., Tran, M.-N., Kim, Y. (2026), "Enabling Service Continuity for Stateful Service Segmentation in Mobile Edge Computing Toward 6G", *IEEE Access*, Vol. 14, pp. 20589–20604. DOI: <https://doi.org/10.1109/ACCESS.2026.3661972>
5. Додонов, О.Г., Кузнецова, М.Г., Горбачик, О.С. (2025), "Моделювання і оцінювання функціональної стійкості інформаційних систем", *Реєстрація, зберігання і обробка даних*, Т. 27, № 1, С. 76–88. DOI: <https://doi.org/10.35681/1560-9189.2025.27.1.335752>
6. Додонов, О.Г., Горбачик, О.С., Кузнецова, М.Г. (2007), "Живучість інформаційно-аналітичних систем: понятійний апарат, моделі аналізу та оцінки", *Реєстрація, зберігання і обробка даних*, Т. 9, № 3, С. 61–72.
7. Kharchenko, V., Sachenko, A., Kochan, V., Fesenko, H. (2016), "Reliability and Survivability Models of Integrated Drone-Based Systems for Post Emergency Monitoring of NPPs", *Proceedings of The International Conference on Information and Digital Technologies*, pp. 127–132. DOI: <https://doi.org/10.1109/DT.2016.7557161>
8. Bindschaedler, L. (2026), "Rebooting Microreboot: Architectural Support for Safe, Parallel Recovery in Microservice Systems", *ARCS 2026: 39th GI/ITG International Conference on Architecture of Computing Systems*, URL: <https://binds.ch/papers/microreboot2026.pdf>
9. Jia, Z., Witchel, E. (2024), "Boki: Towards Data Consistency and Fault Tolerance with Shared Logs in Stateful Serverless Computing", *ACM Transactions on Computer Systems*, Vol. 42, No. 3–4, pp. 1–35. DOI: <https://doi.org/10.1145/3653072>
10. Qi, S., Feng, H., Liu, X., Jin, X. (2025), "Efficient Fault Tolerance for Stateful Serverless Computing with Asymmetric Logging", *ACM Transactions on Computer Systems*, Vol. 43, No. 1–2, pp. 1–43. DOI: <https://doi.org/10.1145/3725985>
11. You, J., Chen, K., Zhao, L., Li, Y., Chen, Y., Du, Y., Wang, Y., Wen, L., Hu, K., Li, K. (2025), "AlloyStack: A Library Operating System for Serverless Workflow Applications", *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 921–937. DOI: <https://doi.org/10.1145/3689031.3717490>
12. Ding, H., Wang, Z., Shen, Z., Chen, R., Chen, H. (2023), "Automated Verification of Idempotence for Stateful Serverless Applications", *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, pp. 887–910, available at: <https://www.usenix.org/conference/osdi23/presentation/ding>
13. Galstyan, N. (2024), "Application-Integrated Record-Replay of Distributed Systems", *Technical Report No. UCB/EECS-2024-4*, University of California, Berkeley, URL: <https://www2.eecs.berkeley.edu/Pubs/TechRpts/2024/EECS-2024-4.pdf>
14. Zhong, X., Zhao, X., Zhang, B., Li, J., Wang, Y., Li, Y. (2025), "A Record–Replay-Based State Recovery Approach for Variants in an MVX System", *Information*, Vol. 16, No. 10, 826. DOI: <https://doi.org/10.3390/info16100826>
15. Ткачов, В. (2026), "Метод відновлення критичних сервісів інформаційної системи на мобільній платформі в умовах керованої деградації", *Автоматизовані системи управління та прилади автоматики*, (188), с. 78–97. DOI: <https://doi.org/10.30837/0135-1710.2026.188.078>
16. Tkachov, V., Ruban, I. (2025), "Integral survivability metric of an information system on a mobile platform under functional cascading and secondary failures", *Innovative Technologies and Scientific Solutions for Industries*, No. 4(34), pp. 78–100. DOI: <https://doi.org/10.30837/2522-9818.2025.4.078>

17. Whitaker, M., Volckaert, B., Al-Naday, M. (2025), "Idempotency in Service Mesh: For Resiliency of Fog-Native Applications in Multi-Domain Edge-to-Cloud Ecosystems", *Proceedings of the 15th International Conference on Cloud Computing and Services Science (CLOSER 2025)*, pp. 182–189. DOI: <https://doi.org/10.5220/0013293900003950>
18. Qi, S., Liu, X., Jin, X. (2023), "Halfmoon: Log-Optimal Fault-Tolerant Stateful Serverless Computing", *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP '23)*, pp. 314–330. DOI: <https://doi.org/10.1145/3600006.3613154>
19. Tošić, A. (2024), "Run-Time Application Migration Using Checkpoint/Restore In Userspace", *Journal of Web Engineering*, Vol. 23, No. 05, pp. 735–748. DOI: <https://doi.org/10.13052/jwe1540-9589.2357>
20. Calagna, A., Yu, Y., Giaccone, P., Chiasserini, C. F. (2023), "Design, Modeling, and Implementation of Robust Migration of Stateful Edge Microservices", *IEEE Transactions on Network and Service Management*, Vol. 21, No. 2, pp. 1877–1893. DOI: <https://doi.org/10.1109/TNSM.2023.3331750>
21. Rong, C., Wang, J.H., Wang, J., Zhou, Y., Zhang, J. (2024), "Live Migration of Video Analytics Applications in Edge Computing", *IEEE Transactions on Mobile Computing*, Vol. 23, No. 3, pp. 2078–2092. DOI: <https://doi.org/10.1109/TMC.2023.3246539>
22. Liu, Z., Unal, M., Parkinson, M.J., Kogias, M. (2025), "DORADD: Deterministic Parallel Execution in the Era of Microsecond-Scale Computing", *Proceedings of the 30th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming (PPoPP '25)*, pp. 282–296. DOI: <https://doi.org/10.1145/3710848.3710872>
23. Tkachov, V. (2026), "Local-process-recovery-figures", *GitHub repository*, URL: <https://github.com/tikey/local-process-recovery-figures>
24. Tkachov, V. (2026), "Source data and generated figures for: A Method for Local Recovery of Processes in Information System on Mobile Platform Using Idempotent Re-Execution and Deterministic State Reconstruction", *Zenodo*. DOI: <https://doi.org/10.5281/zenodo.19565291>
25. Kohli, S., Kharbanda, S., Bruno, R., Carreira, J., Fonseca, P. (2024), "Pronghorn: Effective Checkpoint Orchestration for Serverless Hot-Starts", *Proceedings of the 19th European Conference on Computer Systems (EuroSys '24)*, pp. 298–316. DOI: <https://doi.org/10.1145/3627703.3629556>
26. Takdir, Kitagawa, H., Amagasa, T. (2025), "Local recovery and partial snapshot in distributed stateful stream processing", *Knowledge and Information Systems*, Vol. 67, pp. 9407–9435. DOI: <https://doi.org/10.1007/s10115-025-02509-z>
27. Psarakis, K., Christodoulou, G.C., Fragkoulis, M., Katsifodimos, A. (2025), "Transactional Cloud Applications Go with the (Data)Flow", *Proceedings of the 15th Annual Conference on Innovative Data Systems Research (CIDR '25)*, URL: <https://vldb.org/cidrdb/papers/2025/p25-psarakis.pdf>
28. Zhu, K., Whittaker, M., Petrovic, S., Grandl, R., Ghemawat, S. (2025), "Vive la Différence: Practical Diff Testing of Stateful Applications", *Proceedings of the VLDB Endowment*, Vol. 18, No. 7, pp. 2018–2030. DOI: <https://doi.org/10.14778/3734839.3734841>
29. Nardelli, M., Russo Russo, G. (2024), "Function Offloading and Data Migration for Stateful Serverless Edge Computing", *Proceedings of the 15th ACM/SPEC International Conference on Performance Engineering (ICPE '24)*, pp. 247–257. DOI: <https://doi.org/10.1145/3629526.3649293>
30. Li, T., Chandramouli, B., Burckhardt, S., Madden, S. (2024), "Serverless State Management Systems", *Proceedings of the 14th Conference on Innovative Data Systems Research (CIDR 2024)*, URL: <https://vldb.org/cidrdb/papers/2024/p16-li.pdf>

Received (Надійшла) 15.03.2026

Accepted for publication (Прийнята до друку) 10.04.2026

Publication date (Дата публікації) 29.05.2026

#### Відомості про авторів / About the Authors

**Ткачов Віталій Миколайович** – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, докторант кафедри електронних обчислювальних машин, Харків, Україна;

**Vitalii Tkachov** – Candidate of Technical Sciences, Associate Professor, Kharkiv National University of Radio Electronics, Doctoral candidate at the Department of Electronic Computers, Kharkiv, Ukraine;

e-mail: [vitalii.tkachov@nure.ua](mailto:vitalii.tkachov@nure.ua)

ORCID ID: <https://orcid.org/0000-0002-6524-9937>

## МЕТОД ЛОКАЛЬНОГО ВІДНОВЛЕННЯ ПРОЦЕСІВ ІНФОРМАЦІЙНОЇ СИСТЕМИ НА МОБІЛЬНІЙ ПЛАТФОРМІ

**Предметом дослідження** в статті є локальні порушення внутрішнього стану окремих процесів інформаційної системи на мобільній платформі, які унеможливають коректне продовження виконання без спеціального відновлювального впливу. **Мета роботи** – розроблення методу локального відновлення процесів інформаційної системи на мобільній платформі, який у разі локального порушення внутрішнього стану процесу забезпечує продовження його виконання за допомогою ідемпотентного повторного виконання й детермінованого відтворення мінімально достатнього локального стану в умовах обмежених ресурсів і переривчастої зв'язності. У статті необхідно виконати такі **завдання**: формалізувати локальний стан процесу й ознаки його порушення; встановити класи кроків процесу за допустимістю повторного виконання; розробити правило вибору між мікровідновленням і мікроперезапуском; запропонувати метод детермінованого відтворення мінімально достатнього локального стану процесу; виконати модельне дослідження ефективності запропонованого методу порівняно з повним перезапуском процесу й відкладеним сервісним відновленням. **Досягнуті результати**. Розроблено формалізацію локального стану процесу у вигляді вектора, що містить номер поточного кроку, внутрішні дані процесу, вхідні дані кроку, службові позначки виконання та кратність фіксації результату поза межами процесу. Локальне порушення стану визначено через вектор неузгодженості та скалярний показник локального пошкодження стану процесу. Запропоновано правило локального відновлення, яке формалізує вибір між мікровідновленням і мікроперезапуском з огляду на очікуваний стан процесу після відновлювальної дії, співвідношення між вимогами дії та доступними часом і ресурсами, оцінки дублювання зовнішнього результату, відповідності поточного стану умовам повторного запуску. Розроблено метод детермінованого відтворення мінімально достатнього локального стану процесу, який задає вибір опорної точки повторного запуску й формування сукупності координат локального стану, вхідних даних і службових позначок, необхідних для коректного продовження виконання. Результати модельного дослідження показали, що запропонований метод забезпечує швидше зменшення показника локального пошкодження стану процесу, менше значення накопичувального показника відхилення відновлення, вищу ймовірність успішного повернення процесу до коректного виконання, менший середній час відновлення, менший обсяг даних, потрібних для повторного запуску, і меншу оцінку ймовірності дублювання зовнішнього результату. **Висновки**: запропонований метод є доцільним як окремий процесний механізм у системі забезпечення живучості інформаційної системи на мобільній платформі, оскільки дає змогу скоротити тривалість локальних порушень виконання процесів і підвищити живучість системи без переходу до повного перезапуску або сервісного відновлення в разі, коли локальне відновлення є достатнім.

**Ключові слова**: інформаційна система на мобільній платформі; процес; ідемпотентність; детерміноване відтворення стану; живучість.

### *Бібліографічні описи / Bibliographic descriptions*

Ткачов В. М. Метод локального відновлення процесів інформаційної системи на мобільній платформі. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 72–94. DOI: <https://doi.org/10.30837/0135-1710.2026.189.072>

Tkachov, V. (2026), "A method for local recovery of information system processes on a mobile platform", *Management Information System and Devices*, No. 2 (189), P. 72–94. DOI: <https://doi.org/10.30837/0135-1710.2026.189.072>

Бобровнікова К. Ю., Гурман І. В., Олексюк Д. А.

## ПОРІВНЯЛЬНИЙ АНАЛІЗ МІЖСЕРВІСНОЇ ВЗАЄМОДІЇ З ВИКОРИСТАННЯМ REST І GRPC У СЕРЕДОВИЩІ .NET 8

**Предметом дослідження** є підходи до реалізації синхронної міжсервісної взаємодії в мікросервісній архітектурі із застосуванням REST і gRPC у середовищі .NET 8. **Мета роботи** – провести порівняльний аналіз REST API на базі HTTP/1.1 і JSON і gRPC-сервісу на основі HTTP/2 і Protocol Buffers, а також оцінити їх вплив на продуктивність, обсяг мережевого обміну й використання серверних ресурсів. З огляду на окреслену мету необхідно виконати такі **завдання**: проаналізувати архітектурні розбіжності REST і gRPC; реалізувати два сервіси з однаковою прикладною логікою в середовищі ASP.NET Core 8; забезпечити однакові умови виконання експерименту; виміряти середню затримку відповіді, P95, кількість успішних операцій за секунду, обсяг переданих даних, а також використання CPU і RAM; інтерпретувати досягнуті результати, зважаючи на межі застосування кожного підходу. **Методи дослідження**. Упроваджено порівняльний аналіз архітектурних підходів, експериментальне навантажувальне тестування й кількісне оцінювання продуктивності. Реалізовано два сервіси з ідентичною прикладною логікою та спільним in-memo набором даних без звернення до зовнішньої бази даних. Навантаження генерувалося за допомогою Grafana k6 у режимі constant-arrival-rate. **Результати дослідження**. У межах експерименту gRPC на HTTP/2 з Protocol Buffers продемонстрував нижчі значення середньої затримки та P95, менший обсяг мережевого обміну й вищу пропускну здатність, якщо порівнювати з REST API на HTTP/1.1 з JSON. Загальний обсяг обміну для gRPC виявився приблизно у 2,5 раза меншим, а на верхніх рівнях навантаження середня затримка була нижчою орієнтовно в 1,7–1,8 раза. Крім того, для gRPC зафіксовано більш помірне використання процесорних і пам'яттєвих ресурсів серверного вузла. **Висновки**. Досягнуті результати підтверджують доцільність застосування gRPC насамперед для внутрішньої міжсервісної взаємодії, де критичними є низькі затримки, висока пропускну здатність і ефективне використання мережевих ресурсів. REST доцільно застосовувати в сценаріях, де пріоритетними є сумісність, простота інтеграції, прозорість HTTP-інтерфейсу й зручність використання в публічних API.

**Ключові слова**: мікросервісна архітектура; міжсервісна взаємодія; розподілені системи; продуктивність; затримка відповіді; пропускну здатність.

### Вступ

Сучасний етап розвитку програмної інженерії визначається переходом від монолітних архітектур до розподілених систем, зокрема мікросервісних підходів, які забезпечують гнучкість, масштабованість і незалежність розгортання окремих компонентів. Така архітектурна парадигма дає змогу підвищити адаптивність програмних систем до змін вимог і навантаження, однак водночас суттєво ускладнює організацію взаємодії між окремими сервісами. У мікросервісній архітектурі міжсервісна взаємодія є одним із ключових факторів, що визначає ефективність функціонування системи загалом. Вона безпосередньо впливає на продуктивність, затримку оброблення запитів, надійність передачі даних, а також складність розроблення й супроводу програмного забезпечення. Неефективно спроектовані механізми взаємодії можуть призводити до виникнення вузьких місць, надмірного мережевого навантаження та зниження масштабованості системи.

З огляду на це вибір механізму міжсервісної комунікації доцільно розглядати не лише як питання реалізації API, а як архітектурне рішення, що залежить від зрозумілості інтерфейсу, семантики HTTP, можливостей транспортного рівня та витрат на подання повідомлень.

На практиці для реалізації міжсервісної взаємодії найчастіше застосовуються REST-орієнтовані підходи й протокол gRPC. REST, який у прикладних системах зазвичай реалізується поверх HTTP із використанням ресурсно-орієнтованого підходу й форматів на кшталт JSON, набув значного поширення завдяки простоті реалізації, універсальності та сумісності з різними клієнтськими платформами. Водночас gRPC, що базується на HTTP/2 і за замовчуванням використовує Protocol Buffers як формат серіалізації повідомлень, забезпечує вищу продуктивність, підтримку стримінгових сценаріїв і зменшення накладних витрат на передачу інформації.

Платформа .NET 8 і ASP.NET Core 8 надають сучасні засоби для побудови високопродуктивних

розподілених систем, зокрема підтримку HTTP/2, gRPC і вдосконалення продуктивності вебзастосунків. Подібні можливості для реалізації міжсервісної взаємодії на основі REST, HTTP/2 та gRPC надають також інші сучасні технологічні платформи, зокрема Java/Spring Boot, Quarkus, Go й Node.js. Водночас у межах цієї роботи обрано .NET 8 і ASP.NET Core 8, оскільки вони дають змогу реалізувати REST API і gRPC-сервіси в єдиному технологічному середовищі та зменшити вплив розбіжностей між runtime-платформами, бібліотеками й фреймворками на результати порівняння.

Незважаючи на наявність численних досліджень у сфері мікросервісних архітектур, питання обґрунтованого вибору підходу до міжсервісної взаємодії з огляду на сучасні інструменти розроблення залишається відкритим. Зокрема недостатньо досліджено співвідношення між продуктивністю, складністю реалізації та експлуатаційними характеристиками REST і gRPC у середовищі .NET 8. Отже, науково-практична проблема полягає у визначенні доцільності застосування REST і gRPC для реалізації міжсервісної взаємодії в мікросервісній архітектурі, зважаючи на їх вплив на продуктивність, ефективність передачі інформації та складність розроблення в середовищі .NET 8.

#### **Аналіз сучасних наукових публікацій і визначення проблеми дослідження**

Проблематика мікросервісної архітектури в сучасних дослідженнях розглядається передусім крізь призму складності розподіленого проектування. У працях [1, 2] наголошено, що мікросервіси дають змогу підвищити незалежність розроблення, розгортання й масштабування окремих компонентів, однак водночас ускладнюють визначення меж сервісів, керування контрактами, моніторинг і забезпечення стійкості системи. Отже, ефективність мікросервісної архітектури залежить не лише від декомпозиції, а й від якості взаємодії між її компонентами.

Окремий напрям досліджень пов'язаний із міжсервісною комунікацією як чинника продуктивності та експлуатаційної складності. У роботі [3] доведено, що спосіб взаємодії між сервісами впливає на затримки, пропускну здатність і використання ресурсів. Це дає підстави розглядати комунікаційний механізм не як другорядну деталь реалізації, а як архітектурне рішення, що потребує кількісного оцінювання.

Значну частину публікацій присвячено якості та зрозумілості API. Зокрема в дослідженні [4] доведено, що дотримання правил RESTful-проектування впливає на сприйняття й коректне використання Web API-розробниками. У публікаціях [5–7] сформовано технічну основу для аналізу HTTP-взаємодії, а в роботі [8] зосереджено увагу на тому, що формат серіалізації даних може суттєво впливати на обсяг повідомлень і ефективність комунікації. Отже, у науковій літературі формується підхід, за якого API розглядається не лише з погляду зручності, а й з позиції витрат взаємодії.

Порівнянню комунікаційних технологій у мікросервісних системах присвячені праці [9–12]. У них проаналізовано REST, gRPC та інші підходи з огляду на продуктивність, масштабованість, ресурсні характеристики й практичні межі застосування HTTP API та gRPC. Водночас такі результати істотно залежать від середовища виконання, структури повідомлень, сценарію навантаження й обраних метрик, тому не можуть автоматично переноситися на інші технологічні платформи.

Окремі дослідження, дотичні до окресленої проблеми, розглядають не самі протоколи взаємодії, а умови стабільної роботи сервісів у розподілених середовищах. У статті [13] увагу приділено відновленню критичних сервісів в умовах керованої деградації, що є важливим для ширшого розуміння стійкості сервісно-орієнтованих систем. У праці [14] досліджено прогнозування використання ресурсів у хмарній інфраструктурі, а публікацію [15] присвячено моніторингу й балансуванню навантаження в мікросервісних застосунках. Згадані праці важливі для обґрунтування того, що оцінювання міжсервісної взаємодії має передбачати не лише час відповіді, а й ресурсні й експлуатаційні показники.

Отже, аналіз джерел демонструє, що сучасні дослідження охоплюють загальні принципи мікросервісної архітектури, якість API, транспортні й серіалізаційні чинники, а також експериментальні порівняння REST і gRPC. Водночас залишається необхідність у прикладному порівнянні цих підходів в одному технологічному середовищі, за однакової прикладної логіки й контрольованого навантаження.

**Метою роботи** є проведення порівняльного аналізу підходів REST і gRPC до реалізації міжсервісної взаємодії в мікросервісній архітектурі, дослідження їх впливу на продуктивність і ефективність передачі даних, а також



обґрунтування рекомендацій щодо їх застосування в середовищі .NET 8.

## Матеріали й методи дослідження

### Архітектура міжсервісної взаємодії в мікросервісній системі

Мікросервісна архітектура передбачає побудову програмної системи у вигляді сукупності незалежних сервісів, кожен з яких реалізує окрему бізнес-функцію та взаємодіє з іншими сервісами через мережеві інтерфейси. Такий підхід сприяє масштабованості та гнучкості системи, однак висуває підвищені вимоги до організації міжсервісної взаємодії [1, 2]. Загалом взаємодія між мікросервісами може здійснюватися за двома основними моделями: синхронною та асинхронною. Синхронна взаємодія передбачає безпосередній виклик одного сервісу іншим з очікуванням відповіді, що властиво для REST і gRPC. Асинхронна взаємодія реалізується за допомогою брокерів повідомлень і впроваджується для зменшення зв'язаності сервісів і підвищення відмовостійкості системи [1–3].

У межах синхронної моделі ключовим аспектом є не лише вибір конкретного протоколу, а й визначення способу формалізації контракту між сервісами. REST-орієнтований підхід спирається на ресурсну модель, стандартні HTTP-методи й текстове подання інформації, що робить інтерфейси прозорими для розробників, зручними для тестування

й сумісними з широким колом клієнтських застосунків. Водночас така гнучкість часто супроводжується більш слабкою формалізацією контрактів, потребою в окремому описі API та вищими накладними витратами під час передавання структурованих JSON-повідомлень. Підхід gRPC, навпаки, орієнтований на контрактно визначені віддалені виклики процедур, де структура сервісів і повідомлень задається у .proto-файлах, а клієнтський і серверний код можуть генеруватися автоматично. Це підвищує типобезпечність взаємодії, спрощує узгодження інтерфейсів між сервісами та зменшує ризик помилок, пов'язаних із некоректною інтерпретацією повідомлень. Крім того, використання HTTP/2 і двійкового формату Protocol Buffers створює передумови для зменшення обсягу переданих даних, кращого використання з'єднання й більш стабільної поведінки сервісів за інтенсивної внутрішньої комунікації [6, 8, 10, 11].

У сучасних системах, реалізованих на платформі .NET 8, обидва підходи підтримуються на рівні фреймворку ASP.NET Core. Це дає змогу реалізовувати як REST API, так і gRPC-сервіси в межах єдиного технологічного стеку, що спрощує їх порівняння та інтеграцію.

Подана на рис. 1 схема відтворює гібридний підхід до організації синхронної взаємодії в мікросервісній системі, за якого зовнішній і внутрішній контури комунікації розглядаються як різні відповідно до вимог архітектурні зони.

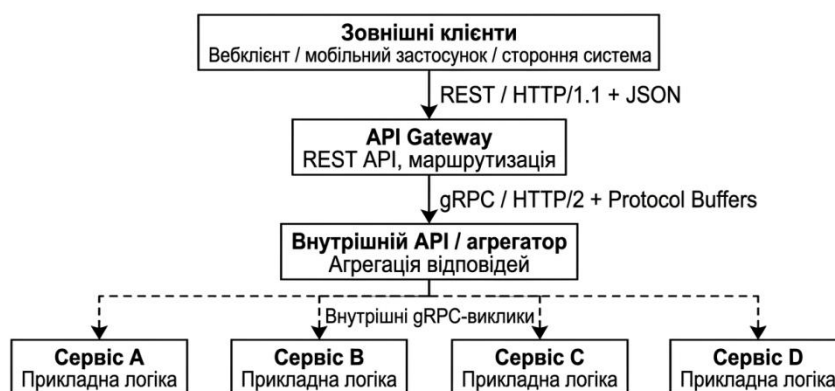


Рис. 1. Архітектура міжсервісної взаємодії в мікросервісній системі з використанням REST і gRPC

Зовнішній контур орієнтований на взаємодію з вебклієнтами, мобільними застосунками та сторонніми системами, тому для нього пріоритетними є сумісність, простота інтеграції, прозорість HTTP-інтерфейсу й зручність тестування. Внутрішній контур, навпаки,

працює в більш контрольованому середовищі, де склад клієнтів і серверів відомий наперед, а основними вимогами стають мінімізація затримок, зменшення обсягу службових даних, типобезпечність контрактів і стабільність поведінки під

навантаженням. У такій архітектурі API Gateway виконує роль межового компонента, який ізолює зовнішніх клієнтів від внутрішньої структури сервісів, спрощує маршрутизацію запитів і дає змогу незалежно еволюціонувати зовнішні та внутрішні інтерфейси. Завдяки цьому REST може залишатися зручним механізмом публічної взаємодії, тоді як gRPC доцільно застосовувати для внутрішніх викликів між сервісами, де переваги контрактної моделі, HTTP/2 та Protocol Buffers мають найбільше практичне значення.

### Реалізація REST-взаємодії в середовищі .NET 8

REST-підхід залишається одним із найпоширеніших способів організації синхронної взаємодії в розподілених і мікросервісних системах завдяки застосуванню стандартних можливостей HTTP, ресурсно-орієнтованої моделі та високої сумісності з різними клієнтськими платформами [4, 5, 10]. У середовищі .NET 8 реалізація REST-сервісів здійснюється засобами ASP.NET Core, що підтримує побудову HTTP API на основі контролерів або Minimal API, маршрутизацію запитів, серіалізацію JSON, ін'єкцію залежностей, middleware-конвеєр, а також механізми автентифікації, авторизації та оброблення помилок. Це сприяє створенню міжсервісних API в межах єдиного технологічного стеку без залучення додаткових фреймворків для базової HTTP-взаємодії.

У мікросервісній архітектурі REST-сервіс зазвичай надає набір кінцевих точок, які відповідають окремим ресурсам або прикладним операціям. Запити між сервісами передаються за допомогою HTTP-методів, а дані найчастіше подаються у форматі JSON. Така модель є зручною для зовнішніх інтеграцій, публічних API та систем, у яких важливими є прозорість інтерфейсу, простота налагодження, можливість тестування запитів стандартними інструментами й підтримка широкого кола клієнтів. У практичній реалізації ASP.NET Core це може бути реалізовано як через класичні контролери, так і Minimal API, що дає змогу скоротити обсяг службового коду для невеликих сервісів або експериментальних реалізацій.

Водночас використання REST у внутрішній міжсервісній взаємодії має низку особливостей, які необхідно брати до уваги під час проєктування. По-перше, контракт REST API часто є менш формалізованим, ніж контракт gRPC, оскільки

структура запитів і відповідей може описуватися окремо, наприклад за допомогою OpenAPI/Swagger, але не завжди безпосередньо застосовується для генерації типобезпечного клієнтського коду. По-друге, текстова природа JSON підвищує зручність читання й діагностики повідомлень, але водночас збільшує обсяг переданих даних і витрати на серіалізацію та десеріалізацію [8]. По-третє, у разі значної кількості внутрішніх синхронних викликів важливими стають повторне використання HTTP-з'єднань, коректне налаштування клієнтських викликів, тайм-аутів, оброблення помилок і контролю каскадних відмов.

Окремим аспектом є еволюція REST API. У мікросервісних системах інтерфейси сервісів змінюються незалежно, тому важливими стають версіювання API, зворотна сумісність контрактів, уніфікована структура помилок, узгоджені коди HTTP-відповідей і спостережуваність запитів. REST добре підтримує інтеграційні сценарії, у яких клієнти можуть бути різнорідними й незалежними від внутрішньої реалізації сервісу. Проте саме ця гнучкість у внутрішніх високонавантажених взаємодіях може перетворюватися на додатковий інженерний тягар, оскільки вимагає окремого контролю контрактів, схем повідомлень, політик повторення запитів і механізмів трасування.

Отже, REST у середовищі .NET 8 є зручним і технологічно зрілим механізмом побудови HTTP API, особливо для публічних інтерфейсів, зовнішніх інтеграцій і сценаріїв, де пріоритетними є простота доступу, сумісність і прозорість обміну. Водночас для інтенсивної внутрішньої міжсервісної взаємодії його ефективність значною мірою залежить від формату даних, версії HTTP, налаштування клієнтських під'єднань і характеру навантаження.

### Реалізація gRPC-взаємодії в середовищі .NET 8

Сервіс gRPC є сучасним підходом до реалізації синхронної міжсервісної взаємодії, який базується на моделі віддаленого виклику процедур і орієнтований на побудову продуктивних, контрактно визначених інтерфейсів у розподілених системах [6, 10, 11]. На відміну від REST, де взаємодія зазвичай організовується навколо ресурсів і HTTP-методів, у gRPC основою є сервісний контракт, описаний у .proto-файлі. У цьому контракті визначаються методи сервісу, типи запитів і відповідей, а також структура повідомлень. Такий підхід забезпечує формалізовану модель взаємодії між сервісами

та зменшує неоднозначність під час інтерпретації переданих даних.

У середовищі .NET 8 реалізація gRPC-сервісів здійснюється засобами ASP.NET Core gRPC. На основі .proto-контракту автоматично генеруються серверні базові класи й клієнтські проксі, що дає змогу працювати з віддаленими викликами як із типізованими методами [10]. Це спрощує супровід міжсервісних інтерфейсів, оскільки помилки невідповідності типів або структури повідомлень можуть бути виявлені ще на етапі компіляції. Для мікросервісної архітектури така властивість є важливою, оскільки зміни контрактів між сервісами мають контролюватися особливо ретельно.

Ключовою технічною особливістю gRPC є використання Protocol Buffers як основного формату серіалізації. На відміну від JSON, який є текстовим і зручним для читання людиною, Protocol Buffers застосовує компактне двійкове подання повідомлень. Це сприяє зменшенню обсягів переданої інформації та витрат на оброблення повідомлень у сценаріях інтенсивної міжсервісної взаємодії. Крім того, gRPC працює поверх HTTP/2, що забезпечує мультиплексування кількох запитів у межах одного з'єднання, більш ефективне використання мережевого каналу й підтримку потокових моделей взаємодії [6, 10].

Практична реалізація gRPC у .NET 8 передбачає не лише опис сервісу в .proto-файлі, а й правильну організацію життєвого циклу викликів. Для внутрішніх сервісних комунікацій важливими є повторне використання gRPC-каналів, контроль часу очікування відповіді, підтримка скасування запитів, передавання службових метаданих і коректне оброблення статусів помилок. На відміну від REST, де помилки зазвичай інтерпретуються через HTTP-коди відповіді, gRPC застосовує власну модель статусів, що дає змогу уніфіковано описувати результати віддалених викликів у різних мовних реалізаціях.

Важливою перевагою gRPC є підтримка кількох моделей взаємодії: унарних викликів, серверного стримінгу, клієнтського стримінгу та двоспрямованого стримінгу. Унарні виклики найближчі до традиційної моделі "запит – відповідь" і ефективні для порівняння з REST API. Стримінгові режими є доцільними для сценаріїв, де потрібно передавати послідовності повідомлень без постійного створення нових запитів, наприклад, для телеметрії, оброблення подій або потокового передавання даних. У межах цієї роботи

основну увагу приділено саме унарним викликам, оскільки вони дають змогу коректніше зіставити gRPC із REST API в однаковій прикладній операції.

Крім переваг, gRPC має й певні обмеження. Його використання є найбільш природним у внутрішньому контурі мікросервісної системи, де можна контролювати версії сервісів, клієнтів, контрактів і транспортного середовища. Для безпосередньої взаємодії з браузерними клієнтами gRPC менш зручний, оскільки стандартні браузерні API не забезпечують повної підтримки звичайного gRPC без додаткових механізмів, як-от gRPC-Web або JSON transcoding [10]. Крім того, контрактна модель .proto вимагає узгодженого підходу до еволюції API: зміни повідомлень мають бути зворотно сумісними, а поля – додаватися й вилучатися з огляду на правила версіювання Protocol Buffers.

Отже, gRPC у середовищі .NET 8 є ефективним механізмом реалізації внутрішньої міжсервісної взаємодії, коли важливими є продуктивність, типобезпечність, компактність повідомлень і стабільність контрактів. Його переваги найбільш помітні в сценаріях інтенсивного обміну інформацією між сервісами, тоді як для публічних API та інтеграції з невідомими клієнтами REST часто залишається більш простим і універсальним рішенням. Саме тому в мікросервісній архітектурі доцільним є не протиставлення REST і gRPC як несумісних технологій, а їх комбіноване впровадження відповідно до межі системи, типу клієнтів і вимог до продуктивності.

### **Порівняльний аналіз REST і gRPC у мікросервісній архітектурі**

Порівняльний аналіз REST і gRPC доцільно проводити не лише з позиції швидкодії, а й з огляду на архітектурну роль кожного підходу в мікросервісній системі. У практичних проєктах ці технології часто розв'язують різні класи завдань: REST переважно використовується як універсальний механізм публічної або міжсистемної інтеграції, тоді як gRPC частіше застосовується у внутрішньому контурі системи, де склад сервісів, клієнтів і контрактів контролюється розробниками. Тому коректне порівняння має зважати не лише на середню затримку чи пропускну здатність, а й на спосіб опису контрактів, модель еволюції API, накладні витрати серіалізації, зручність супроводу й вимоги до клієнтського середовища [9–12].

З погляду архітектурної моделі REST і gRPC мають різну логіку організації взаємодії. REST спирається на ресурсно-орієнтований підхід, у межах якого операції над даними реалізуються за допомогою стандартних HTTP-методів і кінцевих точок API. Така модель добре узгоджується з публічними інтерфейсами, інтеграцією з браузерними клієнтами, сторонніми системами та інструментами тестування. Натомість gRPC орієнтований на віддалений виклик процедур, де основною одиницею взаємодії є не ресурс, а метод сервісу, явно описаний у .proto-контракті. Це робить gRPC ближчим до внутрішньої сервісної взаємодії, у якій важливими є типобезпечність, передбачуваність структури повідомлень і можливість автоматичної генерації клієнтського коду.

Важливою особливістю є спосіб формалізації контрактів. REST API може описуватися за допомогою документації або специфікацій OpenAPI, однак у багатьох реалізаціях контракт існує окремо від клієнтського коду й потребує додаткового контролю узгодженості. Це забезпечує гнучкість, але водночас створює ризик розбіжностей між фактичною поведінкою сервісу й очікуваннями клієнтів. У gRPC контракт є центральним елементом розроблення: структура сервісів і повідомлень визначається у .proto-файлі, на основі якого генеруються серверні та клієнтські компоненти. Завдяки цьому частина помилок інтеграції може виявлятися ще на етапі компіляції, що є суттєвою перевагою для систем із значною кількістю внутрішніх сервісних викликів.

Щодо продуктивності, то gRPC має технічні передумови для кращих результатів у високонавантажених внутрішніх взаємодіях. До таких передумов належать компактне двійкове подання інформації в Protocol Buffers, використання HTTP/2, мультиплексування запитів у межах одного з'єднання й більш ефективна робота з повторюваними службовими даними. REST API, особливо у варіанті HTTP/1.1 + JSON, зазвичай формує більший обсяг повідомлень і потребує додаткових витрат на оброблення текстового формату. Водночас це не означає, що REST є "повільним" у загальному сенсі: за невисокого навантаження, невеликих повідомлень або використання HTTP/2 різниця між підходами може бути менш помітною. Тому твердження про перевагу gRPC варто пов'язувати з конкретною постановкою експерименту, типом повідомлень і профілем навантаження.

З позиції експлуатації REST має перевагу в прозорості та простоті діагностики. HTTP-запити легко перевіряти стандартними інструментами, JSON-повідомлення є читабельними для людини, а коди HTTP-відповідей добре відомі розробникам і адміністраторам. Це спрощує налагодження, інтеграційне тестування й підтримку публічних API. Сервіс gRPC, навпаки, використовує двійковий формат повідомлень і власну модель статусів, тому потребує спеціалізованих інструментів для перегляду запитів, тестування сервісів і трасування взаємодії. Водночас у контрольованому внутрішньому середовищі ця складність компенсується чіткістю контрактів, нижчими мережевими накладними витратами та кращою придатністю до інтенсивних сервісних викликів.

Окремо необхідно брати до уваги питання еволюції API. REST забезпечує високу гнучкість зміни інтерфейсів, але вимагає ретельного керування версіями, сумісністю форматів відповідей і поведінкою клієнтів. У gRPC еволюція API тісно пов'язана з правилами зміни .proto-контрактів: нові поля мають додаватися без порушення сумісності, а вилучення або зміна наявних полів потребує особливої обережності. Отже, REST є більш зручним там, де клієнти різномірні й не завжди контрольовані розробником сервісу, тоді як gRPC краще підходить для середовищ, де команди можуть централізовано керувати контрактами й оновленням клієнтських бібліотек.

З огляду на наведені розбіжності REST і gRPC доцільно розглядати не як несумісні технології, а як інструменти для різних архітектурних зон мікросервісної системи. REST є природним вибором для зовнішнього контуру, публічних API, інтеграцій із вебклієнтами та сторонніми системами. Підхід gRPC зі свого боку є доцільним для внутрішньої міжсервісної взаємодії, де важливими є низькі затримки, компактність повідомлень, типобезпечність контрактів і стабільна поведінка під навантаженням. Узагальнення ключових характеристик цих підходів подано в табл. 1.

Доцільність такого розмежування підтверджується також сучасними дослідженнями продуктивності комунікаційних технологій у мікросервісних системах. Зокрема в роботах [9, 11, 12] продемонстровано, що вибір комунікаційної технології може істотно впливати на затримки, пропускну здатність, масштабованість і використання ресурсів. У цьому разі gRPC зазвичай демонструє



кращі результати у внутрішніх сервісних викликах і сценаріях інтенсивного обміну даними, тоді як

REST зберігає переваги в простоті інтеграції та сумісності з різнорідними клієнтами.

**Таблиця 1.** Загальна порівняльна характеристика REST і gRPC

| Критерій              | REST                                                                                                | gRPC                                                         |
|-----------------------|-----------------------------------------------------------------------------------------------------|--------------------------------------------------------------|
| Протокол              | HTTP/1.1 або HTTP/2                                                                                 | HTTP/2                                                       |
| Формат даних          | JSON, XML та інші текстові формати                                                                  | Protocol Buffers за замовчуванням                            |
| Продуктивність        | Помірна, залежить від реалізації                                                                    | Вища у внутрішніх RPC-сценаріях                              |
| Затримки              | Зазвичай вищі в сценаріях інтенсивного JSON-обміну                                                  | Зазвичай нижчі у внутрішніх RPC-викликах                     |
| Контракти             | Необов'язкові                                                                                       | Обов'язкові (.proto)                                         |
| Складність реалізації | Нижча                                                                                               | Вища на початковому етапі                                    |
| Стримінг              | Можливий через SSE, WebSockets або chunked responses, але не є нативною моделлю класичного REST API | Нативна підтримка client, server та bi-directional streaming |
| Сумісність            | Висока                                                                                              | Обмежена в браузерах без додаткових засобів                  |
| Основне застосування  | Публічні API, зовнішня інтеграція                                                                   | Внутрішня взаємодія між сервісами                            |

Отже, результати порівняльного аналізу свідчать, що REST і gRPC не є несумісними підходами, а доповнюють один одного. REST доцільно застосовувати там, де важливими є універсальність і сумісність, тоді як gRPC є більш ефективним для внутрішньої міжсервісної взаємодії у високонавантажених системах.

#### Експериментальне дослідження ефективності REST і gRPC

Випробування проводилося в локальній мережі 1 Гбіт/с на обчислювальному вузлі з процесором Intel Core i7-12700, 32 GB оперативної пам'яті та SSD-накопичувачем. Як серверне середовище використовувалася Linux-платформа, на якій REST- і gRPC-сервіси були розгорнуті в однакових контейнерах з обмеженням 2 vCPU та 2 GB RAM для кожної реалізації. Навантажувальний клієнт Grafana k6 запускався в тому самому мережевому сегменті, що дало змогу мінімізувати вплив зовнішніх мережевих чинників і зосередити вимірювання на розбіжностях транспортного рівня, серіалізації та оброблення запитів.

Для порівняння REST API на HTTP/1.1 з JSON і gRPC на HTTP/2 з Protocol Buffers було реалізовано два окремі серверні застосунки з однаковою прикладною логікою. Така постановка відповідає типовій схемі зіставлення HTTP API з JSON і gRPC, де gRPC спирається на .proto-контракт, двійкове кодування повідомлень і HTTP/2 [6, 8, 10, 11]. Водночас необхідно зважати, що HTTP API з JSON також можуть працювати поверх HTTP/2, тому в цьому дослідженні використання HTTP/1.1 для REST було свідомим елементом постановки

експерименту, а не універсальною властивістю REST як архітектурного стилю [5–7, 10].

У межах експерименту обидва сервіси реалізовували одну й ту саму операцію GetTelemetryProfile. Запит містив ідентифікатор об'єкта, часовий інтервал, прапорець увімкнення агрегованої статистики й параметр локалізації. Відповідь повертала структурований об'єкт середнього розміру: 14 скалярних полів, вкладений блок стану, масив із 5–7 коротких рядкових міток і блок агрегованих показників. Такий формат обрано свідомо: він достатньо простий, щоб не впроваджувати складну бізнес-логіку, але водночас достатньо репрезентативний за структурою, щоб різниця між JSON і Protocol Buffers була не лише теоретичною, а й у мережевому навантаженні та затримках. Щоб уникнути впливу сторонніх чинників, сервери не зверталися до зовнішньої бази даних. Дані поверталися з попередньо підготовленого in-memory набору, ідентичного для обох реалізацій. Це дало змогу зосередити вимірювання на накладних витратах транспорту, серіалізації, десеріалізації та оброблення HTTP-з'єднань. Для обох сервісів було вимкнено стиснення тіла відповіді, щоб не змішувати ефект протоколу з ефектом компресії. REST-сервіс працював через HTTP/1.1 з keep-alive, а gRPC-сервіс – через HTTP/2.

Обидві реалізації створювалися на ASP.NET Core 8: для REST – Minimal API з System.Text.Json, для gRPC – стандартний стек ASP.NET Core gRPC з кодом, згенерованим із .proto-контракту. Тестування виконувалося в режимі Release; для обох сервісів використовувався один і той самий runtime .NET 8. Така технологічна база є зручною для відтворюваного



порівняння, оскільки одна платформа сприяє зміні лише транспортного й серіалізаційного рівня, не змішуючи ефекти різних runtime-середовищ. У навантажувальному клієнті для REST-сценарію застосовувався стандартний HTTP-клієнт Grafana k6, а для gRPC – модуль k6/net/grpc, який підтримує унарні gRPC-виклики. Режим навантаження задавався за допомогою constant-arrival-rate, тобто як відкрита модель із фіксованою інтенсивністю запуску ітерацій, незалежною від часу відповіді системи. Успішною операцією вважався запит, що завершувався без помилки транспортного рівня й повертав коректну відповідь сервісу в межах вимірювального інтервалу. Для сценаріїв constant-arrival-rate у Grafana k6 цільова інтенсивність задавалася відповідно до кожної точки навантаження, а значення preAllocatedVUs і maxVUs підбиралися так, щоб генератор навантаження не ставав обмежувальним чинником експерименту: для навантаження 100–500 RPS використовувалося preAllocatedVUs = 50, maxVUs = 200, а для 700–1100 RPS – preAllocatedVUs = 150, maxVUs = 400. TLS-шифрування не застосовувалося, оскільки метою експерименту було порівняння накладних витрат транспорту й серіалізації без

додаткового впливу криптографічного оброблення. Під час тестування було вимкнено надлишкове прикладне логування; обидві реалізації запускалися в однакових умовах контейнерного середовища.

Усі наведені нижче числові результати отримано в межах описаного експериментального стенда. Дослідження було виконано серіями за умови навантаження 100, 300, 500, 700, 900 і 1100 запитів за секунду. Для кожної точки проводився 30-секундний прогрів, після чого фіксувався 60-секундний інтервал вимірювання. Кожний сценарій повторювався п'ять разів, а до таблиць заносилися усереднені значення. Основними метриками були: середня затримка відповіді, P95, кількість успішних операцій за секунду, середній ефективний обсяг переданих даних на один виклик, а також середнє використання CPU і RAM на серверному боці. CPU і RAM фіксувалися раз за секунду засобами моніторингу контейнерів із подальшим усередненням за інтервал вимірювання.

#### Затримка й пропускна здатність

У табл. 2 наведено результати для середньої затримки, P95 і кількості успішних операцій за секунду.

Таблиця 2. Результати вимірювання затримки й пропускної здатності

| Навантаження,<br>RPS | REST                              |         |                   | gRPC                              |         |                   |
|----------------------|-----------------------------------|---------|-------------------|-----------------------------------|---------|-------------------|
|                      | середня затримка<br>відповіді, мс | P95, мс | успішних<br>оп./с | середня затримка<br>відповіді, мс | P95, мс | успішних<br>оп./с |
| 100                  | 13                                | 22      | 100               | 9                                 | 14      | 100               |
| 300                  | 19                                | 33      | 298               | 12                                | 19      | 300               |
| 500                  | 29                                | 48      | 492               | 18                                | 27      | 498               |
| 700                  | 42                                | 71      | 669               | 25                                | 39      | 692               |
| 900                  | 68                                | 118     | 824               | 38                                | 59      | 887               |
| 1100                 | 111                               | 192     | 861               | 61                                | 96      | 1058              |

Для наочного подання результатів, що містяться в табл. 2, побудовано діаграми зміни середньої затримки, P95 і кількості успішних операцій за секунду залежно від рівня навантаження. Динаміку середньої затримки для REST і gRPC наведено на рис. 2, значення P95 – на рис. 3, а зміну пропускної здатності – на рис. 4.

Як видно з табл. 2 та рис. 2–4, до рівня приблизно 500 RPS обидва підходи працюють стабільно, однак реалізація gRPC на HTTP/2 з Protocol Buffers уже на цьому етапі демонструє

нижчі значення середньої затримки й вужчий "хвіст" розподілу, відображений через P95. Після 700 RPS різниця між підходами стає більш помітною: REST швидше входить у зону деградації, де приріст цільового навантаження вже не перетворюється на пропорційний приріст кількості успішно оброблених запитів. Для gRPC ця межа настає за вищого рівня навантаження, що є очікуваним для стеку, який використовує більш компактні двійкові повідомлення та HTTP/2-механізми мультиплексування.

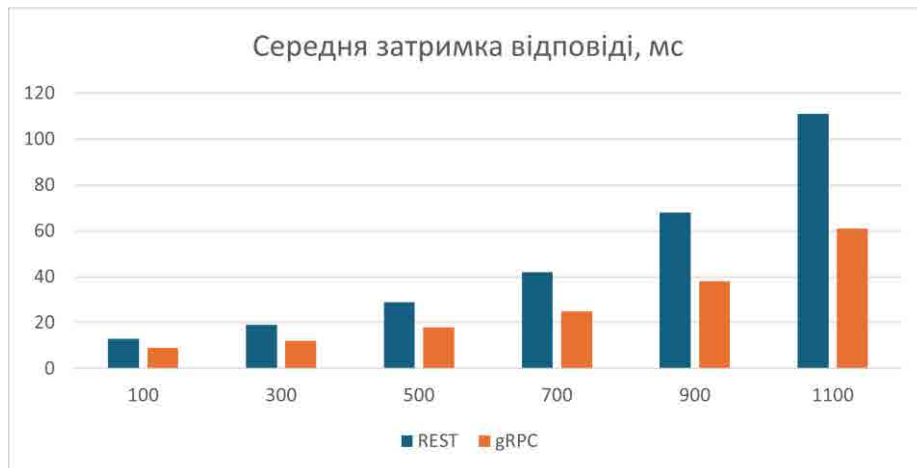


Рис. 2. Динаміка середньої затримки відповіді для REST і gRPC за різних рівнів навантаження

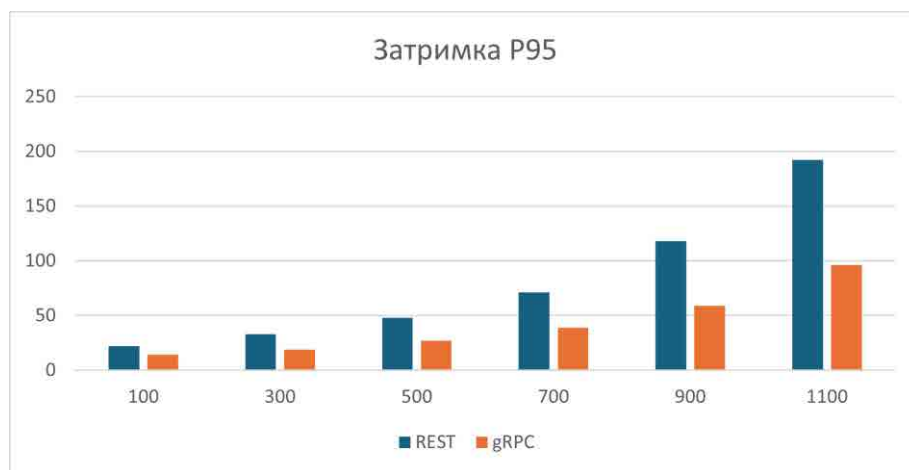


Рис. 3. Динаміка P95 затримки відповіді для REST і gRPC за різних рівнів навантаження



Рис. 4. Кількість успішних операцій за секунду для REST і gRPC за різних рівнів навантаження

#### Порівняння обсягу даних

Окремо оцінено середній ефективний обсяг даних на один виклик. Для такого порівняння важливо зважати, що службова частина мережевого

обміну в REST/HTTP/1.1 і gRPC/HTTP/2 формується по-різному. У REST-запитах заголовки здебільшого передаються як текстові поля, тоді як у HTTP/2 застосовуються механізми кодування та стиснення

заголовків, а їх фактичний внесок залежить від стану з'єднання й повторного використання попередньо переданих значень. Тому в табл. 3 подано не буквальный розмір кожного окремого HTTP-заголовка, а усереднений на один успішний виклик обсяг службової та корисної частини мережевого обміну. Під ефективним обсягом заголовків у цьому дослідженні розуміється середній обсяг службових даних протоколу, що припадає

на один виклик у межах вимірювального сценарію. Такий підхід дає змогу коректніше зіставити REST/JSON і gRPC/Protobuf, оскільки бере до уваги не лише розмір тіла повідомлення, а й практичний внесок протокольних накладних витрат. Очікувано, що Protocol Buffers забезпечує більш компактне двійкове подання структурованих даних порівняно з текстовим JSON [8, 10].

Таблиця 3. Значення середнього обсягу даних на один виклик

| Підхід                   | Ефективний обсяг заголовків запиту, В | Тіло запиту, В | Ефективний обсяг заголовків відповіді, В | Тіло відповіді, В | Загальний обсяг обміну, В |
|--------------------------|---------------------------------------|----------------|------------------------------------------|-------------------|---------------------------|
| REST (HTTP/1.1 + JSON)   | 410                                   | 278            | 430                                      | 932               | 2050                      |
| gRPC (HTTP/2 + Protobuf) | 92                                    | 146            | 108                                      | 472               | 818                       |

Наведені дані відтворюють типову для такого сценарію тенденцію: основний вииграш gRPC формується не лише завдяки меншим накладним витратам на заголовки, а й більш компактному поданню корисного навантаження.

#### Застосування ресурсів CPU і RAM

Оскільки нижча затримка ще не гарантує кращої експлуатаційної ефективності, додатково оцінено витрати CPU і RAM на серверному боці. У табл. 4 подано усереднені значення для характерних точок навантаження.

Таблиця 4. Значення використання серверних ресурсів

| Навантаження, RPS | REST   |         | gRPC   |         |
|-------------------|--------|---------|--------|---------|
|                   | CPU, % | RAM, MB | CPU, % | RAM, MB |
| 300               | 21     | 286     | 16     | 268     |
| 700               | 49     | 431     | 37     | 395     |
| 900               | 66     | 522     | 49     | 451     |
| 1100              | 78     | 603     | 61     | 512     |

Для візуального порівняння застосування серверних ресурсів за різних рівнів навантаження результати табл. 4 продемонстровано у вигляді діаграми (рис. 5).

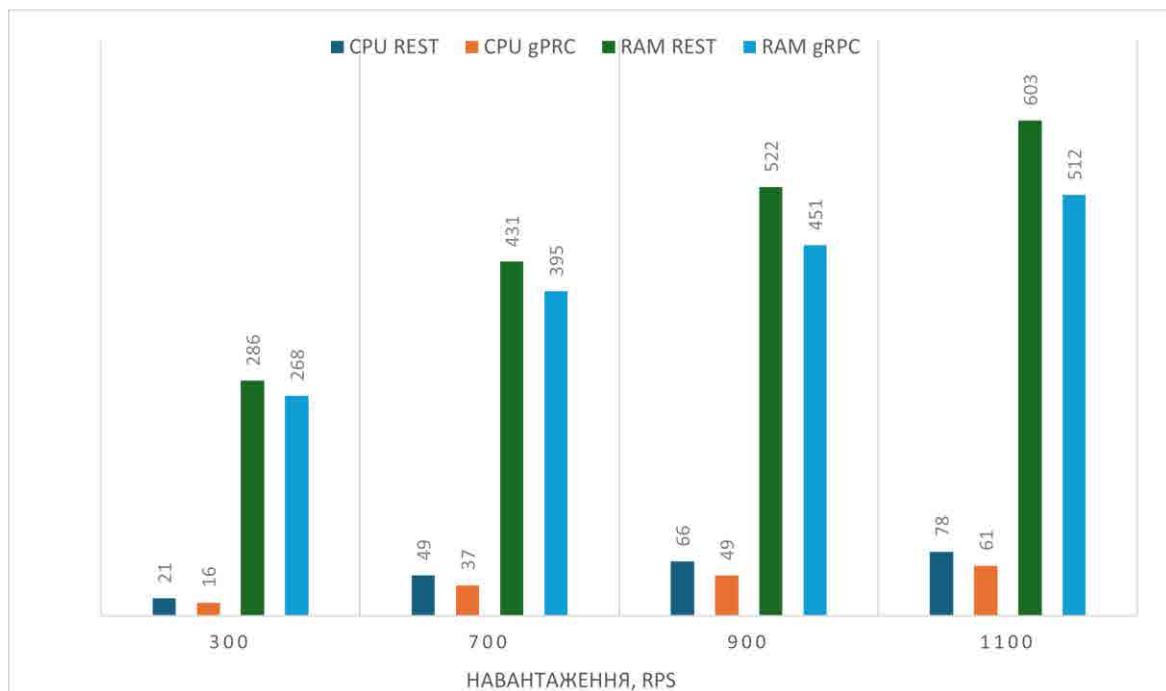


Рис. 5. Використання CPU і RAM серверного контейнера для REST і gRPC за різних рівнів навантаження

Відповідно до отриманих показників, зі зростанням навантаження REST швидше підходить до межі процесорної насиченості. Це узгоджується з очікуванням, що текстова серіалізація JSON і більший обсяг мережевого обміну створюють вищі накладні витрати на оброблення запитів. Для gRPC витрати CPU і RAM також зростають, але повільніше, що добре узгоджується з використанням більш компактного двійкового формату й ефективнішої транспортної взаємодії поверх HTTP/2. Значення RAM варто інтерпретувати як середнє використання пам'яті контейнером у межах вимірювального інтервалу, що містить не лише об'єкти прикладної логіки, а й службові структури runtime, мережеві буфери та механізми оброблення запитів.

### **Інтерпретація результатів**

Проведений експеримент дає змогу дійти певних висновків. По-перше, зіставлення середньої затримки та P95 дає змогу оцінити не лише середній час відповіді, а й стабільність сервісу під навантаженням. По-друге, таблиця обсягу даних демонструє, чи справді виграш пов'язаний із компактністю бінарної серіалізації, а не лише з особливостями реалізації серверного коду. По-третє, CPU і RAM дають змогу відокремити мережеву ефективність від суто обчислювальної вартості оброблення запиту.

У межах цієї постановки експерименту реалізація gRPC на HTTP/2 з Protocol Buffers продемонструвала перевагу над REST API на HTTP/1.1 з JSON у сценарії міжсервісної взаємодії, де важливими є низькі затримки, висока інтенсивність запитів і стабільна поведінка в умовах зростання навантаження. Водночас цей висновок необхідно інтерпретувати в межах заданих умов експерименту. Його не можна автоматично поширювати на всі ймовірні реалізації REST і gRPC, оскільки HTTP API з JSON також здатні працювати поверх HTTP/2, а для браузерних клієнтів стандартні HTTP API часто залишаються більш природним варіантом. Звичайний gRPC у таких сценаріях потребує додаткових механізмів, зокрема gRPC-Web або JSON transcoding [10].

### **Висновки**

У роботі досліджено підходи до реалізації міжсервісної взаємодії в мікросервісній архітектурі з використанням REST і gRPC у середовищі .NET 8. Розглянуто архітектурні особливості синхронного

обміну даними між сервісами, а також проаналізовано практичні аспекти реалізації HTTP API з JSON і gRPC-сервісів на платформі ASP.NET Core 8. Основну увагу зосереджено на впливі транспортного рівня, форматі серіалізації та способі організації викликів на ефективність міжсервісної взаємодії.

У межах роботи проведено експериментальне дослідження, в якому реалізовано два сервіси з однаковою прикладною логікою: REST-сервіс на основі HTTP/1.1 та JSON і gRPC-сервіс на основі HTTP/2 й Protocol Buffers. Для забезпечення коректності порівняння обидва варіанти працювали з однаковим in-memory набором даних, без звернення до зовнішньої бази даних, а навантаження генерувалося засобами Grafana k6 в однакових режимах інтенсивності запитів. Така постановка дала змогу зосередити аналіз саме на відмінностях транспорту, серіалізації та накладних витрат оброблення запитів. Порівняльний аналіз продемонстрував, що REST і gRPC мають різні сфери ефективного застосування. REST властива простота реалізації, висока сумісність, прозорість інтерфейсу та зручність інтеграції, що робить його доцільним для побудови публічних API, зовнішніх інтеграцій і взаємодії з гетерогенними системами. Натомість gRPC є більш придатним для внутрішньої міжсервісної взаємодії, де критичними є низькі затримки, стабільність часу відповіді, висока пропускна здатність і ефективне використання мережевих ресурсів.

За результатами проведеного експерименту встановлено, що в межах запропонованого сценарію реалізація gRPC на HTTP/2 з Protocol Buffers продемонструвала нижчі значення середньої затримки та P95, менший обсяг мережевого обміну й вищу пропускну здатність порівняно з REST API на HTTP/1.1 з JSON. Результати показали, що на верхніх рівнях навантаження середній час відповіді для gRPC був нижчим орієнтовно в 1,7–1,8 раза, тоді як загальний обсяг переданих даних на один виклик виявився меншим приблизно у 2,5 раза, якщо порівнювати з REST. Крім того, для gRPC зафіксовано більш помірне використання процесорних і пам'яттєвих ресурсів серверного вузла. Це дає змогу зробити висновок, що поєднання HTTP/2 і Protocol Buffers дійсно створює передумови для підвищення ефективності внутрішньої сервісної взаємодії.

Водночас досягнуті результати не дають підстав розглядати gRPC як універсальну заміну REST у всіх

сценаріях. У разі невисокого навантаження, потреби в максимально простій інтеграції, широкий клієнтський сумісності або орієнтації на публічний вебінтерфейс REST зберігає свою практичну перевагу. Отже, REST і gRPC доцільно розглядати не як несумісні, а як взаємодоповнювальні підходи, що можуть ефективно комбінуватися в межах однієї мікросервісної архітектури залежно від призначення окремих сервісів і вимог до їх взаємодії. Отже, вибір підходу до реалізації міжсервісної взаємодії має здійснюватися з огляду на вимоги до затримок, пропускну здатності, масштабованості, сумісності та експлуатаційної складності. Практична цінність роботи полягає в тому, що запропонована методика експерименту й досягнуті результати можуть бути використані для подальшого розширення сценаріїв тестування й обґрунтованого вибору механізму міжсервісної взаємодії.

Перспективи подальших досліджень полягають у проведенні повномасштабних експериментів із реальними навантаженнями, у розширенні набору сценаріїв унаслідок різних типів повідомлень і конфігурацій мережевого середовища, а також

у дослідженні поєднання синхронних і асинхронних механізмів взаємодії в мікросервісних системах.

### Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

### Фінансування

Дослідження проводилося без фінансової підтримки.

### Доступність даних

Рукопис не має пов'язаних даних.

### Використання засобів штучного інтелекту

Автори підтверджують, що не застосовували технології ШІ для написання роботи.

## References

1. Dragoni, N., Giallorenzo, S., Lluch Lafuente, A., Mazzara, M., Montesi, F., Mustafin, R., Safina, L. (2017), "Microservices: Yesterday, Today, and Tomorrow", In: *Present and Ulterior Software Engineering*. Cham: Springer, pp. 195–216. DOI: [https://doi.org/10.1007/978-3-319-67425-4\\_12](https://doi.org/10.1007/978-3-319-67425-4_12)
2. Velepucha, V., Flores, P. (2023), "A Survey on Microservices Architecture: Principles, Patterns and Migration Challenges", *IEEE Access*, Vol. 11, pp. 88339–88358. DOI: <https://doi.org/10.1109/ACCESS.2023.3305687>
3. Weerasinghe, L. D. S. B., Perera, I. (2022), "Evaluating the Inter-Service Communication on Microservice Architecture", *7th International Conference on Information Technology Research (ICITR)*, pp. 1–6. DOI: <https://doi.org/10.1109/ICITR57877.2022.9992918>
4. Bogner, J., Kotstein, S., Pfaff, T. (2023), "Do RESTful API design rules have an impact on the understandability of Web APIs?", *Empirical Software Engineering*, Vol. 28, Article 132. DOI: <https://doi.org/10.1007/s10664-023-10367-y>
5. Fielding, R. T., Nottingham, M., Reschke, J. (2022), *HTTP Semantics*. RFC 9110. RFC Editor. DOI: <https://doi.org/10.17487/RFC9110>
6. Thomson, M., Benfield, C. (2022), *HTTP/2*. RFC 9113. RFC Editor. DOI: <https://doi.org/10.17487/RFC9113>
7. Fielding, R. T., Nottingham, M., Reschke, J. (2022), *HTTP/1.1*. RFC 9112. RFC Editor. DOI: <https://doi.org/10.17487/RFC9112>
8. Maltsev, E. Y., Muliarevych, O. (2024), "Beyond JSON: Evaluating Serialization Formats for Space-Efficient Communication", *Advances in Cyber-Physical Systems*, Vol. 9, No. 1, pp. 9–15. DOI: <https://doi.org/10.23939/acps2024.01.009>
9. Niswar, M., Arisandy Safruddin, R., Bustamin, A., Aswad, I. (2024), "Performance evaluation of microservices communication with REST, GraphQL, and gRPC", *International Journal of Electronics and Telecommunications*, Vol. 70, No. 2, pp. 429–436. DOI: <https://doi.org/10.24425/ijet.2024.149562>
10. Microsoft. Compare gRPC services with HTTP APIs. URL: <https://learn.microsoft.com/en-us/aspnet/core/grpc/comparison>
11. Bolanowski, M., Żak, K., Paszkiewicz, A., Ganzha, M., Paprzycki, M., Sowiński, P., Lacalle, I., Palau, C. E. (2022), "Efficiency of REST and gRPC Realizing Communication Tasks in Microservice-Based Ecosystems", In: *New Trends in Intelligent Software Methodologies, Tools and Techniques*. IOS Press, pp. 97–108. DOI: <https://doi.org/10.3233/FAIA220242>



12. Jarmoszewicz, J., Iwanowski, P., Plechawska-Wójcik, M. (2024), "Analysis of the performance and scalability of microservices depending on the communication technology", *Journal of Computer Sciences Institute*, Vol. 33, pp. 323–330. DOI: <https://doi.org/10.35784/jcsi.6499>
13. Tkachov, V. (2026), "Method of restoring critical information system services on a mobile platform under conditions of controlled degradation", *Automated Control Systems and Automation Devices*, No. 188, pp. 78–97. DOI: <https://doi.org/10.30837/0135-1710.2026.188.078>
14. Mykhailichenko, I. V., Liashenko, O. S. (2025), "Resource usage prediction model in cloud computing using Informer architecture", *Automated Control Systems and Automation Devices*, No. 186, pp. 17–28. DOI: <https://doi.org/10.30837/0135-1710.2025.186.017>
15. Glavchev, M. I., Glavchev, D. M., Panchenko, V. I. (2025), "Evaluating the effectiveness of open-source solutions for monitoring and load balancing in microservice applications ", *Automated Control Systems and Automation Devices*, No. 187, pp. 182–199. DOI: <https://doi.org/10.30837/0135-1710.2025.187.182>

Received (Надійшла) 19.03.2026

Accepted for publication (Прийнята до друку) 20.04.2026

Publication date (Дата публікації) 29.05.2026

#### *Відомості про авторів / About the Authors*

**Бобровнікова Кіра Юліївна** – кандидат технічних наук, доцент, Університет економіки і підприємництва, завідувач кафедри математики та інформаційних технологій, Хмельницький, Україна;

**Kira Bobrovnikova** – Candidate of Technical Sciences, Associate Professor, University of Economics and Entrepreneurship, Head of the Department of Mathematics and Information Technologies, Khmelnytskyi, Ukraine;

e-mail: [bobrovnikova.kira@gmail.com](mailto:bobrovnikova.kira@gmail.com)

ORCID ID: <https://orcid.org/0000-0002-1046-893X>

**Гурман Іван Васильович** – кандидат технічних наук, доцент, Університет економіки і підприємництва, доцент кафедри математики та інформаційних технологій, Хмельницький, Україна;

**Ivan Hurman** – Candidate of Technical Sciences, Associate Professor, University of Economics and Entrepreneurship, Associate Professor of the Department of Mathematics and Information Technologies, Khmelnytskyi, Ukraine;

e-mail: [devastator167384@gmail.com](mailto:devastator167384@gmail.com)

ORCID ID: <https://orcid.org/0000-0002-2282-3484>

**Олексюк Дмитро Андрійович** – Університет економіки і підприємництва, старший викладач кафедри математики та інформаційних технологій, Хмельницький, Україна;

**Dmytro Oleksiuk** – University of Economics and Entrepreneurship, Senior Lecturer of the Department of Mathematics and Information Technologies, Khmelnytskyi, Ukraine;

e-mail: [oleksuk.dima@gmail.com](mailto:oleksuk.dima@gmail.com)

ORCID ID: <https://orcid.org/0009-0006-3735-1930>

## COMPARATIVE ANALYSIS OF INTER-SERVICE COMMUNICATION USING REST AND GRPC IN .NET 8 ENVIRONMENT

The subject of the study is approaches to implementing synchronous inter-service communication in a microservice architecture using REST and gRPC in the .NET 8 environment. The aim of the work is to conduct a comparative analysis of a REST API based on HTTP/1.1 and JSON and a gRPC service based on HTTP/2 and Protocol Buffers, as well as to evaluate their impact on performance, network exchange volume, and server resource utilization. To achieve this aim, the following tasks are defined: to analyze the architectural differences between REST and gRPC; to implement two services with identical application logic in the ASP.NET Core 8 environment; to ensure identical experimental conditions; to measure average response latency, P95 latency, the number of successful operations per second, the volume of transmitted data, and CPU and RAM utilization;

and to interpret the obtained results with regard to the application limits of each approach. **Methods.** The study uses comparative analysis of architectural approaches, experimental load testing, and quantitative performance evaluation. Two services with identical application logic and a shared in-memory dataset were implemented without accessing an external database. The workload was generated using Grafana k6 in constant-arrival-rate mode. **Research results.** Within the proposed experimental setup, gRPC over HTTP/2 with Protocol Buffers demonstrated lower average latency and P95 values, a smaller network exchange volume, and higher throughput compared with the REST API over HTTP/1.1 with JSON. The total exchange volume for gRPC was approximately 2.5 times smaller, while at the upper load levels the average latency was about 1.7-1.8 times lower. More moderate CPU and memory resource utilization of the server node was also recorded for gRPC. **Conclusions.** The obtained results confirm the feasibility of using gRPC primarily for internal inter-service communication, where low latency, high throughput, and efficient use of network resources are critical. REST is appropriate in scenarios where compatibility, ease of integration, transparency of the HTTP interface, and usability in public APIs are priorities.

**Keywords:** microservice architecture; inter-service communication; distributed systems; performance; response latency; throughput.

#### *Бібліографічні описи / Bibliographic descriptions*

Бобровнікова К. Ю., Гурман І. В., Олексюк Д. А. Порівняльний аналіз міжсервісної взаємодії з використанням REST і gRPC у середовищі .NET 8. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 95–108. DOI: <https://doi.org/10.30837/0135-1710.2026.189.095>

Bobrovnikova, K., Hurman, I., Oleksiuk, D. (2026), "Comparative analysis of inter-service communication using REST and gRPC in .NET 8 environment", *Management Information System and Devices*, No. 2 (189), P. 95–108. DOI: <https://doi.org/10.30837/0135-1710.2026.189.095>

---

Dmytro Diachenko, Oleg Mikhal

## METHOD FOR ASSESSING THE RISK OF DEFECTS IN SOFTWARE COMPONENTS OF DISTRIBUTED COMPUTER SYSTEMS FOR AUTOMATED TEST PRIORITIZATION

**The subject of research** is methods, models, and tools for assessing the risk of defects in software components of distributed computer systems to support the process of automated test prioritization. **The aim of the work** is to develop a method for assessing the risk of defects in software components of distributed computer systems for automated test prioritization based on machine learning, considering structural characteristics of components, change intensity, and testing results. In accordance with the stated aim, **the following tasks** are to be accomplished: to analyze current approaches to software defect prediction and automated test prioritization in distributed computer systems; to identify informative characteristics of software components that reflect their structural properties, change intensity, and relationships between software components; to develop a method for assessing the risk of defects in software components based on a machine learning model with the formation of an integral risk indicator; to develop an approach to automated test prioritization based on the obtained integral risk indicator; to implement the proposed method and to conduct an experimental evaluation of its effectiveness using classification and ranking metrics. **Methods.** The study employs system analysis methods to investigate the features of functioning of distributed computer systems and the process of automated testing of software components. To construct the feature space, methods for analyzing structural metrics of source code, characteristics of change intensity, relationships between software components, and the construction of derived indicators of complexity and coupling are applied. The development of the defect risk assessment method is carried out using machine learning methods, classification, feature selection, formation of an integral risk indicator, and risk-oriented ranking. For experimental validation of the proposed approach, computational modeling, comparative analysis, and model quality evaluation methods are used, including ROC-AUC, Precision–Recall, F1-score, and Balanced Accuracy metrics. The software implementation is carried out in Python using the Google Colab environment. **Results.** A method for assessing the risk of defects in software components of distributed computer systems has been developed, which ensures the integration of structural, process, and contextual characteristics, as well as the computation of an integral risk indicator and ranking of components for automated test prioritization. Experimental validation has confirmed the effectiveness of the proposed method and its ability to accurately distinguish between defective and non-defective software components, as well as to identify high-risk components for priority testing. **Conclusions.** The results obtained confirm the feasibility of using machine learning methods and an integrated feature space for assessing the risk of defects in software components of distributed computer systems. The proposed method can be used as a basis for improving the justification of decisions on automated test prioritization, timely identification of high-risk components, and further development of intelligent software quality assurance tools.

**Keywords:** software components, distributed computer systems, automated testing, machine learning, defect prediction, integral risk indicator, prioritization, testing, Python, Google Colab.

### 1. Introduction

Modern distributed computer systems form the foundation for a significant number of software platforms that operate through the continuous interaction of a large number of interconnected components. Such systems include cloud services, microservice architectures, enterprise information platforms, data processing systems, and mission-critical hardware-software complexes. They are characterized by a high level of structural complexity, intensive data exchange between components, frequent changes to the software code, and a significant dependence of the system's overall

performance on the reliability of individual software modules. Under these conditions, even local defects in individual components can lead to the disruption of significant parts of the system, reduction in productivity, loss of data consistency, or service failures. This results in increased demands on the organization of automated software testing, which must ensure timely defect detection and the rational allocation of testing resources.

Traditional approaches to automated software testing are largely based on fixed test execution rules, standardized verification procedures, or the uniform distribution of test resources among software components. This approach to testing does not account

for the actual defect risk of individual modules, their structural characteristics, interdependencies with other components, the rate of change, or the results of previous tests. As a result, test resources are often spent on verifying components with a relatively low probability of defects, while high-risk software modules do not receive the appropriate priority. This is particularly critical for distributed computer systems, where the complexity of interactions between components increases the likelihood of defects, and delays in their detection can have significant functional and operational consequences.

One promising approach to improving the effectiveness of automated testing is the application of machine learning methods to assess the risk of defects in software components. The use of such methods allows for the analysis of a set of structural characteristics of the software code, connectivity metrics, change intensity, and test results to form an integrated assessment of defect risk. The resulting metric can be used as a basis for ranking components by potential defect severity and subsequently prioritizing automated testing. This enables a shift from universal testing schemes to a risk-oriented approach, in which testing resources are directed primarily toward the most critical parts of the software system. Existing approaches to software defect prediction are predominantly focused either on the use of individual software code metrics or on the analysis of historical defect data without sufficient integration of the diverse characteristics of software components into a unified assessment framework. At the same time, the task of planning automated testing in distributed computer systems requires an approach that combines the analysis of components' structural properties, the characteristics of their interrelationships, the intensity of changes, and test execution results. This necessitates the development of a method for assessing the defect risk of software components in distributed computer systems, which would ensure the formation of an integrated risk indicator and its use for prioritizing automated testing. Solving this problem is a relevant scientific and applied task aimed at improving testing efficiency, earlier defect detection, and enhancing the overall quality of software for distributed computer systems.

## **2. Analysis of current scientific publications and identification of the research problem**

---

Current research explores various approaches to defect prediction, ranging from classical statistical

models to ensemble algorithms and deep learning methods. Currently, there is a need to identify scientific approaches that can serve as a basis for developing a method for assessing the risk of defects in software components of distributed computer systems, focused not only on predicting the occurrence of defects but also on supporting the prioritization of automated testing. That is why the subsequent analysis of scientific publications is aimed at identifying the achievements and limitations of existing solutions, as well as at justifying the research problem, which lies in the need to construct a method capable of generating an integrated defect risk metric based on a machine learning model and using it for more effective planning of automated testing.

Work [1] examines an approach to software defect prediction within the framework of the Just-In-Time Software Defect Prediction concept [2], which is based on the use of deep learning models. The authors justify the feasibility of early defect detection at the level of individual changes in the source code and demonstrate the possibility of improving software quality through the use of historical data and change metrics. The proposed multi-layer perceptron model achieves a prediction accuracy of 82.08% and identifies the most influential metrics, including the volume of changes, developer experience, and the degree of change distribution across the codebase. However, despite the effectiveness of the approach, the study focuses primarily on classifying changes as either defective or non-defective and does not consider the formation of an integrated defect risk metric for software components as a basis for prioritizing automated testing. Furthermore, the features used primarily reflect the characteristics of code changes without comprehensively accounting for the structural properties of software components and a generalized representation of risk in the context of distributed computer systems. This limits the applicability of the obtained results to test planning tasks and the effective allocation of verification resources.

The paper [3] considers an approach to software defect prediction in which the task of predicting defects in software code is proposed to be interpreted not only as a classification task but as a task of ranking software modules by defect severity. The authors argue that for practical test planning, it is not enough to simply determine whether a module contains a defect, since it is also important to establish the order of module verification given limited testing resources. To this end, the study investigates an approach that allows modules

to be ranked by the number of errors or defect density and the quality of the ranking to be assessed by the average error rate. A comparative analysis of ranking models was also conducted, and the impact of data imbalance and feature selection on prediction quality was analyzed. The results showed that using the number of errors as the target variable yields higher and more stable ranking results than using error density, and that training on unbalanced data and feature selection do not provide universal improvements across all scenarios. Special attention was paid to the use of both static metrics of the program code and metrics characterizing the history of changes and the evolution of software modules. At the same time, the study focuses primarily on improving the accuracy of classifying modules with defects, while the issues of developing a generalized defect risk metric and its use to support the prioritization of automated testing remain insufficiently explored. Furthermore, the study does not pay specific attention to the characteristics of distributed computer systems, for which not only individual characteristics of software modules are important, but also the relationships between software components and their impact on the overall risk of defects. This confirms the relevance of developing an approach focused on assessing the defect risk of software components in distributed computer systems, with the subsequent use of the obtained assessments for more informed prioritization of automated testing.

Article [4] investigates the problem of software defect prediction based on machine learning and deep learning methods. The authors consider the prediction of software component defects as a tool for early detection of errors in software classes with the aim of reducing software maintenance costs and improving its reliability. The study utilizes the large Unified Bug Dataset, which combines data from five open sources and contains information on 47,618 classes and approximately 60 software code metrics reflecting complexity, connectivity, cohesion, size, and other characteristics of the classes' internal quality. The study also compared eight machine learning approaches. Before building the models, the target variable was binarized and the features were standardized, and the results were evaluated using several metrics. The LSTM model showed the best results, confirming the feasibility of using machine learning. However, the study focuses primarily on improving the accuracy of defect class detection based on a set of source code metrics, while the use of prediction results for controlling the automated testing process

remains outside the scope of consideration. Furthermore, the proposed approach is primarily focused on the level of individual classes and does not emphasize the specifics of distributed computer systems, where the relationships between software components and the integrated assessment of defect risk are of significant importance.

Article [5] addresses the problem of software defect prediction based on optimized machine learning models, with a focus on the impact of hyperparameter tuning and feature dimensionality reduction. The authors emphasize that the effectiveness of the models depends significantly not only on the type of algorithm but also on the correct selection of parameters and relevant features used for training. The research methodology includes feature dimensionality reduction using the principal component analysis method, optimization of the number of components and model hyperparameters using random search, as well as a comparison of several machine learning algorithms. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics, which allowed for a comprehensive assessment of the models' ability to detect defective modules. The experimental results showed that the k-nearest neighbors method is the most effective algorithm in most cases, and that hyperparameter optimization and data preprocessing significantly improved prediction accuracy compared to baseline models without tuning. It is worth noting that the study focuses primarily on improving the accuracy of classifying defective modules and does not consider the subsequent use of prediction results in test management or resource allocation processes. Furthermore, the approach is model-specific and does not provide for the integration of results from different algorithms into a single, generalized defect risk metric. This highlights the relevance of developing an approach that would combine defect prediction results with the formation of an integrated risk assessment of software components and ensure their use for intelligent prioritization of automated testing in complex software systems.

Article [6] addresses the problem of software defect prediction in the context of large-scale complex information systems, for which traditional testing approaches are resource-intensive and insufficiently effective.

The authors emphasize that the increasing complexity of systems necessitates a transition to intelligent analysis methods, in particular the use of machine learning for defect prediction based on historical development and testing data. The paper proposes a generalized approach to static defect prediction, which



involves forming a set of features for software modules, building a classification model, and subsequently applying it to determine the risk of defects in new components. Particular attention is paid to the formation of informative metrics that include not only software code characteristics but also development process indicators, specifically the frequency of changes, the complexity of modifications, and organizational aspects of the team's work. The prediction model is implemented as a binary classification task using machine learning methods, specifically logistic regression, which allows for the assessment of the probability of defects in a software module. At the same time, the approach proposed by the authors is primarily focused on classifying modules based on defect risk and does not provide for the formation of a generalized integrated risk indicator that would account for the cumulative impact of diverse characteristics of software components. Furthermore, the prediction results are not directly integrated into test management processes. This justifies the development of an approach that would combine machine learning with the formation of an integrated assessment of software component defect risk and ensure its use to improve testing efficiency in distributed computer systems.

Work [7] is devoted to solving the problem of defect prediction in a dynamic software development environment, specifically in the context of [2], where data arrives sequentially over time. The authors highlight the limitations of traditional approaches, which primarily operate in a static mode and do not account for dynamic changes in data characteristics caused by the evolution of software systems. The paper presents a comprehensive comparison of online and offline approaches to training defect prediction models. The BORB method is proposed, which allows offline models to be adapted to an online scenario through periodic retraining on updated data, taking into account the delay in defect verification. The results of the experimental study showed that offline models in an online scenario can demonstrate slightly better prediction quality compared to classical online methods, although this is achieved at the cost of greater computational complexity. At the same time, the approach is primarily focused on classifying changes as either defective or non-defective and does not provide for the formation of an integrated defect risk metric for software components. Furthermore, the prediction results are not directly integrated into the processes of optimizing and prioritizing automated testing.

This justifies the need to develop a method for assessing defect risk in distributed computer systems that would combine machine learning with the formation of a generalized risk-oriented indicator and ensure its use to improve testing efficiency.

Article [8] addresses the problem of software defect prediction using machine learning methods aimed at improving the quality and reliability of software systems and reducing development costs. The authors emphasize that early detection of defective modules allows for significant optimization of the testing process and reduction of software maintenance costs, which is particularly relevant for modern complex software systems. The paper proposes an approach to generating software project metrics used as input parameters for prediction models. These metrics form a generalized representation of the software project's state and allow for the consideration of the impact of different life cycle phases on defect occurrence. The results of the experimental study demonstrate that the model based on a decision tree regression provides higher prediction accuracy compared to the k-nearest neighbors method. However, the study focuses primarily on predicting the number of defects and does not account for the formation of an integrated risk metric that could be used to support decision-making in the testing or quality management process. Furthermore, the model does not account for dynamic changes in data over time, which can be critical in distributed software systems.

Article [9] addresses the problem of software defect prediction using machine learning methods, with an emphasis on improving accuracy through ensemble approaches. The author emphasizes that the reliability of software systems directly depends on the timely detection of defects, and the application of intelligent methods allows for optimizing the testing process and reducing resource consumption. The paper provides a detailed analysis of approaches to defect prediction, specifically regression and classification models. Particular attention is paid to ensemble learning methods, which allow for improved prediction accuracy by combining several base models. An experimental study was conducted using the open-source NASA PROMISE and GitHub datasets, which contain information about software modules and their defects. In most cases, the ensemble model demonstrated better results compared to individual algorithms and also provided more stable productivity for both small and large datasets. It should be noted that this study focuses primarily on classical machine learning

algorithms and does not consider the capabilities of deep learning or adaptive models, which opens up prospects for further research in this direction.

Article [10] discusses modern approaches to software defect prediction using machine learning methods. The authors emphasize that early defect detection makes it possible to improve software quality and reduce maintenance costs. The paper analyzes the main groups of machine learning methods used in software component defect prediction tasks. It summarizes current research using various datasets and highlights the importance of data preprocessing and the proper selection of metrics for evaluating model quality. The authors note that accuracy, precision, recall, F1-score, and ROC curves are most commonly used to evaluate model performance. Overall, the review confirms the high effectiveness of machine learning methods in defect prediction tasks and demonstrates a trend toward the expanding use of ensemble and deep learning models. It should be noted that this work is primarily review-based and does not propose a specific approach to developing an integrated defect risk metric for software components. Furthermore, the study does not consider the use of prediction results to support the prioritization of automated testing.

Article [11] examines the problem of software defect prediction using traditional and ensemble machine learning methods. The authors emphasize the importance of early defect detection for improving the quality of software systems and reducing maintenance costs. The paper investigates the performance of many machine learning algorithms and applies ensemble approaches. Experimental results obtained on NASA datasets demonstrate the high performance of ensemble models, particularly Gradient Boosting, confirming the feasibility of their use in defect prediction tasks. At the same time, it should be noted that the research is primarily focused on improving classification accuracy and does not involve the development of an integrated defect risk metric for software components, nor its use for prioritizing automated testing in distributed computer systems.

In current research, significant attention is paid to improving the effectiveness of software defect prediction based on machine learning methods, driven by the increasing complexity of software systems and the need to optimize testing processes. One of the key areas is ensuring the quality of training data and its representativeness, as these factors significantly influence the accuracy and generalization ability of models. Article

[12] addresses the limitation of existing defect prediction models caused by insufficient volume and heterogeneity of training samples. A large and diverse dataset, Defectors, is proposed, which covers a significant number of software projects across various domains and enables improved effectiveness in applying machine learning and deep learning models. Particular attention is paid to class balancing and reducing noise in the data, which positively impacts prediction quality. At the same time, the work focuses primarily on building a high-quality training dataset and does not address the integration of the obtained results into the decision-making process regarding testing prioritization. An analysis of current scientific publications has shown that existing approaches to software defect prediction based on machine learning provide a solid theoretical and applied foundation for further research in this field. In most of the works reviewed, the main focus is on improving the accuracy of classifying defective modules, optimizing the feature space, using ensemble algorithms, deep learning, and adapting models to the characteristics of specific datasets. At the same time, such studies are predominantly oriented either toward identifying components with defects or toward ranking modules by the severity of the defects themselves, without fully integrating the prediction results into the decision-making process regarding the organization of automated testing. At the same time, the issues of developing a method capable of combining the structural characteristics of software components, the intensity of changes, the relationships between software components, and test execution parameters into a single integrated defect risk metric remain insufficiently addressed. Furthermore, there is limited coverage of the specifics of applying such approaches specifically to distributed computer systems, for which complex component interactions, data heterogeneity, and the need for well-founded prioritization of automated testing are of significant importance. In this regard, it is advisable to develop a method for assessing the defect risk of software components in distributed computer systems based on machine learning, which will ensure the formation of an integrated risk indicator and its use to improve the effectiveness of automated test prioritization.

---

### 3. Research objectives and tasks

---

The aim of this work is to develop a method for assessing the risk of defects in software components of distributed computer systems in order to prioritize

---

automated testing based on machine learning, taking into account the structural characteristics of the components, the rate of change, and test results. Given this objective, the following tasks must be performed: analyze current approaches to software defect prediction and the prioritization of automated testing in distributed computer systems; identify informative characteristics of software components that reflect their structural properties, change intensity, and interdependencies between software components; develop a method for assessing the risk of software component defects based on a machine learning model with the formation of an integrated risk indicator; develop an approach to prioritizing automated testing based on the obtained integrated risk indicator; implement the proposed method and conduct an experimental evaluation of its effectiveness in terms of classification and ranking metrics.

#### 4. Presentation of the main material

Improving the effectiveness of automated testing in distributed computer systems requires not only identifying defects in individual software components but also quantitatively assessing their risk level. In real-world conditions, testing resources are limited; therefore, the task of ranking components by the expected probability of a defect occurring and identifying the most critical objects for early test control becomes of paramount importance. It is for this purpose that this article proposes a method for assessing the risk of defects in software components of distributed computer systems, which combines machine learning models, feature engineering, and the integration of structural and process characteristics, and justifies the introduction of an integrated risk indicator.

The basic idea behind the method is that each software component is described by a set of quantitative characteristics that reflect both its internal properties and the context of changes and the development history of the corresponding project. Let the set of software components of a distributed system be defined as follows:

$$C = \{c_1, c_2, \dots, c_n\}. \quad (1)$$

Each component  $c_i$  is associated with a multidimensional feature vector

$$\mathbf{x}_i = (\mathbf{x}_i^{(hist)}, \mathbf{x}_i^{(test)}, \mathbf{x}_i^{(str)}, \mathbf{x}_i^{(dep)}, \mathbf{x}_i^{(ops)}), \quad (2)$$

where  $\mathbf{x}_i^{(hist)}$  is a vector of historical characteristics of component changes,  $\mathbf{x}_i^{(test)}$  is a vector of metrics related

to test results or test control parameters,  $\mathbf{x}_i^{(str)}$  is a vector of structural characteristics of the program code,  $\mathbf{x}_i^{(dep)}$  is a vector of characteristics of dependencies and relationships between software components, and  $\mathbf{x}_i^{(ops)}$  is a vector of operational or contextual parameters. This division of the feature space allows for a formal account of the multifactorial nature of software component defects. Unlike classical approaches, which are often limited to code metrics alone, this representation allows for the simultaneous integration of information about structure, development history, the context of changes, and component interactions.

We define the target variable as

$$y_i = \begin{cases} 1, & \text{if the component has a defect,} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

At the first level, the method generates an estimate of the probability of defects in a software component

$$p_i = f(\mathbf{x}_i), \quad (4)$$

where  $f(\cdot)$  – a machine learning model. The value  $p_i$  is interpreted as an estimate of the probability that component  $c_i$  is defective or has an increased susceptibility to defects. Thus, function  $f$  is the core analytical component of the method, and its result is used not as a final decision but as one of the components of the integrated risk assessment. The integral risk indicator for the  $i$ -th component is given by the relation

$$R_i = \alpha p_i + (1 - \alpha) \sum_{j=1}^5 \lambda_j \hat{x}_{ij}, \quad \sum_{j=1}^5 \lambda_j = 1. \quad (5)$$

In this expression,  $R_i$  is the overall risk assessment for a software component,  $\alpha$  is the coefficient determining the contribution of the machine-based prediction  $p_i$  to the overall risk assessment,  $\lambda_j$  are the weighting coefficients for individual groups of features, and  $\hat{x}_{ij}$  is the normalized generalized value of the  $j$ -th group of characteristics for the component  $c_i$ .

The condition  $\sum_{j=1}^5 \lambda_j = 1$  ensures the normalization of the weight scheme and allows  $\lambda_j$  to be interpreted as the relative contribution of the corresponding group of factors to risk formation. The presence of two components in the risk formula is of fundamental importance. The first component  $\alpha p_i$  reflects the contribution of the machine learning model, which accounts for nonlinear dependencies between features

and the target variable. The second component,  $(1-\alpha)\sum_{j=1}^5 \lambda_j \hat{x}_{ij}$ , ensures a structured consideration of group risk factors and enhances the interpretability of the method. As a result, the risk assessment combines both the predictive potential of the model and the analytical transparency of the weighted representation of factors.

To compare the components, we will use the values  $R \in [0,1]$ . However, this approach effectively reduces the method to a standard binary classification followed by ranking; therefore, a more complex risk index structure was introduced, which takes into account not only the probabilistic forecast but also generalized group characteristics. To this end, after filling in the gaps and standardizing the data, all features were divided into three groups: structural, process-related, and contextual. A normalized group index was formed for each group. If  $S_i$  is the structural index,  $P_i$  is the process index, and  $G_i$  is the contextual index, then the integrated risk indicator in its improved form was defined as

$$R_i^{v1} = \alpha p_i + (1-\alpha)(w_1 S_i + w_2 P_i + w_3 G_i), \quad (6)$$

where  $\alpha \in [0,1]$  specifies the contribution of the predicted probability of defects, and the coefficients  $w_1$ ,  $w_2$ ,  $w_3$  determine the relative importance of the structure, process, and context indices, subject to the condition:

$$w_1 + w_2 + w_3 = 1. \quad (7)$$

A situation is possible in which the risk model relies not only on individual raw metrics but also on interaction features. In this case, the component's risk is formed by combining local risk, which reflects the internal properties of the component itself, and context risk, which characterizes the conditions of its development and the operational context. As a result, the final risk index takes the form:

$$R_i^{v2} = \alpha p_i^{stack} + (1-\alpha)(\gamma R_i^{(local)} + (1-\gamma) R_i^{(context)}), \quad (8)$$

where  $p_i^{stack}$  is the defect probability obtained from the model,  $R_i^{(local)}$  is the local structure risk,  $R_i^{(context)}$  is the context risk, and  $\gamma \in [0,1]$  defines the ratio between the local and context components. This scheme best addresses the task of risk-oriented prioritization of automated testing, as it allows not only classifying components as defective or non-defective but also forming an ordered risk scale for planning tests.

Test prioritization is performed by sorting components in descending order of risk:

$$\pi = \text{argsort}(R_i, \downarrow), \quad (9)$$

where  $\pi$  is the ordered sequence of components.

For the Top-k approach [13], the following is used:

$$\text{Class}(c_i) = \begin{cases} \text{low}, & R_i < \tau_1, \\ \text{medium}, & \tau_1 \leq R_i < \tau_2, \\ \text{high}, & R_i \geq \tau_2. \end{cases} \quad (10)$$

where  $\tau$  and  $\tau$  are threshold values that divide a continuous risk scale into three interpretable levels: low, medium, and high. This representation is important from an applied perspective, as it allows not only for ranking components but also for grouping them by criticality. Components in the high class are of the greatest interest for early testing prioritization; components in the medium class can be considered second-priority items for verification; and components in the low class have the lowest expected risk of defects.

To determine the optimal threshold, the following quality function is used:

$$F1(\tau) = \frac{2 \cdot \text{Precision}(\tau) \cdot \text{Recall}(\tau)}{\text{Precision}(\tau) + \text{Recall}(\tau)}, \quad (11)$$

The optimal threshold is determined by:

$$\tau^* = \max_{\tau} F1(\tau). \quad (12)$$

It is proposed to evaluate the effectiveness of the method using the following metrics:

a) area under the ROC curve (ROC-AUC):

$$AUC = \int_0^1 TPR(FPR) dFPR; \quad (13)$$

b) Precision-Recall curve:

$$AP = \int_0^1 \text{Precision}(\text{Recall}) d\text{Recall}; \quad (14)$$

c) F1-score:

$$F1 = \frac{2TP}{2TP + FP + FN}; \quad (15)$$

d) Balanced Accuracy:

$$BA = \frac{TPR + TNR}{2}. \quad (16)$$

We also introduce feature importance as a term in the function:

$$\text{Importance}_j = \frac{1}{K} \sum_{k=1}^K I_{jk}, \quad (17)$$

where  $I_{jk}$  is the importance of the  $j$ -th feature in the  $k$ -th model.

## 5. Justification for the selection, preparation, and integration of datasets

---

To implement the method, it was decided to combine classic open-source software defect datasets with more modern process and project characteristics. The foundation consisted of regenerated PROMISE and BPD datasets in Excel format, as well as an additional dataset, csv\_result-JDT.xls. This data provided a reliable foundation at the software component level, as it contained structural code metrics and a target feature indicating the presence of defects. Specifically, the data included metrics such as cbo, dcc, exportcoupling, importcoupling, nom, wmc, and the defect label. However, structural metrics alone proved insufficient for the method to work effectively. Therefore, to expand the feature space, additional open tables from the SQuaD dataset [14] were incorporated, namely process\_metrics.csv, github\_metrics.csv, and release\_data.csv. The selected tables allowed us to add such important parameters to the structural level as change intensity, number of commits, number of fixes, author activity, number of files with defects, overall repository activity, and release characteristics. The choice of this specific combination of data sources was driven by the desire to build not just a defect classifier, but to develop a method that integrates several types of signals. PROMISE/BPD and JDT provided a detailed description of components, while the SQuaD tables provided a measure of processes and contexts.

Data preparation was a separate, essential stage of implementation. First, all sheets of the Excel file from PROMISE/BPD and the JDT file were read. Column names were normalized to a single format to avoid issues with case, spaces, delimiters, and inconsistent variable names. Next, the project name and release were automatically extracted from the sheet names, which allowed the creation of the project and release fields. At this stage, it was discovered that certain sheets and rows did not contain actual data about components but served a utility or reference function. That is why rows without a class name, sheets of the Main type, as well as other parts unsuitable for modeling were removed from the base dataset. This cleaning allowed us to transition from a mixed source to a component-oriented dataset. After the initial merge, an intermediate dataset named base\_df was formed, in which each row corresponded to a separate software component. Its key fields were

the project name, component name, structural metrics, and defect label.

Next, integration with the SQuaD tables was performed. This stage proved to be the most challenging. It became clear that a direct merge based on the project + release fields did not work, since the project and release identifiers were not consistent across different sources. For example, in the base dataset, the project was labeled as "ant", whereas in the SQuaD tables it appeared as "apache#ant". Similar inconsistencies were observed for "camel" and other projects. To address this issue, a project name normalization procedure was implemented, which mapped designations such as "apache#ant" to the unified value "ant". Another significant problem turned out to be poor-quality or incomplete information about releases in the base set. That is why, instead of merging at the project + release level, a decision was made to perform integration at the project level. As a result, process and project characteristics were attached to all components of the corresponding project as contextual information. Formally, this meant that the same set of process attributes was duplicated for components of a single project. For the purpose of risk assessment, this is acceptable, since the context risk can indeed be shared by components developed under the same conditions.

After integration, the final intermediate dataset modeling\_dataset\_merged\_v2.csv was obtained. A check of feature completeness showed that the basic structural characteristics were fully available; that is, for indicators such as cbo, dcc, exportcoupling, importcoupling, nom, and wmc, the proportion of non-empty values was 1.0. Additional characteristics of processes and projects had a completeness of approximately 0.3649, indicating a partial but meaningful enrichment of the data. Based on this dataset, two final datasets were formed. The first, final\_structural\_dataset.csv, contained only structural features and was used to build the baseline method. The second, final\_extended\_dataset.csv, contained characteristics of structures, processes, and projects and was used for the extended and improved versions of the method.

## 6. Experimental implementation

---

In implementing the proposed method for assessing the risk of defects in software components, the Python programming language was used as one of the most widely used platforms for data analytics and machine learning tasks. The choice of Python is due to the



availability of a well-developed ecosystem of libraries for data processing, model building, and result visualization, including NumPy, pandas, scikit-learn, and matplotlib, which allow for the effective implementation of all stages of data processing: from preprocessing to model quality assessment. The software prototype of the method was implemented in the Google Colab environment, which provides access to computational resources without the need for local infrastructure setup. Using this environment allows working with large datasets, performing complex computational operations for training machine learning models, and provides a convenient mechanism for interactive analysis of results.

Experimental validation was conducted using a fixed split of the data into training and test sets in a 75/25 ratio, with stratification by the target feature. This approach ensured that the ratio of defective to non-defective components was correctly preserved in each subsample. For all models, missing values were imputed with median values, and for scale-sensitive algorithms, feature standardization was also applied. Quality assessment was performed using ROC-AUC, Average Precision, F1-score, Balanced Accuracy, Precision-Recall, as well as top-k analysis, which is

particularly important for the task of test prioritization. A separate threshold selection was performed to maximize the F1-score or Balanced Accuracy. This allowed us not only to compare methods based on probabilities but also to find the optimal operating point for the practical application of the method.

In the first stage of the experiments, a basic version of the method was implemented, which used only structural metrics of the components. The main goal of this stage was to establish a baseline for comparison. The following served as input features: cbo, dcc, exportcoupling, importcoupling, nom, and wmc. Concurrently, an extended version was implemented, in which process and project characteristics were added to the structural features. In practice, this made it possible to verify whether the new information signal actually improves the quality of risk assessment. In the early stages, CatBoost [15] was used for this task, as it is well-suited for tabular data. Figure 1 shows the ROC curves for two approaches to assessing software component defects: the baseline method, which uses only structural metrics, and the extended method, which integrates additional characteristics. The x-axis plots the False Positive Rate (FPR), and the y-axis plots the True Positive Rate (TPR).

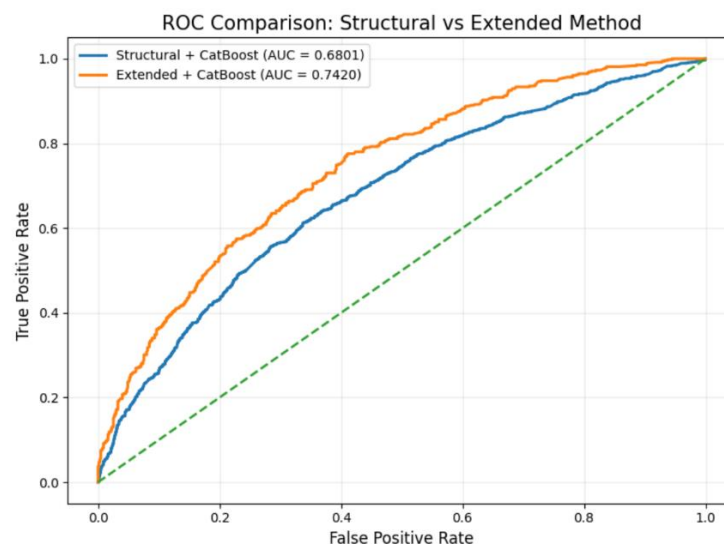


Fig. 1. ROC curves for the structural and extended methods

The extended method consistently outperforms the baseline approach across the entire range of values, as evidenced by the larger area under the ROC curve (AUC = 0.7420 vs. 0.6801). This indicates the model's superior ability to correctly distinguish between components with and without defects, confirming

the validity of integrating structural and process characteristics. Thus, incorporating additional features improves classification quality and ensures more effective defect detection in software components. Fig. 2 presents a comparison of the Precision-Recall curves for the baseline method (structural metrics only)

and the extended method. Recall is plotted on the x-axis, and precision on the y-axis. The extended method demonstrates higher precision values across most of the recall range, as evidenced by the larger area under the curve ( $AUC = 0.6718$  vs.  $0.5062$ ). This means that when detecting component defects, it produces fewer false positive rate (FPR) triggers.

The results obtained indicate a significant improvement in the method's performance under conditions of unbalanced data, which is critically important for test prioritization tasks.

After developing the extended method, the first refinement was implemented. In this version, risk was no longer equated solely with the probability of defects and began to be calculated as a combination of a probabilistic forecast with group indices. To this end, all features were divided into structural, process, and contextual groups, and after standardization, corresponding group indices were formed. Figure 3 shows the ROC curves for the baseline extended method and the improved risk assessment method. It can be seen that both methods demonstrate similar results; however, the improved method has a slightly higher area under the curve.

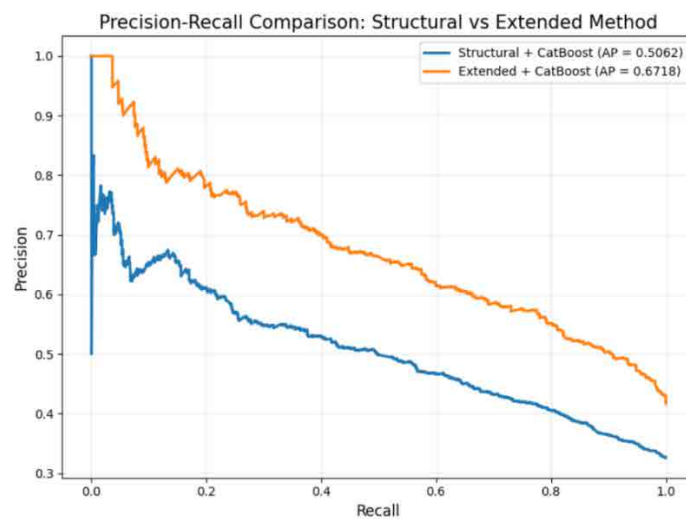


Fig. 2. Precision-Recall for the structural and advanced methods

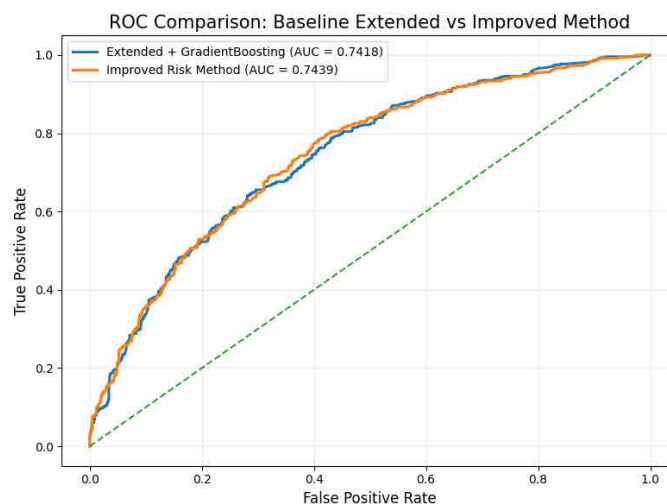


Fig. 3. ROC curves for the extended and improved methods

Figure 4 shows the dependence of the F1-score and Balanced Accuracy metrics on the classification threshold for the improved method. The x-axis plots the threshold values, while the y-axis plots the

corresponding quality metric values. Both metrics reach their maximum values at a threshold of approximately 0.33, as indicated by the corresponding vertical lines. As the threshold is further increased, a significant

decrease in the F1-score is observed, which is associated with an increase in the number of missed defects. Thus, the selected threshold value provides an optimal

balance between classification accuracy and completeness and is suitable for use in the task of test prioritization.

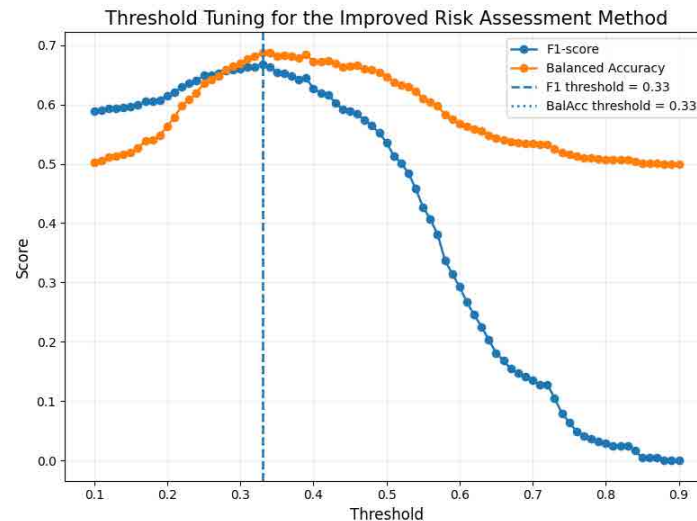


**Fig. 4.** The dependence of the F1-score and Balanced Accuracy metrics on the classification threshold

In practice, this option proved correct but yielded only a very slight improvement compared to the extended version. The difference in ROC-AUC was only a few thousandths, and Average Precision increased only marginally. Thus, aggregating the existing signal without introducing new informative features does little to improve the method. In other words, the first modification was useful more from the perspective of risk interpretation than from the perspective of improving predictive quality. The next step was to develop a second, significantly improved version of the method. It was here that feature interactions, a stack ensemble, and a two-level risk logic were introduced. It was found that the main limitation of the previous version was that additional process characteristics merely duplicated information at the project level but did not interact with the structure of the component itself. To address this shortcoming, new integrated features were developed. In particular, the following characteristics were introduced: complexity\_plus\_coupling, methods\_times\_coupling, coupling\_gap, coupling\_sum, export\_import\_ratio, complexity\_per\_method, wmc\_x\_log\_churn, cbo\_x\_log\_buggy\_files, max\_change\_set\_x\_cbo, age\_x\_complexity\_per\_method. These features allowed us to explicitly account for the interaction between

complexity, coupling, change history, and defect context. They became the primary source of the new information signal. The first layer consisted of Logistic Regression, Random Forest, and Gradient Boosting [16], while the second layer featured a meta-model based on logistic regression, which was trained to combine the probabilities of the base models. Thus, the value  $p_i^{stack}$  was obtained, which served as the basis for the final assessment. To calculate the overall risk, local and contextual risks were introduced. Local risk was formed based on structural and structural-process interactions of features, while contextual risk reflected the background of processes and projects. The final risk index was calculated as a combination of a stack-ensemble forecast and these two components.

The extended method is characterized by the dominance of classical structural metrics, specifically cbo, wmc, and nom, which determine the complexity and connectivity of software components. Dependency metrics (importcoupling, exportcoupling, dcc), which reflect the interaction between modules, also play an important role. Process and historical characteristics, such as max\_change\_set, latest\_release\_month, num\_bugs, and n\_auth, have a secondary influence, indicating the method's predominant focus on static code properties.

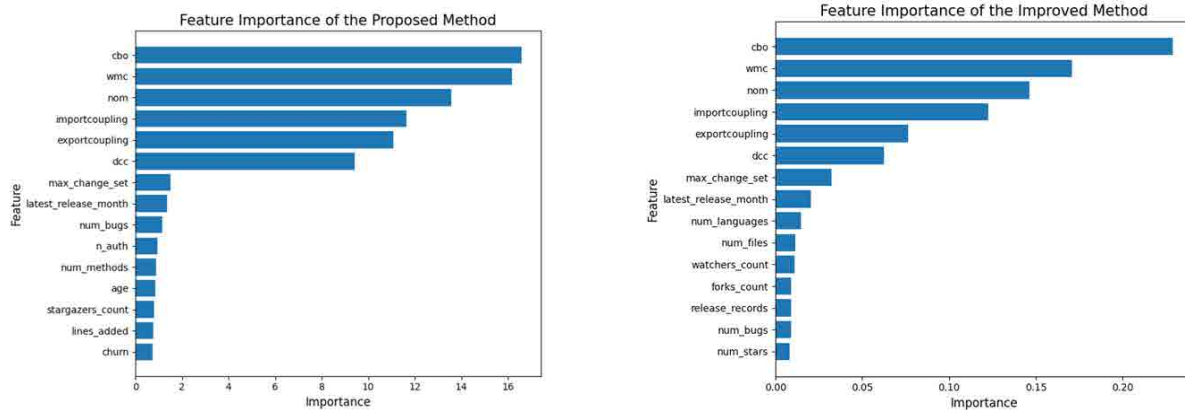


Fig. 5. The Importance of Features for the Extended and Improved (v1) Methods

For the first improved method, structural features remain dominant, but a more balanced distribution of importance is observed across different types of characteristics. Although cbo, wmc, and nom remain key, the contribution of process and repository metrics such as num\_languages, num\_files, watchers\_count, and forks\_count increases. This indicates an expansion of the model's information base and better consideration of the development context, which allows for improved accuracy in defect risk assessment compared to the baseline method. Figure 6 shows a comparison of ROC curves for three methods: the baseline extended method,

the ensemble approach (v1), and the improved method (v2). The x-axis plots the proportion of FPR hits, and the y-axis plots the proportion of TPR hits. The baseline extended method yields an AUC of 0.7347, which corresponds to acceptable classification quality but is inferior to more complex models. Using a stack ensemble improves the quality to  $AUC = 0.7542$ , demonstrating the effectiveness of the ensemble combination of models and a better ability to account for nonlinear dependencies in the data. The improved method v2 demonstrates a similar AUC value of 0.7542, effectively matching the level of the ensemble approach.

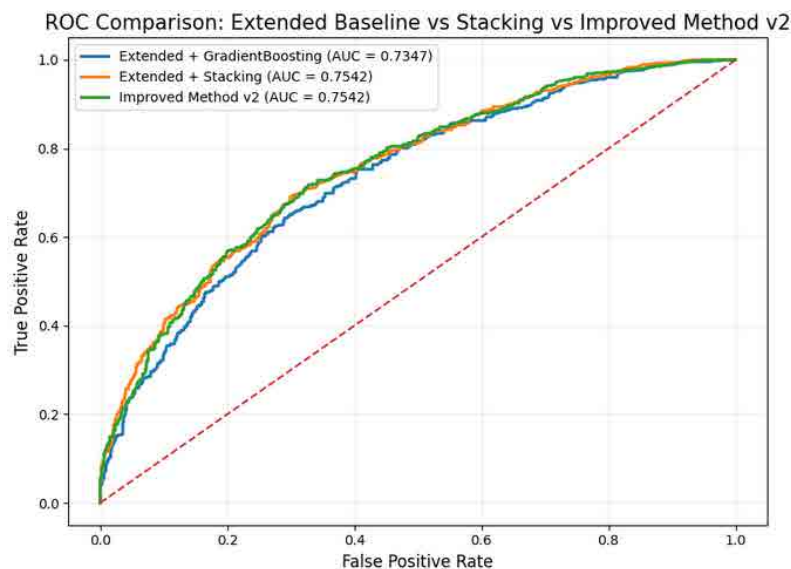


Fig. 6. Comparison of ROC Curves

Compared to previous results (where the AUC for the baseline extended approach was approximately 0.742 and for the first modification of the method was approximately 0.744), it is evident that further refinement has yielded a more significant improvement in

performance. In particular, the improvement relative to the initial structural approach ( $AUC \approx 0.68$ ) is substantial, whereas compared to the previous version of the method, the improvement is incremental. Figure 7 shows the Precision–Recall curves for the baseline

extended approach, the stack ensemble, and the improved method v2. The extended approach has the lowest performance (AP = 0.6646), while the other demonstrates

the best result (AP = 0.6986). The improved method v2 (AP = 0.6940) nearly reaches the ensemble's level and significantly exceeds the baseline.

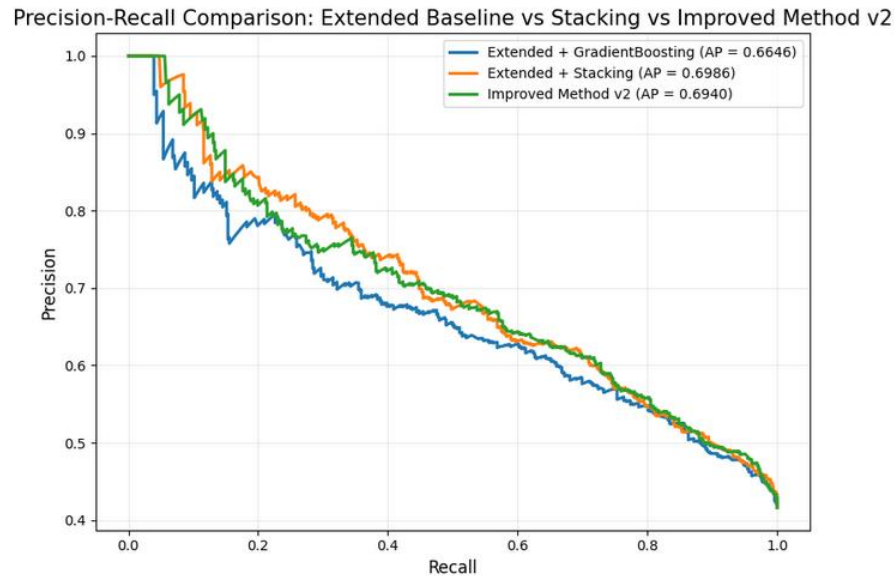


Fig. 7. Comparison of Precision–Recall Curves

Compared to previous versions of the method, there is a steady improvement in performance, particularly in terms of average recall values, which is crucial for defect detection tasks involving imbalanced data.

Figure 8 shows the dependence of the F1-score and Balanced Accuracy on the threshold value for the improved v2 method. The maximum F1-score is achieved at a threshold of approximately 0.33, while Balanced Accuracy is achieved at a higher value, approximately 0.41.

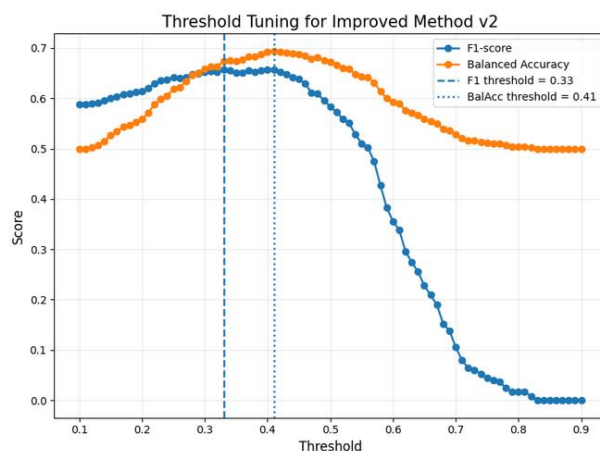


Fig. 8. Error matrix for the improved method

The results indicate a trade-off between classification completeness and accuracy: a lower threshold provides better defect detection (higher F1 score), while a higher threshold improves the balance between classes. Compared to the previous version of the method, the optimal thresholds remain similar, but the Balanced Accuracy curve is more stable, indicating improved model consistency. Figure 9 shows the confusion matrix for the improved v2 method.

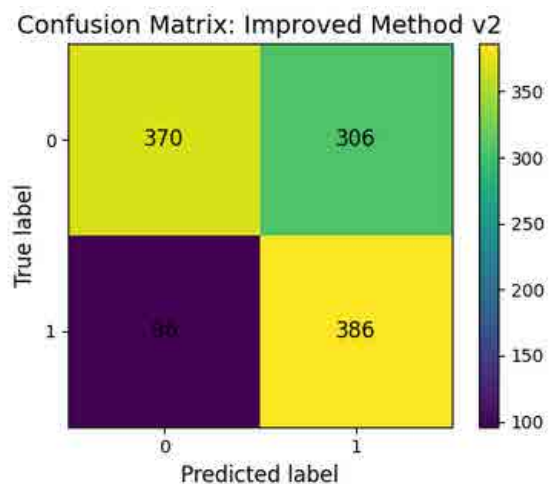


Fig. 9. Error matrix for the improved method



The method correctly classifies 386 defective components and 370 non-defective ones, while having 96 missed defects and 306 false positive rate (FPR) events.

There is a slight decrease in the number of missed defects, which is a positive result for risk detection tasks. At the same time, the number of FPR decisions has increased slightly, indicating a shift in the model toward a more "conservative" strategy – with an emphasis on maximizing defect detection even at the cost of additional

checks. The results obtained demonstrate the method's sufficiently high ability to identify defective components, which is critically important for test prioritization tasks, although the number of FPRs remains significant.

Figure 10 shows the results of Top-k ranking for the improved v2 method. A similar trend to the previous version is observed: as the proportion of selected components increases, the proportion of detected defects rises (to  $\approx 0.49$  at  $k = 0.30$ ), while accuracy gradually decreases.

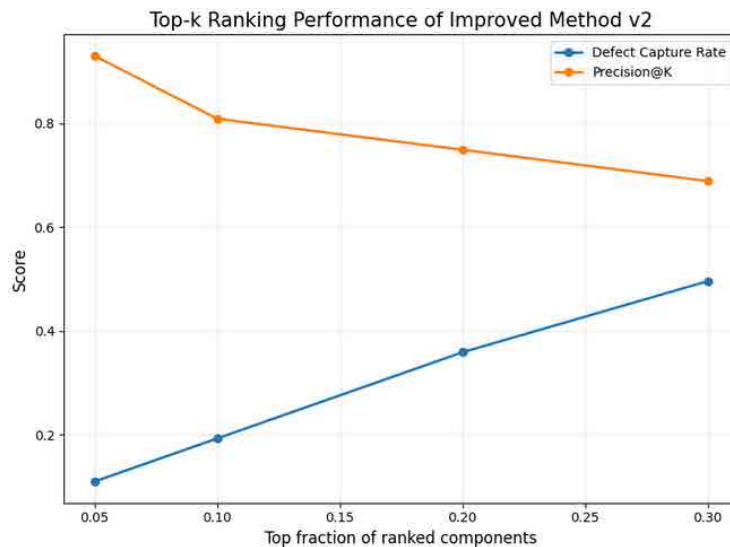


Fig. 10. Top-k ranking results

Compared to the previous version of the method, the improved version demonstrates higher accuracy at small values of  $k$  (particularly for  $k = 0.05 - 0.10$ ), indicating a better concentration of defective components at the top of the ranking. This confirms the increased effectiveness of the method specifically for testing prioritization tasks,

where it is critically important to correctly identify the highest-risk components at early stages.

Fig. 11 shows the distribution of the integral defect risk index for the improved v2 method. The index values are concentrated primarily in the range of 0.2–0.6, with a gradual decrease in frequency for high risk values.

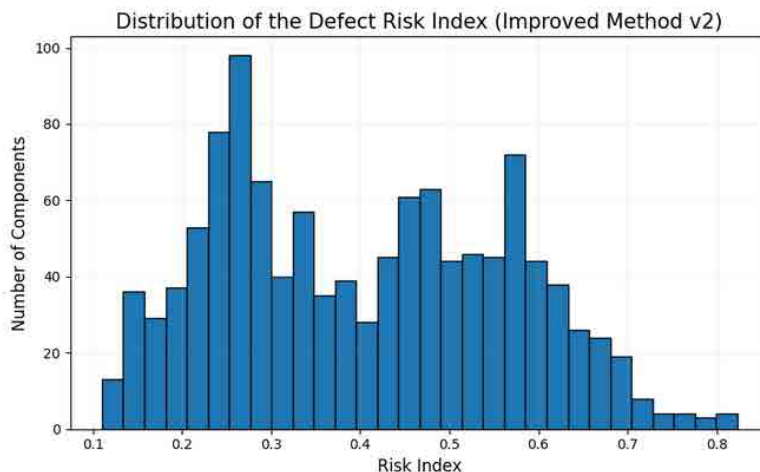


Fig. 11. Distribution of the integral defect risk index

The distribution of the integral defect risk index appears natural and is easily interpreted. Risk values are concentrated primarily in the range of 0.15–0.65, with fewer components in the very low and very high risk zones. The presence of values above 0.6–0.8 indicates the identification of a limited number of high-risk components, which is important for effective test prioritization. This indicates that the index is not degenerate and actually forms a gradation of components by defect level and can be used for planning automated tests.

Values above 0.6–0.8 indicate the identification of a limited number of high-risk components, which is important for effective test prioritization. This indicates

that the index is not degenerate and actually ranks components by defect severity, making it suitable for planning automated tests.

Figure 12 shows an assessment of the importance of features for the improved v2 method. The greatest contribution comes from the aggregated feature methods\_times\_coupling, which reflects the interaction between the complexity and coupling of components. Also significant are cbo\_x\_log\_buggy\_files, complexity\_per\_method, and age\_x\_complexity\_per\_method, which combine historical data with characteristics of structures and processes.

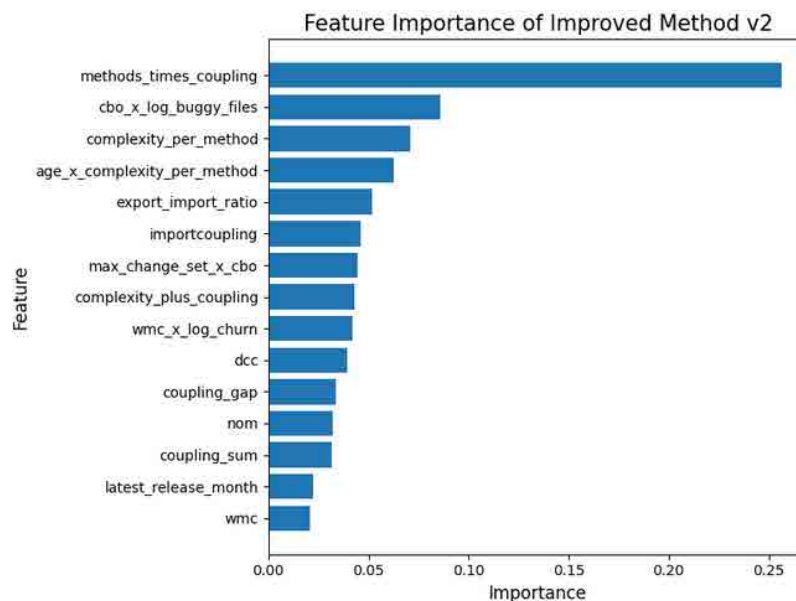


Fig. 12. Assessment of the importance of characteristics

The graph shows that the role of derived and composite attributes is growing, confirming the value of using them to improve the accuracy of defect risk assessment and better reflect complex relationships in the source code.

## 7. Discussion of the results

The results of the experimental study confirm the effectiveness of the developed method for assessing the risk of defects in software components within distributed computer systems. The study was conducted sequentially for several versions of the method, which differed in the composition of features and the method of calculating the integrated risk indicator, while defect probability was estimated using machine learning models.

A comparison of the basic (structural) and extended versions of the method showed that considering only the structural characteristics of the software code does not provide a sufficiently high-quality risk assessment. When process and context characteristics were added, the method's ability to distinguish between components with and without defects significantly improved. This indicates that defects in components are formed under the influence not only of the code's internal complexity and connectivity, but also of the change history, development intensity, and the overall state of the project.

In the first improved version of the method, an integrated risk index was implemented as a combination of the model's probabilistic forecast and generalized characteristics of various groups of features. The results showed that this approach does not lead to a significant

improvement in standard quality metrics compared to the extended version. At the same time, it allows moving from simple classification to a more meaningful representation of results in the form of a continuous risk indicator, which is fundamentally important for component ranking and testing prioritization tasks.

Further development of the method, implemented in version v2, is based on two key elements. First, an ensemble-based defect probability assessment scheme is used, which combines the results of several machine learning models. Second, derived features have been introduced that reflect the interaction between the characteristics of structures, processes, and contexts. This combination has improved the quality of the assessment and provided a more in-depth representation of the factors influencing the occurrence of defects in software components.

Analysis of ROC curves showed that the transition from the extended version of the method to the use of an ensemble model results in an increase in the ROC-AUC value. At the same time, the improved version of the method (v2) demonstrates the same level of generalization ability as the stack ensemble model, confirming the correctness of integrating the ensemble prediction into the method's structure. This means that the proposed risk index formulation scheme does not degrade prediction quality while preserving the advantages of the ensemble approach. A similar trend is observed for the Precision–Recall analysis. The use of an expanded set of features improves quality compared to the structural version of the method, while the application of the ensemble model allows for even better results. The improved version of the method, v2, demonstrates nearly identical values, which indicates the effectiveness of the chosen scheme for integrating the prediction and features.

Analysis of the error matrices shows that the improved versions of the method are geared toward more complete detection of component defects. The reduction in the number of missed defects is achieved at the cost of a certain increase in FPR triggers. This behavior is justified for automated test prioritization tasks, since missing a defective component is more critical than performing additional testing on a defect-free component. Thus, the method implements an evaluation strategy aimed at minimizing the risk of missing defects.

Particular attention should be paid to the component ranking results. The Top-k analysis showed that the largest proportion of defective components is

concentrated at the top of the generated list. The improved version of the method, v2, provides higher accuracy at small values of k, indicating a better ability to identify components with the highest risk. This directly addresses the practical task of determining a limited set of components for priority testing.

The distribution of the integral risk index confirms the method's ability to differentiate components by risk level. The improved v2 version is characterized by a more uniform and stretched distribution of values, allowing for a clearer identification of groups of components with low, medium, and high risk. This creates a foundation for the method's further use in risk classification and decision-making tasks.

Feature importance analysis shows that in the early stages of the method's development, the main contribution came from classical structural characteristics, such as component complexity and connectivity. In the improved v2 version, the greatest weight is given to derived features that combine different types of characteristics. This indicates that the machine learning model is beginning to utilize more complex patterns related to the interaction of factors, rather than just their individual values. It is worth noting that the improvement in evaluation quality is achieved not only by changing the algorithm but also by changing the structure of the feature space. Summarizing the results obtained, we can conclude that the most effective version of the method is the one that combines an ensemble assessment of defect occurrence probability using derived features and an integral risk index. This scheme simultaneously ensures high prediction quality, interpretability of results, and suitability for automated testing prioritization tasks. This demonstrates the advisability of using a comprehensive approach to risk assessment that takes into account both the properties of the program code and the context of its development and evolution.

---

## 8. Conclusions

This paper addresses a scientific and applied problem involving the development of a method for assessing the risk of defects in software components of distributed computer systems for the purpose of prioritizing automated testing. An analysis of current scientific publications has shown that existing approaches are predominantly focused either on the use of individual structural metrics of program code or on the application of machine learning models without proper consideration

---

of process characteristics and contexts, while the issue of integrating heterogeneous factors into a unified risk assessment system remains insufficiently addressed.

Within the scope of this study, a method was developed based on the formation of an integral defect risk indicator for software components using machine learning models, an extended feature space, and their generalization. The proposed scheme involves representing each component as a multidimensional feature vector that encompasses parameters of structures, processes, and contexts, as well as calculating defect probabilities using machine learning models and forming a risk index suitable for ranking components in automated testing tasks.

The use of an integrated risk index allows moving from binary classification to a ranking task, which is more relevant for automated testing practice. The Top-k analysis showed that a significant proportion of components with defects is concentrated at the top of the ranking, which enables effective prioritization of test checks when resources are limited. Feature importance analysis confirmed the feasibility of using derived characteristics that account for the interaction between code complexity, change intensity, and defect context.

The developed method can be used as an analytical support tool for the automated testing process in

distributed computer systems. Prospects for further research include improving risk assessment models, reducing the number of FPR triggers, adapting threshold values to the specifics of particular software projects, and integrating the method into real-world software lifecycle management systems.

---

### Conflict of Interest

The authors of this article declare that they have no conflicts of interest regarding this study, including financial, personal, authorship, or other conflicts that could influence the study and its results presented in this article.

---

### Funding

The study was conducted without financial support.

---

### Data Availability

The manuscript has no associated data.

---

### Use of Artificial Intelligence

The authors confirm that they did not use artificial intelligence technologies in the creation of this work.

## References

1. Dos Santos, R. A. (2024), "Just-In-Time Software Defect Prediction using a deep learning-based model", *New Trends in Computer Sciences*, 2(2), 91–100. DOI: <https://doi.org/10.3846/ntcs.2024.22274>
  2. Jiang, Y., Shen, B., Gu, X., (2025), "Just-in-time software defect prediction via bi-modal change representation learning", *Journal of Systems and Software*. Volume 219. DOI: <https://doi.org/10.1016/j.jss.2024.112253>
  3. Ali, M., Mazhar, T., Al-Rasheed, A., Shahzad, T., Yasin Ghadi, Y., Amir Khan, M., (2024), "Enhancing software defect prediction: a framework with improved feature selection and ensemble machine learning", *PeerJ Computer Science* 10:e1860. DOI: <https://doi.org/10.7717/peerj-cs.1860>
  4. Albattah, W., Alzahrani, M., (2024), "Software Defect Prediction Based on Machine Learning and Deep Learning Techniques: An Empirical Approach", *AI*, 5, 1743–1758. DOI: <https://doi.org/10.3390/ai5040086>
  5. Siswantoro, M., Yuhana, U.L., (2023), "Software Defect Prediction Based on Optimized Machine Learning Models: A Comparative Study". *Teknika*. 12, 2, 166–172. DOI: <https://doi.org/10.34148/teknika.v12i2.634>
  6. Guo, Z., Sun, Y., Fu, B., Cheng, X., (2022), "Research on software defect prediction method based on machine learning oriented to large-scale complex information system", *Proc. SPIE 12161, 4th International Conference on Informatics Engineering & Information Science (ICIEIS2021)*. DOI: <https://doi.org/10.1117/12.2627233>
  7. Cabral, G.G., Minku, L.L., Oliveira, A.L., Pessoa, D.A., Tabassum, S., (2023), "An investigation of online and offline learning models for online Just-in-Time Software Defect Prediction", *Journal of Systems and Software* 28, 121. DOI: <https://doi.org/10.1007/s10664-023-10335-6>
  8. Kaur, G., Pruthi, J., Gandhi, P., (2023), "Machine learning based Software Fault Prediction models," *Karbala International Journal of Modern Science*: Vol. 9. DOI: <https://doi.org/10.33640/2405-609X.3297>
  9. Bayramova, T., (2023), "Software defect prediction using the machine learning methods", *Problems of Information Technology*, 14(2). DOI: <http://doi.org/10.25045/jpit.v14.i2.03>
-

10. Shittu, A.S., Abdulsalami, B. (2023), "Software defect prediction using machine learning approach : a contemporary review", *TechRxiv*, pp. 1–8. DOI: <https://doi.org/10.36227/techrxiv.23648907.v1>
11. Hasan Kabir, S.M., Tanim Rahman, Md., Aunik Hasan Mridul, (2025), "Software Defect Prediction Using Traditional Machine Learning and Ensemble Learning Algorithms", *Smart Wearable Technology*, 1, A9. DOI: <https://doi.org/10.47852/bonviewSWT52025645>
12. Mahbub, P., Shuvo, O., Rahman, M.M.6 (2023) "Defectors: A Large, Diverse Python Dataset for Defect Prediction", *2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*, Melbourne, Australia, pp. 393–397. DOI: <https://doi.org/10.1109/MSR59073.2023.00085>
13. Robredo, M., Esposito, M., Taibi, D., Peñaloza, R., Lenarduzzi, V., (2025), "SQuaD: The Software Quality Dataset – Dataset [Data set]". *Zenodo*. DOI: <https://doi.org/10.5281/zenodo.17566691>
14. Chen, Z., Wang, K., Yang, C., Cao, B., Fan, J., (2025), "Learning Top-k Classification with Label Ranking. In: Taniguchi, T., et al. Neural Information Processing. ICONIP 2025", *Communications in Computer and Information Science*, vol 2755. Springer, Singapore. DOI: [https://doi.org/10.1007/978-981-95-4094-5\\_9](https://doi.org/10.1007/978-981-95-4094-5_9)
15. Hamid, M., Hajje, F., Alluhaidan, A.S., (2025), "Fine tuned CatBoost machine learning approach for early detection of cardiovascular disease through predictive modeling", *Scientific Reports* 15, 31199. DOI: <https://doi.org/10.1038/s41598-025-13790-x>
16. F. Setievi, F., Natalia, J., Tjhang, T.R., Edbert, I. S., Suhartono, D., (2022), "A Comparative Study of Supervised Machine Learning Algorithms for Fake Review Detection," *2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, 2022, pp. 306–312. DOI: <https://doi.org/10.1109/ISRITI56927.2022.10052860>

*Received (Надійшла) 28.02.2026*

*Accepted for publication (Прийнята до друку) 18.04.2026*

*Publication date (Дата публікації) 29.05.2026*

#### *Відомості про авторів / About the Authors*

**Дяченко Дмитро Олександрович** – Харківський національний університет радіоелектроніки, аспірант кафедри електронних обчислювальних машин, Харків, Україна;

**Dmytro Diachenko** – Kharkiv National University of Radio Electronics Kharkiv, Postgraduate Student of the Department of Electronic Computers, Ukraine;

e-mail: [dmytro.diachenko2@nure.ua](mailto:dmytro.diachenko2@nure.ua)

ORCID ID: <http://orcid.org/0009-0006-5751-3511>

**Міхаль Олег Пилипович** – доктор технічних наук, доцент, Харківський національний університет радіоелектроніки, професор кафедри електронних обчислювальних машин, Харків, Україна;

**Oleg Mikhal** – Doctor of Technical Sciences, Associate Professor, Kharkiv National University of Radio Electronics, Professor of the Department of Electronic Computers, Kharkiv, Ukraine;

e-mail: [oleg.mikhal@nure.ua](mailto:oleg.mikhal@nure.ua)

ORCID ID: <http://orcid.org/0000-0002-5977-3177>

### МЕТОД ОЦІНЮВАННЯ РИЗИКУ ДЕФЕКТІВ ПРОГРАМНИХ КОМПОНЕНТІВ РОЗПОДІЛЕНИХ КОМП'ЮТЕРНИХ СИСТЕМ ДЛЯ ПРІОРИТИЗАЦІЇ АВТОМАТИЗОВАНИХ ВИПРОБУВАНЬ

**Предметом дослідження** є методи, моделі й засоби оцінювання ризику дефектів програмних компонентів розподілених комп'ютерних систем для підтримки процесу пріоритизації автоматизованих випробувань.

**Мета** – розроблення методу оцінювання ризику дефектів програмних компонентів розподілених комп'ютерних систем для пріоритизації автоматизованих випробувань на основі машинного навчання з огляду на структурні характеристики компонентів, інтенсивність змін і результати тестування. Відповідно до окресленої мети необхідно виконати такі **завдання**: проаналізувати сучасні підходи до прогнозування дефектів програмного забезпечення й пріоритизації автоматизованих випробувань у розподілених комп'ютерних системах; визначити інформативні



характеристики програмних компонентів, що відтворюють їх структурні властивості, інтенсивність змін і зв'язки між програмними компонентами; розробити метод оцінювання ризику дефектів програмних компонентів на основі моделі машинного навчання з формуванням інтегрального показника ризику; запропонувати підхід до пріоритизації автоматизованих випробувань на основі отриманого інтегрального показника ризику; впровадити запропонований метод і експериментально оцінити його ефективність за показниками класифікації та ранжування.

**Методи:** системний аналіз, машинне навчання, аналіз структурних метрик програмного коду, класифікація, відбір інформативних ознак, формування інтегрального показника ризику й ризик-орієнтованого ранжування. Експериментальну перевірку виконано з використанням методів обчислювального моделювання. **Результати дослідження.** Розроблено метод оцінювання ризику дефектів програмних компонентів розподілених комп'ютерних систем, що забезпечує інтеграцію характеристик структур, процесів і контекстів, а також обчислення інтегрального показника ризику й ранжування компонентів для пріоритизації автоматизованих випробувань. Експериментальна перевірка підтвердила працездатність запропонованого методу та його спроможність ефективно розділяти програмні компоненти з дефектами та без них, а також виділяти компоненти ризику для першочергового тестового контролю. **Висновки.** Досягнуті результати підтвердили доцільність упровадження методів машинного навчання та інтегрованого простору ознак для оцінювання ризику дефектів програмних компонентів розподілених комп'ютерних систем. Запропонований метод може бути використаний як основа для підвищення обґрунтованості рішень щодо пріоритизації автоматизованих випробувань, вчасного виявлення найбільших компонентів ризику й подальшого розвитку інтелектуальних засобів контролю якості програмного забезпечення.

**Ключові слова:** програмні компоненти; розподілені комп'ютерні системи; автоматизовані випробування; машинне навчання; прогнозування дефектів; інтегральний показник ризику; пріоритизація; тестування; Python; Google Colab.

#### Бібліографічні описи / Bibliographic descriptions

Дяченко Д. О., Міхаль О. П. Метод оцінювання ризику дефектів програмних компонентів розподілених комп'ютерних систем для пріоритизації автоматизованих випробувань. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 109–127. DOI: <https://doi.org/10.30837/0135-1710.2026.189.109>

Diachenko, D., Mikhal, O. (2026), "Method for assessing the risk of defects in software components of distributed computer systems for automated test prioritization", *Management Information System and Devices*, No. 2 (189), P. 109–127. DOI: <https://doi.org/10.30837/0135-1710.2026.189.109>

---

Віпшовський Ю. А., Грицюк Ю. І.

## ЛОКАЛІЗАЦІЯ ДЕФЕКТНИХ ЗОН НА ПОВЕРХНЯХ МЕТАЛІВ МЕТОДОМ АНАЛІЗУ КУМУЛЯТИВНИХ ПРОЄКЦІЙ ІНТЕНСИВНОСТІ ПІКСЕЛІВ

**Предмет дослідження** – процес просторової локалізації дефектних зон на зображеннях поверхонь. **Метою** є розроблення обчислювально ефективного методу просторової локалізації дефектних зон на металевих поверхнях, здатного функціонувати в режимі реального часу на стандартному процесорному обладнанні без долучення ресурсомістких моделей машинного навчання. **Завдання:** проаналізувати наявні підходи до дефектоскопії; розробити метод і математичну модель локалізації на основі кумулятивних проєкцій; створити оптимізований алгоритм оброблення матриці пікселів; розробити експериментальне програмне забезпечення; емпірично перевірити запропонований підхід на стандартизованих наборах даних. **Методи дослідження:** цифрове оброблення зображень, математичний апарат аналізу одновимірних кумулятивних проєкцій інтенсивності пікселів, методи теорії множин (для визначення перетину аномальних зон), низькорівневої алгоритмічної оптимізації (прямий доступ до пам'яті, стекова алокація) для прискорення обчислень. **Результати.** Розроблено метод просторової локалізації дефектних зон, що здійснює редукцію розмірності зображення способом його перетворення в систему незалежних векторів. Запропоновано алгоритм із дворівневою ієрархічною верифікацією для усунення проблеми множинних хибних спрацьовувань (фантомних рамок). Створено експериментальне програмне забезпечення, в якому завдяки глибокій оптимізації досягнуто швидкодії оброблення великоформатних зображень (2048 на 1000 пікселів) на рівні 40–50 мс із використанням виключно центрального процесора. Метод успішно формує точні обмежувальні рамки навколо мікродефектів на фонах зі складною металевою текстурою, відсікаючи до 95% однорідного фону. **Висновки.** Запропонований метод є обчислювально ефективною альтернативою сучасним моделям глибокого навчання для систем промислової дефектоскопії. Відмова від двовимірної згортки на користь аналізу одновимірних проєкцій значно зменшує ресурсомісткість. Алгоритм здатний працювати в режимі жорсткого реального часу, забезпечуючи надійну просторову локалізацію без потреби в попередньому навчанні на великих масивах розмічених даних.

**Ключові слова:** кумулятивні проєкції зображень; локалізація дефектних зон поверхонь металів; комп'ютерний зір; аналіз зображень.

### 1. Вступ

Проблема точної просторової локалізації дефектних зон на поверхнях конструкційних матеріалів є однією з ключових у сучасному матеріалознавстві та неруйнівному контролі. Від здатності системи не тільки встановити наявність браку, але й визначити його точні координати залежить якість продукції, можливість автоматичного відбраковування й загальне зниження собівартості її виготовлення. У світовій науковій літературі цій проблемі присвячено численні дослідження з використанням методів комп'ютерного зору й цифрового оброблення зображень. Активно розвиваються як класичні двовимірні підходи, так і високопродуктивні нейромережеві моделі.

Чимало сучасних систем автоматизованого контролю якості поверхонь металів залишаються обчислювально складними, потребують великих обсягів попередньо розмічених даних і спеціалізованих

апаратних пришвидшувачів. Особливо виразно ці недоліки виявляються в разі тривалого оброблення великоформатних зображень у режимі реального часу, де під час стискання зображень для нейромереж може призводити до втрати дрібних просторових ознак мікродефектів.

У цих умовах актуальною є потреба в розробленні підходів до локалізації дефектних зон на поверхнях металів, які поєднують просторову точність із низькою обчислювальною ресурсомісткістю. Перспективним напрямом є використання кумулятивних проєкцій інтенсивності окремих рядків і стовпців зображення, що дає змогу зменшити розмірність оброблюваних даних, перейшовши від аналізу двовимірної матриці пікселів до обчислення одновимірних векторів. Це спрощує структуру системи контролю наявності дефектних зон і підвищує ефективність локалізації широкого класу поверхневих дефектів металів без потреби в тривалому машинному навчанні моделі.

## **2. Аналіз сучасних наукових публікацій і визначення проблеми дослідження**

Проблема автоматизованого виявлення дефектних зон і їх локалізації на металевих поверхнях переживає зміну парадигми. Більшість актуальних досліджень зосереджують увагу на двох напрямках: застосуванні глибоких згорткових нейромереж і впровадженні архітектур на базі візуальних трансформерів [1, 2].

У роботах [3–5] дослідники успішно адаптують сучасні версії алгоритмів сімейства YOLO (YOLOv7, YOLOv8) для локалізації дефектних зон на сталевих листах у реальному часі. Автори праць [6, 7] пропонують моделі семантичної сегментації та трансформери, які дають змогу виділяти дефекти з піксельною точністю, беручи до уваги глобальний контекст зображення. Незважаючи на значну точність локалізації дефектних зон на поверхнях металів, ці підходи мають критичні для промисловості недоліки. Вказані моделі визначаються недостатньою інтерпретованістю алгоритмів прийняття рішень і високою залежністю від обсягів попередньо розмічених навчальних вибірок. Практична імплементація таких систем у виробниче середовище потребує значних апаратних ресурсів, зокрема спеціалізованих графічних пришвидшувачів (GPU). Це зумовлено високою обчислювальною складністю операцій двовимірної згортки в процесі оброблення зображень високої роздільної здатності, через що використання центральних процесорів (CPU) стає неефективним для роботи в режимі реального часу.

Для розв'язання проблеми браку розмічених даних у роботах [8–10] запропоновано методи неконтрольованого навчання, зокрема на базі нормалізаційних потоків і архітектур типу *PatchCore*. Вони навчаються тільки на бездефектних зразках металу. Проте їх недолік полягає в тому, що результатом роботи зазвичай є розмита теплова карта аномалій, а не чітка обмежувальна рамка, що ускладнює подальше автоматизоване оброблення. Крім цього, етап інференсу в таких моделях залишається ресурсомістким.

Альтернативний напрям пов'язаний з удосконаленими класичними методами цифрового оброблення зображень. У дослідженнях [11–13] автори комбінують методи математичної морфології, матриці сірих перерізів і адаптивну сегментацію. Основним недоліком таких підходів є значна обчислювальна складність розрахунку двовимірних

текстурних ознак для кожного вікна сканування. За високих витрат процесорного часу ці методи все одно залишаються вразливими до градієнтного освітлення й мікротекстури металів, що робить їх застосування на високошвидкісних конвеєрах технічно й економічно невиправданим порівняно з більш швидкими алгоритмами редукції розмірності.

На окрему увагу заслуговують дослідження [14], в яких упроваджуються методи інтегральних проєкцій і аналізу одновимірних профілів для промислового бачення в суміжних галузях. Наприклад, ці алгоритми успішно використовуються для локалізації дефектних зон меж дефектів на кераміці або виявлення аномалій зварних швів на рентгенограмах. Однак у контексті металооброблення такі алгоритми здебільшого застосовуються для простих завдань вирівнювання країв стрічки або вимірювання геометрії поверхні досліджуваного металу. Наразі вони майже не адаптовані для повноцінної двовимірної локалізації поверхневих мікрodefektів на фоні складної текстури металів способом перетину проєкційних характеристик.

Сучасні рішення в галузі комп'ютерного зору для металооброблення стикаються з проблемою нестійкості до світлових артефактів. У цьому разі наявні методи поділяються на низькосегментні алгоритми фільтрації зображень, що мають обмежену здатність до узагальнення, і високорівневі моделі глибокого навчання. Останні, попри їх архітектурну гнучкість, є критично залежними від обчислювальних потужностей CPU/GPU та обсягів анованих вибірок, що створює бар'єр для їх оперативної адаптації під конкретні виробничі завдання.

Поза увагою дослідників залишається клас алгоритмів, здатних забезпечити детерміновану просторову локалізацію без етапу тривалого машинного навчання із використанням мінімуму обчислювальних ресурсів. Зокрема недостатньо вивченим залишається підхід до редукції розмірності задачі – переходу від складного двовимірного аналізу матриці пікселів до обчислення системи незалежних одновимірних маргінальних проєкцій інтенсивності пікселів (кумулятивні відображення). Відсутні комплексні дослідження щодо використання математичного перетину дефектних рядків і стовпців для швидкої побудови обмежувальних рамок на однорідних металевих текстурах.

Розроблення й дослідження методу аналізу кумулятивних проєкцій інтенсивностей пікселів

рядків і стовпців зображення, який нівелює обчислювальну надмірність нейромереж і перевершує за надійністю класичні двовимірні фільтри, є невиконаним науково-практичним завданням. Розв'язанню цього питання й присвячено цю статтю.

### 3. Мета й завдання дослідження

**Об'єктом дослідження** є просторова локалізація дефектних зон на поверхнях металів за допомогою аналізу кумулятивних проєкцій інтенсивності пікселів, а **предметом** – алгоритмічні та програмні засоби локалізації дефектних зон на поверхнях металів, які базуються на аналізі одновимірних кумулятивних проєкцій інтенсивності пікселів, що дадуть змогу ефективно визначати просторові координати зон з аномальними інтенсивностями на поверхні.

**Мета дослідження** – розробити обчислювально ефективний метод аналізу кумулятивних проєкцій інтенсивності пікселів рядків і стовпців зображення для просторової локалізації дефектних зон на зображеннях поверхонь металів, що дасть змогу формувати точні координатні рамки пошкоджень без потреби в попередньому машинному навчанні моделі.

Для досягнення окресленої мети визначено такі завдання:

- проаналізувати літературні джерела щодо методів комп'ютерного зору для просторової локалізації дефектних зон поверхонь матеріалів, що дасть змогу виявити недоліки наявних підходів і обґрунтувати вибір методу проєкцій;
- запропонувати метод аналізу кумулятивних проєкцій інтенсивності пікселів рядків і стовпців зображення для просторової локалізації дефектних зон на зображеннях поверхонь металів і математичну модель, що використовує розроблений метод;
- розробити алгоритм оброблення зображень та експериментальне ПЗ для дослідження запропонованого методу із різними вхідними параметрами;
- підготувати набори даних із розміченими еталонними ділянками дефектів металевих поверхонь, які уможливають об'єктивне проведення тестів із досліджуванним методом аналізу кумулятивних проєкцій інтенсивності пікселів;
- експериментально апробувати розроблений метод на сформованих масивах даних для визначення його часової складності та обчислювальної потужності, що дасть змогу встановити межі життєздатності

алгоритму та підтвердити можливість його інтеграції в автоматизовані системи технічного зору, що працюють у режимі реального часу.

### 4. Матеріали й методи дослідження

Для подальшої емпіричної перевірки методу просторової локалізації дефектних зон, математичний апарат якого описано в наступному розділі, було обрано два загальновизнані набори зображень металевих поверхонь [15, 16]: NEU-DET і GC10-DET. Вибір саме цих наборів зображень зумовлений наявністю еталонної розмітки у вигляді просторових координат дефектних зон, що є критично важливим для розрахунку математичної точності локалізації дефектних зон [17].

Набір даних NEU-DET (англ. *Northeastern University Surface Defect Database*) містить зображення дефектів гарячекатаних сталевих смуг. У базі наявні 1800 зображень у градаціях сірого розміром  $200 \times 200$  пікселів, що розподілені на шість класів дефектів: мікротріщини, вкраплення, плями, ямки, окалина й подряпини. Кожне зображення супроводжується файлом анотації у форматі XML, який містить координати обмежувальних рамок дефектів на площині.

Набір даних GC10-DET є більш сучасним і складним масивом зображень сталевих поверхонь, що має 10 типів дефектів, серед яких відколи, зварювальні шви та складки. Зображення в цьому наборі мають значно більшу роздільну здатність і складніше освітлення, що дає змогу перевірити стійкість алгоритму кумулятивних проєкцій до градієнтних шумів. Цей набір також містить відповідну просторову розмітку у вигляді обмежувальних рамок.

### 5. Результати дослідження

#### Математична модель аналізу зображень за допомогою кумулятивних проєкцій

Основою запропонованої моделі просторової локалізації дефектів є зменшення розмірності простору ознак за допомогою алгоритму кумулятивних проєкцій. Модель розглядає зображення поверхні металу як двовимірну матрицю пікселів розмірності  $W \times H$ , де  $W$  – ширина, а  $H$  – висота зображення. Для мінімізації обчислювальних витрат на першому етапі кольорове зображення перетворюється в напівтонове. Яскравість кожного пікселя  $I(x, y)$

обчислюється за базовою формулою усереднення кольорних каналів [18]:

$$I(x, y) = \frac{R(x, y) + G(x, y) + B(x, y)}{3}, \quad (1)$$

де  $R(x, y)$  – інтенсивність червоного кольору пікселя з координатами  $x, y$ ;  $G(x, y)$  – інтенсивність зеленого кольору пікселя з координатами  $x, y$ ;  $B(x, y)$  – інтенсивність синього кольору пікселя з координатами  $x, y$ .

Замість оброблення матриці пікселів як єдиного цілого, запропонована модель розглядає зображення як набір з  $W$  незалежних векторів-стовпців і  $H$  незалежних векторів-рядків. Для кожного такого вектора будується локальна гистограма розподілу інтенсивності пікселів. Математично локальну гистограму для заданого рядка  $x$  можна записати як

$$H_x(v) = \sum_{y=0}^{H-1} \delta(I(x, y), v), \quad (2)$$

де  $v[0, 255]$  – значення інтенсивності пікселя з координатами  $x, y$ ;  $\delta(\cdot)$  – функція Кронекера (становить 1 за умови збігу аргументів і 0 – в іншому

$$tolInt = [MainInt - DetTol \cdot detInt; MainInt + DetTol \cdot detInt], \quad (4)$$

де  $detTol$  – допустиме відхилення від еталонної інтенсивності пікселя, що не вважається дефектом;  $detInt$  – інтервал аналізу інтенсивностей пікселів для зображення, яким можна відсікати чорний або білий фон зображення поверхні металу.

#### Локалізація дефектних зон на поверхнях металів способом перетину кумулятивних проєкцій рядків і стовпців

Процес просторової локалізації дефектних зон зводиться до аналізу наявності пікселів у кожному рядку й стовпці, які виходять за межі інтервалу  $tolInt$ . Якщо в межах вектора (наприклад, стовпця  $y$ ) існують аномальні пікселі, цей стовпець маркується як дефектний і йому присвоюється логічний прапорець  $C_y = 1$ . Аналогічна процедура виконується для рядків із формуванням вектора прапорців  $R_x = 1$ .

Просторова локалізація дефекту (формування обмежувальної рамки – *Bounding Box*) обчислюється як декартів добуток множин індексів дефектних стовпців і рядків:

$$B = \{(x, y) | R_x = 1 \wedge C_y = 1\}. \quad (5)$$

Такий підхід дає змогу швидко локалізувати дефект, обмежуючи його прямокутною ділянкою

разі). Аналогічно формується гистограма  $H_y(v)$  для кожного стовпця  $y$ .

#### Адаптивне визначення фону й порогів аномалій

Для металевих поверхонь властива наявність домінантного фону, що відповідає бездефектній текстурі матеріалу. Модель автоматично визначає цей фон способом побудови глобальної гистограми всього зображення та знаходження її абсолютного максимуму:

$$MainInt = \arg \max_v \sum_{y=0}^{W-1} \sum_{x=0}^{H-1} \delta(I(x, y), v). \quad (3)$$

Значення  $MainInt$  слугує еталоном інтенсивності пікселів для поточного зображення, що робить модель стійкою до глобальних змін освітлення. Для дефектних зон на поверхнях металів, що можуть бути як темнішими (окалина, подряпини), так і світлішими (відблиски на вм'ятинах) за фон, запропонована модель аналізу зображення використовує двостороннє порогове обмеження із заданим параметром толерантності  $DetTol$ :

на площині зображення, що є стандартом для підготовки даних у комп'ютерному зорі.

#### Рекурсивна верифікація меж дефектних зон на поверхнях металів

Безпосередній декартів добуток множин індексів дефектних стовпців і рядків формує сітку можливих перетинів, які є первинними кандидатами на обмежувальні рамки (*Bounding Boxes*). Проте в разі наявності кількох просторово рознесених дефектів на одному зображенні базовий перетин проєкцій створює надлишкову кількість хибних рамок (так званих фантомних перетинів). Наприклад, перетин проєкцій від двох реальних дефектів, розташованих по діагоналі, згенерує чотири прямокутні ділянки, дві з яких будуть порожніми. Для усунення проблеми множинних хибних спрацьовувань і уточнення меж дефекту модель доповнено етапом локальної верифікації за дворівневою ієрархічною схемою.

1. Підтвердження наявності дефекту (відсіювання фантомів). Для кожної ділянки-кандидата  $Bcandidate$ , утвореної глобальним перетином, здійснюється швидка перевірка. Якщо всередині цієї ділянки виявлено хоча б один піксель, інтенсивність якого виходить за межі інтервалу  $tolInt$ , рамка вважається дійсною. Якщо таких пікселів немає – рамка



ідентифікується як фантомна й вилучається з подальшого аналізу.

2. Локальне звуження меж. Оскільки початкова рамка може бути ширшою за сам дефект (через вплив проєкцій від інших аномалій на тих самих осях), алгоритм виконує повторний розрахунок кумулятивних проєкцій, але вже винятково в межах підтвердженої рамки *Bcandidate*. Це дає змогу відкинути порожні краї та максимально точно виділити дефект.

Критерієм зупинки цього уточнювального аналізу є стабілізація координат дефектів: якщо після локального перерахунку проєкцій нові межі рамки збігаються з поточними (тобто розмір обмежувального прямокутника більше не зменшується), алгоритм фіксує ці координати як остаточні та завершує аналіз цієї ділянки. Такий підхід дає змогу швидко локалізувати дефекти, відсіюючи хибні координати й забезпечуючи максимально щільне прилягання обмежувальної рамки до реальних контурів аномалії на площині зображення.

### Особливості програмної реалізації методу побудови кумулятивних проєкцій і підвищення його обчислювальної ефективності

Реалізація розробленого методу побудови кумулятивних проєкцій у вигляді експериментального програмного забезпечення (ПЗ) вимагає подолання обмежень середовища виконання. Звичайний прохід по пікселях з використанням стандартних високорівневих функцій, як-от *GetPixel*, знижує швидкість оброблення зображення до неприйнятних 5–7 с на одне зображення. Для досягнення режиму реального часу замість базового алгоритму було

застосовано три підходи підвищення обчислювальної ефективності програмного коду.

1. Прямий доступ до пам'яті (англ. *Direct Pointer Access*). Зображення жорстко фіксується в оперативній пам'яті (метод *LockBits*), а його оброблення здійснюється через вказівники (*unsafe byte\**). Адресація пікселів виконується за швидкою формулою зміщення:

$$\text{Address} = \text{Pointer}_{\text{base}} + (y \cdot \text{Stride}) + (x \cdot \text{BPP}), \quad (6)$$

де *Stride* – ширина рядка у байтах; *BPP* (*bytes per pixel*) – бітність зображення, тобто кількість байтів, що займає один піксель. Це вилучає накладні витрати на виклики системних API.

2. Стекова алокація (англ. *Stack Allocation*). Формування тисяч локальних гістограм для кожного рядка й стовпця в динамічній пам'яті (*Heap*) неодмінно призводить до критичних затримок через роботу збирача сміття (англ. *Garbage Collector*). Цю проблему повністю усунуто застосуванням оператора *stackalloc*, який виділяє пам'ять під масиви *Hx* та *Hu* безпосередньо в надшвидкому стеку процесорного потоку.

3. Паралелізм на рівні даних (англ. *Data-level parallelism*). Оскільки аналіз кожного рядка й стовпця є математично незалежним процесом, алгоритм розпаралелено за допомогою конструкції *Parallel.For*. Це дає змогу рівномірно завантажити всі доступні ядра центрального процесора.

Завдяки такій програмній оптимізації просторова й часова складність алгоритму побудови кумулятивних проєкцій наближається до лінійної  $O(W \times H)$ , оскільки кожен піксель зчитується з пам'яті мінімально необхідну кількість разів. Структурні розбіжності між базовим і запропонованим пришвидшеним алгоритмами подано на блок-схемі (див. рис. 1).

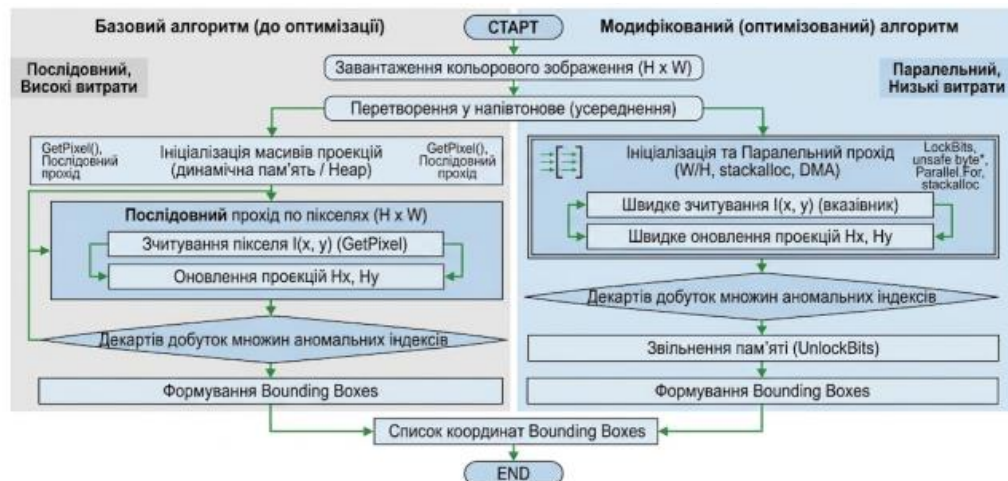


Рис. 1. Блок-схема алгоритмів до модифікації та після

Для оцінювання ефективності запропонованих рішень було проведено серію спостережень натурального експерименту з вимірювання тривалості

оброблення зображень роздільною здатністю  $2048 \times 1000$  пікселів. Результати порівняльного аналізу швидкодії усереднено й подано в табл. 1.

**Таблиця 1.** Порівняння тривалості оброблення зображень ( $2048 \times 1000$  пікселів)

| Метод реалізації                                                               | Мінімальна тривалість оброблення | Середня тривалість оброблення | Максимальна тривалість оброблення |
|--------------------------------------------------------------------------------|----------------------------------|-------------------------------|-----------------------------------|
| Базовий<br>( <i>GetPixel</i> + послідовний прохід)                             | 4871 мс                          | 5234 мс                       | 6412 мс                           |
| Модифікований<br>( <i>LockBits</i> + <i>Stackalloc</i> + <i>Parallel.For</i> ) | 36 мс                            | 44 мс                         | 52 мс                             |

Як видно з отриманих результатів (табл. 1), тривалість локалізації дефекту на зображеннях високої роздільної здатності знижується до десятків мілісекунд, що повністю задовольняє вимоги сучасних швидкісних виробничих ліній.

Окрім підтвердження обчислювальної ефективності, розроблене програмне забезпечення

успішно протестовано на точність локалізації дефектних зон на поверхнях металів. Візуальні результати роботи методу локалізації дефектів на поверхні металів, що демонструють виділені дефектні зони (англ. *Bounding Boxes*) на зразках із тестових наборів даних, подано у вікні розробленої програми (див. рис. 2).

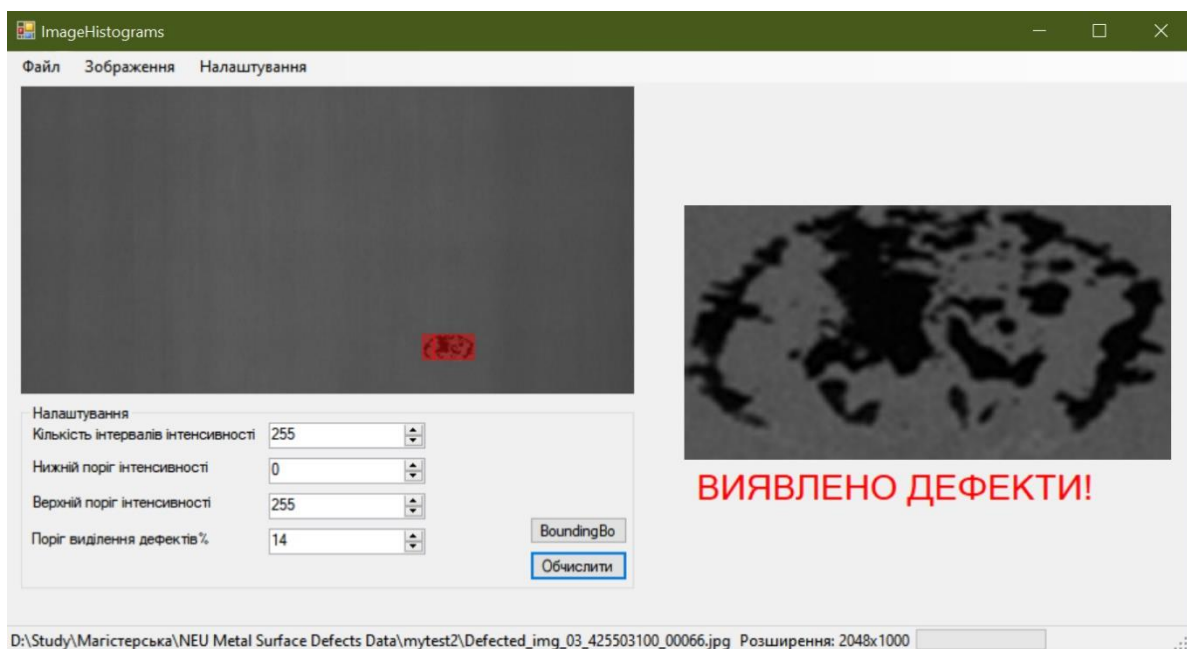


**Рис. 2.** Інтерфейс розробленого програмного забезпечення з результатами локалізації дефектних зон на поверхні металу

Підсумовуючи вищевикладене, можемо констатувати, що запропонований комплексний підхід успішно розв'язує дві ключові проблеми комп'ютерного зору в системах промислової дефектоскопії: точність просторової локалізації та обчислювальну швидкість. Математичний апарат дворівневої ієрархічної верифікації надійно усуває проблему інформаційної надмірності (фантомних рамок), гарантуючи щільне й безпомилкове виділення реальних аномалій на площині зображення. Водночас

застосовані методи низькорівневої програмної оптимізації забезпечили кардинальне прискорення роботи алгоритму (понад 110 разів порівняно з базовою реалізацією).

Досягнуті емпіричні та візуальні результати переконливо доводять, що розроблений метод здатний ефективно функціонувати в режимі жорсткого реального часу, що робить його придатним для безпосереднього впровадження в автоматизовані системи контролю якості на високошвидкісних

виробничих лініях без залучення спеціалізованих графічних прискорювачів.

## 6. Обговорення результатів дослідження

З огляду на досягнуті результати варто зазначити, що використання методів одновимірних проєкцій уже довело свою ефективність у суміжних галузях промисловості, зокрема для детекції аномалій зварних швів [10] і меж дефектів на кераміці [9]. Однак у запропонованій роботі цей математичний апарат уперше адаптовано для повноцінної двовимірної локалізації мікродефектів металу. Проведені спостереження натурного експерименту підтверджують, що запропонований метод просторової локалізації дефектних зон на основі кумулятивних проєкцій інтенсивності пікселів є концептуально працездатним і високоефективним з погляду обчислювальної складності.

Основним досягненням на поточному етапі є надзвичайно висока швидкодія розробленого методу просторової локалізації дефектних зон: завдяки глибокому програмному вдосконаленню алгоритму побудови кумулятивних проєкцій тривалість оброблення великоформатних зображень (роздільною здатністю  $2048 \times 1000$  пікселів) становить у середньому 40–50 мс на стандартному центральному процесорі (CPU). Для порівняння, у недавніх дослідженнях [3, 19, 20] наголошено, що використання сучасних моделей YOLOv8 або гібридних архітектур *Vision Transformer* для зображень аналогічної високої роздільної здатності потребує спеціалізованих тензорних ядер (GPU) з метою досягнення швидкодії на рівні 20–30 кадрів за секунду. Така швидкодія алгоритму побудови кумулятивних проєкцій свідчить про його значний потенціал і підтверджує, що метод безперечно заслуговує на подальше глибоке дослідження щодо його інтеграції у високошвидкісні виробничі лінії в режимі реального часу, де застосування важких нейромережових моделей без спеціалізованих графічних прিশвидшувачів є обчислювально неможливим.

Отримані результати вигідно відрізняються від наявних підходів до локалізації дефектних зон. Нейромережові рішення [2, 20, 21] демонструють високу здатність виділяти обмежувальні рамки (*Bounding Boxes*), але вимагають великих анованих наборів даних і суттєвих апаратних ресурсів для навчання та інференсу. Запропонований

у дослідженні метод досягає мети просторової локалізації дефектних зон за допомогою математичного перетину одновимірних векторів, що значно спрощує програмну реалізацію та знижує вартість упровадження систем комп'ютерного зору на підприємствах.

Як підтверджують фундаментальні огляди [8, 14], класичні методи просторової фільтрації (наприклад, GLCM або морфологія) є вкрай вразливими до градієнтів освітлення. На цьому тлі важливою перевагою запропонованого підходу є гнучкість налаштування порогів чутливості алгоритму. Використання динамічного еталонного фону й допуску відхилення дає змогу адаптуватися до різних умов освітлення на виробництві. Водночас результати проведених спостережень натурного експерименту виявили низку напрямів, які потребують подальшого вивчення. По-перше, глобальне кількісне оцінювання точності локалізації дефектних зон за допомогою спеціалізованих метрик (таких як IoU) на великих стандартизованих наборах даних ще не проводилося. Наразі запропонований метод аналізу кумулятивних проєкцій інтенсивності пікселів розглядається як успішне підтвердження концепції редукції розмірності, а масштабне тестування точності локалізації дефектних зон є пріоритетним завданням для подальших етапів дослідження. По-друге, як і в класичних випробуваннях [6], коректність побудови рамок обмеження дефекту залежить від правильно обраного значення допуску відхилення, що вимагає попереднього калібрування системи під конкретний тип металевої поверхні для уникнення хибних спрацьовувань на плавних градієнтах освітлення.

Зважаючи на практичні результати роботи методу просторової локалізації дефектів на тестових вибірках, варто відзначити істотне зменшення обсягу даних, які містять корисну інформацію про аномалії структури матеріалу. Під час експерименту із великоформатними зображеннями ( $2048 \times 1000$  пікселів) алгоритм не просто підтвердив наявність дефекту, а успішно відсікав до 90–95% площі однорідного фону. Формування компактних обмежувальних рамок виключно навколо пошкоджених зон свідчить про те, що розроблений метод здатний ефективно ізолювати структурні дефекти від природної металевої текстури. Крім цього, під тривалість оброблення кадрів із множинними дефектами алгоритм коректно локалізував просторово рознесені аномалії, розділяючи

їх на окремі незалежні зони інтересу ROI (англ. *Region of Interest*). Такі результати наочно демонструють, що використання кумулятивних проєкцій розв'язує проблему інформаційної надмірності, яка є критичною під час аналізу промислових зображень високої роздільної здатності.

Отже, внаслідок виконаної роботи можна сформулювати наукову новизну й практичну значущість дослідження.

**Наукова новизна** досягнутих результатів полягає в тому, що розроблено метод аналізу кумулятивних проєкцій інтенсивності пікселів зображення для дефектоскопії металів, який, на відміну від наявних ресурсомістких методів двовимірного оброблення матриці пікселів, перетворює зображення в систему незалежних одновимірних кумулятивних проєкцій і використовує їх математичний перетин для формування точних просторових координат поверхневих аномалій без попереднього машинного навчання. Це дало змогу радикально зменшити обчислювальну складність алгоритму побудови кумулятивних проєкцій і забезпечити локалізацію дефектів у режимі реального часу на обладнанні зі стандартною архітектурою.

**Практична цінність** результатів дослідження передбачає, що розроблений і глибоко оптимізований метод може бути безпосередньо інтегрований у промислові системи візуального контролю якості продукції як надшвидкий (40–50 мс) модуль локалізації дефектних зон на поверхнях металів, який дасть змогу миттєво знаходити координати підозрілих зон на великоформатних зображеннях і передавати тільки ці дрібні вирізані фрагменти на подальший аналіз більш складним моделям класифікації, що істотно знижує загальні вимоги до обчислювального обладнання на виробництві.

## 7. Висновки

У процесі дослідження розроблено й апробовано новий метод просторової локалізації дефектних зон на поверхнях металів, що ґрунтується на аналізі кумулятивних проєкцій інтенсивності зображення. Запропонований підхід дає змогу ефективно обчислювати координати проблемних зон і формувати навколо них обмежувальні рамки (*Bounding Boxes*) без потреби у великих навчальних вибірках або використанні складних моделей глибокого навчання.

Розроблений метод просторової локалізації дефектів продемонстрував надзвичайно високу обчислювальну ефективність у процесі тестування на зображеннях великої роздільної здатності. Завдяки глибокій програмній оптимізації, локалізації дефектних зон структур даних у стеку, застосуванню прямого доступу до пам'яті та розпаралелюванню обчислень для зображень великої роздільної здатності (2048 × 1000 пікселів) досягнуто тривалості оброблення одного зображення на рівні 40–50 мс з використанням тільки CPU, тоді як наявні рішення для досягнення схожої швидкодії змушені використовувати GPU. Це свідчить про придатність методу для впровадження на високошвидкісних виробничих лініях у режимі реального часу в умовах використання значно менших обчислювальних ресурсів.

Основними перевагами розробленого методу, на відміну від наявних, є простота програмної реалізації, дуже низькі обчислювальні витрати, незалежність від дорогих графічних пришивидшувачів і здатність до швидкої адаптації до нових типів металевих текстур без етапу тривалого машинного навчання. Метод аналізу кумулятивних проєкцій інтенсивності пікселів може бути органічно інтегрований у промислові автоматизовані системи візуального контролю якості поверхонь металів як базовий модуль локалізації дефектних зон аномалій (*Region Proposal*).

Перспективами подальших досліджень є проведення глобального статистичного аналізу просторової точності локалізації дефектних зон на великих стандартизованих наборах даних із застосуванням метрики перетину площ (IoU). Стратегічно важливим напрямом наукової роботи є використання запропонованого методу кумулятивних проєкцій для автоматичного вирізання локалізованих дефектних зон із великоформатних кадрів. Передача тільки цих невеликих фрагментів на вхід згортковим нейронним мережам дасть змогу швидко й безпомилково класифікувати типи дефектів і цим радикально зменшити загальне навантаження на обчислювальну систему підприємства. Такий гібридний підхід сприятиме створенню економічно доцільних, повністю автоматизованих систем контролю якості матеріалів із високою точністю та безпрецедентною швидкістю.

## Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського



чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

#### Доступність даних

Дані будуть надані за обґрунтованим запитом.

#### Фінансування

Дослідження проводилося без фінансової підтримки.

#### Використання засобів штучного інтелекту

Автори підтверджують, що не застосовували технології ШІ в процесі написання роботи.

#### References

1. Ibrahim, A. A. M. S., Tapamo, J.-R. (2024), "A survey of vision-based methods for surface defects' detection and classification in steel products", *Informatics*, No. 11 (2), article ID 25. DOI: <https://doi.org/10.3390/informatics11020025>
2. Song, K., Yan, Y. (2013), "A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects", *Applied Surface Science*, No. 285, pp. 858–864. DOI: <https://doi.org/10.1016/j.apsusc.2013.09.002>
3. Li, Z., Tian, X., Liu, X., Liu, Y., Shi, X. (2022), "A Two-Stage industrial defect detection framework based on Improved-YOLOV5 and Optimized-Inception-ResnetV2 models", *Applied Sciences*, No. 12 (2), article ID 834. DOI: <https://doi.org/10.3390/app12020834>
4. Zhang, T., Pan, P., Zhang, J., Zhang, X. (2024), "Steel Surface Defect Detection Algorithm based on improved YOLOV8N", *Applied Sciences*, No. 14 (12), article ID 5325. DOI: <https://doi.org/10.3390/app14125325>
5. Lv, X., Duan, F., Jiang, J.-J., Fu, X., Gan, L. (2020), "Deep metallic surface defect detection: The new benchmark and detection network", *Sensors*, No. 20 (6), article ID 1562. DOI: <https://doi.org/10.3390/s20061562>
6. Zhao, L., Zheng, Y., Peng, T., Zheng, E. (2023), "Metal Surface Defect Detection Based on a Transformer with Multi-Scale Mask Feature Fusion", *Sensors*, No. 23 (23), article ID 9381. DOI: <https://doi.org/10.3390/s23239381>
7. Dong, H., Song, K., He, Y., Xu, J., Yan, Y., Meng, Q. (2019), "PGA-NeT: Pyramid Feature Fusion and Global Context Attention Network for Automated Surface Defect Detection", *IEEE Transactions on Industrial Informatics*, No. 16(12), pp. 7448–7458. DOI: <https://doi.org/10.1109/tii.2019.2958826>
8. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P. (2021), "Towards total recall in industrial anomaly detection", *arXiv (Cornell University)*. DOI: <https://doi.org/10.48550/arxiv.2106.08265>
9. Fanwu, M., Tao, G., Di, W., Xiangyi, X. (2025), "Unsupervised surface defect detection using dictionary-based sparse representation", *Engineering Applications of Artificial Intelligence*, vol. 143, article ID 110020. DOI: <https://doi.org/10.1016/j.engappai.2025.110020>
10. Aslam, Y., Santni, N., Ramasamy, S., Ramar, K. (2019), "A Modified Adaptive Thresholding Method using Cuckoo Search Algorithm for Detecting Surface Defects", *International Journal of Advanced Computer Science and Applications*, No. 10 (5), pp. 214–220. DOI: <https://doi.org/10.14569/ijacsa.2019.0100528>
11. Luo, Q., Fang, X., Liu, L., Yang, C., Sun, Y. (2020), "Automated Visual Defect Detection for Flat Steel Surface: a survey", *IEEE Transactions on Instrumentation and Measurement*, No. 69 (3), pp. 626–644. DOI: <https://doi.org/10.1109/tim.2019.2963555>
12. Zhou, C., Lu, Z., Lv, Z., Meng, M., Tan, Y., Xia, K., Liu, K., Zuo, H. (2023), "Metal surface defect detection based on improved YOLOv5", *Scientific Reports*, No. 13 (1), article ID 20803. DOI: <https://doi.org/10.1038/s41598-023-47716-2>
13. Zou, G., Li, T., Li, G., Peng, X., Fu, G. (2019), "A visual detection method of tile surface defects based on spatial-frequency domain image enhancement and region growing", *2019 Chinese Automation Congress (CAC)*, pp. 1631–1636. DOI: <https://doi.org/10.1109/cac48633.2019.8997215>
14. Chen, Z., Feng, X., Liu, L., Jia, Z. (2023), "Surface defect detection of industrial components based on vision", *Scientific Reports*, No. 13 (1), article ID 22136. DOI: <https://doi.org/10.1038/s41598-023-49359-9>
15. Yang, J., Liu, Z. (2024), "A novel real-time steel surface defect detection method with enhanced feature extraction and adaptive fusion", *Engineering Applications of Artificial Intelligence*, No. 138, article ID 109289. DOI: <https://doi.org/10.1016/j.engappai.2024.109289>
16. Ren, Z., Fang, F., Yan, N., Wu, Y. (2022), "State of the art in defect detection based on machine vision", *International Journal of Precision Engineering and Manufacturing-Green Technology*, No. 9, pp. 661–691. DOI: <https://doi.org/10.1007/s40684-021-00343-6>
17. Wang, Z., Zhao, L., Li, H., Xue, X., Liu, H. (2024), "Research on a metal surface defect detection algorithm based on DSL-YOLO", *Sensors*, No. 24 (19), article ID 6268. DOI: <https://doi.org/10.3390/s24196268>
18. Gonzalez, R. C., Woods, R. E. (2018), *Digital Image Processing, Global Edition. 4th Edition*, Pearson Education, New York, 1022 p. available at: <https://www.cl72.org/090imagePLib/books/Gonzales.Woods-Digital.Image.Processing.4th.Edition.pdf>
19. Hu, X., Yang, J., Jiang, F., Hussain, A., Dashtipour, K., Gogate, M. (2023), "Steel surface defect detection based on self-supervised contrastive representation learning with matching metric", *Applied Soft Computing*, No. 145, article ID 110578. DOI: <https://doi.org/10.1016/j.asoc.2023.110578>



20. Wu, L., Ran, Y., Yan, L., Liu, Y., Song, Y., Han, D. (2025), "A dataset for surface defect detection on complex structured parts based on photometric stereo", *Scientific Data*, No. 12, article ID 276. DOI: <https://doi.org/10.1038/s41597-025-04454-6>
21. Zeng, K., Xia, Z., Qian, J., Du, X., Xiao, P., Zhu, L. (2025), "Steel surface defect detection technology based on YOLOv8-MGVS", *Metals*, No. 15 (2), article ID 109. DOI: <https://doi.org/10.3390/met15020109>

Received (Надійшла) 31.03.2026

Accepted for publication (Прийнята до друку) 18.04.2026

Publication date (Дата публікації) 29.05.2026

#### Відомості про авторів / About the Authors

**Віпшовський Юрій Андрійович** – Національний університет "Львівська політехніка", аспірант кафедри програмного забезпечення, Львів, Україна;  
**Yurii Vipshovsky** – National University "Lviv Polytechnic", Postgraduate Student of the Department Software Engineering, Lviv, Ukraine;  
 e-mail: [yurii.vipshovskyi.asp.2025@lpnu.ua](mailto:yurii.vipshovskyi.asp.2025@lpnu.ua)  
 ORCID ID: <https://orcid.org/0009-0007-9873-3753>

**Грицюк Юрій Іванович** – доктор технічних наук, професор, Національний університет "Львівська політехніка", професор кафедри програмного забезпечення, Львів, Україна;  
**Yurii Hrytsiuk** – Doctor of Technical Sciences, Professor, National University "Lviv Polytechnic", Professor of the Department of Software, Lviv, Ukraine;  
 e-mail: [yurii.i.hrytsiuk@lpnu.ua](mailto:yurii.i.hrytsiuk@lpnu.ua)  
 ORCID ID: <https://orcid.org/0000-0001-8183-3466>

## LOCALIZATION OF DEFECTIVE ZONES ON METAL SURFACES VIA CUMULATIVE PIXEL INTENSITY PROJECTION ANALYSIS

**Subject:** the process of spatial localization of defective zones on surface images. Purpose: to develop a computationally efficient method for spatial localization of defective zones on metal surfaces capable of operating in real-time on standard processor hardware without involving resource-intensive machine learning models. **Tasks:** to analyze existing flaw detection approaches; to develop a localization method and mathematical model based on cumulative projections; to create an optimized pixel matrix processing algorithm; to develop experimental software; to conduct empirical validation of the proposed approach on standardized datasets. **Methods:** the research utilizes digital image processing methods, the mathematical apparatus of one-dimensional cumulative pixel intensity projection analysis, set theory methods (to determine the intersection of anomaly zones), and low-level algorithmic optimization techniques (direct memory access, stack allocation) to accelerate computations. **Results:** a spatial localization method for defective zones has been developed that performs image dimensionality reduction by transforming it into a system of independent vectors. An algorithm featuring two-level hierarchical verification is proposed to eliminate the problem of multiple false positives (phantom bounding boxes). Experimental software was created where, through deep optimization, large-format image processing speeds (2048 by 1000 pixels) of 40–50 milliseconds were achieved using only the central processing unit. The method successfully generates accurate bounding boxes around micro-defects on backgrounds with complex metal textures, discarding up to 95 percent of the homogeneous background. **Conclusions:** the proposed method serves as a computationally efficient alternative to modern deep learning models for industrial defectoscopy systems. Shifting from two-dimensional convolution to one-dimensional projection analysis significantly reduces resource consumption. The algorithm is capable of operating in strict real-time, providing reliable spatial localization without requiring prior training on large labeled datasets.

**Keywords:** cumulative image projections; localization of metal surface defects; computer vision; image analysis.

#### Бібліографічні описи / Bibliographic descriptions

Віпшовський Ю. А., Грицюк Ю. І. Локалізація дефектних зон на поверхнях металів методом аналізу кумулятивних проєкцій інтенсивності пікселів. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 128–137. DOI: <https://doi.org/10.30837/0135-1710.2026.189.128>

Vipshovsky, Y., Hrytsiuk, Y. (2026), "Localization of Defective Zones on Metal Surfaces via Cumulative Pixel Intensity Projection Analysis", *Management Information System and Devices*, No. 2 (189), P. 128–137. DOI: <https://doi.org/10.30837/0135-1710.2026.189.128>

Oleksandr Vovk, Oleksandr Hryhoryev, Mariia Hliuza, Ruslan Chebotarov, Mariia Mendieliieva

## USABILITY TESTING OF INSURANCE WEBSITES VIA PROXY USERS

**The subject of the article** is an usability evaluation of web-based insurance calculators using physical proxy users. **The purpose of the study** is to develop and experimentally validate a combined methodology for usability testing of insurance websites using proxy users. To achieve the stated goal, the following **tasks** must be performed: to define a set of usability metrics; to develop a usability testing procedure with physical proxy users; to evaluate the usability of web-based insurance calculator using the proposed testing approach; to analyze identified usability problems; to assess the strengths and limitations of the proposed approach. **Research methods:** The study proposes usability testing methodology that combines quantitative and qualitative usability metrics. Mathematical modeling, normalization and aggregation of usability metrics, experimental evaluation based on scenarios using physical proxy users, comparative analysis, and sensitivity analysis were applied. **Results achieved.** The obtained result confirms the feasibility of applying the combined usability testing methodology, since both quantitative and qualitative usability components make an independent contribution to the integral evaluation of usability. As a result of experiment, it was found that the objective usability component of the insurance calculator interface was lower than the subjective one, which confirms the presence of underlying usability issues related to interaction efficiency, despite an overall general positive perception of the interface. The obtained value ( $\Delta UI \approx 0.061$ ) of integral usability index range confirms the reliability of the proposed methodology. **Conclusions.** The proposed combined usability testing methodology allows to identify hidden problems in the efficiency of interaction with the interface. The methodology can be applied even in cases where real end users are not available for testing, which makes it suitable for data-limited systems.

**Keywords:** usability; software testing; user experience; web design; insurance website; UX metrics; proxy users.

### 1. Introduction

The effectiveness of interaction between human and any software extremely depends on its convenience and ease of use. Those factors can largely affect website's demand and its success in the market. Usability is important because it helps create an intuitive interface, which forms a positive user experience (UX). Consequently, this affects customer satisfaction and website traffic, which may increase a company's revenue.

According to the International Association of Insurance Supervisors (IAIS) 2017 report, InsurTech is a branch of Fintech that transforms the insurance industry through digitalization. This involves implementing digital solutions to optimize processes of signing and servicing insurance contracts, claims settlement, and customer support during insurance payouts [1].

On the one hand, the impact of InsurTech on accident insurers is significant, as most of personal accident insurance products are relatively simple and targeted at a wide range of consumers [2]. On the other hand, insurance websites are often characterized by high cognitive load due to the need to input and work with insurance data that is becoming increasingly complex and voluminous [3], especially for long-term products or those with high premiums. Consequently,

tasks associated with interacting with insurance websites are becoming more complex, which in turn requires greater cognitive effort from the end user [4]. However, it is important to consider that the structure of the interface of any website or application can influence the usability of the software and user cognitive load.

Most insurance companies in Ukraine utilize digital technologies to optimize their operations. This enhances customer convenience through a simplified insurance process that enables online policy purchases using websites or mobile applications. Such digitalization also helps to automate internal workflows, which increases operational speed and reduces error rates in claims settlement [5]. However, the low level of digital literacy among a segment of the Ukrainian population encourages the use of intermediaries (brokers) when selecting and purchasing insurance products and contracts online.

Currently, not all insurers are capable of maintaining innovation-driven operations due to the challenging environment in Ukraine, driven by economic instability, which has led to decreased solvency and a decline in demand for insurance services. Furthermore, updated regulatory requirements have significantly impacted the industry landscape: the number of insurance companies in Ukraine decreased from 210 as of December 31, 2020, to 65 as of December 31, 2024.

The current market consists of about 55 non-life (accident) insurers and 10 life insurance companies [6].

Car insurance, health insurance, and property insurance remain the most popular categories of insurance services in Ukraine [7]. The market leaders in terms of insurance premiums are large companies such as ARX, TAS Insurance Group, VUSO, Oranta, INGO, USK, PZU Ukraine, and others, which collectively account for more than half of the market's insurance transactions. Additionally, insurance digital comparison services and platforms are widely used in Ukraine to help users identify the most accessible insurance products based on parameters such as policy cost, coverage, and company ratings [8]. Examples of such platforms include TrustLife, Finance.ua Insurance, HotlineFinance, and others.

Digital solutions continue to expand in the insurance industry through personalized approaches, self-service platforms, and behavioral pricing models, which are aimed at increasing customer satisfaction, trust, and purchase intention [9, 10]. Nevertheless, around 90% of insurance products in Germany are still sold through traditional physical sales channels [11]. They are represented by insurance agents or brokers and bancassurance channels linked to corporate sector. The share of online sales remains limited, except certain products as car insurance [12].

Since insurance products are complex and German customers are inherently risk-averse, this underscores the role of intermediaries in insurance coverage selection [11]. Furthermore, insurance websites foster to the adaptation of sales to digital expectations of customers by providing quick access to information and online interaction. Moreover, such platforms operate with sensitive customers' data, which highlights the importance of creating a convenient and secure environment for end users.

Comprehensive assessment of insurance websites can be complicated due to a problem of insufficient consistency between subjective and objective usability metrics. Additionally, there is no formalized testing approach to verify usability of such applications without direct user participation.

## **2. Analysis of recent studies and publications**

Usability plays a key role in success of web and digital solutions because it determines the effectiveness of user interaction with such products and can directly contribute

to the company's competitive advantage [13, 14]. According to Jakob Nielsen, usability consists of five main components such as learnability, efficiency, memorability, errors, and satisfaction [15]. Additionally, principles of minimizing cognitive load and using persistent and clear navigation contribute to increased usability [16]. UX can be evaluated using structured questionnaires combined with usability testing, accounting for psychological factors of decision-making [17, 18].

According to ISO 9241-11 standard, usability is defined as efficiency, effectiveness, and users' satisfaction in a given context of use [19]. Findings in insurance digital services show that easy and intuitive navigation and modern interface design are extremely important for UX and users' trust in the system's reliability and security [20, 21]. This highlights the importance of prioritizing interface simplicity and high-quality design. Many of the usability recommendations from J. Nielsen and Nielsen Norman Group, from breaking down content of software application into blocks to optimizing system response time, are aimed at minimizing user cognitive load.

Usability evaluation methods can be divided into data-driven, user-based, feedback methods [22, 23]. Data driven methods use machine learning algorithms to analyze user behavior data but need to understand user needs to promote effective user interaction and optimize user interface design. User-based methods provide detailed interaction analysis but can be limited in sample size and resources. On the one hand usability feedback methods are simple, but on the other hand results can be affected by self-reporting bias.

Insurance websites provide users an opportunity to obtain information about available insurance products and to purchase them quickly and remotely, with minimal time spent. Online insurance policies usually have the ability to submit claims online. The main functions of an insurance website or mobile application include information search by client, product purchase, policy servicing, claim settlement, and loss adjustment [24].

Moreover, insurance websites often use calculators for the automated calculation of policy costs (insurance premiums). During the conclusion of insurance contracts (both life and non-life), the process of underwriting or client risk assessment occurs automatically or semi-automatically. This process is based on data entered by user (e.g., gender, age, health status, driving experience, etc.), which serves as the basis for determining the final policy price.

Users of financial and insurance web applications typically perform complex tasks when selecting and paying products. Disruptions in these workflows may look like confusing navigation, unclear instructions, prolonged system response time, failure to display insurance contracts, or slow system performance [25]. Such issues can lead to decreased user satisfaction and task abandonment. Factors that reduce user confidence and trust and hinder effective use of website include complex and confusing website's interface, as well as problems with visual design components, such as small font sizes and non-intuitive layouts, which are especially critical for elderly users.

Since it is important for insurance websites to maintain long-term relationships and effectively manage large customer base, strengthening customer loyalty becomes a potential source of a company's market and financial success. Thus, customer satisfaction with insurance websites is closely linked to their usability, where ease of navigation and aesthetic appeal play a crucial role in this relationship.

Although digital insurance applications provide benefits to customers and companies, the handling of sensitive data poses risks [26]. Moreover, regulatory requirements in Germany, such as *Versicherungsvertragsgesetz (VVG)*, oblige companies to ensure transparency, security, and usability [27]. It is possible to involve proxy users, who are physical participants with characteristics or experiences similar to target audience, to participate in usability evaluation when the access to end users is limited [28, 29]. At the same time, strict data protection policies limit the use of traditional testing methods, and methodological guidelines in such cases remain limited.

### 3. Purpose and objectives of the study

Given the difficulties associated with engaging real end users for usability evaluation in related areas, such as digital insurance services, alternative approaches to testing are needed.

The aim of this study is to develop and experimentally validate a methodology for usability testing of insurance websites using physical proxy users.

The object of the study is a web-based insurance calculator, and the subject of the study is its usability evaluation using proxy users.

To achieve this goal, the following research questions are addressed:

1. To define a set of usability metrics.
2. To develop a usability testing procedure with physical proxy users.
3. To evaluate the usability of web-based insurance calculator using the proposed testing approach.
4. To analyze identified usability problems.
5. To assess the strengths and limitations of the proposed approach.

### 4. Research materials and methods

Insurance plays a substantial role in financial stability and digital platforms allow insurance contracts to be processed in real-time. However, web-based insurance interfaces are often complex due to personal data requirements and integration with internal systems, which increases the risk of user errors [15]. In order to improve interaction in such systems, informative prompts and simplified forms for user data entry can be implemented. Usability testing for web-based insurance applications was conducted using a combined testing approach, which allows for the assessment of both the effectiveness and the perception of the interface by physical proxy users.

The study employs a combined usability testing methodology for insurance websites, which integrates both objective and subjective data collection methods and includes the following stages, as shown in Figure 1:

*Stage 1.* Respondents selection and group competence index calculation.

*Stage 2:*

- interface preliminary analysis;
- identification of key functional blocks of a website.

*Stage 3:*

- defining a set of subjective evaluation metrics and usability hypotheses;
- subjective evaluation (questionnaire);
- finding problem areas of the interface.

*Stage 4:*

- selection of test scenarios for key functional blocks of the interface;
- objective evaluation (moderated real-time usability testing);
- confirmation and classification of identified usability issues.

*Stage 5.* Integration and analysis.

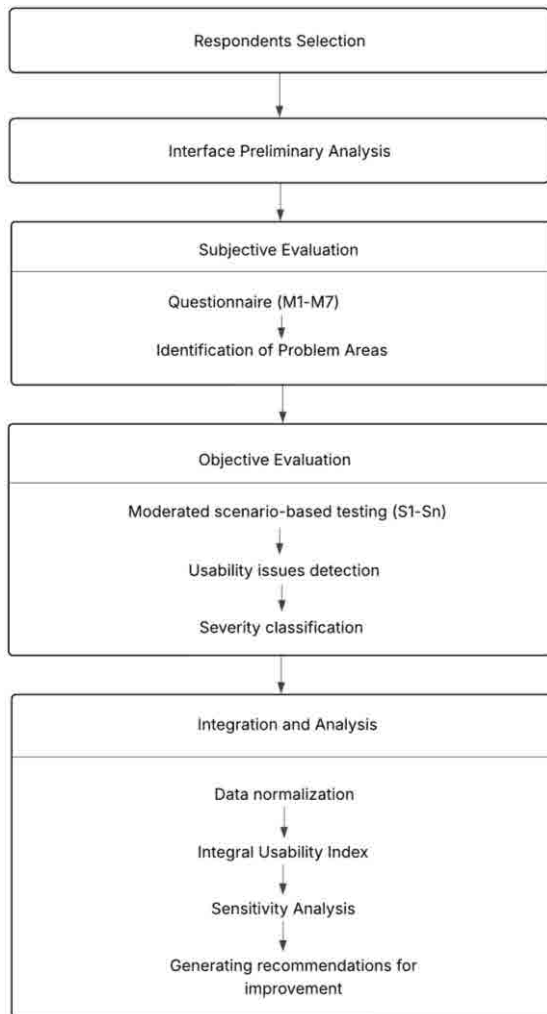


**Fig. 1.** Proposed combined methodology for usability evaluation of insurance website

**Stage 1** involves selecting respondents for usability testing. Considering that access to real users of insurance websites and platforms may be restricted due to corporate security policies or the EU General Data Protection Regulation (GDRP), it becomes necessary to engage physical proxy users. These participants must be individuals who possess experience or specific characteristics, which are similar to the target audience of the insurance website. For instance, they should have basic digital literacy and be at least minimally interested in quick conclusion of an insurance policy. To ensure more accurate usability evaluation results, the respondent group should represent different levels of experience with online insurance contracts (low, medium, and high). Additionally, the representativeness of proxy users can be enhanced by including respondents with a direct connection to the specific insurance product. For example, involving a small proportion of car owners among proxy

users for car insurance testing, or individuals aged 21–50 who prioritize health and quality medical services for health insurance evaluations.

Since both subjective and objective usability evaluation of the interface will be conducted, it is required to involve two groups of respondents. The first group will be used for subjective evaluation and it should be larger because such type of research will focus on satisfaction metrics (questionnaire) and requires a broader sample size to make a confident decision. This is especially relevant in cases when it is necessary to determine whether there is a significant difference in designs [17] of a website or mobile application. The second group consists of a smaller number of participants in order to objectively confirm usability issues identified during the subjective evaluation. In accordance with J. Nielsen's findings, a sample size of up to 8 respondents is sufficient to identify approximately 90% of usability problems.

To enhance the accuracy of result interpretation, a group competence index (GCI) was introduced, which takes into account the distribution of website interface knowledge among participants. The index was determined as a weighted average on an ordinal scale with uniform rank coding of competence levels (1 – low level, 2 – medium level, 3 – high level), according to the following formula:

$$GCI = \frac{\sum_{i=1}^3 C_i l_i}{n}, \quad (1)$$

where  $C_i$  – competence code for a group of respondents, by proficiency level;  $l_i$  – number of respondents in a group of a certain level of competence;  $n$  – total number of respondents in all groups.

**Stage 2** involves identifying key usability test tasks based on the insurance type (life or non-life), such as policy calculation, plan selection, and contract submission. The corresponding key interface components are defined for these tasks (e.g., calculator forms, personal accounts, and contract terms). A preliminary analysis allows to identify main usability issues of the interface and to formulate usability testing objectives, which can help to enhance user interaction with the product. In order to achieve this, a usability specialist (UX analyst) or a designer may be engaged to conduct an expert evaluation of the interface from a perspective of a new user in a system.

**Stage 3** focuses on developing a respondent questionnaire, which includes up to seven subjective



evaluation metrics (M1-M7). These metrics cover key aspects of user experience with the interface (provided that relevant metrics are available on the insurance website):

M1 – Overall convenience (ease of use, simplicity, and interface clarity, including ease of finding information on pages).

M2 – Interface design (including visual appeal, readability, and comprehensibility of key interface elements).

M3 – Ease of navigation (how easy it is to navigate between different sections of the site, find specific elements on the page, perform backward/forward transitions, etc.).

M4 – Functionality (interface content accessibility, system performance and available functionality).

M5 – Evaluation of error messages.

M6 – Interaction and feedback (user action outcomes, including visual changes and/or audio signals).

M7 – Cognitive load (whether design elements create unnecessary noise such as poor layout, complex paths to a goal, absent or implicit feedback from a system, which interferes with focusing on a task).

It is necessary to define corresponding usability hypotheses for each subjective metric. These hypotheses detail metrics and set the scope for specific problem areas of the website.

Each subjective metric is rated on a 5-point Likert scale, where 1 – is bad/strongly disagree and 5 – is very good/agree. Additionally, each subjective metric can consist of a different number of sub-questions  $i$  (from 2 to 6), the scores for which are aggregated into component subjective metrics, which are determined as the average values of the answers to the corresponding sub-questions of the questionnaire.

For each component subjective metric, variance and weighting coefficients are calculated to evaluate consistency and ensure the comparability of results:

$$Var(M_j) = \frac{\sum_{i=1}^n (M_{ij} - \bar{M}_j)^2}{n-1}, \quad (2)$$

$$w_j = Var(M_j) / \sum_{j=1}^7 Var(M_j), \quad (3)$$

where  $\bar{M}_j$  – average metric  $M$  value across its sub-questions;  $j$  – metric index,  $j = \overline{1,7}$ ;  $n$  – number of sub-questions for  $j$  metric;  $Var(M_j)$  – variance of  $j$ -th subjective metric;  $w_j$  – variance-based weights.

Component subjective metrics M1-M7, which were measured using a Likert scale from 1 to 5, are linearly normalized to the range [0;1] for comparability with objective results:

$$M_{j_{norm}} = \frac{M_j - M_{\min}}{M_{\max} - M_{\min}} = \frac{M_j - 1}{4}, \quad (4)$$

where  $M_j$  – metric  $M$  value  $M$ ;  $M_{\min}$  – minimum metric value;  $M_{\max}$  – maximum metric value.

As a next step, the integral subjective component of usability is calculated as an average value of normalized metrics:

$$UI_{subj} = \frac{\sum_{j=1}^n M_{j_{norm}}}{n}. \quad (5)$$

Upon completion of the subjective evaluation, interface elements that negatively impact user experience are identified. These findings can subsequently be used to develop objective test scenarios.

**Stage 4** involves the selection of usability test scenarios for key functional blocks of the interface (Si). The scenarios should not be too time-consuming to perform. Test developer (moderator) must prioritize which interface functions and potential problem areas require immediate verification, for example:

S1 – simple insurance product calculation;

S2 – new contract generation;

S3 – finding elements and verifying validations.

Moderator provides participants of the moderated usability testing with a general task statement, the final goal, and the necessary steps to initiate the test. Test participants are not restricted in their actions, but they must adhere to the main objective of the test scenario. Participants are allocated approximately 15–20 minutes to complete test tasks or scenarios, after which moderator immediately analyzes and records the results (e.g., in a spreadsheet).

According to the ISO 9241-11 standard, the following objective metrics are used (depending on the scenario objective and key functions): task completion time, effectiveness, number of errors in scenario and navigation errors. Objective metrics are grouped by type within scenarios, according to the presence of the metric in the test scenario, where  $T$  – time (s),  $E$  – number of errors,  $R$  – binary value of scenario success (0 – failure, 1 – success),  $N$  – binary value of navigation errors (0 – absent, 1 – present).

Objective metrics of varying scales and types are normalized to a unified range to ensure comparative

analysis across scenarios. Normalized values are calculated for every scenario featuring these specific.  $T$  and  $E$  values are estimated using the minimization principle:

$$X_{norm} = 1 - \frac{X - X_{min}}{X_{max} - X_{min}}, \quad (6)$$

where  $X = (T, E)$ .

Next, the aggregated metrics  $T$ ,  $E$ ,  $R$ ,  $N$  are calculated for all scenarios using the mean value method for the set of scenarios where such metric is encountered:

$$X_{avg} = \frac{\sum_{i=1}^k X_{i_{norm}}}{k}, \quad (7)$$

where  $X_{norm} = (T_{norm}, E_{norm}, R, N)$ ;  $k$  – number of scenarios where the metric is present.

The aggregated error rate  $E_{total}$  is calculated using the formula:

$$E_{total} = wE_{avg} + (1-w)N_{avg}, \quad (8)$$

where  $E_{avg}$  – aggregated value of  $E$  metric across all scenarios in which it is present;  $N_{avg}$  – aggregated value of  $N$  metric;  $w$  – weights for values  $E_{avg}$  and  $N_{avg}$ .

It was applied  $w=0.6$  because correct execution of scenario is mandatory, but navigation and usability are equally important components for the end user.

Then, the integral objective component of usability is calculated:

$$UI_{obj} = \frac{T_{avg} + E_{total} + R_{avg}}{3}. \quad (9)$$

After moderated usability testing is completed, it is needed to determine if usability issues identified during subjective evaluation have been objectively confirmed and any previously undiscovered issues are recorded. The identified problems are classified and assigned severity and priority levels, as shown in Table 1, depending on their technical impact on the task and end users and its frequency of occurrence (severity), and the specified priority by manager of addressing the identified issues (priority). Then, recommendations are formed for further validation through re-testing for the identified usability problems.

At **Stage 5**, to overcome the inconsistency of disparate metrics, an integral usability index is introduced, which allows aggregation of individual measures into a single comparable criterion.

**Table 1.** Classification of identified usability issues

| Severity  | Description                                                                              | Priority  | Description                                                                              |
|-----------|------------------------------------------------------------------------------------------|-----------|------------------------------------------------------------------------------------------|
| Critical  | Critical issue that requires immediate correction because it disrupts core functionality | Now!      | Problem should be fixed immediately due to the criticality of the problem                |
| Serious   | Significant impact, but does not affect the basic logic of the site                      | High      | The fix is important but not urgent.                                                     |
| Important | Important but not critical issue that can be fixed in the longer term                    | Middle    | Fix can be postponed even though the problem needs to be addressed                       |
| Light     | Minimal impact on core functionality                                                     | Low       | Low priority – the problem is not significant, the main functionality works correctly    |
| –         | –                                                                                        | Won't fix | The problem will not be fixed because its resolution depends on the customers' decision. |

The index includes objective and subjective components of the assessment. To formulate the integral usability index, a simple linear model was used with the parameter  $\lambda$ , which regulates the contribution of objective and subjective metrics into UX. According to ISO/IEC 25010 requirements for functional correctness and reliability of applications, it was assumed that the objective component  $UI_{obj}$  is more significant, while subjective component  $UI_{subj}$  is less significant:

$$UI = \lambda UI_{obj} + (1-\lambda) UI_{subj}, \quad \lambda = 0.5. \quad (10)$$

The value of parameter  $\lambda$  is determined for a specific insurance website, however, in the general

case,  $\lambda = 0.5$  it is used to balance the interface's attractiveness and efficiency. The consistency between objective and subjective evaluation of UX was analyzed based on the  $\Delta$  value:

$$\Delta = |UI_{obj} - UI_{subj}|. \quad (11)$$

## 5. Research results

This section presents the results of an experimental study of the proposed combined usability testing methodology for web-based insurance calculator. The object of the study is the eVorsorge GmbH (eVo) insurance calculator website, which is designed to

calculate the cost of insurance services based on entered user data and save contacts in system. These features are implemented through the functionality of the insurance calculator (Tarifrechner), which has step-by-step navigation between pages. To enhance visual aesthetics and take into account customer feedback, a new calculator interface, designed as Maklerrechner, was developed using the React library. The interface supports validation with visual display of errors (see Figure 2).

All actions in Maklerrechner are integrated into a single-page interface with side navigation and restriction of available fields in accordance with the terms of the contract, as shown in Figure 3. The target audience is represented by brokers, employees of insurance companies and private and corporate clients. The Maklerrechner version of insurance website was used as a test platform for usability testing based on typical user scenarios.



Fig. 2. Visual presentation of Maklerrechner interface error messages

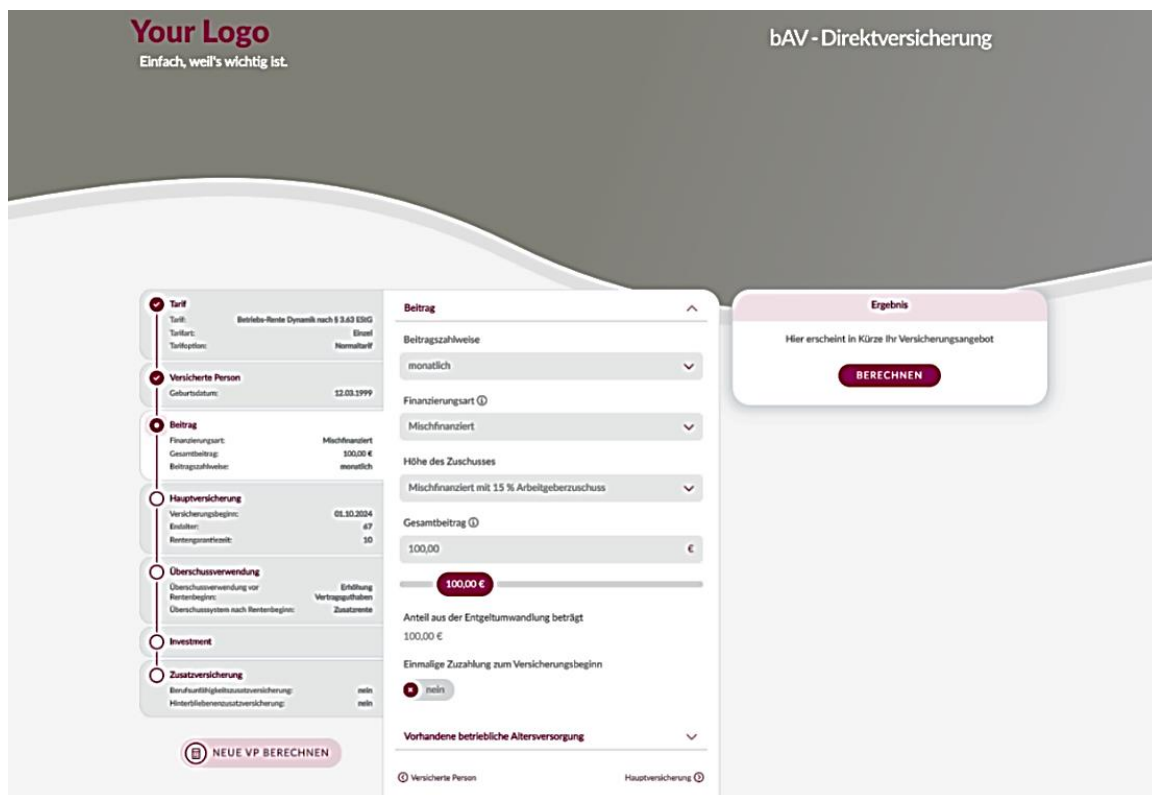


Fig. 3. Visual design of the developed Maklerrechner interface

The main actions of specialized users (e.g. insurance brokers) are given in the algorithm on Figure 4. The key

steps of the algorithm are filling in user data and calculating the contract premium.

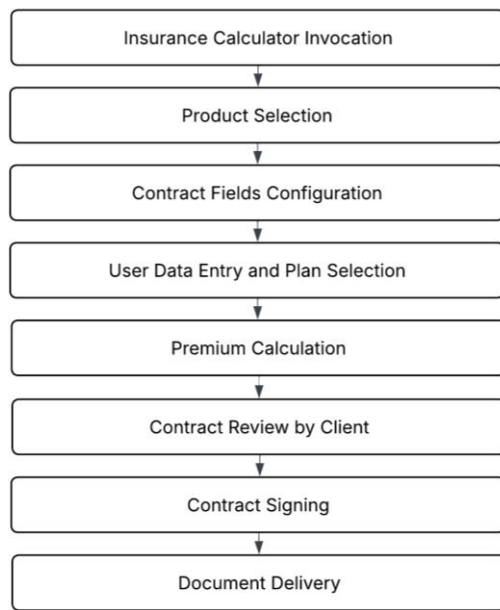


**Fig. 4.** Key actions of specialized users of the insurance website

There were restrictions on access to real users for usability testing due to eVorsorge company security policies and the GDPR data privacy law that led to use of physical proxy users. eVorsorge insurance company employees were divided into two groups of physical proxy users, who participated in the experimental study, in which typical scenarios of the insurance calculator usage were completed (e.g., creating a new insurance contract, simple contract premium calculation).

Test participants with different levels of experience in Maklerrechner interface were used to evaluate the interface. First group of the proxy users ( $n = 36$ ) was used for subjective evaluation of the current state of Maklerrechner interface and identification of its main problems using a questionnaire. The second group ( $n = 8$ ) was used for objective analysis of the identified UX problems through moderated real-time testing. The characteristics of the first and second groups of respondents are presented in Table 2.

**Table 2.** Characteristics of physical proxy users

| Competence Level                                              | Competence Code ( $C$ ) | Number of respondents ( $l_i$ ) | Group Competence Index ( $GCI$ ) |
|---------------------------------------------------------------|-------------------------|---------------------------------|----------------------------------|
| Tarifrechner – Subjective evaluation ( $n = 36$ participants) |                         |                                 |                                  |
| Low                                                           | 1                       | 0                               | 2.833                            |
| Medium                                                        | 2                       | 6                               |                                  |
| High                                                          | 3                       | 30                              |                                  |
| Maklerrecher – Subjective evaluation ( $n = 36$ participants) |                         |                                 |                                  |
| Low                                                           | 1                       | 14                              | 1.944                            |
| Medium                                                        | 2                       | 10                              |                                  |
| High                                                          | 3                       | 12                              |                                  |
| Tarifrechner – Objective assessment ( $n = 8$ participants)   |                         |                                 |                                  |
| Low                                                           | 1                       | 0                               | 3                                |
| Medium                                                        | 2                       | 0                               |                                  |
| High                                                          | 3                       | 8                               |                                  |
| Maklerrecher – Objective assessment ( $n = 8$ participants)   |                         |                                 |                                  |
| Low                                                           | 1                       | 3                               | 2.125                            |
| Medium                                                        | 2                       | 1                               |                                  |
| High                                                          | 3                       | 4                               |                                  |

As can be seen from the data in Table 2, *GCI* for the Tarifrechner interface was higher than *GCI* for Maklerrechner in both groups of participants (2.833 and 3 for first and second group correspondingly). *GCI* for new Maklerrechner interface was higher for the second group of proxy users ( $n = 8$ ) and scored 2.125 out of 3, which indicates a predominance of users with medium level of competence. Thus, obtained interface usability values should be interpreted in the context of moderate competence.

To evaluate subjective perception of the interface, subjective metrics M1-M7, which cover key aspects of

UX with the insurance calculator website, and a Likert scale from 1 to 5 points were used.

Usability testing hypotheses for the corresponding subjective metrics were formulated as follows (see Table 3):  $H_1$  – Maklerrechner interface is easy to use and perform tasks;  $H_2$  – Insufficient readability of tooltips;  $H_3$  – Simplified sidebar and location and visibility of buttons allow easy transitions between sections;  $H_4$  – Maklerrechner interface is intuitive and provides sufficient operational speed;  $H_5$  – The system provides direct navigation to the specific field containing

the error and enables batch closing of errors, which simplifies the process of their correction;  $H_6$  – Tooltip pinning upon clicking the Info Button enables easy

interaction with the hint text;  $H_7$  – Maklerrechner interface outperforms Tarifrechner in terms of usability and efficiency.

**Table 3.** Aggregated and normalized subjective metric values

| Subjective Metric | Description                            | Avg   | Variance | Weight       | Normalized Value |
|-------------------|----------------------------------------|-------|----------|--------------|------------------|
| M1                | Overall convenience                    | 4.387 | 0.018    | <b>0.027</b> | 0.847            |
| M2                | Interface design                       | 4.473 | 0.032    | <b>0.046</b> | 0.868            |
| M3                | Ease of navigation                     | 4.275 | 0.019    | <b>0.027</b> | 0.819            |
| M4                | Functionality                          | 4.070 | 0.146    | 0.210        | 0.768            |
| M5                | Evaluation of error messages           | 4.130 | 0.168    | 0.242        | 0.783            |
| M6                | Evaluation of Info Button              | 4.347 | 0.167    | 0.241        | 0.837            |
| M7                | Comparison with Tarifrechner interface | 4.127 | 0.145    | 0.208        | 0.782            |
| <b>Mean</b>       |                                        |       |          |              | <b>0.815</b>     |

Obtained experiment results in Table 3 show that subjective metrics M1 (Overall convenience), M2 (Interface design) and M3 (Ease of navigation) have the highest weight values: 0.027, 0.046 and 0.027 correspondingly. These metrics also have low variance values, which indicates their reliability and significance in evaluating the usability of Maklerrechner interface.

The integral subjective component value of usability  $UI_{subj}$  was 0.815, which shows a positive UX rating

of the interface as intuitive and easy to use, and also demonstrates consistency in positive ratings across most key aspects of interface. After survey completion, an analysis of the obtained results was conducted, as presented in Table 4. The results of the subjective usability evaluation were used to develop objective test scenarios, which allowed to focus UX evaluation on most important usability issues.

**Table 4.** Usability problems found (subjective evaluation)

| No. | Competence level | UX problem                                                                                                                | Interpretation of the Problem                                   |
|-----|------------------|---------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|
| 1   | Low              | It was more difficult to find certain elements on the page                                                                | There are navigation issues for new users                       |
| 2   | All levels       | Errors in application work                                                                                                | There are problems with the interface or functional component   |
| 3   | High             | Automatic closure of error messages results in a slower interaction speed                                                 | Users expect more control over interactions with error messages |
| 4   | Low              | Maklerrechner interface and navigation structure are unintuitive and complex, making it difficult for users to comprehend | It is more difficult for new users to work with Maklerrechner   |
| 5   | Middle           | Not very clear error messages in the Maklerrechner compared to standart Tarifrechner                                      | Inconvenience of error messages compared to Tarifrechner        |
| 6   | All levels       | In Maklerrechner it is more difficult to find the desired element on the page                                             | The page structure needs improvement                            |
| 7   | All levels       | Inconvenience of using the Info Button                                                                                    | Info Button work needs improvement                              |

Objective usability assessment was performed by a separate group of 8 participants ( $n = 8$ ), which allowed to record quantitative metrics of the effectiveness and accuracy of task performance. Objective usability assessment in this study was based on moderated real-time testing using 4 types of tasks or scenarios designed to validate previously identified UX problems. Moderated testing took place online on the Zoom platform for each participant separately.

The following objective usability metrics were used: task completion time  $T$ , effectiveness or

successful completion  $R$ , number of user errors in scenario  $E$  and navigation errors  $N$ . Each test scenario S1-S4 contained a different set of objective metrics, depending of the purpose of the scenario. Results of the objective assessment are presented in Table 5. After normalizing the values of objective metrics, the aggregation of objective metrics across all scenarios was performed using the Formula 7.

The results of aggregated normalized values of objective metrics across all scenarios are shown in Table 6.



**Table 5.** Grouping of mean values of objective metrics by type and scenario

| Scenario | Description                                | Type of Objective Metric |       |              |              |
|----------|--------------------------------------------|--------------------------|-------|--------------|--------------|
|          |                                            | $T, s$                   | $E$   | $R, \{0,1\}$ | $N, \{0,1\}$ |
| S1       | Simple insurance product calculation       | 64                       | 0.375 | 1            | 0.5          |
| S2       | Creating a new contract                    | 64                       | 0.125 | 1            | –            |
| S3       | Finding elements and verifying validations | –                        | –     | 0.875        | 0.5          |
| S4       | Using Zurück zur Berechnung button         | –                        | –     | –            | 0.625        |

**Table 6.** Normalized values of objective assessments across all scenarios

| Objective Metric | Description                   | Normalized Value |
|------------------|-------------------------------|------------------|
| T                | Time                          | 0,301            |
| E                | Errors                        | 0,723            |
| R                | Effectiveness (Resultativity) | 0,958            |
| Sum $\Sigma$     |                               | 1,982            |

Usability testing hypotheses were tested using a subjective questionnaire and the objective assessment confirmed them. Usability testing hypothesis  $H_1$  (convenience of the new interface) was confirmed: the average survey rating when comparing interfaces was 4.387. The hypothesis  $H_2$  (Insufficient readability of the text in the tooltips) was rejected: the average readability of the text was rated as 4.47 out of 5. The hypothesis  $H_3$  (simplified sidebar for navigation and buttons) was rejected: navigation and ease of finding elements on Maklerrechner page were rated at 3.76. The hypothesis  $H_4$  (interface ensures fast and intuitive task execution) was confirmed: the average survey rating was 4.070. The hypothesis  $H_5$  (simplified work with validation of errors) was rejected: the error message clarity value compared to the one in old Tarifrechner interface was 3.88. The hypothesis  $H_6$  (Info Button ease of use) was confirmed: the average survey rating was 4.35. The hypothesis  $H_7$  (effectiveness of the new interface) was confirmed: the average survey rating was 4.13, but there were certain elements of the new interface and structure that turned out to be more complex.

Objective usability analysis of the interface identified a number of UX problems that were not captured during subjective usability evaluation. In particular, the appearance of some interface elements did not meet user expectations (not all required fields are validated on contract form), some elements were

misleading (red text under a tooltip) or redundant (a calendar for entering date of birth). Moderated testing also demonstrated that participants found it more convenient when click on a validation message automatically takes them to the corresponding field with error.

Overall integral usability index  $UI$  was calculated using the Formula 10 and its value was 0.738, which shows user satisfaction with Maklerrechner interface, positive perception of its design, navigation and functionality (see Table 7).

Consistency of objective and subjective usability evaluation was measured through  $\Delta$ , which was 0.154 (scale [0;1]). Comparison of integral usability index  $UI$  value calculated at different  $\lambda$  is presented in Table 8.

**Table 7.** Integral usability index

| Component                                       | Normalized Value |
|-------------------------------------------------|------------------|
| Subjective component                            | 0.815            |
| Objective component                             | 0.661            |
| <b>Integral usability index <math>UI</math></b> | <b>0.738</b>     |

**Table 8.** Values of integral usability index across different  $\lambda$  values

| $\lambda$ | $UI$  |
|-----------|-------|
| 0.3       | 0.768 |
| 0.5       | 0.738 |
| 0.7       | 0.707 |

Integral usability index range was calculated using the following formula:

$$\Delta UI = \max(UI) - \min(UI), \quad (12)$$

$$\Delta UI = 0.768 - 0.707 = 0.061.$$

Obtained integral usability index range value was 0.061 and it had moderate sensitivity to choice of  $\lambda$  value in range [0;1].

Identified UX problems were classified by severity and priority and are presented in Table 9.

**Table 9.** Severity assessment and prioritization of identified UX issues

| UX problem                                                                                        | Severity  | Priority | Proposed Solution                                                                                                     |
|---------------------------------------------------------------------------------------------------|-----------|----------|-----------------------------------------------------------------------------------------------------------------------|
| Inconvenient navigation                                                                           |           |          |                                                                                                                       |
| Small size of main navigation buttons                                                             | Light     | Low      | Refine button visual appearance and enlarge their hit areas to enhance clickability and reduce interaction errors     |
| Poor Berechnen button placement                                                                   | Important | Middle   | Include Berechnen button at the bottom of the final data entry page, analogous to the main navigation controls        |
| Poor Zurück zur Berechnung button placement                                                       | Serious   | High     | Add the button to the left panel after Neue VP Berechnen button                                                       |
| Incorrect work of Info-Button                                                                     |           |          |                                                                                                                       |
| Unable to copy text from tooltip                                                                  | Important | Middle   | Add message capture when clicking a button                                                                            |
| Inconsistency of the visual appearance of elements with user expectations                         |           |          |                                                                                                                       |
| Field requirement                                                                                 | Light     | Low      | Redesign the visual appearance of the elements and add distinct labeling to highlight required inputs                 |
| Incorrect display of fields                                                                       | Light     | Low      |                                                                                                                       |
| Inconvenient work of error messages                                                               |           |          |                                                                                                                       |
| Inconvenient display of error messages, especially with large amounts of incorrectly entered data | Important | Middle   | Add the ability to automatically take user to the corresponding field with error when a validation message is clicked |
| Irrelevant interface elements                                                                     |           |          |                                                                                                                       |
| Non-functional data entry slider                                                                  | Light     | Low      | Add a Beitrag value gradation with maximum and minimum values                                                         |
| Redundant calendar for entering date of birth                                                     | –         | –        | Won't fix                                                                                                             |

## 6. Discussion of research results

The simple linear model of integral usability index that was presented in the previous section, is applied to Maklerrechner web-based insurance calculator for a case that an objective component of usability  $UI_{obj}$  was taken as more significant according to ISO/IEC 25010 requirements to software applications, with  $\lambda = 0.5$  value in range  $[0;1]$  and moderate level of GCI for physical proxy users.

The results obtained in Tables 3 and 5 demonstrate how subjective and objective components of usability and integral usability index behave under such conditions.

The Table 7 and calculated value of  $\Delta = 0.154$  show a discrepancy between objective and subjective usability evaluation, which indicates that combined usability testing methodology shows practical effectiveness by identifying hidden interaction usability problems, which are not reflected in subjective user feedback, in a much more precise manner than subjective and objective usability testing approaches used separately by other researchers.

The Table 8 and range of integral usability index  $\Delta UI \approx 0.061$ , which show moderate sensitivity to the  $\lambda$  parameter, confirms reliability of proposed combined usability testing methodology. Thus,  $\lambda = 0.5$  was chosen as a balanced value.

The proposed combined usability testing approach can be applied in insurance web applications, financial

platforms, and other applications with restricted access to real end users. The proposed testing methodology can be relevant for UX analysts, QA engineers, designers, developers, and software development teams, providing a balanced evaluation with both quantitative and qualitative metrics.

A limitation of the study is the analysis of a single interface version. Validation of the proposed improvements through re-testing may be considered in the future.

## 7. Conclusion

An analysis of usability evaluation methods for web-based insurance applications was conducted. As a result, two main types of usability evaluation were identified as subjective, which used a questionnaire that cover key aspects of UX for the interface, and objective, which used a moderated real-time testing. Key usability metrics were identified for each type of usability evaluation. That made it possible to enable the identification of interface elements that cause UX problems and subsequently design objective test scenarios to check performance metrics of user interaction with the interface. Unlike some publications, which focus only on the one type usability evaluation, this study focuses on combined usability testing approach that is less researched.

The ability to use physical proxy users for usability testing of insurance web-based application was analyzed due to the insurance company security policies and

GDPR restrictions. It was found that test participants could be divided into two groups to provide subjective (36 participants) and objective evaluation (8 participants) of Maklerrechner interface. GCI value for both Tarifrechner and Maklerrechner interfaces was higher than 1.90 out of 3, which demonstrated that proposed usability testing methodology can be successfully applied in the context of moderate competence.

The proposed testing methodology includes sequential stages as respondent selection, preliminary analysis, subjective stage, objective stage and integration stage. This approach made it possible to compare obtained subjective and objective results of usability evaluation between each other using normalization of values.

The integral usability index allows to evaluate overall level of usability of the interface. Simple linear model of integral usability index ensures and regulate the contribution of objective and subjective metrics into UX. The proposed combined usability testing approach found a discrepancy ( $\Delta = 0.154$ ) between objective and subjective evaluation, which shows their inconsistency. The range of integral usability index ( $\Delta UI \approx 0.061$ ) indicates moderate sensitivity to the parameter  $\lambda$ , which confirms proposed testing methodology reliability. Objective metrics received lower values comparing to subjective ones, which shows the presence of hidden usability efficiency issues despite an overall positive perception of the interface. As a balanced solution, the  $\lambda = 0.5$  value was set, which enables equal

contribution from both types of metrics and reduces evaluation bias.

The use of proxy users allows the proposed usability testing methodology to be applied even in cases where real end users are unavailable, which makes it suitable for systems with limited data access (e.g., insurance applications). The proposed methodology combines quantitative and qualitative usability metrics and provides a versatile and effective tool for usability assessment.

### Conflict of Interest

The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship, or other, that could influence the study and its results presented in this article.

### Funding

The study was conducted without financial support.

### Data Availability

Data are available from the authors upon reasonable request.

### Use of Artificial Intelligence

The authors confirms that they did not use artificial intelligence technologies in the creation of the this work.

## References

1. Прокопчук, О. Т., Харенко, А. О., Гумен, О. В., Мирошніченко, М. М., Клименко, В. О., Жарун, Р. Ф. (2024), "Цифрові трансформації та інноваційні рішення на ринку страхових послуг України", *Збірник наукових праць Уманського національного університету садівництва*, No. 105(2), pp. 106-115. DOI: <https://doi.org/10.32782/2415-8240-2024-105-2-106-115>
2. Bian, W., Ge, T., Ji, Y., Wang, X. (2023), "How is Fintech reshaping the traditional financial markets? New evidence from InsurTech and insurance sectors in China", *China Economic Review*, No. 80, 102004. DOI: <https://doi.org/10.1016/j.chieco.2023.102004>
3. Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V. (2022), "Efficient Framework for Forecasting Auto Insurance Claims Utilizing Machine Learning Based Data-Driven Methodologies", *International Research Journal of Economics and Management Studies*, No. 1(2), 10-56472. DOI: <https://doi.org/10.56472/25835238/IRJEMS-V1I2P107>
4. Darejeh, A., Marcusa, N., Mohammadi, G., Sweller, J. (2024), "A critical analysis of cognitive load measurement methods for evaluating the usability of different types of interfaces: guidelines and framework for Human-Computer Interaction", *arXiv preprint arXiv:2402.11820*. DOI: <https://doi.org/10.48550/arXiv.2402.11820>
5. Федорович, І. М. (2023), "Основні тренди та напрями розвитку страхового ринку України", *Інвестиції: практика та досвід*, No. 3, pp. 45-49. DOI: <https://doi.org/10.32702/2306-6814.2023.3.45>
6. Kuzheliev, M., Nechyporenko, A. (2025), "Вплив FinTech на діяльність страхових компаній на фінансовому ринку України", *Європейський науковий журнал Економічних та Фінансових інновацій*, No. 1(15), pp. 430-438. DOI: <https://doi.org/0.32750/2025-0138>

7. Житар, М. (2024), "Тенденції розвитку страхового ринку України в умовах воєнного стану", *Економіка та суспільство*, No. 61. DOI: <https://doi.org/10.32782/2524-0072/2024-61-24>
8. Motashko, T. P. (2023), "Comparison websites in the insurance distribution", *Market Infrastructure*, No. 71. DOI: <https://doi.org/10.32782/infrastructure71-42>
9. Eckert, C., Neunsinger, C., Osterrieder, K. (2022), "Managing customer satisfaction: digital applications for insurance companies", *Geneva Pap Risk Insur Issues Pract*, No. 47, pp. 569-602. DOI: <https://doi.org/10.1057/s41288-021-00257-z>
10. Windari, J., Oktavia, T. (2025), "Evaluation User Experience of Interest in Using Digital Insurance Platforms in Indonesia," *2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM), Bangkok, Thailand*, pp. 1-6. DOI: <https://doi.org/10.1109/IMCOM64595.2025.10857589>
11. Schnichels, S. (2025), "Customer-agent-relationship in the German insurance industry within augmenting digitalization of its society" (Doctoral dissertation, Sheffield Hallam University and Munich Business School), available at: <https://shura.shu.ac.uk/id/eprint/36390> (last accessed 10.02.2026).
12. Gesamtverband der Versicherer. available at: <https://www.gdv.de/gdv> (last accessed 10.02.2026).
13. Вовк, О. В., Глюза, М. П. (2023), "Основи UX-Research", *Інформаційні технології в сучасному світі: дослідження молодих вчених*, С. 94.
14. Вовк, О. В., Глюза, М. П., Григор'єв, О. В. (2024), "Алгоритми оцінки якості Usability веб-сайту", *Поліграфічні, мультимедійні та web-технології*, Т. 2, С. 71–72.
15. Вовк, О. В., Глюза, М. П. (2023), "Usability-тестування як ефективний показник успішності веб-продуктів", *Науковий простір: актуальні питання, досягнення та інновації*, С. 348–350.
16. Krug, S. (2018), *Don't make me think!: Web & Mobile Usability: Das intuitive Web*. MITP-Verlags GmbH & Co. KG.
17. Geisen, E., Bergstrom, J. R. (2017), *Usability testing for survey research*. Morgan Kaufmann.
18. Weinschenk, S. (2011), *100 things every designer needs to know about people*. Pearson Education.
19. Noviar, H., Widodo, S. (2024), "Comparative Analysis of Usability in Duolingo and Cake Mobile Applications Based on ISO 9241-11 Using Partial Least Square (PLS)", *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 2, pp. 1–16.
20. Ichsan, A. A., Santoso, H. B. (2025), "Supporting IFRS 17 Readiness: A Usability Case Study of Tugu's Core Insurance System in Indonesia", *Sriwijaya International Journal of Dynamic Economics and Business*, pp. 299–318.
21. Von Kalckreuth, N., Kopka, M., Schmid, C., Kratzer, C., Reptuschenko, A., Feufel, M. A. (2025), "Trustworthiness of the electronic health record in Germany: an exploratory, user-centered analysis", *Frontiers in Digital Health*, No. 7, 1473326.
22. Liu, Y., Xu, Y., Zhou, S. (2024), "Enhancing User Experience through Machine Learning-Based Personalized Recommendation Systems: Behavior Data-Driven UI Design", *TechRxiv*. DOI: <https://doi.org/10.36227/techrxiv.173337559.97825928/v1>
23. Sharma, S, Kumar, B. A. (2025), "A systematic review of user-based usability testing practices in self-care mHealth apps". *DIGITAL HEALTH*, No. 11. DOI: <https://doi.org/10.1177/20552076251374184>
24. Кузьо, Н. Є., Косар, Н. С. (2022), "Тенденції розвитку та перспективи активізації збуту суб'єктів страхового ринку України", *Менеджмент та підприємництво в Україні: етапи становлення і проблеми розвитку*, No. 4(2), pp. 253–265. DOI: <https://doi.org/10.23939/smeu2022.02.253>
25. Genne, S. (2023), "Optimizing user experience in high-traffic financial web applications using analytics", *International Journal of Engineering & Extended Technologies Research (IJEETR)*, No. 5(5), pp. 7231–7241. DOI: <https://doi.org/10.15662/IJEETR.2023.0505008>
26. Fadilah, M. D., Rochimah, S. (2023), "Security Evaluation of Insurance Portal Agency Information System Based on ISO/IEC 25010 Quality Standard Utilizing OWASP ZAP", *3rd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), Denpasar, Bali, Indonesia*, pp. 352–357, DOI: <https://doi.org/10.1109/ICICyTA60173.2023.10428701>
27. Gesetz über den Versicherungsvertrag. available at: <https://www.gesetze-im-internet.de/> (last accessed 10.02.2026).
28. Matin, S., Asma, E., Allen, E., Mangwiro, L., Mjumira, R., Valle, M., Acemyan, C., Oden, M., Kawaza, K., Kortum, P., Richards-Kortum, R. (2023), "Proxy users for usability testing of medical devices for use in sub-saharan africa", *International Symposium on Human Factors and Ergonomics in Health Care*, No. 12(1), pp. 208–212. Sage CA: Los Angeles, CA: SAGE Publications.
29. Moreno de Oliveira, G., Nascimento Carvalho, A., Lamego, B.S., Monteiro, I.T., Gonçalves, E.J.T., Basilio, A.F. (2022), "Reporting the application of user experience tools and proxy users in an industrial process based on double diamond", *International Conference on Human-Computer Interaction*, pp. 57-74. DOI: [https://doi.org/10.1007/978-3-031-05897-4\\_5](https://doi.org/10.1007/978-3-031-05897-4_5)

Received (Надійшло) 10.03.2026

Accepted for publication (Прийнята до друку) 31.03.2026

Publication date (Дата публікації) 29.05.2026

*Відомості про авторів / About the Authors*

**Вовк Олександр Володимирович** – кандидат технічних наук, Харківський національний університет радіоелектроніки, доцент кафедри медіасистем та технологій, Харків, Україна;

**Oleksandr Vovk** – Phd (Technical Sciences), Kharkiv National University of Radio Electronics, Associate Professor of the Department of Media Systems and Technologies, Kharkiv, Ukraine;

e-mail: [oleksandr.vovk@nure.ua](mailto:oleksandr.vovk@nure.ua)

ORCID ID: <https://orcid.org/0000-0001-9072-1634>

**Григор'єв Олександр Вікторович** – кандидат технічних наук, Харківський національний університет радіоелектроніки, професор кафедри медіасистем та технологій, Харків, Україна;

**Oleksandr Hryhoryev** – Phd (Technical Sciences), Kharkiv National University of Radio Electronics, Associate Professor of the Department of Media Systems and Technologies, Kharkiv, Ukraine;

e-mail: [oleksandr.hryhoryev@nure.ua](mailto:oleksandr.hryhoryev@nure.ua)

ORCID ID: <https://orcid.org/0000-0001-6467-7983>

**Глюза Марія Павлівна** – Харківський національний університет радіоелектроніки, здобувач вищої освіти (2-ий рівень, магістр), кафедра медіасистем та технологій, Харків, Україна;

**Mariia Hliuza** – Kharkiv National University of Radio Electronics, Student at the second (Master's) level of the Department of Media Systems and Technologies, Kharkiv, Ukraine;

e-mail: [mariia.hliuza@nure.ua](mailto:mariia.hliuza@nure.ua)

ORCID ID: <https://orcid.org/0009-0007-1310-262X>

**Чеботарьов Руслан Ігорович** – Харківський національний університет радіоелектроніки, аспірант кафедри медіасистем та технологій, Харків, Україна;

**Ruslan Chebotarov** – Kharkiv National University of Radio Electronics, Postgraduate Student of the Department of Media Systems and Technologies, Kharkiv, Ukraine;

e-mail: [ruslan.chebotarov@nure.ua](mailto:ruslan.chebotarov@nure.ua)

ORCID ID: <https://orcid.org/0000-0002-1191-775X>

**Менделєва Марія Василівна** – Харківський національний університет радіоелектроніки, асистент кафедри медіасистем та технологій, Харків, Україна;

**Mariia Mendielieva** – Kharkiv National University of Radio Electronics, Assistant of the Department of Media Systems and Technologies, Ukraine;

e-mail: [mariia.mendielieva@nure.ua](mailto:mariia.mendielieva@nure.ua)

ORCID ID: <https://orcid.org/0009-0002-4282-3147>

## ТЕСТУВАННЯ ЮЗАБІЛІТІ СТРАХОВИХ ВЕБСАЙТІВ ЗА ДОПОМОГОЮ ПРОКСІ-КОРИСТУВАЧІВ

**Предметом дослідження** є оцінювання зручності застосування цифрових страхових вебкалькуляторів за допомогою проксі-користувачів. **Мета роботи** – розроблення й експериментальна перевірка комбінованої методики тестування юзабіліті страхових вебсайтів за допомогою проксі-користувачів. З огляду на окреслену мету необхідно виконати такі **завдання**: визначити набір метрик зручності використання; розробити процедуру тестування зручності застосування за участю проксі-користувачів; оцінити зручність використання вебсайту страхового калькулятора за допомогою запропонованого підходу; проаналізувати виявлені дефекти зручності використання; оцінити переваги й недоліки запропонованого підходу. **Методи дослідження**. У роботі запропоновано комбіновану методику юзабіліті тестування, яка поєднує кількісні та якісні показники юзабіліті. Застосовано математичне моделювання, нормалізацію та агрегацію метрик юзабіліті, експериментальне оцінювання на основі сценаріїв користувачів, порівняльний аналіз і аналіз чутливості. **Результати роботи**. Дослідження підтвердило доцільність застосування комбінованої методики тестування, оскільки кількісні та якісні показники роблять незалежний внесок в інтегральну оцінку юзабіліті. Результати тестування виявили, що об'єктивні показники інтерфейсу страхового калькулятора є нижчими за суб'єктивні, що підтверджує наявність прихованих проблем ефективності



взаємодії за умови загального позитивного сприйняття інтерфейсу. Розмах інтегрального індексу ( $\Delta UI \approx 0.061$ ) підтверджує надійність методики. **Висновки.** Запропонована методика комбінованого юзабіліті тестування дає змогу виявляти приховані проблеми ефективності взаємодії з інтерфейсом. Методика може бути застосована навіть у разі, якщо реальні користувачі недоступні для тестування. Це робить її придатною для систем з обмеженим доступом до даних.

**Ключові слова:** юзабіліті; тестування програмного забезпечення; досвід користувачів; вебдизайн; страховий вебсайт; метрики досвіду користувачів; проксі-користувачі.

#### *Бібліографічні описи / Bibliographic descriptions*

Вовк О. В., Григор'єв О. В., Глюза М. П., Чеботарьов Р. І., Менделєва М. В. Тестування юзабіліті страхових вебсайтів за допомогою проксі-користувачів. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 138–152. DOI: <https://doi.org/10.30837/0135-1710.2026.189.138>

Vovk, O., Hryhoryev, O., Hliuza, M., Chebotarov, R., Mendieliya, M. (2026), "Usability testing of insurance websites via proxy users", *Management Information System and Devices*, No. 2 (189), P. 138–152. DOI: <https://doi.org/10.30837/0135-1710.2026.189.138>

---

Даценко С. С., Кучук Г. А.

## МІЖВИБІРКОВА УЗАГАЛЬНЕНІСТЬ ДЛЯ ВИЯВЛЕННЯ ФЕЙКОВИХ НОВИН ІЗ ВИКОРИСТАННЯМ КОНТРАСТИВНОГО АДАПТАЦІЙНОГО НАВЧАННЯ

Потреби сучасного промислового бізнесу дедалі частіше стикаються з викликами, пов'язаними з поширенням фейкових новин, які можуть суттєво впливати на репутацію, довіру клієнтів і фінансові показники промислових компаній. Наявні системи виявлення демонструють майже ідеальну точність на окремих еталонних наборах даних, проте зазнають катастрофічної невдачі за умов застосування до інформації з різних джерел, що викликає серйозні сумніви щодо готовності до впровадження в реальних умовах. **Предмет дослідження** – методи виявлення фейкових новин у структурно різноманітних наборах даних. **Метою** є розроблення й валідація підходу до виявлення фейкових новин у наборах даних зі структурно різноманітним вмістом із використанням контрастивного адаптаційного навчання, що дасть змогу створити єдину модель на основі трансформера. **Завдання дослідження:** встановлення базових показників для окремих наборів даних за допомогою мовної моделі DeBERTa-v3-base на трьох різних наборах даних фейкових новин; побудова систематичної матриці перенесення між наборами даних для кількісного оцінювання невдач узагальнення; надання пропозиції та оцінювання контрастивного об'єднаного адаптаційного навчання як підходу адаптації до домену; дослідження з абляцією контрастивних гіперпараметрів. **Метод дослідження.** Мовна модель DeBERTa-v3-base налаштовується на трьох наборах даних фейкових новин ISOT, LIAR та WELFake. Комбінована навчальна задача інтегрує класифікацію за крос-ентропією з контрольованою контрастивною втратою по об'єднаних наборах даних з використанням збалансованого вибіркового відбору з різних наборів даних. **Результати.** Матриця перенесення між наборами даних свідчить про катастрофічну невдачу узагальнення: модель, яка досягає 100% F1 на ISOT, демонструє лише 23,9% на WELFake та 40,3% на LIAR. Запропонована контрастивна об'єднана модель створює єдину модель, яка одночасно досягає 97,5% F1 на ISOT, 57,4% на LIAR і 98,4% на WELFake з абсолютним покращенням F1 до 97,9% порівняно з базовими переносами. **Висновки.** Абляційний аналіз підтверджує, що контрастивна задача має особливу користь у складніших доменах. Контрастне об'єднане адаптаційне навчання створює моделі, придатні для розгортання в середовищах новин із декількома джерелами.

**Ключові слова:** виявлення фейкових новин; міжвибіркова узагальненість; трансформерна мовна модель; трансферне навчання; оброблення природної мови.

### Вступ

У сучасному інформаційному середовищі промисловий бізнес дедалі частіше стикається з викликами, пов'язаними з поширенням фейкових новин, які можуть суттєво впливати на репутацію, довіру клієнтів і фінансові показники промислових компаній. У відповідь на ці загрози формується цілий спектр інструментів для виявлення й нейтралізації дезінформації, що активно інтегруються в корпоративні процеси [1]. Компанії впроваджують системи моніторингу медіа, які дають змогу в реальному часі відстежувати згадки про бренд і оперативно реагувати на підозрілий контент. Важливе значення відіграють технології штучного інтелекту, зокрема рішення на основі Natural Language Processing, які аналізують тональність, структуру й достовірність текстів. Окрім цього, використовуються інструменти аналізу соціальних мереж, що допомагають виявляти аномальні патерни поширення інформації

та координувати кампанії дезінформації [2]. Значну увагу приділяють також автоматизованим системам оцінювання ризиків, що сигналізують про потенційні загрози для бренду ще до того, як вони набудуть широкого розголосу [3]. Бізнес активно співпрацює із зовнішніми платформами й експертними організаціями, щоб підвищити точність перевірки фактів. Водночас інтеграція таких інструментів у внутрішні комунікаційні стратегії дає змогу підвищити рівень інформаційної стійкості компаній. Не менш важливим є використання систем управління репутацією, які поєднують аналітику, автоматизацію та кризовий менеджмент [4]. Усе це формує новий підхід до корпоративної безпеки, де інформаційні ризики розглядаються нарівні з фінансовими чи операційними. Отже, інструменти для виявлення фейкових новин стають невід'ємною частиною стратегічного управління бізнесом і логічно продовжують ширший контекст трансформацій інформаційного середовища.

Методи машинного навчання продемонстрували значні успіхи у виявленні фейкових новин під час оцінювання на окремих наборах даних. Традиційні алгоритми, як-от LightGBM і XGBoost, досягають показників F1, що перевищують 99% на широко використовуваному наборі даних ISOT [5], тоді як моделі глибокого навчання, зокрема BERT і RoBERTa, повідомляють про аналогічно ефективні результати [6]. У роботі підтверджено ці висновки й продемонстровано, що LightGBM з функціями TF-IDF досягає 99,8% F1 на ISOT, перевершуючи складніші архітектури глибокого навчання за умов обмежених ресурсів [5]. Подальше дослідження гібридних нейронних архітектур (BiLSTM-CNN, BiLSTM-GRU, BiLSTM-GNN) на наборі даних FEVER досягло точності 79,5% зі значними проблемами перенавчання [7].

Однак високі результати на окремих наборах даних створюють оманливу уяву про можливості виявлення. Моделі, навчені на одному наборі даних, вивчають статистичні закономірності, специфічні для конструкції цього набору даних (стилі письма, властиві для джерела, тематичні упередження й артефакти колекції), а не узагальнювальні показники обману [8]. Унаслідок розгортання в реальних умовах, де новинний контент різко різниться за джерелом, стилем, довжиною та предметною галуззю, ці моделі зазнають катастрофічних невдач. Це фундаментальне обмеження залишилося без достатньої уваги в сучасній літературі, до того ж переважну більшість досліджень з виявлення фейкових новин проведено тільки на одному наборі даних [9].

## 1 Огляд літератури й формулювання проблеми

### 1.1 Традиційні підходи

#### й підходи на основі глибокого навчання

Галузь автоматизованого виявлення фейкових новин швидко розвивалася, проходячи шлях від методів, що ґрунтуються на правилах, традиційного машинного навчання до складних архітектур глибокого навчання. Ранні підходи, основані на контенті, спиралися на лінгвістичну інженерію ознак – аналіз стилю письма, метрик читабельності, шаблонів пунктуації та індикаторів настрою [10]. Ці вручну створені ознаки в поєднанні з методами градієнтного підсилення, як-от LightGBM і XGBoost, довели надзвичайну ефективність на еталонних

наборах даних, часто перевершуючи більш складні нейронні архітектури [5, 11].

Підходи, основані на глибокому навчанні, дали змогу автоматично виявляти відповідні ознаки в необробленому тексті. Рекурентні архітектури (LSTM, BiLSTM, GRU) беруть до уваги послідовні залежності [12], тоді як згорткові нейронні мережі виокремлюють локальні шаблони  $n$ -грам [13]. Гібридні моделі, що поєднують ці парадигми, продемонстрували конкурентоспроможну продуктивність [7, 14]. Поява попередньо навчених моделей *Transformer* ознаменувала зміну парадигми: BERT [15], RoBERTa [16] і DeBERTa [17] досягають найсучасніших результатів завдяки двоспрямованому контекстному розумінню, водночас механізми розплетеної уваги DeBERTa досягають кращої продуктивності з меншою кількістю параметрів, ніж BERT-large.

### 1.2 Оцінювання на різних наборах даних

#### і зсув домену

Незважаючи на високі результати на окремих наборах даних, узагальнення між наборами даних залишається критично нерозв'язаним питанням. Проблема зсуву домену, коли розподіл даних навчання  $P_{source}(X, Y)$  принципово відрізняється від цільового розподілу  $P_{target}(X, Y)$ , гостро відчувається у виявленні фейкових новин, оскільки різні набори даних відтворюють різні аспекти обману [8].

Автори публікації [18] провели одне з небагатьох систематичних оцінювань наборів даних і виявили, що в моделях, навчених на ISOT, падіння точності становило 37–42% за умови тестування на інших наборах даних, до того ж моделі *FakeNewsNet* досягли лише випадкової продуктивності на невидимих даних. Нещодавно Груенштейдл і Кіранне [19] провели перехресне оцінювання даних і моделей, використовуючи попередньо навчені моделі сімейства BERT на європейських наборах даних новин, і підтвердили значне зниження продуктивності на невидимих доменах. Однак це дослідження було суто діагностичним – автори оцінювали наявні моделі без пропозиції методів покращення перехресного узагальнення наборів даних. Описаний в роботі [19] експеримент розширює цей напрям завдяки побудові контрольованої матриці передачі з єдиною архітектурою та впровадженню контрастивного об'єднаного адаптаційного навчання як конкретного рішення.

Кілька факторів впливають на погане перенесення між наборами даних:

- упередженість джерела, коли справжні та фейкові новини взяті із систематично різних джерел;
- тематична упередженість, коли навчальні дані охоплюють обмежені предметні галузі;
- структурна упередженість, коли набори даних принципово різняться за текстовими характеристиками.

### 1.3 Контрастивне навчання для адаптації домену

Контрастивне навчання стало потужною технікою для навчання представлень, стійкого до змін у розподілі даних. Фреймворк SimCLR [20] продемонстрував високу переносимість навчання ознак, тоді як *Supervised Contrastive Learning* (SupCon) [21] розширив ці можливості, даючи змогу використовувати інформацію про мітки. У NLP контрастивні цілі застосовувалися до вбудовування речень [22], класифікації за кількома спробами [23] й адаптації домену [24]. Основним вагомим висновком є те, що контрастивна втрата порівняння заохочує модель вивчати незалежні від домену ознаки способом побудови позитивних пар зі зразків, що мають однакову мітку в різних наборах даних.

Тоді як наявні методи контрастивного навчання для виявлення фейкових новин зосереджені на крос-модальному узгодженні між текстом і зображеннями [25, 26], такий підхід застосовує порівняльне навчання до адаптації доменів між наборами даних, що є принципово іншою метою.

## 2 Мета й завдання дослідження

Предметом дослідження є методи виявлення фейкових новин у структурно різноманітних наборах даних. Метою – розроблення й валідація підходу до виявлення фейкових новин у наборах даних зі структурно різноманітним вмістом із використанням контрастивного адаптаційного навчання, що дасть змогу створити єдину модель на основі трансформера. У такий спосіб буде подоланий розрив між високими показниками на еталонних наборах даних і готовністю до впровадження в реальних умовах.

Для досягнення окресленої мети необхідно виконати такі завдання:

- 1) встановити базові показники DeBERTa-v3-base для трьох структурно різноманітних наборів даних фейкових новин (ISOT, LIAR, WELFake) з метою

кількісного оцінювання меж продуктивності для окремих наборів даних;

- 2) побудувати систематичну матрицю перенесення між наборами даних для кількісного оцінювання невдач у генералізації;

- 3) запропонувати й оцінити контрастивне об'єднане адаптаційне навчання як підхід адаптації до домену;

- 4) дослідити абляцію контрастивних гіперпараметрів (вага втрат і температура) для визначення оптимальних конфігурацій.

## 3 Методологія

### 3.1 Набори даних

Для перевірки запропонованого підходу використовувались три структурно різноманітні набори даних, що подають різні аспекти проблеми виявлення фейкових новин.

Набір даних *ISOT Fake News Dataset* [27] містить 44 889 статей зі справжньою інформацією, отриманою з *Reuters*, і фейковими новинами з ненадійних джерел, позначеними організаціями з перевірки фактів. Статті мають середню довжину – 2252 символи зі збалансованими бінарними мітками (47,7% реальні, 52,3% фейкові).

Набір даних *LIAR Dataset* [28] містить 12 789 стислих політичних заяв від *PolitiFact*. Оригінальні мітки з шести класів бінаризовані (pants-fire + false + badly-true → fake; half-true + mostly-true + true → real). Твердження мають середню довжину – лише 134 символи, що робить це принципово іншим завданням, яке вимагає розуміння правдивості політичних тверджень, а не стилістичного аналізу. Набір даних *WELFake Dataset* [29] містить новинні статті, об'єднані з чотирьох джерел: *Kaggle*, *McIntire*, *Reuters* і *BuzzFeed Political*. Оскільки у WELFake є статті *Reuters*, що перетинаються зі справжніми новинами ISOT, тому було виконано систематичну дедуплікацію з використанням 200-символьних текстових відбитків і вилученням 38 285 статей, що перетинаються. Дедуплікований набір даних містить 25 346 статей (46,4% справжніх, 53,6% фейкових) із середньою довжиною 34 56 символів. Усі набори даних були попередньо оброблені за допомогою уніфікованого конвеєра: вилучення URL-адрес і електронної пошти, очищення шаблонів *Reuters*, нормалізація пробілів і фільтрація мінімальної довжини (20 символів). Кожен набір

даних був розділений на навчальний, валідаційний і тестовий набори за допомогою стратифікованого розподілу 70/15/15 з фіксованим випадковим початковим числом.

У табл. 1 подано основні характеристики трьох наборів даних після попереднього оброблення.

**Таблиця 1.** Характеристики набору даних після попереднього оброблення

| Набір даних   | Зразки | Класи | Середня довжина | Вихідний ресурс             |
|---------------|--------|-------|-----------------|-----------------------------|
| ISOT          | 44,889 | 2     | 2,252 chars     | Reuters + flagged sources   |
| LIAR (binary) | 12,789 | 2     | 134 chars       | PolitiFact statements       |
| WELFake       | 25,346 | 2     | 3,456 chars     | Multi-source (deduplicated) |

### 3.2 Мовна модель DeBERTa-v3-base

Було використано DeBERTa-v3-base [17] як базову мовну модель. Трансформерна модель DeBERTa запроваджує розплатану увагу, подаючи кожен токен за допомогою окремих векторів вмісту й позиції, а також удосконалений декодер з маскою, який зважає на абсолютну позицію в шарі декодування. Модель DeBERTa-v3 додатково підвищує ефективність завдяки попередньому навчанню з виявленням замієнених токенів у стилі ELECTRA. Базова модель містить приблизно 86 млн параметрів із прихованим виміром 768 та 12 шарами *Transformer*. Представлення токенів [CLS] є агрегованим представленням послідовності для класифікації.

Для стандартного адаптаційного навчання (експерименти 1–2) архітектура моделі містить базову мережу DeBERTa, шар відсіву ( $p = 0,1$ ) та лінійну класифікаційну головку, навчену із втратою перехресної ентропії.

Для навчання на перехресних наборах даних (експеримент 3) архітектура була розширена за допомогою проєкційної головки для контрольованого контрастивного навчання. Представлення [CLS] надходить до двох паралельних головок: головки класифікації ( $768 \rightarrow \text{num\_classes}$ , втрата перехресної ентропії) та головки проєкції ( $768 \rightarrow 256 \rightarrow 128$ , LayerNorm + ReLU, L2-нормалізовані вбудовування для втрат SupCon).

Контрастивна втрата з наставником [21] визначається як

$$L_{\text{SupCon}} = \sum_i \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \left( \frac{\exp\left(\frac{z_i \cdot z_p}{\tau}\right)}{\sum_{a \neq i} \left(\frac{z_i \cdot z_a}{\tau}\right)} \right),$$

де  $P(i)$  – набір зразків, що мають ту саму мітку, що й анкер  $i$ ;  $z$  – L2-нормалізовані проєкції;

$\tau$  – параметр температури. Загальна мета полягає в  $L_{\text{total}} = L_{\text{CE}} + \beta \cdot L_{\text{SupCon}}$ , де  $\beta$  контролює вагу контрастиву. Найважливіше те, що позитивні пари будуються через межі наборів даних, що змушує модель навчатися представляти класи, незалежно від домену.

### 3.3 Реалізація підходу виявлення фейкових новин із використанням контрастивного адаптаційного навчання

У запропонованому підході реалізовано алгоритм збалансованого пакетного відбору, який буде кожну партію з приблизно рівними частинами від кожної комбінації (набір даних, мітка), забезпечуючи позитивні пари між наборами даних у кожній партії.

У всіх проведених експериментах використовувався оптимізатор AdamW із косинусною кривою зміни швидкості навчання й навчанням із змішаною точністю. Повний набір гіперпараметрів подано в табл. 2.

Апаратне забезпечення, що використовувалося в експериментах, мало такі характеристики: NVIDIA GeForce RTX 5070 Ti (16 GB VRAM), Windows 11.

Загалом було визначено чотири типи проведених експериментів, призначених для поетапного оцінювання проблеми узагальнення й запропонованого рішення.

В експерименті типу 1 модель DeBERTa-v3-base проходить незалежне адаптаційне навчання на кожному наборі даних для встановлення базових показників ефективності в межах домену, зокрема з 6-класовою оцінкою LIAR для аналізу детальної класифікації правдивості.

В експерименті типу 2 кожна модель з експерименту 1 оцінюється на тестових наборах усіх інших бінарних наборів даних без додаткового навчання, створюючи матрицю перенесення  $3 \times 3$ , яка кількісно оцінює узагальнення між наборами даних.



Таблиця 2. Гіперпараметри навчання

| Параметр                          | Значення                    |
|-----------------------------------|-----------------------------|
| Оптимізатор                       | AdamW                       |
| Швидкість навчання                | $2 \times 10^{-5}$          |
| Зниження ваги                     | 0.01                        |
| Графік LR                         | Косинусний з 10% розігрівом |
| Розмір партії                     | 16                          |
| Максимальна кількість епох        | 5 (3 для абляції)           |
| Терпіння до раннього зупинення    | 2                           |
| Змішана точність                  | fp16                        |
| Максимальна довжина послідовності | 512 (256 для абляції)       |
| Контрастивна вага ( $\beta$ )     | 0.5                         |
| Температура ( $\tau$ )            | 0.07                        |
| Розмірність проєкції              | 128                         |

В експерименті типу 3 всі три бінарні набори даних об'єднуються для спільного навчання з контрастивною архітектурою, де збалансований вибірник забезпечує наявність позитивних пар між наборами даних у кожному пакеті.

Експеримент типу 4 орієнтований на проведення систематичної абляції над контрастивною вагою  $\beta \in \{0, 0; 0, 3; 0, 5; 1, 0\}$  і температурою  $\tau \in \{0, 05; 0, 07; 0, 2\}$ , де  $\beta = 0, 0$  – об'єднане навчання лише за CE без контрастивної мети, ізолюючи внесок об'єднання даних від контрастивного сигналу.

Для доповнення кількісного аналізу вбудовування токенів [CLS] як з базової, так і з контрастної моделей проєктуються у два виміри за допомогою t-SNE для якісного аналізу навчених просторів представлення. Усі експерименти аналізують точність, прецизію, відкликання, показник F1 та AUC-ROC, до того ж F1 слугує основним метричним показником порівняння.

Таблиця 3. Базові результати для одного набору даних

| Набір даних    | Точність | Прецизія | Відтворюваність | F1    | AUC-ROC |
|----------------|----------|----------|-----------------|-------|---------|
| ISOT           | 1.000    | 1.000    | 1.000           | 1.000 | 1.000   |
| WELFake        | 0.995    | 0.996    | 0.996           | 0.996 | 1.000   |
| LIAR (binary)  | 0.590    | 0.611    | 0.205           | 0.307 | 0.633   |
| LIAR (6-class) | 0.205    | 0.042    | 0.205           | 0.070 | 0.496   |

Результати проведених експериментів типу 1 демонструють суттєву різницю в продуктивності на різних наборах даних. Так, датасети ISOT і WELFake досягають майже ідеальної класифікації:

Як приклад, що підтверджує ефективність запропонованого підходу, наведено пряме порівняння: модель, навчена виключно на наборі ISOT, досягає показника F1 на рівні 0,5% за умови перенесення на набір WELFake, тоді як контрастивна об'єднана модель, навчена на тих самих даних, але з використанням запропонованої багатозадачної мети, досягає показника F1 на рівні 98,4% на наборі WELFake, що становить абсолютне поліпшення на 97,9 процентних пунктів порівняно з єдиною уніфікованою моделлю.

#### 4 Результати моделювання

Експерименти типу 1 проводилися з метою отримання базових показників для кожного окремого набору даних. У табл. 3 наведено результати тестування за п'ятьма визначеними параметрами в межах домену для моделі DeBERTa-v3-base, налаштованої окремо на кожному наборі даних.

DeBERTa отримує 100% балів на ISOT і 99,5% на WELFake, що узгоджується з попередніми результатами, відповідно до яких LightGBM досягає 99,8% F1 на ISOT [5]. Ці тестові

набори фактично досягли реальної межі можливостей.

На противагу цьому, бінарний алгоритм LIAR досягає лише 59,0% точності з 30,7% F1, що демонструє невід'ємну складність визначення правдивості тверджень із стислих політичних заяв (медіана 134 символи). Варіант із 6 класами має точність 20,5% (трохи вище 16,7% випадкової базової лінії). У цьому разі модель деградує до прогнозування одного класу більшості. Ці результати підтверджують, що фейкові новини залишаються невиявленими навіть за допомогою найсучасніших трансформерів.

Унаслідок проведення експериментів типу 2 було сформовано матрицю перенесення між наборами даних, що є основним емпіричним внеском цього дослідження. Результати експериментів узагальнено в табл. 4 та подано у вигляді теплової карти на рис. 1. У матриці перенесення між наборами даних наведені на основній діагоналі узагальнені результати оцінки F1 демонструють продуктивність у межах домену.

**Таблиця 4.** Матриця перенесення між наборами даних (оцінка F1)

| Тренування / тест | ISOT         | LIAR         | WELFake      |
|-------------------|--------------|--------------|--------------|
| ISOT              | <b>1.000</b> | 0.403        | 0.239        |
| LIAR              | 0.057        | <b>0.307</b> | 0.005        |
| WELFake           | 0.105        | 0.522        | <b>0.996</b> |

**Таблиця 5.** Продуктивність контрастивної об'єднаної моделі

| Набір даних   | Точність | Прецизія | Відтворюваність | F1    | AUC-ROC |
|---------------|----------|----------|-----------------|-------|---------|
| ISOT          | 0.974    | 0.998    | 0.953           | 0.975 | 1.000   |
| LIAR (binary) | 0.639    | 0.601    | 0.549           | 0.574 | 0.672   |
| WELFake       | 0.982    | 0.978    | 0.990           | 0.984 | 0.997   |

Контрастивна об'єднана модель демонструє значне покращення продуктивності порівняно з усіма базовими моделями перенесення між наборами даних (див. рис. 2). Найбільш разючі покращення спостерігаються саме там, де базові моделі зазнавали найбільших невдач: LIAR→WELFake покращується з 0,005 до 0,984 (+97,9% абсолютного F1), LIAR→ISOT – з 0,057 до 0,975 (+91,8%), а WELFake→ISOT – з 0,105 до 0,975 (+87,0%). Ця єдина модель зберігає високу ефективність у межах домену: 97,5% F1 на ISOT (–2,5% порівняно зі спеціалізованим базовим показником), 98,4% на WELFake (–1,2%) та 57,4% на LIAR

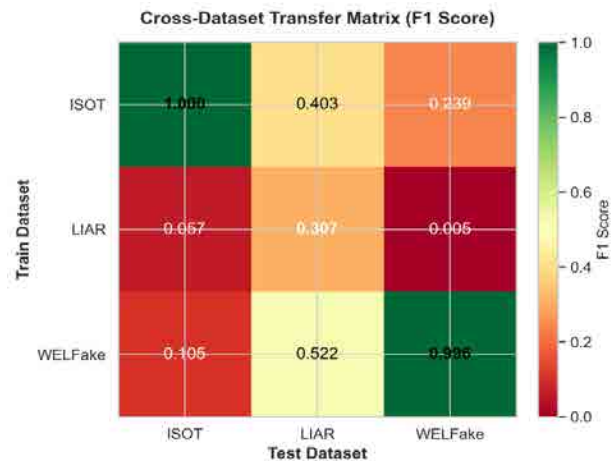


**Рис. 1.** Теплова карта матриці перенесення між наборами даних

Отримана узагальнена матриця перенесення свідчить про катастрофічну невдачу узагальнення. Показник F1 моделі ISOT знижується зі 100% до 40,3% на наборі LIAR і до 23,9% на наборі WELFake. Модель LIAR дає 5,7% F1 на ISOT і 0,5% на WELFake – гірше, ніж випадковий вибір. Погіршення якості перенесення коливається від 47,5% (WELFake→LIAR) до 98,5% (LIAR→WELFake).

У процесі проведення експериментів типу 3 здійснено адаптаційне навчання контрастивної об'єднаної моделі. У табл. 5 подано узагальнені результати щодо продуктивності контрастивної об'єднаної моделі, оціненої для кожного набору даних окремо.

(+26,7 процентних пунктів порівняно з базовим показником для одного набору даних).

Експерименти типу 4 були орієнтовані на дослідження абляції. Абляція в аналізі фейкових новин є важливим методом оцінювання впливу окремих ознак на ефективність моделі. Вона передбачає послідовне видалення певних компонентів, як-от текстові характеристики, джерело інформації чи заголовок, щоб визначити їх значущість. Завдяки цьому дослідники можуть зрозуміти, які саме фактори найбільше впливають на точність виявлення неправдивих повідомлень. Абляційний аналіз також дає змогу встановити зайві або малоефективні ознаки,

що не мають суттєвого впливу на результат. Це сприяє оптимізації моделей, роблячи їх простішими, швидшими та більш інтерпретованими.

Отже, використання абляції підвищує прозорість і надійність систем автоматичного розпізнавання фейкових новин.

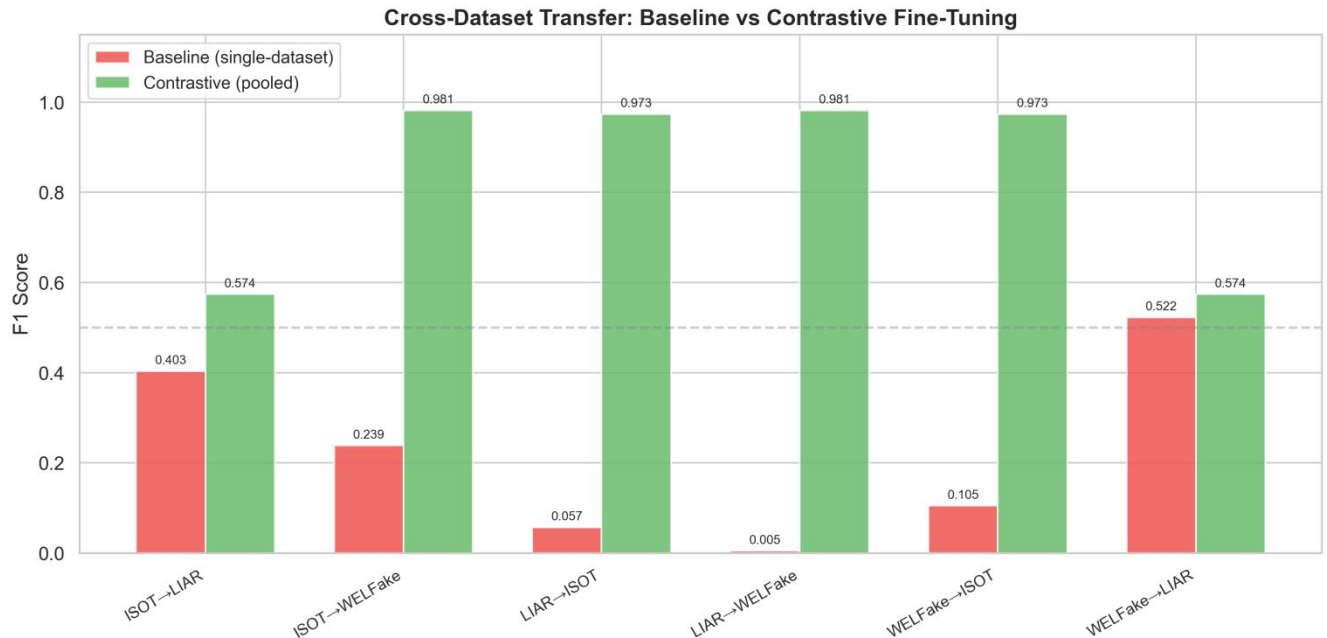


**Рис. 2.** Порівняння показника F1 між наборами даних для всіх пар перенесення: червоні стовпці позначають базові результати перенесення для одного набору даних; зелені стовпці – результати контрастивної об’єднаної моделі

У табл. 6 і 7 подано результати систематичної абляції залежно від контрастного коефіцієнта  $\beta$  і температури  $\tau$  відповідно.

**Таблиця 6.** Абляція: контрастна вага  $\beta$  (оцінки F1)

| $\beta$       | ISOT  | LIAR  | WELFake | Середнє |
|---------------|-------|-------|---------|---------|
| 0.0 (CE only) | 0.973 | 0.585 | 0.978   | 0.845   |
| 0.3           | 0.971 | 0.573 | 0.978   | 0.841   |
| 0.5           | 0.975 | 0.607 | 0.973   | 0.852   |
| 1.0           | 0.973 | 0.574 | 0.980   | 0.842   |

**Таблиця 7.** Абляція: температура  $\tau$  (оцінки F1, де  $\beta = 0,5$ )

| $\tau$ | ISOT  | LIAR  | WELFake | Середнє |
|--------|-------|-------|---------|---------|
| 0.05   | 0.972 | 0.566 | 0.975   | 0.838   |
| 0.07   | 0.974 | 0.566 | 0.977   | 0.839   |
| 0.20   | 0.974 | 0.606 | 0.974   | 0.851   |

На підставі аналізу досягнутих результатів можна зробити висновки щодо параметрів абляції. Методи ISOT і WELFake не чутливі до контрастних гіперпараметрів: за умови зміни цих параметрів показник F1 змінюється менше ніж на 0,7% для всіх налаштувань. LIAR демонструє значну чутливість – значення  $\beta = 0,5$  забезпечує найкращий показник

LIAR F1 (0,607), що на 2,2% вищий, ніж за умови об’єднання лише за CE ( $\beta = 0,0$ ,  $F1 = 0,585$ ). Щодо температури, то  $\tau = 0,2$  дає найкращий результат LIAR (0,606). Тенденції абляції зображено на рис. 3. Бачимо, що показники ISOT і WELFake залишаються стабільними, тоді як LIAR демонструє значну чутливість: найкращі результати досягаються за умов  $\beta = 0,5$  і  $\tau = 0,2$ .

Простір вбудовування було проаналізовано візуально. На рис. 4 подано приклад унаочнення у форматі t-SNE вбудовувань [CLS] для базової моделі ISOT (ліворуч) і контрастивної моделі з об’єднаним пулом (праворуч).

На наведеному прикладі бачимо, що в разі застосування базової моделі, навченої тільки на ISOT, набори даних і класи перемішані, а за умови використання контрастивної об’єднаної моделі набір LIAR утворює окремий кластер, тоді як ISOT та WELFake демонструють покращене розділення реальних і фальшивих даних. Отже, базова модель створює безладний простір вбудовування, в якому зразки з усіх наборів даних і класів перемішані між собою. Це підтверджує, що модель, навчена на ISOT, не має корисної структури представлення для інших

наборів даних. Контрастивна модель створює значно більш впорядкований простір: вислови з набору LIAR утворюють окремий кластер, що відображає їх різні

текстові характеристики, тоді як ділянки ISOT і WELFake демонструють більш чітке розмежування між справжніми й фальшивими даними.

#### Hyperparameter Ablation Study

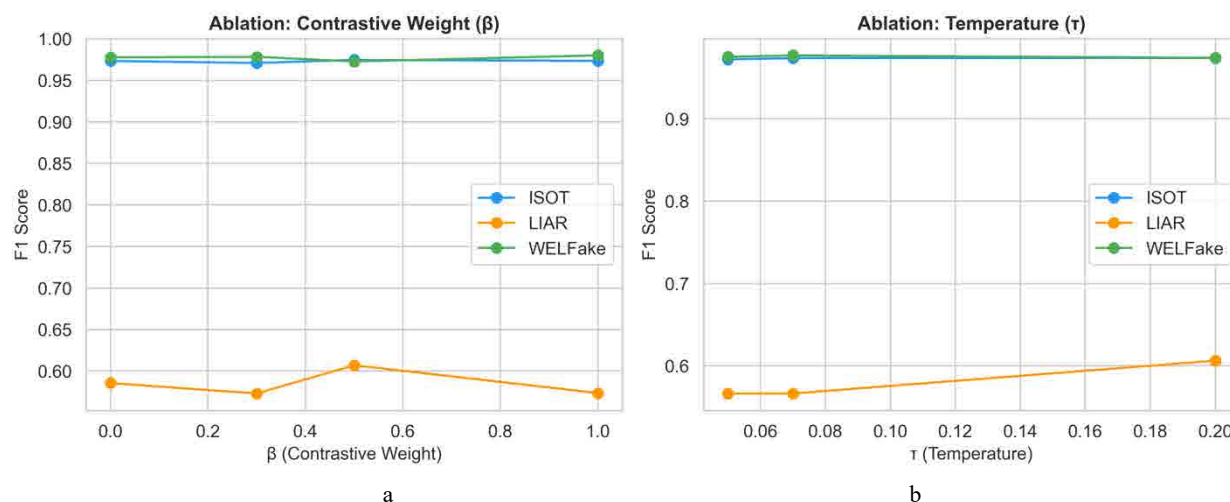


Рис. 3. Абляція залежно від контрастного коефіцієнта  $\beta$  (a) і температури  $\tau$  (b)

#### t-SNE: Embedding Space Comparison

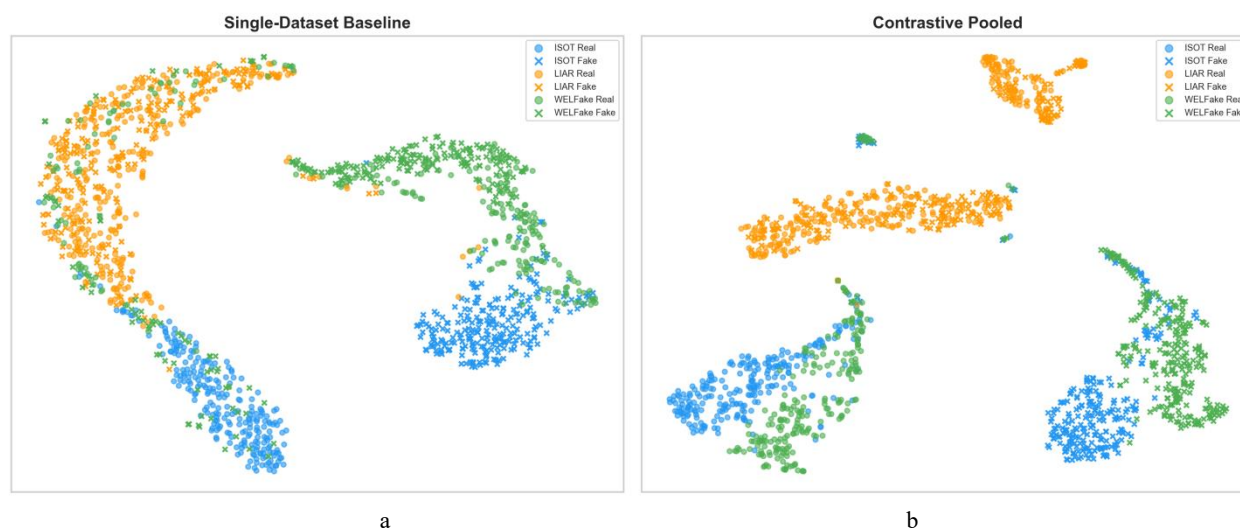


Рис. 4. Візуалізація вбудовувань [CLS] за допомогою t-SNE (кола – справжні дані; хрестики – фальшиві дані):  
a – базова модель, навчена виключно на наборі ISOT; b – контрастивна об'єднана модель

#### Обговорення результатів дослідження

Отримані результати надають кількісні докази того, що контрольні показники для окремих наборів даних створюють глибоко оманливу картину можливостей виявлення. Перехід від 100% F1 (внутрішньодоменний) до 0,5% F1 (міждоменний) для LIAR→WELFake є фундаментальним режимом відмови, а не просто зниження продуктивності. Тому оцінювання наборів даних має стати стандартною практикою в дослідженнях,

присвячених виявленню фейкових новин. Ступінь невдачі перенесення навіть між двома наборами даних повного тексту (ISOT→WELFake: 0,239 F1) свідчить про те, що моделі навчаються розпізнавати джерело, а не виявляти обман. Статті *Reuters* мають певні стилістичні особливості, які слугують майже ідеальним показником "справжності" в ISOT, але ці особливості відсутні в джерелах, що не належать до *Reuters*.

Важливим висновком є те, що об'єднане навчання виключно на даних CE ( $\beta = 0,0$ ) вже

забезпечує більшу частину покращення результатів на інших наборах даних, тоді як контрастивне навчання надає додаткове покращення для складних доменів, як-от LIAR. Це свідчить про те, що доступ до різноманітних навчальних прикладів є основним чинником узагальнення, тоді як контрастивна задача допомагає вдосконалити репрезентації саме для тих доменів, де класифікаційний сигнал є найслабшим.

Результати дослідження дають змогу розширити підхід до корпоративної безпеки, де інформаційні ризики розглядаються нарівні з фінансовими чи операційними, а запропоновані інструменти для виявлення фейкових новин можуть стати невід'ємним складником стратегічного управління бізнесом. Сучасна галузева практика навчання моделей на єдиному наборі даних та їх впровадження, як це роблять соціальні мережі, агрегатори новин і системи модерації контенту, у цьому разі є принципово хибною. Запропонована матриця перенесення кількісно оцінює масштаб ризику: система, що демонструє 100% точність під час внутрішнього оцінювання, може працювати майже на рівні випадковості, коли стикається з новинами із джерел, не поданих у її навчальних даних.

Запропонований підхід контрастивного об'єднаного адаптаційного навчання відповідає сучасним потребам, створюючи єдину модель, яку можна впровадити в новинах різних сфер. Замість того, щоб підтримувати окремі моделі для різних типів контенту (великі статті, стислі політичні заяви, новини з різних джерел), організації можуть застосувати одну уніфіковану модель з високою ефективністю у всіх протестованих сферах. Це зменшує інженерну складність, витрати на інфраструктуру й ризик виникнення сліпих зон, пов'язаних із конкретною сферою, у процесах модерації контенту.

Методологія дедуплікації в запропонованому підході також має практичну цінність для організацій, що створюють навчальні набори даних із багатьох публічних джерел. Наприклад, виявлення 38 285 статей, що дублюються між двома широко використовуваними еталонами, демонструє, що забруднення наборів даних є реальною проблемою, яка може непомітно завищувати показники продуктивності.

Крім того, результати абляції дають практичні рекомендації щодо впровадження: контрастна вага  $\beta = 0,5$  і температура  $\tau = 0,2$  забезпечують оптимальну міждоменну генералізацію, а висновок про те,

що лише об'єднане навчання дає змогу отримати більшу частину переваг, означає, що організації, які не мають ресурсів для створення інфраструктури контрастного навчання, все одно можуть досягти значного поліпшення генералізації за допомогою простого об'єднання даних із різних джерел.

Проведене дослідження обмежується наборами даних англійською мовою, оскільки багатомовна генералізація створює додаткові виклики. Дедуплікація WELFake скорочує набір даних до статей, які не належать до *Reuters*, що потенційно змінює рівень складності. Бінаризація LIAR зводить до нуля значущі градації, що потенційно може спричинити шум міток. Контрастивний підхід вимагає одночасного доступу до декількох наборів даних під час навчання, що не завжди є можливим в умовах обмеженої конфіденційності.

## Висновки

У статті розглянуто систематичне дослідження перехресного узагальнення наборів даних у виявленні фейкових новин, продемонстровано, що оцінка одного набору даних створює принципово спотворену картину прогресу. Побудована матриця перенесення виявляє катастрофічну невдачу узагальнення: досягнення 100% F1 на базі DeBERTa-v3 падає до 0,5% F1 за умови перенесення на різні набори даних із середньою деградацією, що перевищує 75%. Для розв'язання цієї проблеми запропоновано контрастивне об'єднане адаптаційне навчання. Метод поєднує перехресну ентропійну класифікацію з контрольованим контрастивним навчанням одночасно на кількох наборах даних, що надає такі досягнення: 97,5% F1 на ISOT, 57,4% на LIAR та 98,4% на WELFake – із покращенням до 97,9% абсолютного F1 порівняно з базовими перенесеннями. Аналіз абляції підтверджує, що контрастна мета особливо корисна для більш складних доменів класифікації.

**Наукова новизна.** У роботі вперше застосовано контрольоване контрастивне навчання саме для міжнаборової адаптації доменів у виявленні фейкових новин на основі тексту. На відміну від попередніх контрастивних підходів щодо міжмодального вирівнювання між текстом і зображеннями, а також від діагностичних досліджень між наборами даних, які виявляють проблему узагальнення без пропозиції рішень, це дослідження поєднує кількісне оцінювання



проблеми за допомогою контрольованої матриці перенесення  $N \times N$  із конкретною стратегією навчання, що породжує єдину модель, ефективну в різних новинних доменах. Додатковим методологічним внеском є виявлення й вилучення статей, що дублюються між наборами даних, тобто усунення забруднення даних, яке має значення для всіх майбутніх досліджень із використанням цих бенчмарків.

**Практична цінність** запропонованого підходу виходить за межі академічних бенчмарків: інструменти виявлення фейкових новин, здатні до узагальнення між доменами, стають невід'ємною частиною стратегічного управління бізнесом і логічно вписуються в ширший контекст трансформацій сучасного інформаційного середовища.

**Напрями подальшої роботи** передбачають розширення фреймворку на додаткові бенчмарки й багатомовні набори даних, вивчення більш складних методів адаптації предметної галузі (змагальне навчання, метанавчання), дослідження взаємодії між упровадженням лінгвістичних ознак і порівняльним навчанням, а також вивчення того, як продуктивність масштабується залежно від кількості об'єднаних наборів даних.

#### Конфлікт інтересів

Автори заявляють, що не мають конфлікту інтересів, зокрема фінансового, особистого,

авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

#### Фінансування

Це дослідження було проведено без фінансової підтримки.

#### Доступність даних

Дані будуть надані за обґрунтованим запитом. Набори даних ISOT, LIAR і WELFake, використані в цьому дослідженні, є загальнодоступними у відповідних оригінальних джерелах [27–29]. Дедуплікований варіант WELFake та експериментальний код будуть надані за обґрунтованим запитом.

#### Використання засобів штучного інтелекту

DeepL Translator (DeepL GmbH, неймережева модель машинного перекладу на основі трансформерів) використовувався для створення попередніх перекладів розд. 1 та 3 з англійської на українську, а також для перекладу анотації з української на англійську. Весь перекладений текст автори перевірили вручну й відредагували для забезпечення термінологічної точності та стилістичної узгодженості з мовою перекладу.

#### References

1. Herasymov, S., Tkachov, A., Bazarnyi, S. (2024), "Complex method of determining the location of social network agents in the interests of information operations", *Advanced Information Systems*, Vol. 8, No. 1, pp. 31–36. DOI: <https://doi.org/10.20998/2522-9052.2024.1.04>
2. Podorozhniak, A., Liubchenko, N., Oliinyk, V., Roh, V. (2023), "Research application of the spam filtering and spammer detection algorithms on social media and messengers", *Advanced Information Systems*, Vol. 7, No. 3, pp. 60–66. DOI: <https://doi.org/10.20998/2522-9052.2023.3.09>
3. Allcott, H., Gentzkow, M. (2017), "Social media and fake news in the 2016 election", *Journal of Economic Perspectives*, Vol. 31, No. 2, pp. 211–236. DOI: <https://doi.org/10.1257/jep.31.2.211>
4. Bakir, V., McStay, A. (2018), "Fake News and The Economy of Emotions", *Digital Journalism*, Vol. 6, No. 2, pp. 154–175. DOI: <https://doi.org/10.1080/21670811.2017.1345645>
5. Datsenko, S. (2025), "Comparative Study of Machine Learning Approaches for Fake News Detection on Social Platforms", *2025 IEEE 6th KhPI Week on Advanced Technology (KhPIWeek)*, pp. 1–5. DOI: <https://doi.org/10.1109/KhPIWeek61436.2025.11288402>
6. Christodoulou, C., Salamanos, N., Leonidou, P., Papadakis, M., Sirivianos, M. (2023), "Identifying misinformation on YouTube through transcript contextual analysis with Transformer models", *arXiv preprint*, arXiv: 2307.12155. DOI: <https://doi.org/10.48550/arXiv.2307.12155>
7. Datsenko, S. (2025), "Neural architecture comparison for fact verification on FEVER dataset", *Control, Navigation and Communication Systems*, Vol. 3, No. 81m pp. 72–75. DOI: <https://doi.org/10.26906/SUNZ.2025.3.072>

8. Shen, Y., Liu, Q., Guo, N., Yuan, J., Yang, Y. (2023), "Fake news detection on social networks: A survey", *Applied Sciences*, Vol. 13, No. 21, article 11877. DOI: <https://doi.org/10.3390/app132111877>
9. Capuano, N., Fenza, G., Loia, V., Nota, F. D. (2023), "Content-Based Fake News Detection With Machine and Deep Learning: a Systematic Review", *Neurocomputing*, Vol. 530, pp. 91–103. DOI: <https://doi.org/10.1016/j.neucom.2023.02.005>
10. Nickerson, R. S. (1998), "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises", *Review of General Psychology*, Vol. 2, No. 2, pp. 175–220. DOI: <https://doi.org/10.1037/1089-2680.2.2.175>.
11. Ozbay, F. A., Alatas, B. (2020), "Fake news detection within online social media using supervised artificial intelligence algorithms", *Physica A: Statistical Mechanics and Its Applications*, Vol. 540, article 123174. DOI: <https://doi.org/10.1016/j.physa.2019.123174>
12. Abualigah, L., Al-Ajlouni, Y. Y., Daoud, M. S., Altalhi, M., Migdady, H. (2024), "Fake news detection using recurrent neural network based on bidirectional LSTM and GloVe", *Social Network Analysis and Mining*, Vol. 14, article 40. DOI: <https://doi.org/10.1007/s13278-024-01198-w>
13. Sastrawan, I. K., Bayupati, I. P. A., Arsa, D. M. S. (2021), "Detection of fake news using deep learning CNN-RNN based methods", *ICT Express*, Vol. 8, pp. 396–408. DOI: <https://doi.org/10.1016/j.ict.2021.10.003>
14. Nasir, J. A., Khan, O. S., Varlamis, I. (2021), "Fake news detection: A hybrid CNN-RNN based deep learning approach", *International Journal of Information Management Data Insights*, Vol. 1, No. 1, article 100007. DOI: <https://doi.org/10.1016/j.jjimei.2020.100007>
15. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019), "BERT: Pre-training of deep bidirectional transformers for language understanding", *Proceedings of NAACL-HLT*, pp. 4171–4186. DOI: <https://doi.org/10.18653/v1/N19-1423>
16. Liu, Y., Ott, M., Goyal, N. et al. (2019), "RoBERTa: A Robustly Optimized BERT Pretraining Approach", *arXiv preprint*, arXiv: 1907.11692. DOI: <https://doi.org/10.48550/arXiv.1907.11692>
17. He, P., Gao, J., Chen, W. (2023), "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing", *Proceedings of ICLR*. DOI: <https://doi.org/10.48550/arXiv.2111.09543>
18. Hoy, N., Koulouri T. (2022), "Exploring the Generalisability of Fake News Detection Models", *Proceedings of IEEE International Conference on Big Data*, pp. 5731–5740. DOI: <https://doi.org/10.1109/BigData55660.2022.10020583>
19. Gruensteidl, M. N., Kirrane, S. (2026), "A Comparison and Critical Reflection of Information Disorder Detection Techniques: Performing a Cross-Data and Cross-Model Evaluation", *Information Fusion*, Vol. 127, Part B, article 103806. DOI: <https://doi.org/10.1016/j.inffus.2025.103806>
20. Chen, T., Kornblith, S., Norouzi, M., Hinton, G. (2020), "A Simple Framework for Contrastive Learning of Visual Representations", *Proceedings of ICML*. DOI: <https://doi.org/10.48550/arXiv.2002.05709>
21. Khosla, P., Teterwak, P., Wang, C. et al. (2020), "Supervised Contrastive Learning", *Proceedings of NeurIPS*, Vol. 33, pp. 18661–18673. DOI: <https://doi.org/10.48550/arXiv.2004.11362>
22. Gao, T., Yao, X., Chen, D. (2021), "SimCSE: Simple Contrastive Learning of Sentence Embeddings", *Proceedings of EMNLP*, pp. 6894–6910. DOI: <https://doi.org/10.18653/v1/2021.emnlp-main.552>
23. Gunel, S., Du, J., Conneau, A., Stoyanov, V. (2021), "Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning", *Proceedings of ICLR*. DOI: <https://doi.org/10.48550/arXiv.2011.01403>
24. Bhattacharjee, A., Kumarage, T., Moraffah, R., Liu H. (2023), "ConDA: Contrastive Domain Adaptation for AI-generated Text Detection", *Proceedings of IJCNLP-AACL*, pp. 598–610. DOI: <https://doi.org/10.18653/v1/2023.ijcnlp-main.40>
25. Yan, F., Zhang, M., Wei, B., Ren, K., Jiang, W. (2024), "SARD: Fake news detection based on CLIP contrastive learning and multimodal semantic alignment", *Journal of King Saud University – Computer and Information Sciences*, Vol. 36, article 102160. DOI: <https://doi.org/10.1016/j.jksuci.2024.102160>
26. Shen, X., Huang, M., Hu, Z., Cai, S., Zhou, T. (2024), "Multimodal Fake News Detection with Contrastive Learning and Optimal Transport", *Frontiers in Computer Science*, Vol. 6, article 1473457. DOI: <https://doi.org/10.3389/fcomp.2024.1473457>
27. ISOT Research Lab (2024), "ISOT Fake News Dataset", University of Victoria, available at: <https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/> (last accessed 20.04.2026).
28. Wang, W. Y. (2017), "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection", *Proceedings of ACL*, pp. 422–426. DOI: <https://doi.org/10.18653/v1/P17-2067>
29. Verma, P. K., Agrawal, P., Amorim, I., Prodan, R. (2021), "WELFake: Word Embedding Over Linguistic Features for Fake News Detection", *IEEE Transactions on Computational Social Systems*, Vol. 8, No. 4, pp. 881–893. DOI: <https://doi.org/10.1109/TCSS.2021.3068519>

Received (Надійшла) 22.03.2026

Accepted for publication (Прийнята до друку) 16.04.2026

Publication date (Дата публікації) 29.05.2026

*Відомості про авторів / About the Authors*

**Даценко Сергій Сергійович** – Національний технічний університет "Харківський політехнічний інститут", аспірант кафедри комп'ютерної інженерії та програмування, Харків, Україна;

**Serhii Datsenko** – National Technical University "Kharkiv Polytechnic Institute", Postgraduate Student of the Computer Engineering and Programming Department, Kharkiv, Ukraine;

e-mail: sergdacenko@gmail.com

ORCID ID: <https://orcid.org/0000-0001-9514-0433>

**Кучук Георгій Анатолійович** – доктор технічних наук, професор, Національний технічний університет "Харківський політехнічний інститут", професор кафедри комп'ютерної інженерії та програмування, Харків, Україна;

**Heorhii Kuchuk** – Doctor of Technical Sciences, Professor, National Technical University "Kharkiv Polytechnic Institute", Professor of Computer Engineering and Programming Department, Kharkiv, Ukraine;

e-mail: kuchuk56@ukr.net

ORCID ID: <http://orcid.org/0000-0002-2862-438X>

## CROSS-DATASET GENERALIZATION FOR FAKE NEWS DETECTION USING CONTRASTIVE FINE-TUNING

The needs of modern industrial business are increasingly challenged by the spread of fake news, which can significantly affect the reputation, customer trust, and financial performance of industrial companies. Existing detection systems demonstrate nearly perfect accuracy on individual benchmark datasets; however, they suffer catastrophic failure when applied to information from different sources, raising serious concerns about their readiness for real-world deployment. The subject of the study is methods for fake news detection in structurally diverse datasets. The aim of the study is to develop and validate an approach for fake news detection in datasets with structurally diverse content using contrastive fine-tuning, enabling the creation of a unified transformer-based model. The objectives of the study include: establishing baseline performance indicators for individual datasets using the DeBERTa-v3-base language model on three different fake news datasets; constructing a systematic cross-dataset transfer matrix to quantitatively evaluate generalization failures; proposing and evaluating contrastive joint fine-tuning as a domain adaptation approach; and conducting ablation studies of contrastive hyperparameters.

**Research method.** The DeBERTa-v3-base language model is fine-tuned on three fake news datasets: ISOT, LIAR, and WELFake. The combined training objective integrates cross-entropy classification with supervised contrastive loss on merged datasets using balanced sampling across different datasets. **Results.** The cross-dataset transfer matrix demonstrates catastrophic generalization failure: a model achieving 100% F1 on ISOT attains only 23.9% on WELFake and 40.3% on LIAR. The proposed contrastive joint model produces a unified model that simultaneously achieves 97.5% F1 on ISOT, 57.4% on LIAR, and 98.4% on WELFake, with an absolute F1 improvement of up to 97.9% compared to baseline transfers. **Conclusions.** Ablation analysis confirms that the contrastive objective is particularly beneficial in more challenging domains. Contrastive joint fine-tuning enables the development of models suitable for deployment in multi-source news environments.

**Keywords:** fake news detection; cross-dataset generalization; transformer language model; transfer learning; natural language processing.

*Бібліографічні описи / Bibliographic descriptions*

Даценко С. С., Кучук Г. А. Міжвибіркова узагальненість для виявлення фейкових новин із використанням контрастивного адаптаційного навчання. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 153–164. DOI: <https://doi.org/10.30837/0135-1710.2026.189.153>

Datsenko, S., Kuchuk, H. (2026), "Cross-Dataset Generalization for Fake News Detection Using Contrastive Fine-Tuning", *Management Information System and Devices*, No. 2 (189), P. 153–164. DOI: <https://doi.org/10.30837/0135-1710.2026.189.153>

Кириченко І. В., Демчук В. Г., Луценко В. В.

## РОЗРОБЛЕННЯ Й ДОСЛІДЖЕННЯ ПІДХОДУ ДО ВИЯВЛЕННЯ АНОМАЛІЙ У ПОТОКОВИХ КОНВЕЄРАХ ДАНИХ НА ОСНОВІ ОПЕРАЦІЙНОЇ ТЕЛЕМЕТРІЇ

**Об'єктом вивчення** є процес операційного моніторингу, діагностики й забезпечення функціональної надійності розподілених поточкових конвеєрів даних у режимі реального часу; **предметом** – методи й алгоритми автоматизованого виявлення аномалій у поточкових системах на основі аналізу багатовимірних часових рядів операційної телеметрії (метрик продуктивності та використання ресурсів) інфраструктурних компонентів Apache Spark Structured Streaming і Apache Kafka. **Мета дослідження** полягає в розробленні та експериментальній валідації легковагового підходу проактивного виявлення аномалій у швидкісних конвеєрах даних, яка функціонує виключно на основі метапоказників інфраструктури без ресурсомісткої інспекції корисного навантаження, для мінімізації часу реакції на інциденти й усунення додаткових затримок в обробленні даних. **Досягнуті результати.** У процесі дослідження розроблено архітектуру системи моніторингу й сформовано 14-вимірний вектор простору ознак, який передбачає нормалізовані значення системних метрик, швидкості їх зміни (градієнти) й синтетичні безрозмірні коефіцієнти (ефективність оброблення, нормований лаг). Для класифікації станів системи застосовано ансамблевий алгоритм машинного навчання. Експериментальне моделювання типових збоїв (сплески затримки, падіння пропускної здатності, аномальний лаг) на хмарному кластері AWS підтвердило високу ефективність підходу. Запропонований багатовимірний підхід підвищив точність виявлення інцидентів (F1-score) з 0.62 (показник класичного Rule-based-методу на основі статичних порогів) до 0.92 за рівня хибних спрацювань (FPR) лише 0.8%. Середній час виявлення аномалії (MTTD) було скорочено зі 115 с до 25 с. Обчислювальні накладні витрати мікросервісу моніторингу становили менше ніж 1.5% процесорного часу кластера. **Висновки.** Експериментально доведено, що аналіз багатовимірної операційної телеметрії за допомогою методів машинного навчання є високоефективним проксі-індикатором "здоров'я" конвеєрів даних. Запропонований підхід успішно розв'язує питання важкої валідації, притаманної традиційним інструментам перевірки якості даних, і повністю відповідає парадигмі сучасної Shift-Left-архітектури. Рішення забезпечує глибоку спостережуваність із нульовим впливом на продуктивність, є надійною першою лінією проактивного захисту й створює технологічне підґрунтя для реалізації механізмів автоматичного самовідновлення інфраструктури.

**Ключові слова:** потокове оброблення даних; конвеєри даних; виявлення аномалій; операційна телеметрія; машинне навчання; Isolation Forest; Apache Kafka; Apache Spark; Shift-Left-архітектура; спостережуваність систем.

### Вступ

У сучасному ландшафті цифрової трансформації значення аналітики в реальному часі перейшла з категорії конкурентної переваги в статус фундаментальної вимоги для виживання бізнесу. Здатність організацій обробляти, аналізувати й приймати рішення на основі даних у момент їх генерації (протягом мілісекунд або секунд) визначає ефективність критичних операційних процесів. Відповідно до показників IDC за 2025 р., понад 63% корпоративних сценаріїв використання потребують оброблення інформації протягом лічених хвилин, щоб вона залишалася корисною для бізнесу [1]. У таких доменах, як фінансовий моніторинг, затримка в аналізі транзакції може призвести до незворотних втрат: шахрайські операції мають блокуватися до завершення авторизації, що вимагає наскрізної затримки менше ніж 100 мс. Аналогічно у сферах

динамічного ціноутворення для ритейлу або сервісів замовлення поїздок, а також у персоналізації клієнтського досвіду в реальному часі будь-яке відставання аналітичного конвеєра безпосередньо корелює із втраченим доходом або деградацією лояльності користувачів [2].

Однак забезпечення надійності та якості даних у таких швидкісних потоках стикається з серйозним технологічним бар'єром. Традиційні методи перевірки якості даних, подані такими інструментами, як Great Expectations, AWS Deequ, Soda або Apache Griffin, виявляються надто "важкими" для сучасних стрімів. Основна проблема полягає в тому, що ці фреймворки базуються на парадигмі сканування самих даних – вони розраховують статистичні профілі, перевіряють розподіли й валідують кожне поле безпосередньо в потоці або в мікробатчах. Такий підхід неминуче збільшує затримку оброблення, оскільки вимагає додаткових обчислювальних ресурсів і доступу

до вмісту повідомлень, що часто стає bottleneck для всього конвеєра. У системах, де критичним є кожен мілісекундний інтервал, упровадження важких валідаторів вмісту нівелює переваги потокового оброблення [1, 3].

### Аналіз останніх досліджень і публікацій

Розвиток сучасних цифрових екосистем призвів до безпрецедентного зростання обсягів даних, що генеруються в режимі реального часу. Для оброблення таких масивів інформації галузевим стандартом стало використання розподілених систем, де Apache Kafka є магістраллю для транспортування повідомлень, а Apache Spark – основним рушієм для їх оброблення й аналізу. Проте складність і динамічність таких архітектур роблять їх вразливими до аномалій на різних рівнях конвеєра: від інфраструктурних збоїв до логічних помилок у потоковому обробленні. У цьому розділі систематизовано наявні підходи, інструменти й виявлені обмеження, визначено дослідницьку прогалину, яку покликаний заповнити запропонований підхід.

Класичні статистичні методи – Z-score та ковзне середнє – залишаються базовими інструментами виявлення відхилень у метричних потоках. Zhong et al. (2023) [5] у систематичному огляді методів виявлення аномалій у часових рядах у контексті AIOps розглядають ковзне вікно з метою обчислення локальних параметрів розподілу як стандартний підхід для виявлення аномалій у KPI-метриках хмарних сервісів. Автори зазначають, що Z-score й похідні від нього методи добре адаптуються до поступового дрейфу даних завдяки ковзному обчисленню середнього й стандартного відхилення, проте фундаментально обмежені одновимірним аналізом і не беруть до уваги кореляції між метриками різних компонентів розподіленої системи. Огляд демонструє, що базові статистичні детектори оперують статичними порогами й не здатні моделювати складні поведінкові патерни, властиві гетерогенним стрімінговим конвеєрам.

Аналогічний підхід із використанням алгоритму CUSUM (Cumulative Sum Control Chart) для виявлення змін у метриках інфраструктури хмарних і граничних середовищ досліджено в роботі Skaperas et al. (2024) [6]. Автори експериментально оцінюють CUSUM і баєсівські детектори точок змін

на реальних тестових стендах граничних обчислень і виявляють принципові компроміси: CUSUM демонструє обчислювальну ефективність і придатність для онлайн-виявлення, проте вимагає індивідуального налаштування параметрів для кожного типу метрики. Це унеможливляє автоматичне масштабування на гетерогенні конвеєри даних із динамічними навантаженнями, властивими для Apache Kafka і Spark.

У масштабному порівняльному дослідженні Schmidl et al. (2022) [7] оцінено алгоритми виявлення аномалій у часових рядах на понад 1 тис. наборах даних, зокрема з методами на основі прогнозування за алгоритмом Хольта–Вінтерса. Попри задовільну точність на регулярних і сезонних робочих навантаженнях, дослідження демонструє, що прогнозні методи суттєво деградують за умови різких нерегулярних змін, властивих для подійно-орієнтованих архітектур стрімінгових систем. Крім того, аналіз проводиться переважно в офлайн-режимі, що принципово обмежує застосовність методів у сценаріях виявлення аномалій у реальному часі.

Значного поширення набули методи машинного навчання без учителя. Алгоритм Isolation Forest застосовувався в низці сучасних досліджень завдяки своїй обчислювальній ефективності та стійкості до шуму. Islam et al. (2024) [8] запропонували промислове дослідження аномалій на реальних хмарних системах IBM Cloud, упроваджуючи методи машинного навчання, зокрема з Isolation Forest, до масиву телеметричних даних із багатьма ознаками. Ключовим виявленим обмеженням є необхідність повторного перенавчання або перелаштування моделі за умови зміни операційних патернів – типової ситуації для промислових Spark-кластерів під час планових і позапланових змін конфігурації.

Метод LOF (Local Outlier Factor) та його потокові варіанти розглянули Zamanzadeh Darban et al. (2024) [9] у ґрунтовному огляді методів глибокого навчання для виявлення аномалій у часових рядах. Автори систематизують як класичні методи (разом із LOF та його інкрементальними модифікаціями), так і підходи глибокого навчання, аналізуючи їх застосовність до потокових сценаріїв. Огляд демонструє, що методи на основі густини, зокрема LOF, стикаються з проблемами масштабованості в багатовимірних просторах ознак і не моделюють кореляційних залежностей між метриками різних рівнів системи – критично важливий аспект для архітектур типу Kafka + Spark.



Автоенкодери та їх варіанти (VAE) активно застосовуються для виявлення аномалій у багатовимірних метричних потоках. *Jacob i Diao* (2025) [10] пропонують підхід DIVAD (*Domain-Invariant VAE for Anomaly Detection*) – архітектуру варіаційного автоенкодера з механізмом інваріантності до доменного зсуву – й оцінюють його на бенчмарку Exathlon, побудованому на реальних метриках Spark-кластерів, DIVAD досягає покращення пікового F1-показника на 20% порівняно з найкращими конкурентами, проте вимагає значних обчислювальних ресурсів для онлайн-інференсу й не тестувався на метриках, притаманних для Kafka-брокерів. *Wang et al.* (2024) [11] розробляють FCVAE – частотно-розширений варіаційний автоенкодер для моніторингу вебсистем – і демонструють переваги частотного підходу для метрик із гетерогенними часовими патернами, але визнають обмеження в одночасному моделюванні довгоперіодичних і короткоперіодичних тенденцій.

Apache Kafka є де-факто стандартом побудови поточкових конвеєрів даних, і дослідження, присвячені його операційним властивостям, активно розвиваються. *Mayer et al.* (2020) [12] пропонують спеціалізовану архітектуру моніторингу для Apache Kafka на основі JMX-метрик і проводять комплексний бенчмарк продуктивності брокера, вимірюючи пропускну здатність і затримку під різними конфігураціями. У дослідженні автори документують набір ключових метрик брокера, необхідних для вчасного виявлення деградації, та наголошують на ризику перетворення Kafka на вузьке місце архітектури в разі недостатнього моніторингу. Утім запропонована система обмежується порогами, налаштованими вручну, і не бере до уваги контекст downstream-споживачів (зокрема Spark Structured Streaming), що є критичним для розуміння першопричини відхилення.

Питання латентності та пропускну здатності конвеєра Kafka – Spark як джерела операційних метрик досліджено в роботі *Fedorovych et al.* (2024) [13]. Автори фіксують ключові метрики конвеєра: пропускну здатність, затримку на рівні брокера й воркерів Spark, вплив кількості Kafka-партицій на продуктивність Spark-задач. Результати демонструють, що мікробатч-режим забезпечує більш стабільну затримку ( $\approx 197$  мс із Kafka-джерелом) порівняно з безперервним режимом у сценаріях високого навантаження, а перевищення

пропускну здатності Kafka над можливостями Spark-воркерів є індикатором накопичення затримок. Однак підхід є описовим і не пропонує алгоритму автоматичного виявлення аномалій.

Дослідження операційних метрик Apache Spark концентруються переважно на характеристиці та бенчмаркуванні конвеєрів. *Jacob et al.* (2021) [14] систематично документують набір метрик Spark-кластера в межах бенчмарку Exathlon: драйвер надає 243 метрики Spark UI (затримку планування, статистику поточкових даних), кожен виконавець – 140 метрик (JVM heap memory, GC overhead, task duration, shuffle read / write). Бенчмарк охоплює шість типів аномальних подій (resource contention, process failures, misbehaving inputs) і забезпечує еталонний набір даних для оцінювання алгоритмів. Проте детекція проводиться в офлайн-режимі на завершених трасах виконання, а не в реальному часі в процесі роботи конвеєра.

Метрика тривалості мікробатчів і scheduling delay в Spark Structured Streaming як індикатори навантаженості конвеєра детально схарактеризовані в роботі *Fedorovych et al.* (2024) [13]. Результати бенчмарку демонструють, що стійке перевищення затримки оброблення над тривалістю мікробатч-інтервалу є надійним провісником деградації конвеєра, а кількість партицій Kafka безпосередньо визначає здатність Spark рівномірно розподіляти навантаження між виконавцями. Підхід є реактивним і не зважає на upstream-стан Kafka, що не дає змоги відрізнити перевантаження споживача від затримок на рівні продюсера.

Аномально високий час GC є передвісником події OOM (*OutOfMemory*). Бенчмарк Exathlon фіксує GC overhead серед ключових метрик виконавців, що корелюють з аномальними подіями в Spark-кластері, зокрема з вичерпанням пам'яті та resource contention [14]. Результати демонструють, що метрики пам'яті JVM у поєднанні з shuffle-метриками формують інформативний сигнатурний набір для класифікації типів аномалій у поточковому обробленні, проте методологія залишається ретроспективною без компоненти автоматичного виявлення аномалій у реальному часі.

Для реалізації методик виявлення аномалій дослідники використовують екосистему інструментів, що забезпечує збір, зберігання та візуалізацію даних у режимі реального часу.

Prometheus є найбільш поширеною платформою збору метрик у середовищах cloud-native. *Pragathi et al.* (2024) [15] досліджують розгортання моніторингового стека Prometheus + Grafana + Node Exporter у Kubernetes-кластері й порівнюють можливості Prometheus і NetData. Prometheus демонструє найкращу інтеграцію з Kubernetes і підтримує Pull-модель збору метрик з інтервалом 15 с, проте базова система сповіщень (Alertmanager) оперує лише порогами й не підтримує складні поведінкові патерни або ML-основані правила виявлення аномалій.

Grafana як платформа візуалізації та побудови дашбордів розглядається в тій самій роботі [15] в контексті операційного моніторингу розподіленої інфраструктури. Автори систематизують найкращі практики побудови дашбордів і наголошують на принциповому обмеженні: Grafana призначена для відображення метрик, а не для автономного виявлення аномалій. Наявні механізми алертингу Grafana оперують порогами або простими арифметичними виразами PromQL і не моделюють кореляцій між метриками різних підсистем.

Інтеграція Prometheus і Kafka у межах єдиного конвеєра моніторингу описана в роботі *Fedorovych et al.* (2024) [13]. Автори будують пайплайн, де Kafka водночас транспортує оброблені дані та приймає результати вимірювань (Kafka sink). Це дає змогу збирати наскрізні метрики затримки від брокера до Spark-виконавця. Незважаючи на практичну цінність підходу, система залишається реактивною: метрики збираються й використовуються для діагностики продуктивності, але не для автономного виявлення аномалій через ML-основані або адаптивні алгоритми.

Apache Spark має вбудовану систему метрик, а бенчмарк Exathlon документує повний набір метрик виконавців. *Jacob et al.* (2021) [14] описують 140 метрик на рівні executor – heap memory, GC overhead, task duration, shuffle metrics – та оцінюють їх придатність для виявлення аномалій на еталонному наборі даних. Дослідження зупиняється на рівні збору й характеристики метрик; компонента автоматичного виявлення аномалій у режимі реального часу відсутня.

Інтегровані рішення, що одночасно охоплюють Kafka і Spark в єдиному контексті виявлення аномалій, у літературі висвітлено недостатньо. Найближчою до цієї теми є робота *Jacob* і

*Diao* (2025) [10], де запропоновано метод DIVAD, оцінений на бенчмарку Exathlon (дані Spark-кластера). Автори зосереджують увагу виключно на метриках Spark-виконавців, тоді як Kafka-брокер поза контекстом розгляду, а аналіз метрик брокера й кореляція зі станом Spark-задач відсутні. *Fedorovych et al.* (2024) аналізують конвеєр Kafka – Spark комплексно з погляду продуктивності, але не пропонують компоненти виявлення аномалій.

Підхід до кореляційного аналізу метрик у мікросервісних архітектурах, релевантний для розглядуваної задачі, описали *Pham et al.* (2024) [16]. Автори проводять масштабну оцінку 21 методу визначення першопричини на основі каузальних графів між метриками різних сервісів, тестуючи їх на двох загальноприйнятих бенчмарках мікросервісів. Результати демонструють, що жоден метод не домінує в усіх сценаріях, а ефективність на синтетичних датасетах не відтворює реальну поведінку у виробничих системах. Метод не адаптований до особливостей stateful-обчислень Spark і брокерської семантики Kafka.

Унаслідок проведеного огляду літературних джерел визначено суттєву прогалину: незважаючи на значну кількість робіт з виявлення аномалій у часових рядах і операційного моніторингу стрімінгових систем, жодне з розглянутих досліджень не пропонує комплексного підходу, що водночас:

- об'єднує метрики двох рівнів потокового конвеєра – брокера повідомлень (Kafka: consumer lag, throughput, partition metrics) і обчислювального рушія (Spark: executor metrics, job duration, shuffle metrics) – в єдиній моделі виявлення аномалій;

- забезпечує виявлення аномалій у режимі реального часу (low-latency inference) без необхідності накопичення великої передісторії або ресурсоемного перенавчання;

- зважає на крос-компонентні кореляції між метриками Kafka і Spark для розрізнення причинно-наслідкових ланцюжків деградацій (upstream bottleneck vs. downstream processing overload);

- адаптується до динамічної зміни патернів навантаження без ручного перелаштування порогів.

Статистичні підходи [5–7] обмежені статичними порогами, переважно офлайн-обчисленнями й одновимірним аналізом. ML-методи [8, 9] не тестувалися на специфічних комбінованих метриках Kafka / Spark і вимагають стабільних патернів навантаження. VAE-основані методи [10, 11]

демонструють значні обчислювальні вимоги для онлайн-інференсу й не охоплюють крос-системних метрик. У роботах з моніторингу Kafka [12, 13] і Spark [14] кожену платформу розглянуто переважно ізольовано. Інструментальні рішення [15] обмежуються порогами й візуалізацією без автономного виявлення аномалій і без кореляційного аналізу між рівнями конвеєра. Підходи до аналізу першопричин не адаптовані до особливостей об'єднаного середовища Kafka + Spark [16].

Отже, запропонований у цій роботі підхід, що ґрунтується на спільному аналізі операційних метрик Spark і Kafka за допомогою адаптивних алгоритмів виявлення аномалій у режимі реального часу, є оригінальним рішенням, що безпосередньо закриває виявлену прогалину.

#### Мета й постановка завдань

У межах цього дослідження під якістю даних у потоковому конвеєрі розуміється здатність системи забезпечувати вчасну, повну й безперервну доставку й оброблення поточкових повідомлень без втрати, деградації або критичного зростання затримки.

Об'єктом вивчення є моніторинг "здоров'я" та функціональної цілісності конвеєрів передачі даних за допомогою аналізу метапоказників або операційної телеметрії самої системи. Замість того, щоб заглядати всередину кожного пакета даних, запропоновано оцінювати стан системи за непрямими ознаками – поведінкою інфраструктурних компонентів Apache Spark і Apache Kafka. Телеметрія, що передбачає метрики пропускну здатності, затримки, інтенсивності запитів і використання ресурсів, формує унікальний цифровий відбиток нормальної роботи пайплайну. Будь-яка аномалія в даних – чи то різка зміна схеми, чи то сплеск обсягу, чи то деградація якості на джерелі – неминуче викликає відлуння в операційних метриках системи оброблення [4].

Метою цього дослідження є розроблення легкого підходу виявлення аномалій, який здатний функціонувати в режимі реального часу, аналізуючи виключно телеметрію Spark і Kafka без утручання в самі дані. Такий підхід дає змогу ідентифікувати проблеми в конвеєрах із мінімальними накладними витратами, забезпечуючи високу спостережуваність і швидко реакцію на інциденти без збільшення затримки оброблення.

#### Матеріали й методи дослідження

Apache Kafka – це розподілена платформа потокової передачі подій із відкритим вихідним кодом, розроблена інженерами LinkedIn і передана під управління Apache Software Foundation (2011 р.). Нині Kafka є стандартом у побудові платформ потокової передачі даних масштабу підприємства, охоплюючи понад 80% компаній зі списку Fortune 100 [4].

Ключова концепція Kafka основана на моделі журналу подій: записи зберігаються у впорядкованих незмінних послідовностях – топіках, які розбиваються на партиції. Такий підхід забезпечує паралельне оброблення завдяки розподілу навантаження між партиціями, масштабованість унаслідок горизонтального нарощування кількості брокерів, відмовостійкість через реплікацію партицій між брокерами, а також гарантію черговості повідомлень у межах однієї партиції [4].

З погляду операційного моніторингу Kafka надає набір ключових JMX-метрик: UnderReplicatedPartitions (кількість партицій із реплікацією нижчою від заданого рівня), consumer\_group\_lag (відставання групи споживачів від позиції producer), RequestHandlerAvgIdlePercent (відсоток незайнятості обробників запитів), NetworkProcessorAvgIdlePercent і OfflinePartitionsCount. Аномальні значення перелічених метрик є першими сигналами про деградацію продуктивності конвеєра [17].

Загальну архітектуру потокового конвеєра на основі Apache Kafka й Apache Spark Structured Streaming зображено на рис. 1 [18].

*Перший стовпець (ліворуч) – джерела даних:*

1. комірка 1: IoT / Sensors (~100M events/day) – Дані телеметрії та показники з фізичних пристроїв і датчиків, що генеруються у великих обсягах;
2. комірка 2: Фінансові події (транзакції, FIX) – Інформація про банківські перекази, платежі або торгові операції на біржах;
3. комірка 3: Веблоги (разом) – Записи про дії користувачів на вебсайтах (історія кліків, перегляди сторінок, помилки);
4. комірка 4: Мобільні додатки – застосунки – Аналітика поведінки користувачів, краш-репорти та події з мобільних програм;
5. комірка 5: Системні метрики (сервери, мережа) – Технічні показники навантаження та стану інфраструктури (CPU, пам'ять, мережевий трафік).

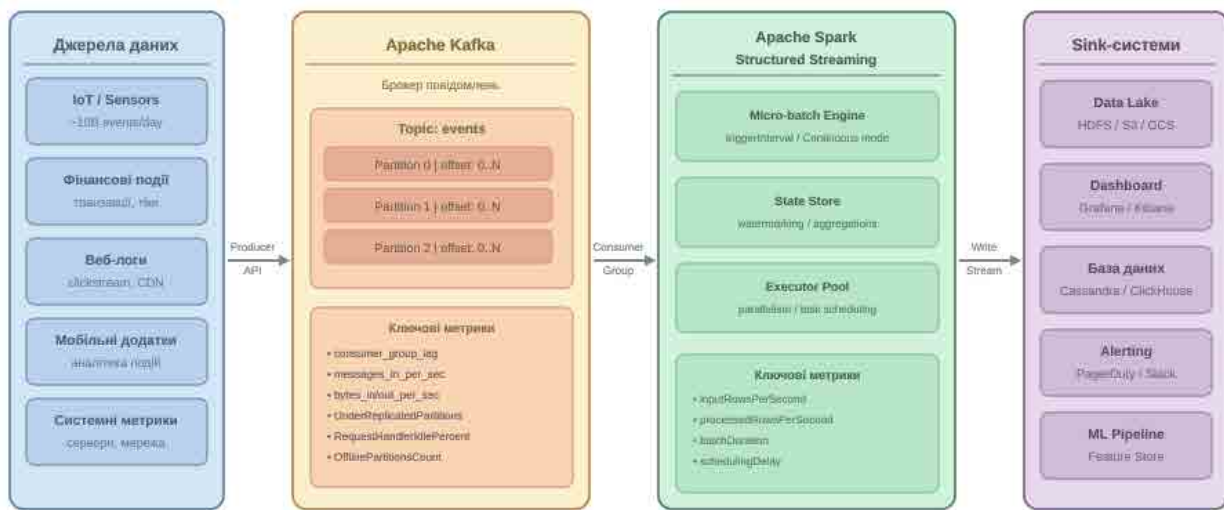


Рис. 1. Загальна архітектура потокового конвеєра на основі Apache Kafka й Apache Spark Structured Streaming

#### Другий стовпець – Apache Kafka:

1. комірка 1: Topic: events (Partition 0, 1, 2) – Логічний канал (топик), у який записуються всі події, розділений на партиції для забезпечення паралельного читання та масштабованості;

2. комірка 2: Ключові метрики – Набір найважливіших показників для моніторингу стану брокера та швидкості обміну повідомленнями (відставання консьюмерів, пропускну здатність, статус реплікації партицій).

#### Третій стовпець – Apache Spark Structured Streaming:

1. комірка 1: Micro-Batch Engine (trigger interval / continuous mode) – Рушій, що обробляє безперервний потік даних невеликими порціями через задані інтервали або в режимі реального часу;

2. комірка 2: State Store (watermarking / aggregations) – Механізм збереження проміжних результатів, який дозволяє робити агрегації за часовими вікнами та коректно обробляти події, що запізнилися (watermarking);

3. комірка 3: Executor Pool (parallelism / task scheduling) – Пул робочих процесів, які розподіляють між собою завдання для паралельної та швидкої обробки даних;

4. комірка 4: Основні метрики – Показники продуктивності самої обробки: швидкість читання та запису рядків за секунду, тривалість батчу та затримка планування.

#### Четвертий стовпець (праворуч) – Sink-система:

1. комірка 1: Data Lake (HDFS / S3 / GCS) – Сховище для довгострокового збереження великих обсягів оброблених (або сирих) даних у вигляді файлів;

2. комірка 2: Dashboard (Grafana / Kibana) – Інструменти для побудови графіків та візуалізації аналітики в режимі реального часу;

3. комірка 3: База даних (Cassandra / ClickHouse) – Системи керування базами даних (часто NoSQL або колоночні), оптимізовані для швидкого доступу та аналітичних запитів;

4. комірка 4: Alerting (PagerDuty / Slack) – Системи автоматичного сповіщення команд про критичні відхилення, інциденти або досягнення певних тригерів;

5. комірка 5: ML Pipeline (Feature Store) – Інфраструктура машинного навчання, куди дані надходять як розраховані ознаки для тренування або роботи моделей.

Сучасний бізнес дедалі більше покладається на безперервні потоки даних для підтримки прийняття рішень. Перехід від традиційного пакетного оброблення до потокового зумовлений потребою в миттєвій реакції на події. У фінансовому секторі це полягає в упровадженні систем реального часу, як-от FedNow у США або SEPA Instant у Європі, де розрахунки й кліринг відбуваються миттєво, що вимагає відповідної інфраструктури моніторингу з нульовою толерантністю до затримок [17].

Реальна цінність аналітики в реальному часі полягає в переході від реактивного стилю управління ("що сталося вчора?") до проактивного ("що відбувається зараз і як реагувати?"). Проте надійність цих систем прямо залежить від стабільності конвеєрів даних. Якщо пайплайн починає працювати некоректно або затримувати дані, бізнес отримує хибну інформацію або застарілі індикатори, що може



спричинити катастрофічні наслідки в автоматизованих системах прийняття рішень [4, 17, 18].

Проблема традиційних DQ-інструментів, як-от Great Expectations або Soda, полягає в їх архітектурній орієнтації на статистику. Вони розглядають дані як набір записів у таблиці, до яких можна застосувати SQL-запити для перевірки повноти, унікальності або валідності.

Для виконання перевірки за допомогою Great Expectations необхідно [19, 20]:

1) отримати вибірку або повний набір даних із потоку;

2) завантажити ці дані в пам'ять процесу валідації (Pandas, Spark DataFrame);

3) виконати набір правил (Expectations), що часто передбачає обчислювально складні операції, зокрема розрахунок квантилів або перевірку на відповідність регулярним виразам.

У контексті Spark Structured Streaming, де мікробатчі можуть запускатися кожні кілька сотень мілісекунд, додавання такого циклу валідації до кожного батчу призводить до експоненціального зростання часу оброблення. Навіть якщо інструмент використовує механізми push-down (передачу запитів на рівень бази даних або Spark-двигуна), саме сканування рядків створює навантаження на введення / виведення та процесор, що конкурує з основною бізнес-логікою за ресурси кластера.

Іншою критичною вадою традиційних DQ-інструментів є їх "сліпота" до стану інфраструктури. Вони можуть повідомити, що в потоці з'явилися порожні значення, але вони не знають, чи є це наслідком помилки в джерелі даних, чи результатом деградації продуктивності одного з брокерів Kafka, що призвело до часткової втрати повідомлень. Такий розрив між якістю даних та "здоров'ям" системи змушує команди інженерів витрачати години на ручну кореляцію метрик із різних систем моніторингу [19].

У великих розподілених системах, як-от Spark і Kafka, телеметрія забезпечує спостережуваність, що дає змогу розуміти внутрішній стан системи на основі її зовнішніх виходів.

Перевага використання телеметрії для виявлення аномалій полягає в її легкості. Метрики – це зазвичай числові значення (лічильники, гістограми, таймери), збір яких практично не впливає на продуктивність

основної системи. Наприклад, відстеження швидкості вхідних повідомлень у Spark або лагу споживача в Kafka не потребує читання вмісту повідомлень – ці дані вже доступні в пам'яті керівних процесів.

У цій роботі звернено увагу на чотири категорії телеметрії, які в сукупності дають змогу ідентифікувати аномалії даних без їх безпосереднього аналізу.

- Пропускна здатність: кількість записів і обсяг даних у байтах за одиницю часу.

- Затримка: час оброблення батчу, час перебування повідомлення в черзі, наскрізна затримка.

- Споживання ресурсів: завантаження CPU, використання пам'яті, мережева активність під час оброблення конкретних обсягів даних.

- Стан черг і відставання: кількість необроблених повідомлень у Kafka.

Другий порядок інсайтів полягає в тому, що аномалії даних (наприклад, раптова зміна формату з JSON на XML або поява аномально довгих рядків) завжди проявляються як зміни в часі оброблення (addBatch duration у Spark) або зміні коефіцієнта стиснення в Kafka. Отже, системна телеметрія слугує проксі-індикатором якості даних.

В умовах промислової експлуатації поточкові конвеєри неминуче стикаються з різноманітними відхиленнями від нормального режиму роботи. Під аномалією в контексті потокового конвеєра розуміємо статистично значуще відхилення значень однієї або декількох операційних метрик системи від їх базової лінії, яке свідчить про деградацію продуктивності або порушення цілісності даних.

Аномалії в потокових системах мають складний, багатопараметричний характер: жодна окрема метрика зазвичай не дає повної картини про стан системи. Наприклад, зростання batchDuration у Spark може бути спричинене як збільшенням вхідного навантаження (що позначиться на зростанні inputRowsPerSecond), так і внутрішніми проблемами виконавця (Executor) – GC-паузами або вичерпанням пам'яті, що не відображено в метриках Kafka. Класифікацію основних типів аномалій подано в табл. 1 [21].

Характеристичну картину аномалій в операційних метриках потокового конвеєра проілюстровано на рис. 2, де показано два типових відхилення: сплеск затримки й падіння пропускної здатності з одночасним зростанням споживчого лагу.



Таблиця 1. Класифікація аномалій у потокових конвеєрах даних

| Тип аномалії                                   | Прояв / симптоми                                                                        | Причини                                                         | Вплив на систему                                              |
|------------------------------------------------|-----------------------------------------------------------------------------------------|-----------------------------------------------------------------|---------------------------------------------------------------|
| Сплески затримки (Latency Spikes)              | Різке зростання <code>batchDuration</code> , <code>schedulingDelay</code> > норми       | Перевантаження Еxecutor, GC-паузи JVM, мережеві збої            | Порушення SLA, деградація якості сервісу, накопичення лагу    |
| Падіння пропускної здатності (Throughput Drop) | <code>processedRowsPerSecond</code> < <code>inputRowsPerSecond</code> , зростання черги | Back-pressure, вичерпання ресурсів Еxecutor, back-pressure      | Затримка даних, переповнення буферів Kafka, consumer lag      |
| Аномальний лаг споживача (Consumer Lag)        | <code>consumer_group_lag</code> перевищує допустимий поріг                              | Повільний Spark еxecutor, різке зростання вхідного навантаження | Критична затримка реакції системи, переповнення Kafka-топиків |
| Втрата повідомлень (Message Loss)              | Розрив між <code>producer_offset</code> та <code>consumer_offset</code>                 | Некоректна конфігурація <code>acks</code> , збоїв брокера Kafka | Неповні або некоректні дані в downstream-системах             |

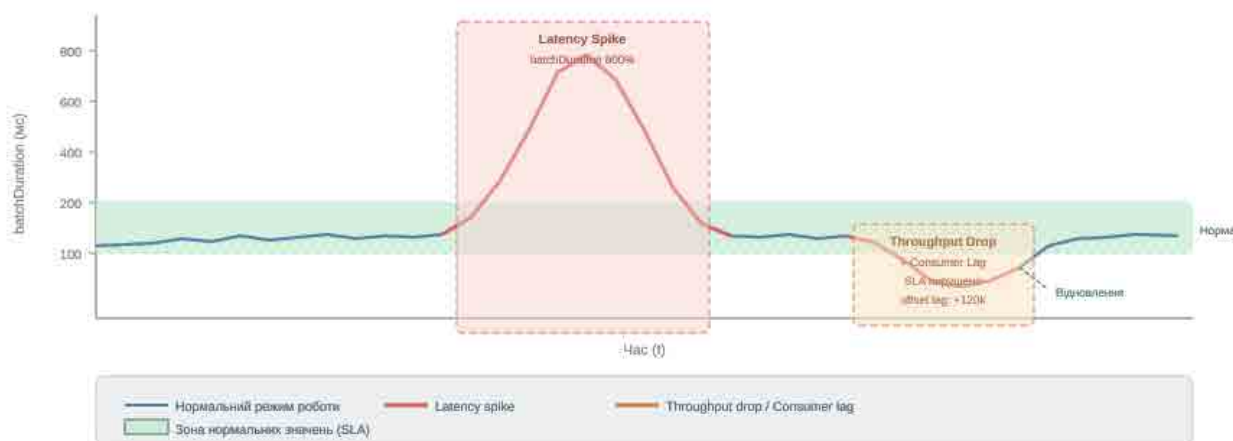


Рис. 2. Характеристичні аномалії в операційних метриках потокового конвеєра даних

Аномалії в потокових конвеєрах безпосередньо впливають на операційні показники організацій, що їх використовують. За інформацією *Gartner* [2], простій критичної інформаційної системи обходиться підприємствам у середньому \$5 600 USD за хвилину або понад \$300 000 за годину. У контексті фінансових платіжних систем навіть короткочасне збільшення латентності оброблення транзакцій з 100 мс до 1 с призводить до конверсійних втрат у розмірі 7% [3].

Порушення угод про рівень обслуговування (SLA) у потокових системах має двоякий характер. По-перше, технічний SLA містить гарантії щодо затримки (наприклад, p99-латентність < 200 мс), пропускної здатності (наприклад, мінімум 500000 подій/с) та доступності системи (наприклад, 99.9% час роботи). По-друге, бізнес SLA – гарантії щодо вчасності та повноти аналітичних звітів, сповіщень і бізнес-рішень, що генеруються на основі потокових даних.

Раннє виявлення аномалій (проактивний моніторинг) дає змогу скоротити MTTD (*Mean Time To Detect*) і MTTR (*Mean Time To Recover*), що

безпосередньо корелює зі зниженням сукупних витрат на ліквідацію наслідків інцидентів і підтримання рівня клієнтської задоволеності.

#### Підхід виявлення аномалій у потокових конвеєрах даних

Запропонований підхід реалізує конвеєрну архітектуру з п'яти послідовних етапів: потокова передача даних через Apache Kafka, оброблення й агрегація метрик у Apache Spark Streaming, збір і попереднє оброблення телеметрії модулем Metric Collector, формування вектора ознак у блоці Feature Engineering і виявлення аномалій алгоритмом Isolation Forest із подальшою генерацією діагностичного сповіщення (рис. 3).

Принципова відмінність від наявних рішень полягає у використанні виключно внутрішніх метрик самої обчислювальної інфраструктури – без залучення зовнішніх агентів якості даних, розмічених навчальних наборів або апіорних знань про структуру аномалій [19, 21].

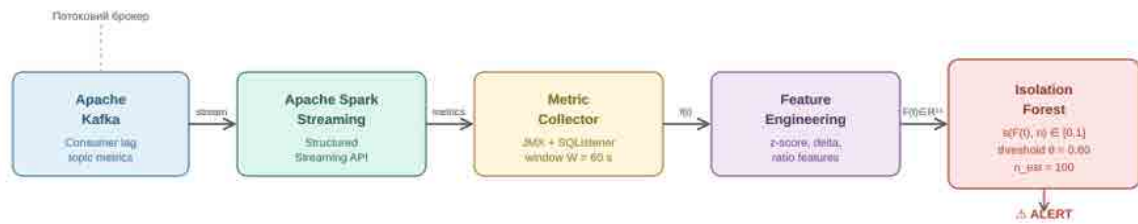


Рис. 3. Виявлення аномалій за допомогою конвеєрної архітектури

Метрики збираються через два незалежних канали. По-перше, метрики Apache Kafka отримуються через JMX Exporter (порт 9309) у форматі Prometheus з інтервалом 15 с; ключовою метрикою є `avgOffsetsBehindLatest` – затримка споживача, який є інтегральним показником відставання оброблення від надходження даних [12]. По-друге, метрики Apache Spark Structured Streaming надходять через `StreamingQueryListener` API: метод `onQueryProgress()` викликається після завершення кожного мікропакета й передає об'єкт `StreamingQueryProgress` із повним

набором статистик поточного циклу [13]. Системні метрики CPU та Memory збираються агентом `node_exporter` (Prometheus) з того самого інтервалу [15].

Критерії відбору метрик:

- доступність через стандартні інтерфейси без додаткової інструментації;
- чутливість до характерних режимів відмови;
- взаємна інформативність за умови кореляційного аналізу.

Детальна характеристика кожної метрики подана в табл. 2 [14, 22].

Таблиця 2. Набір операційних метрик системи виявлення аномалій

| Позначення                          | Джерело            | Фізичний зміст і значення у виявленні аномалій                                                                               | Тип аномалії, що виявляється    |
|-------------------------------------|--------------------|------------------------------------------------------------------------------------------------------------------------------|---------------------------------|
| <code>inputRowsPerSecond</code>     | Spark SQLListener  | Швидкість надходження рядків з Kafka у Spark Streaming. Раптовий спад сигналізує про зупинку Producer або розрив топіка.     | Message loss, Producer failure  |
| <code>processedRowsPerSecond</code> | Spark SQLListener  | Фактична швидкість оброблення. Якщо $x_2 < x_1$ тривало, система накопичує чергу – ознака back-pressure.                     | Back-pressure, Executor failure |
| <code>numInputRows</code>           | Spark SQLListener  | Кількість рядків у поточному мікропакеті. Різкий стрибок вказує на burst-навантаження.                                       | Load burst, skew                |
| <code>batchDuration [мс]</code>     | Spark SQLListener  | Тривалість оброблення мікропакета. Перевищення тривалості тригерного інтервалу призводить до накопичення черги.              | Latency spike, GC pause         |
| <code>triggerExecution [мс]</code>  | Spark durationMs   | Час виконання тригера планувальника. Зростання відбиває затримку планування або перевантаження Driver.                       | Driver overload                 |
| <code>avgOffsetsBehindLatest</code> | Kafka consumer JMX | Середній Kafka consumer lag (відставання споживача від останнього offset). Зростання понад 5000 – критичний сигнал.          | Consumer lag, throughput drop   |
| CPU usage [%]                       | System metrics     | Завантаженість CPU вузлів кластера. Стале значення $> 85\%$ вказує на ресурсне голодування Executor.                         | CPU overload                    |
| Memory usage [%]                    | System metrics     | Рівень використання пам'яті. Перевищення $> 90\%$ призводить до GC-паузи й неминучого зростання <code>batchDuration</code> . | JVM GC spike, OOM               |

Система оперує вектором із восьми базових операційних метрик  $n = 8$ , що охоплюють три виміри поведінки конвеєра: продуктивність оброблення (метрики Spark Streaming), стан споживача (метрика Kafka) й ресурсне споживання (системні метрики).

Ключовою мотивацією для комбінування метрик Spark і Kafka є те, що жодна окрема метрика не несе достатньої інформації для однозначної ідентифікації

аномалії. Наприклад, зростання `batchDuration` ( $x_4$ ) може бути зумовлене як GC-паузою JVM (супроводжується зростанням `Memory usage`), так і back-pressure через перевантаження Kafka Consumer (супроводжується зростанням  $x_6$ ). Розрізнення цих сценаріїв вимагає одночасного аналізу декількох метрик, що обґрунтовує необхідність багатовимірного підходу [10, 16].

Формування вектора ознак і підготовка даних має подання у вигляді багатовимірної часової ряду. Потік операційних метрик моделюється як дискретний багатовимірний часовий ряд із фіксованим кроком  $\Delta t = 15$  с. У кожний момент часу  $t$  стан конвеєра описується вектором спостережень:

$$X_{(t)} = [x_1(t), x_2(t), \dots, x_n(t)]^T \in \mathbb{R}^n, \quad t \in \{1, 2, \dots, T\}, \quad (1)$$

де  $n = 8$  – кількість метрик. Відповідно, за  $T$  кроків формується матриця спостережень  $X \in \mathbb{R}^{n \times T}$ , яка слугує вхідними даними для подальшого оброблення [10].

**Вектор ознак.** Для кожного моменту  $t$  формується вектор ознак:

$$F_{(t)} = [\bar{x}_1, \dots, \bar{x}_n, \Delta x_1, \dots, \Delta x_n, \eta(t), \lambda(t)]^T \in \mathbb{R}^d, \quad (2)$$

$$d = 2n + 2 = 18,$$

де  $\bar{x}_i$  – нормалізовані значення;  $\Delta x_i$  – першопохідні (швидкість зміни);  $\eta$  – коефіцієнт ефективності оброблення;  $\lambda$  – нормований лаг споживача. Розмірність вектора  $d = 2n + 2 = 18$  [9].

**Z-score нормалізація.** Оскільки метрики мають суттєво різні масштаби й одиниці вимірювання, застосовується z-score-нормалізація за рухомим baseline-вікном тривалістю 24 год:

$$\bar{x}_i(t) = \frac{x_i(t) - \mu_i}{\sigma_i}, \quad x_i \in (-\infty, +\infty), \quad (3)$$

де  $\mu_i$  та  $\sigma_i$  – вибіркові середнє та стандартне відхилення метрики  $x_i$ , обчислені за baseline-вікном. Нормалізація усуває ефект різномасштабності й дає змогу Isolation Forest рівноважно розглядати всі виміри простору ознак. Параметри  $\mu_i$ ,  $\sigma_i$  оновлюються кожні 6 год за умови відсутності зафіксованих аномалій у відповідний проміжок часу [10].

**Ознаки першої похідної (delta features).** Поряд зі значеннями метрик система обчислює швидкість їх зміни:

$$\Delta x_i(t) = x_i(t) - x_i(t-1), \quad t \geq 2. \quad (4)$$

Ознаки  $\Delta x_i(t)$  є чутливими до градієнту процесу й дають змогу виявляти поступові деградації на ранніх стадіях, зокрема повільне зростання batchDuration або затримання протягом кількох годин, яке не перевищує абсолютних порогів, але демонструє стійку позитивну тенденцію [7].

**Ознаки співвідношень.** Для моделювання структурних залежностей між метриками додаються дві безрозмірні ознаки. Коефіцієнт ефективності

оброблення  $\eta(t)$  визначає відношення фактичної продуктивності до вхідного навантаження:

$$\eta(t) = \frac{\text{processedRowsPerSecond}(t)}{\text{inputRowsPerSecond}(t)}, \quad \eta \in (0, 1]. \quad (5)$$

Значення  $\eta(t) \approx 1.0$  відповідає нормальному режиму;  $\eta(t) < 0.7$  є операційним визначенням back-pressure – стану, за якого Spark не встигає обробляти вхідний потік і черга Kafka починає зростати [22].

Нормований Kafka lag  $\lambda(t)$  впроваджується для порівняння величини відставання з поточним вхідним потоком:

$$\lambda(t) = \frac{\text{avgOffsetsBehindLatest}(t)}{\text{inputRowsPerSecond}(t) + \varepsilon}, \quad \varepsilon \rightarrow 0. \quad (6)$$

Висока величина  $\lambda(t)$  за одночасно низького  $\eta(t)$  є сильним сигналом системного перевантаження, тоді як високий  $\lambda(t)$  у разі нормального  $\eta(t)$  може вказувати на короткочасний burst або неправильне партиціонування [13].

Для виявлення аномалій застосовується алгоритм Isolation Forest (IF) – ансамблевий метод навчання без учителя, що ґрунтується на принципі ізоляції.

Обґрунтування вибору IF:

- не потребує розміченого навчального набору;
- має лінійну обчислювальну складність  $O(n \cdot t)$ ;
- є стійким до "прокляття розмірності" порівняно з методами на основі відстані (LOF, kNN);
- здатний виявляти як точкові аномалії, так і контекстні аномалії у просторі ознак  $\mathbb{R}^{18}$  [8].

Принцип роботи Isolation Forest: ансамбль із  $T_{est}$  випадкових дерев ізоляції будується способом рекурсивного розбиття простору ознак уздовж випадково обраної осі та випадково обраного порогового значення. Аномальні спостереження, розташовані в розріджених ділянках простору, потрапляють у листові вузли за значно меншу кількість розбиттів (мала глибина ізоляції  $h(x)$ ), ніж нормальні спостереження в щільних кластерах (велика  $h(x)$ ). Оцінка аномальності обчислюється як

$$s(x, n) = 2^{-\frac{E[h(x)]}{c(n)}}, \quad c(n) = 2H(n-1) - \frac{2(n-1)}{n}, \quad (7)$$

де  $H(i)$  –  $i$ -те гармонічне число;  $E[h(x)]$  – середня глибина ізоляції в ансамблі.

Значення  $s(x, n) \rightarrow 1$  відповідає аномальному спостереженню (мала середня глибина ізоляції),  $s(x, n) \rightarrow 0.5$  – невизначеності,  $s(x, n) \rightarrow 0$  –

нормальному спостереженню. На рис. 4 показано типовий розподіл anomaly score для нормальних та аномальних спостережень у задачі моніторингу потокового конвеєра [23].

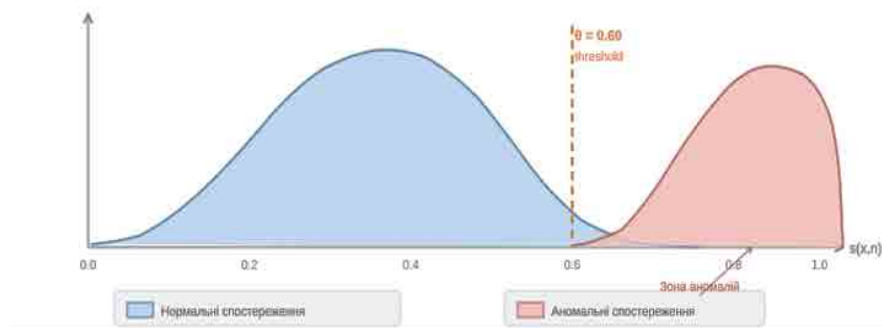


Рис. 4. Розподіл anomaly score Isolation Forest для нормальних і аномальних спостережень; вертикальна пунктирна лінія позначає поріг  $\theta = 0.60$

Параметри навчання:  $n\_estimators = 100$  (кількість дерев ансамблю, достатня для стабілізації оцінки дисперсії),  $contamination = 0.01$  (апріорна частка аномалій у baseline, що відповідає реальним даним продуктивних систем [23]),  $max\_samples = 256$  (розмір субвибірки для кожного дерева, що забезпечує різноманітність ансамблю). Модель навчається впродовж перших семи діб роботи конвеєра (warm-up period) і перенавчається щотижня в офплік-вікні (02:00–04:00) [8].

Бінарне рішення щодо аномальності знімку метрик у момент  $t$  сформулюється через порогову функцію:

$$A_t = \begin{cases} 1, & \text{аномалія, якщо } s(F(t), n) > \theta, \\ 0, & \text{норма, інакше,} \end{cases} \quad (8)$$

де  $\theta$  – поріг аномальності.

Значення  $\theta = 0.60$  обрано як компроміс між хибнопозитивним рівнем і рівнем пропуску аномалій (false negative rate), верифікований на анованому наборі тестів із 42 інцидентами трьох класів (latency spike, throughput drop, consumer lag burst). Налаштування  $\theta$  для конкретного середовища розгортання виконано за методом ROC-аналізу на основі оцінки з кутом нахилу кривої:  $\theta^* = \operatorname{argmax}(Recall - \alpha \cdot FPR)$ , де  $\alpha = 2.0$  позначає операційну вартість хибного спрацювання [6].

Рівень серйозності аномалії визначається лінійно залежно від перевищення порогу:  $severity = (s(F(t), n) - \theta) / (1 - \theta) \in [0, 1]$  за умови

$A_t = 1$ . Значення  $severity < 0.3$  визначається як WARNING,  $severity \geq 0.3$  як CRITICAL, що відповідає різному регламенту реагування операційної команди [15].

Для зниження частоти хибнопозитивних спрацювань у моменти штатних перезапусків конвеєра або планових пікових навантажень упроваджується механізм мінімального часу аномалії: сигнал генерується лише за умови  $A_t = 1$  протягом двох або більше послідовних знімків (тобто  $\geq 30$  с). Такий підхід дає змогу фільтрувати короточасні артефакти в значеннях метрик без втрати чутливості до стійких аномальних режимів [6].

У табл. 3 подано покроковий опис алгоритму системи виявлення аномалій. Алгоритм виконується у вигляді потокового мікросервісу з фіксованим циклом опитування  $\Delta t = 15$  с і затримка від виникнення аномалії до генерації алерту не перевищує 30 с (один або два цикли залежно від тривалості аномалії) [21].

1. Latency spike – це аномалія типу "сплеск затримки", яка визначається різким зростанням  $x_4$  (batchDuration) за відносно стабільного вхідного потоку  $x_1$ . У просторі ознак цей патерн виражено у великих значеннях  $\bar{x}_4$  та  $\Delta x_4$  за умови помірних значень  $\bar{x}_6$ . Зростання batchDuration понад trigger\_interval призводить до того, що наступний мікропакет починається із затримкою, внаслідок чого Kafka consumer lag  $\bar{x}_6$  починає накопичуватися – формується ланцюгова реакція деградації. Isolation

Forest ідентифікує такий патерн через нетипову комбінацію  $(\bar{x}_4, \Delta x_4, \bar{x}_6)$ , яка відсутня в baseline [14].

2. Back-pressure (throughput drop). Цей патерн визначається зниженням  $\eta(t)$  нижче ніж 0.7 за умови одночасного зростання  $\lambda(t)$ . Spark Structured Streaming за замовчуванням обмежує швидкість споживання з Kafka через параметр maxOffsetsPerTrigger: якщо обчислювальна потужність Еxecutor вичерпана, processedRowsPerSecond ( $x_2$ ) падає нижче, ніж inputRowsPerSecond ( $x_1$ ), і черга Kafka починає накопичуватися. У просторі ознак цей режим відповідає одночасно малому  $\eta(t)$ , великому  $\lambda(t)$  і помірному зростанню  $\bar{x}_4$  [22].

3. CPU overload. Тривале перевантаження CPU ( $\bar{x}_7 > 85\%$ ) безпосередньо впливає на продуктивність

JVM і може спричинити GC-паузи, виражені в стрибкоподібному зростанні batchDuration. Особлива властивість: одночасне зростання  $\bar{x}_7$  і  $\bar{x}_4$  за відносно стабільного  $\lambda(t)$  [14].

4. Consumer lag burst – аномальний сплеск лагу споживача. Цей тип аномалії виявляється у швидкому зростанні avgOffsetsBehindLatest / consumer\_group\_lag за умов, коли швидкість надходження повідомлень перевищує здатність Spark-конвеєра їх обробляти [14].

Без системних метрик цей сценарій важко відрізнити від аномалії типу skewed partitioning, де зростання batchDuration пов'язане з нерівномірним розподілом задач між Еxecutor, а не з ресурсним голодуванням.

Таблиця 3. Покроковий алгоритм системи виявлення аномалій

| Крок | Назва          | Операція                                                                                             | Деталі реалізації                                                                                                     |
|------|----------------|------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| 1    | Збір метрик    | Отримати $X(t)$ з Kafka JMX та Spark StreamingQueryListener                                          | Інтервал збору 15 с; Prometheus scrape + onQueryProgress callback; зберігати в pandas DataFrame із timestamp-індексом |
| 2    | Sliding window | Агрегувати $X(t)$ у вікні $W = 60$ с                                                                 | rolling(window=4, min_periods=4).agg(['mean', 'std', 'max']); крок = 15 с $\rightarrow$ 4 знімки на вікно             |
| 3    | Нормалізація   | Обчислити $\bar{x}_i(t) = (x_i - \mu_i) / \sigma_i$ по baseline 24 год                               | sklearn.preprocessing.StandardScaler, fit на перших 24 год; часткове дооновлення кожні 6 год                          |
| 4    | Похідні ознаки | Обчислити $\Delta x_i(t) = x_i(t) - x_i(t-1)$                                                        | df.diff(1); для першого кроку $\Delta x_i = 0$                                                                        |
| 5    | Ratio-ознаки   | $\eta(t) = x_2 / x_1$ ; $\lambda(t) = x_6 / (x_1 + \varepsilon)$                                     | np.divide з epsilon=1e-6; кліпувати $\eta \in [0, 1]$                                                                 |
| 6    | Вектор ознак   | $F(t) = [\bar{x}_1 \dots \bar{x}_6, \Delta x_1 \dots \Delta x_6, \eta, \lambda] \in \mathbb{R}^{14}$ | pd.concat по осі стовпців; dtype float32                                                                              |
| 7    | Навчання IF    | Навчити IsolationForest на baseline $F_0$ (7 діб)                                                    | $n\_estimators = 100$ , $contamination = 0.01$ , $max\_samples = 256$ ; joblib.dump для збереження моделі             |
| 8    | Scoring        | $s(t) = model.score\_samples(F(t))$                                                                  | sklearn повертає від'ємне; перетворення: $s\_pos = -s + 0.5$ ; $s\_pos \in [0, 1]$                                    |
| 9    | Рішення        | $A_t = 1$ , якщо $s(t) > \theta = 0.60$ , інакше 0                                                   | Емпіричний поріг $\theta$ ; налаштування через precision-recall на annotated subset                                   |
| 10   | Алерт          | Генерувати сповіщення з діагностикою                                                                 | Логувати аномалію; Grafana annotation API; $severity = f(s(t) - \theta)$                                              |

### Експеримент

Для емпіричної валідації запропонованого підходу розгорнуто ізольований тестовий конвеєр даних у хмарному середовищі Amazon Web Services (AWS), що імітує типову промислову архітектуру

потокowego оброблення в парадигмі Shift-Left. Інфраструктура передбачала:

- Apache Kafka (v.3.4.0): керований кластер із 3 брокерів (тип інстансів m5.xlarge), 1 топік transactions\_stream із 12 партиціями та фактором реплікації 3;



- Apache Spark (v.3.4.1): кластер під управлінням YARN, що містив 1 Driver-вузол і 3 Worker-вузли (тип r5.xlarge, 4 vCPU, 32 GB RAM кожен). Мікробатч тригер (trigger interval) застосунок Structured Streaming було налаштовано на 500 мс;

- система моніторингу та інференсу: Prometheus для збору JMX і Node Exporter метрик (інтервал  $\Delta t = 15$  с). Модуль розрахунку ознак і алгоритм Isolation Forest було розгорнуто як окремий мікросервіс на базі Python (бібліотека scikit-learn).

Джерелом даних був синтетичний генератор подій, який емулював потік фінансових транзакцій (формат JSON, середній розмір повідомлення становив 1.2 КБ). Базове навантаження було налаштовано на рівні 25000 повідомлень за секунду з нормально розподіленою дисперсією  $\pm 15\%$ , що імітувало природні коливання трафіку.

Оскільки в промислових системах аномалії є відносно рідкісними подіями, для об'єктивного оцінювання алгоритму експеримент було поділено на дві фази. Перша фаза (Warm-up period) тривала сім діб: конвеєр працював у нормальному режимі без збоїв для накопичення "чистих" baseline-даних і навчання ансамблю Isolation Forest відповідно до методології.

Друга фаза (Testing phase) тривала 72 год (три доби), протягом яких у систему було штучно інжектровано 45 аномальних інцидентів (по 15 кожного класу за інформацією з табл. 1).

1. Latency Spikes (сплески затримки): імітувалися способом короткочасного утилізування процесорного часу (до 100% CPU) на Executor-вузлах за допомогою утиліти stress-ng та ініціації примусових GC-пауз.

2. Throughput Drop (Back-pressure): імітувалися впорскуванням "важких" повідомлень (вкладені масиви великого розміру), що збільшували час десеріалізації в Spark, не змінюючи в цьому разі вхідну швидкість потоку ( $x_i$ ).

3. Consumer Lag Burst (аномальний лаг): імітувалися мікросплесками генерації даних до 70000 подій/с на 3–5 хв, що значно перевищувало обчислювальну ємність кластера.

Для оцінювання якості застосовувалися класичні метрики: Precision (точність), Recall (повнота), F1-Score та FPR (False Positive Rate). Для визначення операційної ефективності вимірювався MTTD (Mean Time To Detect) – середній час від фактичної ін'єкції аномалії до генерації алерту ( $A_i = 1$ ).

Запропонована багатовимірна модель (Multivariate Isolation Forest) порівнювалася з класичним підходом (Rule-based Baseline), що використовував статичні порогові моніторингу (алерт генерувався, якщо CPU > 90%, batchDuration > 2000 мс або consumer\_lag > 5000 протягом 30 с).

## Результати

Експериментальне тестування підтвердило гіпотезу про те, що операційна телеметрія дає змогу ефективно ідентифікувати деградацію потокового конвеєра даних. За умови заданого порогу аномальності  $\theta = 0.60$  запропонована система успішно виявила 42 з 45 штучно згенерованих інцидентів. Зведені результати порівняння підходів виявлення аномалій подано в табл. 4.

Таблиця 4. Ефективність виявлення аномалій: порівняння запропонованого підходу з базовим (Rule-based)

| Підхід виявлення аномалій        | Precision | Recall | F1-Score | FPR (хибні спрацювання) | Середній час виявлення (MTTD) |
|----------------------------------|-----------|--------|----------|-------------------------|-------------------------------|
| Rule-based (статичні пороги)     | 0.76      | 0.53   | 0.62     | 4.2%                    | 115 с                         |
| Multivariate IF (запропонований) | 0.91      | 0.93   | 0.92     | 0.8%                    | 25 с                          |

Запропонований підхід на основі Isolation Forest продемонстрував значну перевагу над базовим підходом (F1-Score 0.92 проти 0.62).

- Найбільший розрив у продуктивності спостерігався в сценаріях Throughput Drop (Recall = 0.93 для IF). Оскільки вхідний потік залишався стабільним, а швидкість оброблення падала, ознака ефективності оброблення  $\eta(t)$

миттєво опускалася нижче ніж 0.7. Статичні правила пропускали ці інциденти майже 2 хв, очікуючи, поки лаг досягне жорсткого порогу.

- У сценаріях Latency Spikes завдяки ознакам першої похідної  $\Delta x_4$  система фіксувала стійкий градієнт зростання batchDuration до того, як він перетинав абсолютні критичні межі.

Середній час виявлення (MTTD) для запропонованого підходу становив 25 с. Зважаючи на вікно агрегації ( $\Delta t = 15$  с) і механізм фільтрації (підтвердження протягом двох тактів), це означає, що алгоритм розпізнає структурні зміни в метриках майже миттєво після їх появи. Статичні правила вимагали в середньому 115 с для фіксації проблеми. Обчислювальні накладні витрати процесу вилучення телеметрії становили  $< 1.5\%$  процесорного часу кластера, що повністю підтверджує легкість розробленого підходу порівняно з перевіркою корисного навантаження.

### Обговорення

Високий показник F1-Score (0.92) і низький рівень хибних спрацювань (FPR = 0.8%) стали можливими виключно завдяки 14-вимірному вектору ознак  $F(t)$ . Як продемонстрував експеримент, ізольований аналіз окремих метрик не дає змоги відрізнити легітимний сплеск трафіку від системної деградації. Упровадження синтетичних безрозмірних ознак – коефіцієнта ефективності  $\eta(t)$  і нормованого лагу  $\lambda(t)$  – надало алгоритму Isolation Forest здатність до контекстного розуміння. Система навчилася розрізняти, що зростання черги Kafka є нормою під час пропорційного сплеску вхідного трафіку (Load Burst), але є критичною аномалією, якщо трафік незмінний, а пропускна здатність впала.

Традиційні системи перевірки якості даних, орієнтовані на пакетне оброблення, вимагають десеріалізації кожного повідомлення. За швидкості 25000 подій/с це створює експоненційне навантаження на I/O і процесор. Запропонований підхід, аналізуючи телеметрію системи (JMX та Listener метрики), виконується асинхронно (out-of-band). Її обчислювальна складність є константою  $O(1)$  залежно від обсягу корисного навантаження. Це повністю усуває вплив моніторингу на наскрізну затримку оброблення (end-to-end latency).

Незважаючи на високу швидкість реакції, цей підхід має концептуальні обмеження.

По-перше, підхід є сліпим до прихованих семантичних аномалій у даних. Якщо генератор почне надсилати некоректні Kafka-записи, де певні поля мають неправильні значення (зберігаючи

правильну схему та правильні типи), обчислювальне навантаження на Spark залишиться стабільним, і телеметрія не зафіксує відхилень. Отже, розроблений підхід не замінює інструменти Data Quality, а діє як Tier-1 для захисту інфраструктури.

По-друге, алгоритм вразливий до проблеми холодного старту: після кожного архітектурного оновлення або масштабування кластера (scale-out) модель генеруватиме хибні сповіщення й потребуватиме періоду перенавчання (warm-up) для адаптації до нової норми.

### Висновки

#### й перспективи подальшого розвитку

У зв'язку з масовим переходом бізнесу на архітектури реального часу (Real-Time Analytics) забезпечення безперервності роботи потокових конвеєрів даних стає фундаментальною вимогою. У запропонованій роботі розв'язано науково-прикладне завдання автоматизованого виявлення аномалій у розподілених потокових системах на базі Apache Spark і Apache Kafka. Зокрема, виконано певні завдання.

1. Розроблено легковаговий (lightweight) підхід до моніторингу, оснований на вилученні телеметрії. Сформовано 14-вимірний простір ознак, що містить нормалізовані значення метрик продуктивності, швидкості їх зміни (градієнти) та структурні співвідношення між входом і обробленням  $\eta, \lambda$ .

2. Доведено високу точність виявлення. Експериментальна апробація за допомогою алгоритму Isolation Forest на моделі промислового стенда довела його ефективність. Якщо порівнювати з традиційними статичними правилами, запропонований підхід підвищив точність класифікації (F1-score) з 0.62 до 0.92.

3. Мінімізовано час реакції. Застосування машинного навчання до операційних метрик дало змогу скоротити середній час виявлення інциденту (MTTD) з 115 с до 25 с за умови збереження нульового впливу на швидкість оброблення даних у конвеєрі. Це повністю відповідає сучасній парадигмі Shift-Left архітектури.

Запропонований підхід органічно інтегрується в сучасні парадигми DataOps. Перспективи подальшого дослідження лежать у площині переходу від "виявлення" до "самовідновлення" (auto-remediation). Інтеграція виходів моделі з планувальниками

ресурсів (наприклад, Kubernetes HPA) дасть змогу конвеєрам автоматично масштабувати Executor-вузли для запобігання перевантаженням. Крім того, перспективним є дослідження моделей глибокого навчання (Deep Learning), зокрема LSTM-Autoencoders, для кращого врахування довгострокових сезонних залежностей складних систем.

#### Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

#### Фінансування

Дослідження проводилося без фінансової підтримки.

#### Доступність даних

Рукопис не має пов'язаних матеріалів у сховищі даних.

#### Використання засобів штучного інтелекту

Автор підтверджує, що не застосовував технології штучного інтелекту для написання роботи.

#### References

- Goedegebuure, A., Kumara, I., Driessen, S. (2024), "Data Mesh: A Systematic Gray Literature Review", *ACM Computing Surveys*, Vol. 57, pp. 1–36. DOI: <https://doi.org/10.1145/3687301>
- Sulaiman, S., Rizwan, M. (2025), "Decentralizing Data for Larger Organizations: A Data Mesh Approach", *Journal of Computational Analysis & Applications*, Vol. 34, pp. 177–186. DOI: <https://doi.org/10.48047/jocaaa.2025.34.11.15>
- Borodii, I., Osukhivska, H. (2026), "Research on the efficiency of data loading and storage in Data Lakehouse architectures for the formation of analytical data systems", *Information Technology: Computer Science, Software Engineering and Cyber Security*, No. 4, pp. 28–36. DOI: <https://doi.org/10.48550/arXiv.2604.21449>
- Sudhanshubhai, P. J. (2026), "The Evolution of Shift-Left Testing in Modern Software Development", *Journal of Computational Analysis & Applications*, Vol. 35, pp. 1091–1100. DOI: <https://doi.org/10.48047/jocaaa.2026.35.01.88>
- Zhong, Z. (2023), "A Survey of Time Series Anomaly Detection Methods in the AIOps Domain", *arXiv preprint arXiv:2308.00393*, pp. 1–36. DOI: <https://doi.org/10.48550/arXiv.2308.00393>
- Skaperas, S., Koukis, G., Kapetanidou, I. A., Tsaoussidis, V., Mamatas, L. (2024), "A Pragmatical Approach to Anomaly Detection Evaluation in Edge Cloud Systems", *Proc. IEEE INFOCOM Workshops*, pp. 1–6. DOI: <https://doi.org/10.48550/arXiv.2401.07717>
- Schmidl, S., Wenig, P., Papenbrock, T. (2022), "Anomaly Detection in Time Series: A Comprehensive Evaluation", *Proc. VLDB Endow.*, vol. 15, no. 9, pp. 1779–1797. DOI: <https://doi.org/10.14778/3538598.3538602>
- Islam, M. S., Rakha, M. S., Pourmajidi, W., Sivaloganathan, J., Steinbacher, J., Miranskyy, A. (2024), "Anomaly Detection in Large-Scale Cloud Systems: An Industry Case and Dataset", *arXiv preprint arXiv:2411.09047*, pp. 1–12. DOI: <https://doi.org/10.48550/arXiv.2411.09047>
- Darban, Z. Z., Webb, G. I., Pan, S., Aggarwal, C. C., Salehi, M. (2024), "Deep Learning for Time Series Anomaly Detection: A Survey", *ACM Comput. Surv.*, vol. 57, no. 2, Art. no. 39, pp. 1–42. DOI: <https://doi.org/10.1145/3691338>
- Jacob, V., Diao, Y. (2025), "Unsupervised Anomaly Detection in Multivariate Time Series across Heterogeneous Domains", *arXiv preprint arXiv:2503.23060*, pp. 1–21. DOI: <https://doi.org/10.48550/arXiv.2503.23060>
- Wang, Z., et al. (2024), "Revisiting VAE for Unsupervised Time Series Anomaly Detection: A Frequency Perspective", in *Proc. ACM Web Conf. (WWW)*, pp. 1–10. DOI: <https://doi.org/10.48550/arXiv.2402.02820>
- Mayer, R., Mayer, L., Laich, L. (2020), "How Fast Can We Insert? An Empirical Performance Evaluation of Apache Kafka", in *Proc. IEEE Int. Conf. Cloud Eng. (IC2E)*, pp. 186–196. DOI: <https://doi.org/10.48550/arXiv.2003.06452>
- Fedorovych, I., Osukhivska, H., Lutsyk, N. (2024), "Performance Benchmarking of Continuous Processing and Micro-Batch Modes in Spark Structured Streaming", in *Proc. ITAP, CEUR Workshop Proc.*, vol. 3896, paper 5, pp. 1–11.
- Jacob, V., Song, F., Stiegler, A., Diao, Y., Tatbul, N. (2021), "Exathlon: A Benchmark for Explainable Anomaly Detection over Time Series", *Proc. VLDB Endow.*, vol. 14, no. 11, pp. 2613–2626. DOI: <https://doi.org/10.14778/3476249.347630>

15. Pragathi, B. C., Maddirala, H., Sneha, M. (2024), "Implementing an Effective Infrastructure Monitoring Solution with Prometheus and Grafana", *Int. J. Comput. Appl.*, vol. 186, no. 38, pp. 7–14. DOI: <https://doi.org/10.5120/ijca2024923873>
16. Pham, L., Ha, H., Zhang, H. (2024), "Root Cause Analysis for Microservice System Based on Causal Inference: How Far Are We?", *Proc. 39th IEEE/ACM Int. Conf. Autom. Softw. Eng. (ASE)*, pp. 1–13. DOI: <https://doi.org/10.48550/arXiv.2408.13729>
17. Vysotska, V., Kyrychenko, I., Demchuk, V. (2025), "Competency module of shift-left architecture in big data", in *Proc. PhD Workshop on Artificial Intelligence in Computer Science at CoLInS*, pp. 1–15. DOI: <https://doi.org/10.31110/COLINS/2025-3/003>
18. Bhardwaj, A. K. (2026), "Integrating Security Early A Shift-Left Model for DevSecOps in Modern Software Pipelines", *2026 6th International Conference on Image Processing and Capsule Networks (ICIPCN)*. DOI: <https://doi.org/10.1109/ICIPCN67432.2026.11438475>
19. Vysotska, V., Kyrychenko, I., Demchuk, V., Gruzdo, I. (2024), "Holistic Adaptive Optimization Techniques for Distributed Data Streaming Systems", *CEUR Workshop Proc.*, vol. 3624, pp. 1–13. DOI: <https://doi.org/10.31110/COLINS/2024-2/009>
20. Mandala, N. R. (2021), "ETL in Data Lakes vs. Data Warehouses", *ESP J. Eng. Technol. Adv.*, vol. 1, pp. 224–230. DOI: <https://doi.org/10.56472/25832646/JETA-V1I2P123>
21. Vysotska, V., Kyrychenko, I., Demchuk, V. (2025), "Adaptive Issue Detection in Holistic Optimization of Distributed Data Streaming Systems", in *Proc. MoDaST 2025: Modern Data Science Technologies Doctoral Consortium*, Lviv, Ukraine, pp. 1–17. URL: <https://ceur-ws.org/Vol-4005/paper15.pdf>.
22. Sharma, M. (2025), "Streaming Queries: Enabling Real-Time Elastic Scaling in Modern Applications", *Journal of Computer Science and Technology Studies*, Vol. 3, pp. 319–326. DOI: <https://doi.org/10.32996/jcsts.2025.7.3.36>
23. Chandola, V., Banerjee, A., Kumar, V. (2009), "Anomaly detection: A survey", *ACM Comput. Surv.*, vol. 41, no. 3, Art. no. 15, pp. 1–58. DOI: <https://doi.org/10.1145/1541880.1541882>

Received (Надійшло) 18.03.2026

Accepted for publication (Прийнята до друку) 12.05.2026

Publication date (Дата публікації) 29.05.2026

*Відомості про авторів / About the Authors*

**Кириченко Ірина Віталіївна** – кандидат технічних наук, Харківський національний університет радіоелектроніки, доцент кафедри програмної інженерії, Харків, Україна;

**Iryna Kyrychenko** – PhD (Technical Sciences), Kharkiv National University of Radio Electronics, Associate Professor of the Department of Software Engineering, Kharkiv, Ukraine;  
e-mail: [iryna.kyrychenko@nure.ua](mailto:iryna.kyrychenko@nure.ua)  
ORCID ID: <https://orcid.org/0000-0002-7686-6439>

**Демчук Вадим Геннадійович** – Харківський національний університет радіоелектроніки, аспірант кафедри програмної інженерії, Харків, Україна;

**Vadym Demchuk** – Kharkiv National University of Radio Electronics, Postgraduate Student of the Department of Software Engineering, Kharkiv, Ukraine;  
e-mail: [vadym.demchuk@nure.ua](mailto:vadym.demchuk@nure.ua)  
ORCID ID: <https://orcid.org/0000-0003-3700-2344>

**Луценко Віталій Володимирович** – Харківський національний університет радіоелектроніки, студент кафедри програмної інженерії, Харків, Україна;

**Vitalii Lutsenko** – Kharkiv National University of Radio Electronics, student of the Department of Software Engineering, Kharkiv, Ukraine;  
e-mail: [vitalii.lutsenko@nure.ua](mailto:vitalii.lutsenko@nure.ua)  
ORCID ID: <https://orcid.org/0009-0003-8965-5758>

## RESEARCH OF AUTOMATED DATA QUALITY TESTING AND ANOMALY DETECTION METHODS IN DATA PIPELINES

**The object** of study is the process of real-time operational monitoring, diagnostics, and ensuring the functional reliability of distributed streaming data pipelines. **The research** focuses on methods and algorithms for automated anomaly detection in streaming systems, based on the analysis of multivariate time series of operational telemetry (performance metrics and resource utilization) from Apache Spark Structured Streaming and Apache Kafka infrastructure components. **The aim** of the work is to develop and experimentally validate a lightweight methodology for proactive anomaly detection in high-speed data pipelines. This methodology operates exclusively on infrastructure meta-indicators, without resource-intensive payload inspection, to minimize incident response time and eliminate additional data processing latency. **Results Achieved.** During the study, a monitoring system architecture was developed, and a 14-dimensional feature space vector was formed. This vector includes normalized system metrics, their rates of change (gradients), and synthetic dimensionless coefficients (processing efficiency, normalized lag). An ensemble machine learning algorithm was applied to classify system states. Experimental modeling of typical failures (latency spikes, throughput drops, anomalous lag) on an AWS cloud cluster confirmed the approach's high efficiency. The proposed multivariate model increased incident detection accuracy (F1-score) from 0.62 (the performance of a classic rule-based method using static thresholds) to 0.92, with a false positive rate (FPR) of only 0.8%. The mean time to detect (MTTD) anomalies was reduced from 115 to 25 seconds. The computational overhead of the monitoring microservice accounted for less than 1.5% of the cluster's CPU time. **Conclusions.** It has been experimentally demonstrated that analyzing multivariate operational telemetry using machine learning is a highly effective proxy for data pipeline health. The proposed methodology successfully addresses the "heavy validation" problem inherent in traditional data quality tools and fully aligns with the modern Shift-Left architecture paradigm. The solution provides deep, zero-impact observability, serves as a reliable first line of proactive defense, and establishes a technological foundation for implementing infrastructure auto-remediation mechanisms.

**Keywords:** streaming data processing; data pipelines; anomaly detection; operational telemetry; machine learning; Isolation Forest; Apache Kafka; Apache Spark; Shift-Left architecture; system observability.

### Бібліографічні описи / Bibliographic descriptions

Кириченко І. В., Демчук В. Г., Луценко В. В. Розроблення й дослідження підходу до виявлення аномалій у потокових конвеєрах даних на основі операційної телеметрії. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 165–181. DOI: <https://doi.org/10.30837/0135-1710.2026.189.165>

Kyrychenko, I., Demchuk, V., Lutsenko, V. (2026), "Research of automated data quality testing and anomaly detection methods in data pipelines", *Management Information System and Devices*, No. 2 (189), P. 165–181. DOI: <https://doi.org/10.30837/0135-1710.2026.189.165>

---



Matvii Kuchapin, Kyrylo Smelyakov, Anastasiya Chupryna, Sergiy Luchenko

## COMPREHENSIVE ANALYSIS OF METHODS AND TOOLS FOR HYBRID IMAGE ANNOTATION IN COMPUTER VISION SYSTEMS

The **subject matter** of the study is the methods, algorithms, and software tools for annotating visual data in computer vision systems within the Data-Centric AI paradigm. The study analyzes the processes of structuring unstructured information where labeling quality determines the accuracy of artificial intelligence models, considering the limitations of manual methods and risks of systematic errors in full automation. The **goal** of the work is to perform a comprehensive analysis of modern image annotation methods and tools in computer vision systems and a quantitative evaluation of the effectiveness of hybrid Human-in-the-Loop strategies to improve the efficiency of forming high-quality datasets within the Data-Centric AI paradigm. The following **tasks** were addressed in the article: a systematization of annotation types from image classification to panoptic segmentation and 3D scenes was conducted; a review of the toolset based on the Segment Anything Model and GroundingDINO was performed; comparative evaluations of manual, automatic, and hybrid labeling scenarios based on accuracy ( $mIoU$ ) and labor intensity were carried out; issues regarding operator trust and interaction ergonomics were identified. The following **methods** are used – systematic comparative analysis of hybrid Human-in-the-Loop strategies, cross-domain synthesis of active learning and interactive segmentation studies, and formalization of acceleration, quality, and manual labor indicators. The following **results** were obtained: the hybrid pipeline provides a 5.4x process acceleration. The formation time of a semantic mask for an object is reduced from 65.0 to 12.0 seconds while maintaining quality at  $mIoU = 0.94$ , with only a 0.02 loss relative to the reference standard. The hybrid scenario was found to be optimal within the threshold range of  $0,82 < Q_{min} \leq 0,94$ , covering the broadest class of practical tasks from training production models to medical diagnostics. A direct dependence of annotator performance on the reliability of automatic prompts was identified. **Conclusions:** hybrid annotation is the optimal strategy for creating Ground Truth in critical domains such as autonomous driving and medicine, providing a balance between speed and accuracy. The proposed formalization of the optimization problem with  $Q_{min}$  threshold constraint enables informed selection of the annotation scenario for a specific domain. Future research involves improving synthetic data generation to overcome the domain gap and developing adaptive interfaces to reduce cognitive load.

**Keywords:** computer vision; image annotation; active learning; segmentation; synthetic data.

### 1. Introduction

The exponential growth in the volume of visual data is the foundation of modern computer vision (CV). However, in its unstructured form, this information has limited value without appropriate semantic interpretation [1]. Data structuring has become a key challenge, making annotation the most complex and costly stage of artificial intelligence (AI) development. Despite significant progress in the development of neural network architectures, within the Data-Centric AI paradigm, the quality of training datasets is recognized as the primary factor limiting model productivity [2].

Poor annotation quality has critical consequences for the effectiveness of systems in achieving accurate predictions. This illustrates the fundamental principle of Data Science – "Garbage In, Garbage Out" [3]. The situation is complicated by the fact that traditional manual methods scale poorly and are subjective, while full automation carries the risk of systematic errors [4].

Therefore, the most promising solution is hybrid Human-in-the-Loop approaches, which combine intelligent systems with expert verification [5].

In this context, the aim of this study is to conduct a comprehensive analysis of modern methods and tools for image annotation in computer vision systems and to quantitatively evaluate the effectiveness of hybrid Human-in-the-Loop strategies for improving the efficiency of creating high-quality datasets within the Data-Centric AI paradigm. The main focus is on researching hybrid strategies that allow for the effective combination of the computational capabilities of machine learning algorithms with expert verification, thereby mitigating the shortcomings of both purely manual and fully automated approaches. The analysis aims to systematize existing annotation types and software tools, as well as to quantitatively justify the effectiveness of implementing intelligent pipelines for creating high-quality training datasets in modern computer vision systems.

The scientific novelty of the study lies in the following:

- a systematic cross-domain comparative analysis of hybrid annotation systems from various subject areas, such as medical diagnostics, security systems, synthetic art classification, and object detection, was conducted, and a formalized criterion for selecting the optimal annotation scenario based on a quality threshold constraint was formulated, which allowed us to determine the limits of effective application of each approach depending on domain requirements;
- a formalized weighted effectiveness metric is proposed which, unlike known metrics, accounts for the criticality of accuracy for a specific domain through a parametric coefficient, allowing for a well-founded selection of the optimal annotation scenario for domains with varying quality requirements;
- based on a synthesis of independent research results, the influence of ergonomic factors in operator-system interaction on the effectiveness of hybrid annotation has been identified and systematized, in particular, the effect of negative framing, where the operator's bias regarding the unreliability of AI leads to the rejection of correct suggestions, which allows for the formulation of practical requirements to ensure the transparency of algorithmic recommendations and the communication of the model's confidence level when designing interfaces for hybrid annotation systems.

## 2. Analysis of the Current State of Research on Hybrid Annotation Systems

An analysis of scientific publications in recent years indicates a steady trend toward the implementation of Human-in-the-Loop (HITL) approaches, where the human factor is combined with machine learning

algorithms to optimize annotation processes. Current research focuses on resolving the dilemma between annotation cost and training data quality. Hybrid systems demonstrate high adaptability across various domains, ranging from medical diagnostics and security systems to digital art analysis.

A key achievement of this period was the demonstration of the effectiveness of active learning and interactive segmentation methods. These approaches allow for a 50–87% reduction in the volume of manual annotation without sacrificing model accuracy, delegating routine tasks to algorithms and complex cases to experts.

To conduct a detailed study of the effectiveness of hybrid approaches, five recent papers were selected and analyzed based on the following criteria:

- the presence of quantitative results demonstrating the effectiveness of the HITL approach;
- coverage of various subject areas to ensure cross-domain representativeness;
- publication in peer-reviewed journals from 2023–2025.

The sample covers various subject areas, allowing for an assessment of the methods' versatility. Summary data on the models used, metrics, and key results of these studies are presented in Table 1.

A review of the literature shows that combining active learning with interactive annotation is the dominant strategy for conserving resources. The paper "HAL-IA: A Hybrid Active Learning framework using Interactive Annotation for medical image segmentation" [6] proposes a hybrid strategy where the model automatically selects the most ambiguous regions, and an expert refines them by clicking on superpixels. This approach allows for achieving accuracy metrics close to those of full annotation while maintaining high Dice/IoU scores, using significantly less annotated data, and minimizing manual intervention.

**Table 1.** A comparative review of modern hybrid annotation systems

| Study type                                       | Problem and data                                          | Key results                                                                                                                      |
|--------------------------------------------------|-----------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|
| Active Learning and Interactive segmentation [6] | Medical segmentation (ultrasound, CT, X-ray). 4 datasets. | High accuracy (Dice, IoU) with fewer clicks and annotated images; reduced annotation costs.                                      |
| HITL Hybrid decision making [7]                  | Smuggling detection in X-ray images (security).           | Improved recall (threat detection) compared to human performance; increased system throughput.                                   |
| Human-in-the-Loop and Active Learning [8]        | Classification: "Real vs. Artificial Art".                | Accuracy of ~98.65% using 87.5% less labeled data; rapid adaptation to new generative models.                                    |
| HITL with confidence estimation [9]              | Emotional engagement assessment (DAiSEE video dataset).   | A correlation was found between model reliability and annotator performance; AI errors cause frustration and reduce consistency. |
| Progressive active learning [10]                 | Object detection on floor plans.                          | mAP = 0.833 when trained on 500 labeled and 4,500 unlabeled images; significant reduction in manual labor.                       |

Similar effectiveness is confirmed in the paper "The power of progressive active learning in floorplan images for energy assessment" [10], where a progressive learning strategy was applied for object detection in floor plans. Starting training with just 500 labeled images, the model gradually learned on a dataset of 4,500 unlabeled examples, achieving an mAP accuracy of  $\sim 0.833$ , which is comparable to the results of full manual annotation.

A telling example is the classification of synthetic art in the paper "Hybrid intelligence approach for detecting synthetic art" [8], where the AL approach reduced the need for labels by 87.5%. Moreover, the system demonstrated high adaptability when images from new generative models appeared; accuracy initially dropped to 75%, but after several feedback iterations, it recovered to 98%. This indicates that selectively involving humans in the most informative and complex samples is significantly more effective than linear, continuous labeling.

A second critically important research area is the optimization of "human-machine" interaction and the ergonomics of decision-making. In the scientific study "Application of human-in-the-loop hybrid augmented intelligence approach in security inspection system" [7], in the context of security tasks such as X-ray screening, the superiority of hybrid strategies over fully automated or manual methods was demonstrated. In particular, the "deviation-priority" mode, where the machine focuses on detecting suspicious objects and a human verifies them, allowed for maximizing the completeness of threat detection, thereby expanding the system's "safety zone". At the same time, the alternative "clearance-priority" mode ensured a balance between reliability and speed of inspection.

However, the success of such collaboration directly depends on psychological factors, particularly the operator's trust in the algorithm. The study "Human-in-the-Loop Annotation for Image-Based Engagement Estimation: Assessing the Impact of Model Reliability on Annotation Accuracy" [9] revealed a direct correlation between the reliability of the model's prompts and the quality of the annotator's work. Experiments showed that when working with an unreliable model, users felt frustrated and demonstrated inconsistency in their decisions. Most interesting is the identified effect of negative framing. If users were biased regarding the system's unreliability, they tended to reject even correct AI prompts. Thus, algorithm transparency and clear communication of the model's confidence level

are no less important than its mathematical accuracy, as they establish the necessary level of trust for productive collaboration.

Thus, the analysis confirms that the key to the effectiveness of modern annotation systems lies in the harmonious combination of algorithmic optimization through active learning and the deep integration of ergonomic principles. Technological accuracy must be complemented by transparency in interaction, as it is the operator's trust and cognitive comfort that become decisive factors in achieving high-quality annotation in hybrid environments.

### 3. Research Objectives and Tasks

---

The objective of this research is a comprehensive analysis of modern methods and tools for image annotation in computer vision systems to improve the efficiency of creating high-quality datasets within the Data-Centric AI paradigm. Unlike existing works that examine the effectiveness of hybrid approaches in individual domains, this study focuses on a unified cross-domain analysis with the formalization of a criterion for selecting the optimal annotation scenario, taking into account the criticality of the subject domain.

To achieve this goal, the following main tasks must be addressed:

- systematize annotation types and methods, covering various levels of detail from image classification to panoptic segmentation and 3D scene labeling;
- review modern software platforms and tools, in particular those based on the fundamental models Segment Anything Model and Grounding DINO;
- conduct a comparative analysis of the effectiveness of manual, automatic, and hybrid annotation approaches in terms of spatial accuracy (mIoU) and computational complexity;
- identify current scientific challenges in the field of data preparation, including issues related to the adaptation of synthetic domains and the reduction of cognitive load on annotators.

### 4. Methodological Foundations and Tools for Visual Data Annotation

---

#### 4.1. Formalization of the Annotation Task and Selection of Annotation Strategies

In the context of building intelligent systems, annotation is defined as the process of enriching "raw"

data with metadata, which transforms unstructured visual signals into a format suitable for algorithmic processing and knowledge extraction [11]. For visual data, this process bridges the semantic gap between the low-level pixel representation of an image and the high-level semantics of a scene, with the level of detail ranging from image classification to pixel segmentation and 3D space labeling [12].

Within the data-driven paradigm, the quality, completeness, and representativeness of annotated data are recognized as the primary factor in model performance [13]. Annotations form a ground truth label that defines target values for calculating the loss function during training and serves as a reference standard during validation. Errors in labels, such as annotation noise, reduce the model's ability to generalize, forcing it to learn spurious correlations instead of true features [13], which makes annotation a key iterative component of the data preparation pipeline [14].

To establish a sound methodological framework for comparing annotation strategies, it is advisable to formalize the key components of this process. Formally, annotation can be defined as a mapping  $A: X \rightarrow Y$ , where  $X$  is the space of input images, and  $Y$  is the space of labels. For the task of semantic segmentation, each image  $x \in X$  has dimension  $H \times W$ , and the corresponding label  $y \in Y$  is a matrix of the same dimension, where each element  $y(i, j) \in \{1, \dots, C\}$  determines the pixel's membership in one of the  $C$  semantic classes. An ideal mapping  $A^*$  generates a ground-truth annotation, but any real annotation process  $A$  introduces a certain level of noise  $\eta$ , formalized as  $A(x) = A^*(x) + \eta$ , where  $\eta$  depends on the chosen strategy. For manual annotation  $\eta$  is due to the operator's subjectivity; for automatic annotation, to systematic model errors; and for hybrid annotation, to a combination of both, taking into account corrections.

The relationship between the quality of the annotation and the model's productivity is formalized using the empirical risk function. For a model  $f_\theta$  with parameters  $\theta$ , the empirical risk is defined as:

$$R(\theta) = \frac{1}{N} \sum_{i=1}^N L(f_\theta(x_i), y_i) \quad (1)$$

where  $L$  – loss function.

In the presence of annotation noise  $\eta$ , the true learning risk takes the form of:

$$\tilde{R}(\theta) = \frac{1}{N} \sum_{i=1}^N L(f_\theta(x_i), y_i + \eta_i) \quad (2)$$

According to studies on the impact of noisy labels [13], systematic noise  $\eta$  causes a shift in the parameter optimum  $\theta^*$ , resulting in the model learning false correlations instead of the true features of the objects. This relationship defines the fundamental requirement of minimization  $\|\eta\|$  to ensure training quality, which justifies the need for expert verification of automatic labels in hybrid pipelines.

The problem of selecting the optimal annotation strategy can be formulated as a conditional optimization problem. Minimize the total cost  $C_{\text{total}} = C_{\text{auto}} + C_{\text{human}}$  subject to the constraint  $Q(\hat{A}) \geq Q_{\min}$ , where  $Q$  is the annotation quality metric (e.g., mIoU), and  $Q_{\min}$  is the minimum acceptable quality threshold for the target domain. In a fully manual scenario  $C_{\text{auto}} = 0$ , and  $C_{\text{human}} = \max$ . In a fully automatic scenario  $C_{\text{human}} = 0$ , but the constraint  $Q \geq Q_{\min}$  may be violated due to systematic model errors. The Human-in-the-Loop hybrid scenario minimizes  $C_{\text{total}}$ , ensuring compliance with the quality constraint through targeted expert correction of only problematic areas. The parameter  $Q_{\min}$  is domain-dependent: for critical applications, such as autonomous driving or medical diagnostics, it approaches the values of the ground-truth manual annotation, whereas for pre-filtering tasks, a significantly lower threshold is permitted.

#### 4.2. Multidimensional Typology of Annotation Processes

The annotation of visual information is a multi-component and multi-faceted process, which in modern research is typically classified according to three interrelated dimensions: the type of source data, the level of detail of visual primitives, and the degree of automation in the annotation process.

From the perspective of information theory, the validity of this classification is determined by differences in the information capacity of various types of annotations. Image classification requires only  $\log_2(C)$  bits of information per sample, where  $C$  is the number of classes. The bounding box adds spatial coordinates, increasing the amount of information to

$\log_2(C) + 4\log_2(R)$  bits, where  $R$  is the resolution of the coordinate grid. Pixel segmentation requires  $H \times W \times \log_2(C)$  bits, which is orders of magnitude higher than the previous levels. This exponential growth in information capacity theoretically justifies the nonlinear increase in annotation complexity and explains why segmentation tasks are the prime candidates for automation.

The first vector defines the specific spatiotemporal characteristics of the information. The baseline here consists of 2D images. These are static arrays of pixels where annotation focuses exclusively on spatial features within the frame plane [15]. Video data imposes significantly higher processing requirements, as the addition of a temporal dimension necessitates ensuring temporal coherence and tracking unique object identifiers across a sequence of frames [16]. The highest level of complexity is characterized by 3D data, such as LiDAR point clouds, where annotation occurs in three-dimensional space, requiring the determination of physical dimensions and the precise orientation of the object, which is critical for autonomous navigation systems [17].

Along with the nature of the input data, an important classification parameter is the level of detail in the scene description, i.e., the type of annotation primitives. The most aggregated level is image-level annotation, where only a general category is established without specifying the localization of objects. For object detection tasks, the standard approach involves the use of bounding boxes – rectangular regions that can be either axially aligned or oriented at an arbitrary angle to more accurately capture the object's geometry [15]. The highest spatial accuracy is provided by polygonal or pixel segmentation methods, where each pixel is associated with a specific class or a specific instance of an object, allowing for the most detailed delineation of its shape [18]. A separate group consists of key points used to describe the human skeletal structure or mechanisms, as well as polylines used to mark road infrastructure elements, trajectories, and other linear objects.

The third defining vector of the typology is the annotation methodology, which ranges from fully manual approaches to complete process automation. Traditionally, manual annotation is considered the "gold standard" due to its high accuracy and expert oversight; however, it is expensive, labor-intensive, and

has limited scalability. In response to these limitations, a semi-automated approach known as "Human-in-the-loop" has emerged, combining automatic preliminary tag generation by algorithms with subsequent human verification. This approach significantly reduces the amount of manual work and can accelerate the annotation process by a factor of 10–20 [19].

The most modern and rapidly growing field is synthetic annotation, in which labeled data is procedurally generated in artificial or simulated environments. This allows for the automatic generation of annotations with perfect accuracy and full control over scene parameters. At the same time, a key challenge of this approach remains the gap between the simulated and real domains, which requires the use of domain adaptation and style alignment methods to ensure the correct transfer of models to real-world conditions [20].

#### 4.3. Main Tasks and Formats of Image Annotation

Image annotation covers a wide range of tasks that form the basis of computer vision models, from simple classifiers to autonomous navigation systems. These tasks differ significantly in terms of annotation effort, which directly influences the choice of annotation strategy in the context of the optimization problem formalized above  $C_{\text{total}}$ .

Image classification is a fundamental task in which models assign a global label (single-class or multi-class) to the entire image without localizing individual objects [21]. From an annotation perspective, this task is characterized by minimal labor intensity and the lowest sensitivity to noise  $\eta$ , since it does not require spatial detail.

Object detection requires determining not only the category but also the spatial localization in the form of bounding boxes: horizontal (COCO, PASCAL VOC formats [22]) or oriented at an arbitrary angle for aerial photography and satellite image analysis tasks [23]. The accuracy of the bounding boxes directly affects the quality of detector training, and the degree of object overlap in dense scenes significantly complicates both manual and automatic annotation.

Image segmentation provides the highest spatial accuracy by performing classification at the pixel level. We distinguish between semantic (class of each pixel), instance (identification of individual instances), and panoptic segmentation, which combines both approaches [18, 24, 25]. Segmentation tasks are the



most labor-intensive; creating a single pixel mask requires significantly more time than bounding box annotation, making them prime candidates for implementing hybrid Human-in-the-Loop strategies.

Pose and keypoint estimation involves annotating skeletal models of objects, ranging from 17 keypoints in the COCO format to 106 for facial analysis, which is used in augmented reality systems, sports analytics, and gesture recognition [26]. A separate area is the description of images in natural language, which is critical for training CLIP, BLIP-2, and GPT-4V models and for creating assistive systems [27].

The presented typology of tasks demonstrates that the labor intensity of annotation increases non-linearly with the level of detail, ranging from seconds for classification to minutes for pixel-level segmentation of a single image. This effect has a theoretical basis in the relationship between the complexity of the label and the minimum sample size required to train the model. According to statistical learning theory, more complex label spaces require a greater number of annotated examples to achieve an equivalent level of generalization, which places a double burden on resources. Each sample is both more expensive to annotate and requires a larger quantity. This increase directly determines the feasibility of using automated and hybrid strategies, where minimization  $C_{\text{total}}$  becomes critical precisely for tasks with a high label complexity component  $C_{\text{human}}$ .

#### **4.4. The current state of annotation technologies and tools**

The modern visual data annotation ecosystem is undergoing a clear shift from labor-intensive manual annotation processes to comprehensive, automated workflows driven by artificial intelligence algorithms. This transformation is driven by the rapid increase in demand for large-scale training datasets and the need to eliminate the bottleneck of slow and expensive manual annotation. While the industry previously relied heavily on crowdsourcing approaches, the "human-in-the-loop" paradigm is now becoming dominant, within which the annotator's role shifts from creating labels to that of an expert verifier and data curator [2]. The introduction of automatic pre-tagging reduces manual work by 90–95%, leaving only the most complex or ambiguous cases for human review. An important component of ensuring annotation quality is measuring the level of agreement among annotators, which is done using metrics such as Cohen's kappa or Fleiss's kappa [28]. Using these

metrics helps reduce the subjectivity of annotation and improve its consistency in large-scale projects.

The application of deep active learning approaches has become a key tool for improving the efficiency of the annotation process. Unlike the traditional practice of annotating the entire dataset, this approach involves a phased, iterative selection of only those samples that are most valuable for further training of the model. The most common strategies include uncertainty-based sampling methods, where priority is given to examples with the lowest model confidence in its own prediction, as well as diversity-based sampling, which aims to cover a wide range of features and ensure the representativeness of the feature space. According to the results of a number of studies, Active Learning is capable of achieving an accuracy level comparable to that of a fully labeled dataset using only 15–20% of the initial volume of annotated data, which allows for a significant reduction in the financial and time costs of annotation [29].

The theoretical basis for the effectiveness of active learning is the principle of maximizing information gain at each iteration; for annotation, a sample is selected that minimizes the entropy of the posterior distribution of model parameters. Formally, the uncertainty-based sampling strategy implements approximate optimization of mutual information between the model parameters and the label of the selected sample. This theoretically guarantees faster training convergence compared to random sampling and explains the empirically observed 80–85% reduction in annotation volume.

A revolutionary shift in labeling practices was made possible by the emergence of large visual foundation models, which were trained on massive multimodal collections of billions of "image-text" pairs. One of the most significant developments is the Segment Anything Model (SAM) from Meta AI, which introduced the concept of "promptable segmentation" – segmentation activated via interactive or text prompts. SAM is capable of generating highly accurate object masks in real time, while demonstrating exceptional generalization ability to new domains without additional training, i.e., operating in zero-shot mode [30].

Modern automated annotation pipelines increasingly combine multiple models into a single pipeline. For example, Grounding DINO is used for open-vocabulary object detection based on text instructions. The resulting bounding boxes are then passed to SAM to generate accurate segmentation masks. This approach already enables the implementation of fully automated

annotation systems that do not require pre-training on target domain classes [30]. In tasks requiring an immediate response, models from the YOLO family (YOLOv8, v9, v10) continue to play a key role; thanks to their high speed and accuracy, they are widely used for automatic pre-tagging, leaving humans with a minimal amount of work to verify and refine the results [31].

One of the key areas of development in modern annotation technologies is aimed at overcoming the problem of data scarcity by creating synthetic datasets. The use of highly realistic graphics engines, such as Unreal Engine or Unity, as well as specialized simulation platforms, including NVIDIA Omniverse and CARLA, enables the creation of complex scene configurations with automatically generated and highly accurate annotations. This approach is particularly important in the area of autonomous transportation, where collecting real-world data on rare or hazardous road situations is often impractical or risky [20].

## 5. Research Results

### 5.1 Formalization of the task and annotation scenarios

The research focuses on the task of semantic segmentation of objects in 2D images, as one of the most representative tasks for modern computer vision systems. Let there be a set of images  $D$  for which we need to construct a set of semantic masks  $Q$ .

$$D = \{I_i\}_{i=1}^N, \quad M = \{M_i\}_{i=1}^N \quad (3)$$

where  $I_i$  is a single input image from the sample;  $M_i$  is the corresponding semantic mask of the

input image;  $N$  is the total number of images in the dataset.

The analysis considers three basic scenarios for generating annotations  $M$ :

- $S_m$  fully manual annotation, which is considered the gold standard;
- $S_{auto}$  fully automatic annotation based on YOLO/SAM models;
- $S_{hyb}$  hybrid annotation with expert verification and correction.

The hybrid approach involves an iterative process in which the results of automatic generation are subject to expert review, refinement of object geometry, and correction of erroneous or missing segments. A generalized diagram of the hybrid pipeline is shown in Fig. 1.

The diagram shown illustrates a hybrid annotation pipeline in which the automatic generation of preliminary masks using YOLO and SAM neural networks is integrated with expert review in a human-in-the-loop mode. After the initial processing of complex 2D images of street scenes, a specialist performs a thorough correction of object geometry, removes false detections, and adds missing elements to achieve maximum accuracy. An important part of the process is an iterative feedback cycle, which allows for the rapid fine-tuning of system parameters and a gradual increase in its level of autonomy. The final output of the pipeline is high-quality Ground Truth-standard semantic masks, which provide an ideal balance between data preparation speed and reliability.

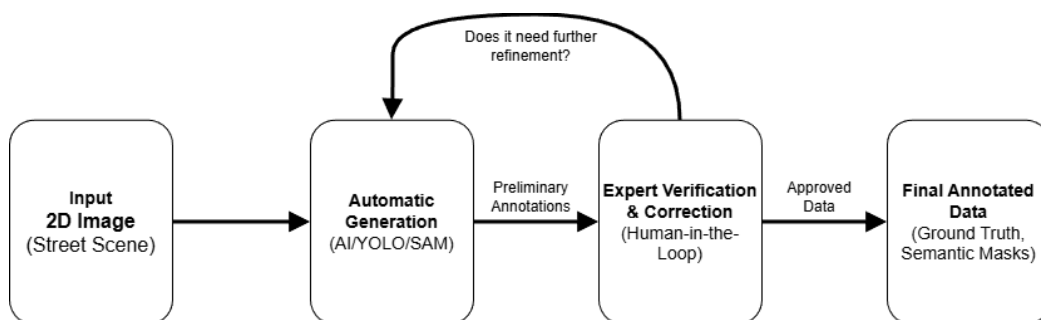


Fig 1. Block diagram of a hybrid image annotation pipeline

This approach mitigates the shortcomings of purely automated labeling, ensuring the creation of accurate datasets for autonomous driving systems while significantly optimizing time expenditure.

### 5.2. System of Quality and Labor-Intensity Metrics

A comprehensive assessment of the effectiveness of each scenario is based on two complementary groups of indicators: quality metrics and labor intensity metrics.

To quantitatively assess the quality of an annotation, the standard metric of mean Intersection over Union (*mIoU*) is used:

$$mIoU = \frac{1}{C} \sum_{c=1}^C \frac{|P_c \cap G_c|}{|P_c \cup G_c|} \quad (4)$$

where  $P_c$  is the set of pixels assigned to class  $c$  in the generated or corrected mask;  $G_c$  is the corresponding set of pixels in the ground truth annotation;  $C$  is the number of classes.

The metric *mIoU* characterizes the spatial consistency of segmentation masks and allows for a quantitative assessment of the accuracy of object boundary reproduction.

To formalize the labor intensity of the annotation process, the average time required to generate a mask for a single object is introduced:

$$T = T_{gen} + T_{corr} \quad (5)$$

where  $T_{gen}$  is the time required for automatic generation of the initial mask;  $T_{corr}$  is the time required for expert verification and correction.

It is worth noting that the component  $T_{corr}$  is not a constant and is directly dependent on the reliability of the automatic prediction. In the case of low model confidence, the cognitive load on the operator increases, which can lead to frustration and reduced consistency of actions. Thus, the effectiveness of the hybrid scenario is determined by the algorithm's ability to

minimize the need for a complete mask reconstruction, limiting itself to its refinement.

To quantitatively assess the extent of manual intervention, the relative manual labor coefficient is used:

$$R = T_{corr} / T_{man} \quad (6)$$

where  $T_{man}$  is the average time required to fully annotate a single object manually.

Thus, the  $R$  metric reflects the proportion of human involvement in the hybrid scenario relative to fully manual annotation.

### 5.3. Quantitative analysis and comparison of scenarios

To derive numerical estimates, we used statistical aggregation of results from recent studies in the areas of interactive segmentation and progressive learning [6, 8, 10]. The validity of this approach is determined by the fact that the selected studies use a single quality metric (*mIoU*), a common task of semantic segmentation of 2D images, and comparable architectural solutions based on the YOLO and SAM models, which ensures methodological consistency of the sample. Time values are normalized to a single object, allowing the results of different pipelines to be compared regardless of dataset size and hardware configuration. It should be noted that the presented metrics are indicative: specific values may vary depending on the complexity of the domain, hardware, and the skill level of the annotators. The unified metrics are presented in Table 2.

**Table 2.** Comparison of the effectiveness of image annotation scenarios

| Annotation scenario  | Average time per object (sec) | Annotation quality ( <i>mIoU</i> ) | Manual labor |
|----------------------|-------------------------------|------------------------------------|--------------|
| Manual               | ~65.0                         | 0.96 (Baseline)                    | 100%         |
| Automated (YOLO/SAM) | ~0.5                          | 0.82                               | 0            |
| Hybrid               | ~12.0                         | 0.94                               | 15–20%       |

To quantify the performance gain, an acceleration factor relative to fully manual annotation is introduced:

$$K_{speed} = T_{man} / T_{hyb} \quad (7)$$

For the values in Table 2, the acceleration factor is  $K_{speed} = 65/12 \approx 5.4$ . Thus, using the hybrid scenario allows the time required to generate a single semantic mask to be reduced by more than five times.

We define the loss of quality relative to manual annotation as:

$$\Delta IoU = mIoU_{man} - mIoU_{hyb} \quad (8)$$

For generalized data, the quality loss is  $\Delta IoU = 0.96 - 0.94 = 0.02$ .

This level of error is lower than typical values of inter-rater variability, which, according to [28], range from 0.0001 to 0.0003 0.03–0.05 s for semantic segmentation tasks when measured using Cohen's kappa coefficient. This allows us to conclude that the hybrid method provides annotation quality identical to or higher than that of a group of independent experts, but with better consistency of object boundaries due to the algorithmic stability of the YOLO and SAM models.

To analyze the trade-off between quality and time expenditure, a generalized performance metric is introduced:

$$E = mIoU / T \quad (9)$$

Accordingly, for each scenario, we have the following metrics:  $E_{man} = 0.96/65$ ,  $E_{auto} = 0.82/0.5$ ,  $E_{hyb} = 0.94/12$ .

Although the " $S_{auto}$ " scenario formally has the highest performance metric ( $E$ ) due to its speed, it does not meet the requirements for creating a ground truth reference. For a more accurate assessment, it is advisable to introduce a weighting factor for accuracy criticality,  $n$ , which transforms the performance metric into the following form:

$$E_n = mIoU^n / T, \quad (10)$$

where, for  $n > 1$ , the advantage of the hybrid scenario becomes evident, as it provides the necessary accuracy threshold for training models in critical domains such as autonomous driving or medical diagnostics.

Although the fully automatic scenario achieves the highest value of the indicator  $E$  due to its minimal

processing time, it does not guarantee sufficient segmentation accuracy for tasks where correct object boundary geometry and semantic consistency of masks are critical. The hybrid scenario, in turn, provides an optimal balance between speed and quality, combining a significant reduction in annotation time with high performance metrics  $mIoU$ .

The results show that in hybrid mode, the operator performs only targeted corrections of problematic segmentation areas, specifically removing artifacts at object boundaries, refining the geometry of small elements, and correcting omissions in complex scene areas. A visual comparison of the results of automatic and hybrid segmentation using a complex urban scene as an example is shown in Fig. 2, where the reduction in noise and improvement in the spatial coherence of the masks following expert correction are clearly demonstrated.

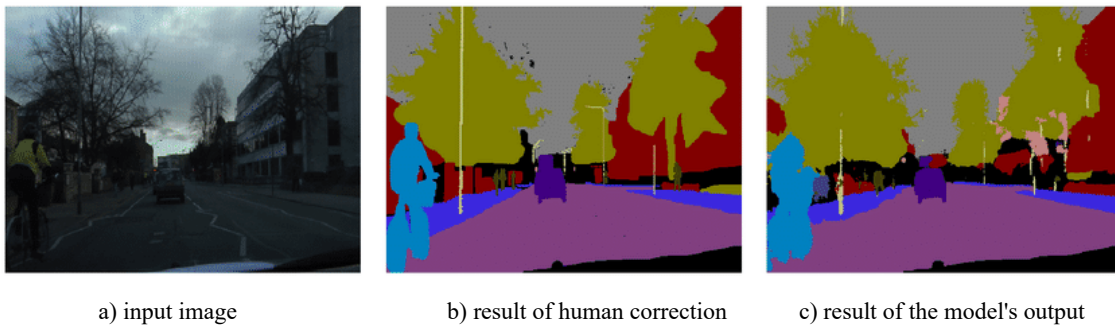


**Fig. 2.** Comparison of automatic and hybrid segmentation:

a) input image; b) result of human correction; c) model output (contains noise) [32]

Quantifying the extent of manual labor is of particular practical importance. The value of the coefficient  $R = 0.15-0.2$  indicates that in the hybrid scenario, manual labor is reduced by 80–85% compared to fully manual annotation. This result is consistent with experimental observations from recent studies, which report the possibility of reducing the volume of manual annotation by 80–90% when active learning and interactive segmentation are applied.

The results in Table 2 take on a non-trivial character when interpreted through the conditional optimization problem formalized in Section 4.1:  $\min C_{total}$  subject to the constraint  $Q(\hat{A}) \geq Q_{min}$ . The threshold  $Q_{min}$  is a domain-dependent parameter, and it is this parameter that determines the set of feasible scenarios for each specific application. Comparisons of admissibility based on aggregated data are presented in Table 3.

**Table 3.** Admissibility of annotation scenarios depending on domain requirements

| Application type                   | $Q_{min}$ | $S_{man}$ (0.96) | $S_{auto}$ (0.82) | $S_{hyb}$ (0.94) | Optimal scenario      |
|------------------------------------|-----------|------------------|-------------------|------------------|-----------------------|
| Pre-filtration, prototyping        | 0.70      | ✓                | ✓                 | ✓                | Automatic (T ~ 0.5 s) |
| Training of general-purpose models | 0.85      | ✓                | ✗                 | ✓                | Hybrid (T ~ 12 s)     |
| Medical diagnostics                | 0.90      | ✓                | ✗                 | ✓                | Hybrid (T ~ 12 s)     |
| Autonomous driving (Ground Truth)  | 0.95      | ✓                | ✗                 | ✗                | Manual (T ~ 65 s)     |

An analysis of Table 3 reveals a non-trivial result that refutes the naive assumption of a universal advantage of the hybrid approach. The advantage of the hybrid scenario is not universal but arises only within the range  $0,82 < Q_{\min} \leq 0,94$ . At  $Q_{\min} \leq 0,82$ , a fully automatic approach is sufficient, while at  $Q_{\min} > 0,94$ , only manual annotation ensures the required level of quality. It is precisely this range that covers the vast majority of real-world industrial and research tasks, as confirmed by the practice of creating datasets in the Waymo Open Dataset, nuScenes, and COCO projects, where the

threshold quality requirements fall within the range 0.85–0.93. Thus, the Human-in-the-Loop hybrid scenario is optimal for the broadest class of practical tasks, ranging from training production models to medical diagnostics, but not for the extreme cases of the spectrum.

To quantitatively assess the impact of the parameter  $n$  on the scenario rankings, the values of the weighted efficiency index  $E_n = mIoU^n / T$  were calculated for each scenario at different levels of criticality, as shown in Table 4.

**Table 4.** Sensitivity analysis of the weighted efficiency index ( $E_n$ ) to the criticality parameter

| $n$ | $E_n$ (manual) | $E_n$ (automatic) | $E_n$ (hybrid) | Rating                       |
|-----|----------------|-------------------|----------------|------------------------------|
| 1   | 0.015          | 1.640             | 0.078          | auto $\gg$ hybrid $>$ manual |
| 2   | 0.014          | 1.345             | 0.074          | auto $\gg$ hybrid $>$ manual |
| 5   | 0.013          | 0.741             | 0.061          | auto $>$ hybrid $>$ manual   |
| 10  | 0.010          | 0.275             | 0.045          | auto $>$ hybrid $>$ manual   |
| 20  | 0.007          | 0.038             | 0.024          | auto $>$ hybrid $>$ manual   |
| 30  | 0.005          | 0.005             | 0.013          | hyb $>$ auto $\approx$ man   |

An analysis of Table 4 reveals two patterns. First, as  $n$  increases, the performance of the automatic scenario degrades significantly faster than that of the hybrid scenario: when  $n=30$ , the value of  $E_n$  (automatic) decreases by  $\sim 330$  times compared to  $n=1$ , whereas  $E_n$  (hybrid) decreases by only  $\sim 6$  times. This is due to the nonlinear nature of the penalty  $mIoU^n$ . The difference between  $mIoU = 0,82$  and  $mIoU = 0,94$  becomes decisive when raised to a high power. Second, even with moderate values of  $n$ , the weighted metric  $E_n$  does not shift the absolute ranking in favor of the hybrid scenario, indicating the need to supplement the quantitative analysis with a threshold constraint  $Q_{\min}$ , as shown in Table 3. It is the combination of the threshold constraint  $Q_{\min}$  with the metric  $E_n$  that provides a comprehensive criterion for selecting an annotation strategy.

Thus, the analysis conducted allows us to formulate a well-founded criterion for selecting an annotation strategy: for a specific application with the threshold requirement  $Q_{\min}$ , the optimal scenario is the one with the minimum time  $T$  among those that satisfy the constraint  $mIoU \geq Q_{\min}$ . According to the aggregated data, the hybrid scenario is optimal for the widest range of practical tasks ( $0.82 < Q_{\min} \leq 0.94$ ), providing

a fivefold speedup relative to the manual approach while reducing the amount of manual labor to 15–20%. At the same time, the results of the sensitivity analysis (Table 4) demonstrate that the indicator  $E_n$  is not a self-sufficient criterion and requires supplementation with the threshold constraint  $Q_{\min}$ , which underscores the practical significance of the proposed formalization of the optimization problem.

At the same time, the study has several limitations. First, the quantitative analysis is based on the aggregation of results from independent studies with different experimental conditions, which makes it impossible to statistically test the significance of differences between scenarios. Second, the analysis is limited to the task of semantic segmentation of 2D images. For other types of annotation (3D point clouds, video tracking), quantitative ratios may differ significantly. Third, the values of time metrics depend on hardware configuration and the skill level of operators, which limits the direct transfer of results to specific production environments. However, these limitations do not diminish the value of the obtained results, since the proposed formalization of the optimization problem with a threshold constraint is invariant with respect to specific numerical values and remains methodologically valid for any domains and configurations. The quantitative estimates presented in Tables 2–4 should be considered representative



indicative values demonstrating general patterns of the relationship between speed and quality for the three classes of scenarios.

## 6. Conclusions

Image annotation is a critically important step in the data preparation process for computer vision systems. In the context of the Data-Centric AI paradigm, it is the quality of the annotations, rather than the architectural complexity of the models, that determines the accuracy and reliability of predictions. The annotation process provides a semantic interpretation of "raw" pixel data, forming the ground truth annotations necessary for training and validating algorithms.

The formalization of the annotation task as a conditional optimization problem and the comparative analysis conducted indicate that the modern technological landscape is undergoing a gradual shift in approaches, moving from exclusively manual labor toward the implementation of hybrid Human-in-the-Loop systems. Such systems synergistically combine automated pre-labeling using fundamental models, notably SAM and YOLO, with subsequent expert verification. A quantitative analysis of scenarios demonstrated that the hybrid approach provides a fivefold acceleration of the process from 65 to 12 seconds per object while maintaining high quality. The recorded accuracy loss of 0.02 relative to the manual ground truth falls within typical inter-annotator variations, confirming the reliability of the hybrid method for creating professional datasets.

The economic feasibility of implementing such pipelines lies in their ability to reduce the amount of

manual labor to. However, the effectiveness of such collaboration depends on the reliability of automatic prompts: low model confidence exponentially increases correction time and the cognitive load on the operator. For critical areas, such as autonomous driving or medical diagnostics, it is rational to use a weighted performance metric that prioritizes accuracy over speed.

The effectiveness of hybrid systems is determined not only by algorithms but also by the operator's trust in AI and the ergonomics of the interfaces. Future research aims to improve algorithms for generating synthetic samples to address the "domain gap" problem and to develop adaptive interfaces that minimize annotator fatigue.

## Conflict of Interest

The authors declare that they have no conflicts of interest regarding this study, including financial, personal, authorship, or other conflicts that could influence the study and its results presented in this article.

## Funding

The study was conducted without financial support.

## Data Availability

Data will be provided upon reasonable request.

## Use of Artificial Intelligence

The authors confirm that they did not use artificial intelligence technologies in the creation of this work.

## References

1. Song, H. et al. (2020), "Weighted Topic Model Learned From Local Semantic Space for Automatic Image Annotation", *IEEE Access*, Vol. 8, pp. 76411–76422. DOI: <https://doi.org/10.1109/ACCESS.2020.2989200>
2. Monarch, R. M. (2021), "Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI", *New York*, 424 p.
3. Montezuma, D. et al. (2022), "Annotating for Artificial Intelligence Applications in Digital Pathology: A Practical Guide for Pathologists and Researchers", *United States & Canadian Academy of Pathology*, Vol. 36, pp. 100086. DOI: <https://doi.org/10.1016/j.modpat.2022.100086>
4. Demrozi, F. et al. (2023), "A Comprehensive Review of Automated Data Annotation Techniques in Human Activity Recognition", *Cornell University arXiv*. DOI: <https://doi.org/10.48550/arXiv.2307.05988>
5. Sun, Q. et al. (2025), "DocSpiral: A Platform for Integrated Assistive Document Annotation through Human-in-the-Spiral", *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vol. 3, pp. 267–274. DOI: <https://doi.org/10.18653/v1/2025.acl-demo.26>

6. Li, X. et al. (2023), "HAL-IA: A Hybrid Active Learning framework using Interactive Annotation for medical image segmentation", *Medical Image Analysis*, Vol. 88, pp. 102862. DOI: <https://doi.org/10.1016/j.media.2023.102862>
7. Huang, Y. et al. (2025), "Application of human-in-the-loop hybrid augmented intelligence approach in security inspection system", *Frontiers in Artificial Intelligence*, Vol. 8. DOI: <https://doi.org/10.3389/frai.2025.1518850>
8. Ramanathan, A. S., Oyelere, S. S., Baruah, N. (2025), "Hybrid intelligence approach for detecting synthetic art", *Human-Intelligent Systems Integration*, Vol. 7, pp. 325–340. DOI: <https://doi.org/10.1007/s42454-025-00081-z>
9. Yadnakudige Subramanya, S. et al. (2025), "Human-in-the-Loop Annotation for Image-Based Engagement Estimation: Assessing the Impact of Model Reliability on Annotation Accuracy", *Human-Computer Interaction. HCII 2025. Lecture Notes in Computer Science*, Vol. 15770, pp. 169–186. DOI: [https://doi.org/10.1007/978-3-031-93864-1\\_12](https://doi.org/10.1007/978-3-031-93864-1_12)
10. Al-Turki, D. et al. (2023), "The power of progressive active learning in floorplan images for energy assessment", *Scientific reports*, Vol. 16238. DOI: <https://doi.org/10.1038/s41598-023-42276-x>
11. Sager, C., Janiesch, C., Zschech, P. (2021), "A survey of image labelling for computer vision applications", *Journal of Business Analytics*. DOI: <https://doi.org/10.48550/arXiv.2104.08885>
12. Bachani, V. et al. (2024), "Image Segmentation Survey: Classical and Deep Learning Methods", *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pp. 1–6. DOI: <https://doi.org/10.1109/ICECET61485.2024.10698602>
13. Song, H. et al. (2023), "Learning From Noisy Labels With Deep Neural Networks: A Survey", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 34, pp. 8135–8153. DOI: <https://doi.org/10.1109/TNNLS.2022.3152527>
14. Whang, S., Roh, Y., Song, H. (2023), "Data collection and quality challenges in deep learning: a data-centric AI perspective", *The VLDB Journal* 32, pp. 791–813. DOI: <https://doi.org/10.1007/s00778-022-00775-9>
15. Zou, Z. et al. (2023), "Object Detection in 20 Years: A Survey", *Proceedings of the IEEE*, Vol. 111, pp. 257–276. DOI: <https://doi.org/10.1109/JPROC.2023.3238524>
16. Ciaparrone, G. et al. (2020), "Deep learning in video multi-object tracking: A survey", *Science Direct, Neurocomputing*, Vol. 381, pp. 61–88. DOI: <https://doi.org/10.1016/j.neucom.2019.11.023>
17. Nagiu, A. S. et al. (2024), "3D Object Detection for Autonomous Driving: A Comprehensive Review", *2024 6th International Conference on Computing and Informatics (ICCI)*, pp. 01–11. DOI: <https://doi.org/10.1109/ICCI61671.2024.10485120>
18. Minaee, S. et al. (2022), "Image Segmentation Using Deep Learning: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44, pp. 3523–3542. DOI: <https://doi.org/10.1109/TPAMI.2021.3059968>
19. Wu, X. et al. (2022), "A survey of human-in-the-loop for machine learning", *Science Direct, Future Generation Computer Systems*, Vol. 135, pp. 364–381. DOI: <https://doi.org/10.1016/j.future.2022.05.014>
20. Mumuni, A., Mumuni, F., Gerrar, N. K. (2024), "A Survey of Synthetic Data Augmentation Methods in Machine Vision", *Springer Nature Link, Machine Intelligence Research*, Vol. 21, pp. 831–869. DOI: <https://doi.org/10.1007/s11633-022-1411-7>
21. Chen, Z. M. et al. (2019), "Multi-Label Image Recognition With Graph Convolutional Networks", *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5172–5181. DOI: <https://doi.org/10.1109/CVPR.2019.00532>
22. Padilla, R. et al. (2021), "A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit", *Electronics*, Vol. 10, pp. 279. DOI: <https://doi.org/10.3390/electronics10030279>
23. Ajmera, F. et al. (2021), "Survey on Object Detection in Aerial Imagery", *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, pp. 1050–1055. DOI: <https://doi.org/10.1109/ICICV50876.2021.9388517>
24. Gu, W., Bai, S., Kong, L. (2022), "A review on 2D instance segmentation based on deep neural networks", *Science Direct, Image and Vision Computing*, Vol. 120, pp. 104401. DOI: <https://doi.org/10.1016/j.imavis.2022.104401>
25. Elharrouss, O. et al. (2021), "Panoptic Segmentation: A Review", *Cornell University arXiv*. DOI: <https://doi.org/10.48550/arXiv.2111.10250>
26. Chen, Y., Tian, Y., He, M. (2020), "Monocular human pose estimation: A survey of deep learning-based methods", *Science Direct, Computer Vision and Image Understanding*, Vol. 192, p. 102897. DOI: <https://doi.org/10.1016/j.cviu.2019.102897>
27. Stefanini, M. et al. (2021), "From Show to Tell: A Survey on Image Captioning", *Cornell University arXiv*. DOI: <https://doi.org/10.48550/arXiv.2107.06912>
28. Roh, Y., Heo, G., Whang, S. E. (2021), "A Survey on Data Collection for Machine Learning: A Big Data – AI Integration Perspective", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 33, pp. 1328–1347. DOI: <https://doi.org/10.1109/TKDE.2019.2946162>
29. Gu, F. et al. (2021), "A Survey on Deep Learning for Human Activity Recognition", *Association for Computing Machinery*, Vol. 54, p. 34. DOI: <https://doi.org/10.1145/3472290>

30. Kirillov, A. et al. (2023), "Segment Anything", 2023 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003. DOI: <https://doi.org/10.1109/ICCV51070.2023.00371>
31. Terven, J. et al. (2023), "A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS", *Machine Learning and Knowledge Extraction*, Vol. 5, p. 1680–1716. DOI: <https://doi.org/10.3390/make5040083>
32. Corbière, C. et al. (2019), "Addressing Failure Prediction by Learning Model Confidence", *Computer Vision and Pattern Recognition*, arXiv. DOI: <https://doi.org/10.48550/arXiv.1910.04851>

*Received (Надійшла) 19.02.2026*

*Accepted for publication (Прийнята до друку) 15.04.2026*

*Publication date (Дата публікації) 29.05.2026*

### *Відомості про авторів / About the Authors*

**Кучапін Матвій Юрійович** – Харківський національний університет радіоелектроніки, аспірант кафедри програмної інженерії, Харків, Україна;

**Matvii Kuchapin** – Kharkiv National University of Radio Electronics, Postgraduate Student at the Software Engineering Department, Kharkiv, Ukraine;

e-mail: [matvii.kuchapin@nure.ua](mailto:matvii.kuchapin@nure.ua)

ORCID ID: <http://orcid.org/0009-0006-1953-2893>

**Смеляков Кирило Сергійович** – доктор технічних наук, професор, Харківський національний університет радіоелектроніки, завідувач кафедри програмної інженерії, Харків, Україна;

**Kyrylo Smelyakov** – Doctor of Technical Sciences, Professor, Kharkiv National University of Radio Electronics, Head at the Department of Software Engineering, Kharkiv, Ukraine;

e-mail: [kyrylo.smelyakov@nure.ua](mailto:kyrylo.smelyakov@nure.ua)

ORCID ID: <http://orcid.org/0000-0001-9938-5489>

**Чуприна Анастасія Сергіївна** – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, доцент кафедри програмної інженерії, Харків, Україна;

**Anastasiya Chupryna** – Candidate of Technical Sciences, Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor of the Department of Software Engineering, Kharkiv, Ukraine;

e-mail: [anastasiya.chupryna@nure.ua](mailto:anastasiya.chupryna@nure.ua)

ORCID ID: <http://orcid.org/0000-0003-0394-9900>

**Лученко Сергій Васильович** – кандидат технічних наук, Національний аерокосмічний університет "Харківський авіаційний інститут", старший викладач кафедри інженерії програмного забезпечення, Харків, Україна;

**Sergiy Luchenko** – Candidate of Technical Sciences, National Aerospace University "Kharkiv Aviation Institute", Senior Lecturer of the Department of Software Engineering, Kharkiv, Ukraine;

e-mail: [s.luchenko@khai.edu](mailto:s.luchenko@khai.edu)

ORCID ID: <http://orcid.org/0009-0006-9606-5774>

## КОМПЛЕКСНИЙ АНАЛІЗ МЕТОДІВ ТА ІНСТРУМЕНТІВ ГІБРИДНОГО АНОТУВАННЯ ЗОБРАЖЕНЬ У СИСТЕМАХ КОМП'ЮТЕРНОГО ЗОРУ

**Предметом дослідження** є методи, алгоритми та програмні інструменти анотування візуальних даних у системах комп'ютерного зору в межах парадигми Data-Centric AI. У статті проаналізовано процеси структурування неструктурованої інформації, де якість розмітки визначає точність моделей штучного інтелекту. Виявлено обмеження ручних методів і ризики систематичних помилок за умови повної автоматизації. **Мета дослідження** – комплексний

аналіз сучасних методів та інструментів анотування зображень у системах комп'ютерного зору та кількісне оцінювання доцільності гібридних стратегій Human-in-the-Loop для підвищення ефективності формування якісних наборів даних у межах парадигми Data-Centric AI. У статті необхідно виконати такі **завдання**: систематизувати типи анотувань від класифікації до паноптичної сегментації та розмітки 3D-сцен; розглянути інструментарій на основі моделей Segment Anything Model і GroundingDINO; здійснити порівняльне оцінювання ручного, автоматичного й гібридного сценаріїв за показниками точності (mIoU) та трудомісткості; визначити проблеми довіри оператора до алгоритмічних підказок та ергономіки взаємодії. **Методи**: систематичний порівняльний аналіз гібридних стратегій Human-in-the-Loop, крос-доменний синтез результатів досліджень активного навчання та інтерактивної сегментації, формалізація показників прискорення, якості й відносного обсягу ручної праці. **Досягнуті результати**. Доведено, що гібридний конвеєр (YOLO/SAM та експертна корекція) забезпечує прискорення процесу в 5,4 раза. Час формування семантичної маски об'єкта скорочується з 65 до 12 с за умови збереження якості  $mIoU = 0,94$ , де втрата щодо еталона становить лише 0,02. Установлено, що гібридний сценарій є оптимальним у діапазоні порогових вимог  $0,82 < Q_{min} \leq 0,94$ , що охоплює найширший клас практичних завдань – від навчання виробничих моделей до медичної діагностики. Виявлено пряму залежність якості роботи анотатора від надійності автоматичних підказок, що підтверджує важливість прозорості алгоритмів. **Висновки**. Гібридне анотування є оптимальною стратегією для створення Ground Truth у критичних доменах (автономне водіння, медицина), що забезпечує баланс швидкості й точності. Запропонована формалізація задачі оптимізації з пороговим обмеженням  $Q_{min}$  дає змогу обґрунтовано обирати сценарій анотування для конкретного домену. Перспективи подальших досліджень полягають у вдосконаленні методів генерації синтетичних даних у симульованих середовищах і розробленні адаптивних інтерфейсів для зниження когнітивного навантаження на експертів.

**Ключові слова:** комп'ютерний зір; анотування зображень; активне навчання; сегментація; синтетичні дані.

#### Бібліографічні описи / Bibliographic descriptions

Кучапін М. Ю., Смеляков К. С., Чуприна А. С., Лученко С. В. Комплексний аналіз методів та інструментів гібридного анотування зображень у системах комп'ютерного зору. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 182–195. DOI: <https://doi.org/10.30837/0135-1710.2026.189.182>

Kuchapin, M., Smelyakov, K., Chupryna, A., Luchenko, S. (2026), "Comprehensive analysis of methods and tools for hybrid image annotation in computer vision systems", *Management Information System and Devices*, No. 2 (189), P. 182–195. DOI: <https://doi.org/10.30837/0135-1710.2026.189.182>

Ivan Kushnir, Valentyna Kyrii

## SOFTWARE DEVELOPMENT EFFORT ESTIMATION MODEL BASED ON PRE-TRAINED ZERO-SHOT MODELS

Project management of software development involves planning, execution, and control of work to ensure adherence to budget and deadlines. The subject of the study is effort estimation, which is one of the most critical tasks – it allows managers to minimize risks and optimally allocate resources. With the emergence of artificial intelligence (AI) and the continuation of digital transformation, the application of natural language processing (NLP) methods and large language models (LLM) has become crucial. These tools can analyze textual requirements and user stories to predict the necessary effort. They enable semantic structuring of complex textual requirements and improve estimation accuracy. Using zero-shot models eliminates the need for additional training on target data, saving resources and ensuring greater adaptability. The goal of this work is to develop a model for effort estimation based on software requirements presented in various forms. This is achieved through the following tasks: to develop an effort estimation model for software development based on modern pre-trained zero-shot models, and to conduct an experimental study to evaluate the accuracy of the proposed models. The study used a public dataset and AI models with varying architectures, training methods, and licensing terms. The model evaluation was performed using metrics of mean absolute error, mean relative error, and root mean square deviation. As a result, an effort estimation model for software development was built, which in the experiment showed that the accuracy of effort prediction largely depends on the nature of the dataset and the style of formulating user stories. Compared to generalized models from previous studies, training project-specific models significantly increases prediction accuracy. The proposed effort estimation model based on zero-shot vectorization effectively automates the evaluation of user stories, facilitating resource planning at various stages of the software development lifecycle. The obtained results open the door to the flexible, scalable application of NLP based on large language models in software development project management.

**Keywords:** project management; software development; effort estimation; artificial intelligence; large language models; text embedding.

### 1. Introduction

Software development project management involves the systematic planning, execution, and control of software development activities to ensure the delivery of high-quality products within budget and on schedule. Quantifying the resources and time required to complete a project is one of the core processes in project management. Project managers need accurate estimates to plan their activities, minimize risks, and achieve desired outcomes. An inaccurate estimate can lead to project failure, so this critical process must be approached with care.

Project management theory has evolved in various directions. These include significant engineering achievements and the use of advanced technologies. Project management can be described as the strategic application of knowledge, combined with the experience, tools, and methods necessary to achieve desired goals by managing various resources and risks along the path of plan execution [1]. The rapid digital revolution, which was partly driven by the 2020 pandemic crisis, has led to radical changes in project management practices. The widespread use of artificial intelligence (AI) is now

a driving force, as it takes on routine tasks and increases efficiency, while allowing project managers to focus on the strategic aspects of management. The integration of AI into project operations is taking place across all critical functions, such as resource planning and risk management, as well as team communication. This is prompting project managers to explore new opportunities for using AI in the workplace [2].

Artificial intelligence is generally defined as the simulation by machine systems of human intellectual processes, including the abilities to learn, think logically, make decisions, and solve problems [3–4]. It encompasses highly specialized applications designed for specific functions, as well as general-purpose systems capable of performing any intellectual tasks that a human can perform [5–6]. These systems have a significant impact on software development project management, particularly in the area of estimating development effort.

Despite noticeable progress in the development of project management across various areas, the issue of effort estimations during software development remains challenging. At its core, it determines the success of a project, as deadlines, budget allocation, and the consistency of the final product's quality depend on the



accuracy of these estimates [7–8]. From a practical standpoint, it can be agreed that traditional methods of preliminary labor cost estimation boil down, for the most part, to three approaches: expert estimates, algorithmic models, and methods based on the analysis of historical data.

The latter group includes, in particular, the analogy-based estimation method, which uses information about completed projects to forecast future labor requirements [9, 10]. The adoption of agile methodologies, with their iterative nature and user-story-oriented structure, creates additional challenges for applying traditional estimation paradigms, shifting the focus to the analysis of individual tasks rather than the project as a holistic system [11]. Such dynamics complicate the adaptation of the analogy-based estimation method within agile project management approaches, where the constant evolution of requirements and informal, "noisy" textual specifications – particularly user stories – cause significant difficulties [12]. Textual requirements, which are characterized by domain specificity, terminological ambiguity, and often combine natural language with source code or quantitative metrics, require the use of complex semantic processing methods to identify relevant labor cost indicators [13].

The use of natural language processing (NLP) methods in combination with AI allows for the systematization of text processing. NLP effectively identifies key concepts, establishes semantic relationships between them, and gradually transforms unstructured descriptions into semantically coherent representations [14, 15]. Word embedding models are among the most widely used in addressing these challenges. In such models, each word is converted into a numerical vector generated based on large text corpora, which allows for the consideration of hidden semantic relationships between terms. This is precisely what enables machine learning models to interpret the content of specifications and accompanying textual documentation more accurately [16]. It is also worth noting the impact of large language models (LLMs) built on transformer architectures and attention mechanisms. Such models form contextual representations not only of sentences but also of documents as a whole. This improves the accuracy of extracting key ideas from large text corpora, which is particularly important when working with software requirements [17, 18].

The aim of this study is to develop a model for estimating the effort required for software development

using pre-trained LLM zero-shot models, based on requirements presented in various forms.

## 2. Literature Review and Problem Statement

---

Estimating the effort required for software development projects is a relevant area of scientific research and a practical task.

In [19, 20], current trends in approaches to effort estimations in software development projects were analyzed. The authors highlight machine learning and artificial intelligence methods as key trends in data analysis and the construction of resource cost prediction models. The paper highlighted the features of natural language processing (NLP) models as tools for analyzing textual requirements when effort estimations.

The authors of [21] conducted a comprehensive review of existing studies that use machine learning methods to forecast labor costs in software development. The paper highlighted the high effectiveness of such models based on public and private datasets. However, these models were largely based on algorithmic approaches to estimation, which in turn require large amounts of historical data and demonstrate low effectiveness when working with textual requirements, as noted in [19].

Studies [22, 25] examined the combination of NLP and artificial intelligence methods for estimating resource costs in software development projects. The authors used the codebase as input data in combination with classical estimation models, such as COSMIC and COCOMO. Although the results demonstrate significant effectiveness, it should be noted that the codebase becomes available primarily in the later stages of the development lifecycle, whereas text-based requirements, as artifacts of the early phases, allow for cost estimates to be obtained at earlier stages of project planning.

In [11], a deep learning model is proposed for predicting effort using story point estimates. The described approach combines deep learning algorithms for generating text embeddings with their subsequent integration into prediction models. The authors note the significant potential of applying machine learning methods to the analysis of textual requirements. Among the key challenges, they highlighted the search for and preparation of relevant data for training models. It is this work that laid the

foundation for a number of subsequent studies and is often used as a benchmark.

Expanding on this analysis, the authors [13] conducted a comparative analysis of various text vectorization models, specifically the context-free word2vec and the contextualized BERT models. The contextualized models demonstrated better results, providing higher accuracy in estimating effort in story points. In addition, it was found that the selection of model parameters can significantly affect prediction results.

In the study [24], the FastText model was used for text data vectorization followed by effort classification. The authors treat estimation not as a regression task but as a classification task, clustering the data and applying semi-supervised noise filtering. Afterward, text requirements are transformed into vectors using FastText, and a classification layer is implemented to determine the task complexity level (XS, S, M, L, XL). When using experts for data cleaning, this solution demonstrates good results; however, it does not account for other approaches to text vectorization and is limited to a fixed number of classes, which is a drawback compared to other prediction models.

In [25], the application of pre-trained FastText, GPT-2, and SBERT models in combination with machine learning algorithms, specifically Random Forest and Support Vector Machine (SVM) methods, is analyzed. This work compiles and analyzes existing state-of-the-art models [11, 13, 26]. The authors once again confirmed the high effectiveness of combining modern text vectorization methods with classical machine learning algorithms and developed a model that performs on par with existing ones and even shows better results in some metrics. However, the authors remain focused on using pre-trained models primarily on domain-specific datasets, which limits their scope of application compared to zero-shot models. They also do not consider models capable of processing queries submitted in other formats, such as graphical, multimedia, or formalized.

Thus, thanks to existing research, a set of models has already been established [11, 13, 25, 26] that can serve as a benchmark for future work. At the same time, it should be emphasized that these works leave a significant gap in the systematic evaluation of various pre-trained zero-shot vectorization models. They are mostly focused either on large, finely tuned models or on classical vectorization methods without zero-shot properties. However, the use of zero-shot models can

facilitate the widespread adoption of effort estimation models built on their basis, as these models will not be domain-dependent. Also, most modern works consider only textual representations of requirements when forecasting the required resource volumes, whereas software development projects are not limited in their choice of how to represent requirements.

---

### 3. Research objectives and tasks

The objective of this work is to develop a model for effort estimations based on software requirements presented in various forms. This will allow for the effective estimation of human resource costs for software development at various stages of the project lifecycle.

To achieve this goal, the following objectives were formulated:

- to develop a model for effort estimations for software development based on modern pre-trained zero-shot models;
- conduct an experimental study to evaluate the accuracy of the proposed model, implemented using various pre-trained zero-shot models, on public datasets.

---

### 4. Materials and methods

This work uses a set of text-based requirements presented in [11, 13, 25]. This text dataset contains descriptions of requirements in the form of user stories. Labels corresponding to the development effort for each story, measured in story points, are attached to this dataset. It is important to note that all text requirements in this corpus are presented in English.

The dataset was collected from large open-source project management systems; it contains 23,313 requirements distributed across 16 projects, including well-known projects such as Apache, Appcelerator, DuraSpace, Atlassian, Moodle, Lsstcorp, Mulesoft, Spring, and Talendforge.

Despite the challenges of accurately measuring the actual labor costs required to implement requirements, the authors of the dataset were able to estimate the time taken to complete project phases based on changes in requirement statuses. Specifically, effort was recorded as the time interval from the "in progress" status to the update to "resolved". In [11], statistical analysis using Spearman's and Pearson's correlation tests demonstrated a significant relationship between assigned story points and actual recorded effort.

---

To cover a wide range of architectural and methodological trade-offs, seven competing text vectorization models were selected, identified based on recent large-scale comparative studies [27, 28]. These models differ in vector dimension, training methods, and licensing terms, which allows us to investigate the impact of these factors on the effectiveness of solving applied tasks.

Let us consider the models in order of the dimension of the resulting vector representations of text.

The group with the highest vector dimension includes three models. The Gemini Embedding model (gemini-embedding-001) from Google [29] generates 3,072-dimensional vectors and is built on the large Gemini language model, demonstrating consistently high results in multilingual and English-language benchmarks. It was evaluated on the Massive Multilingual Text Embedding Benchmark (MMTEB) [27], which covers over 100 tasks in more than 250 languages, consistently outperforming previous state-of-the-art solutions in classification, similarity measurement, clustering, ranking, and search. OpenAI's text-embedding-3-large model [30] also generates 3,072-dimensional vectors and is the company's most productive solution in the field of text vectorization, improving MIRACL scores from 44.0% to 54.9% and MTEB scores from 62.3% to 64.6% compared to the smaller text-embedding-3-small version. The reduced model generates 1,536-dimensional vectors – placing it at the lower end of this group – and provides a more computationally efficient and cost-effective option, while maintaining performance at 44.0% on MIRACL and 62.3% on MMTEB [27, 31].

The medium-dimensionality group consists of two models with 1024-dimensional vectors. The Qwen3-Embedding-0.6B model [32] is an open-source solution built on the large language model of the Qwen3 family. It combines scaled unsupervised pre-training with supervised fine-tuning on data covering various domains and languages. The model demonstrates high performance on the multilingual MMTEB tests and in code search tasks [27]. The multilingual-e5-large-instruct model [33] was evaluated simultaneously; it is an open-source, instruction-enhanced vectorization model. It was trained on one billion multilingual text pairs using a contrastive approach and further fine-tuned on labeled datasets. The E5-large-instruct variant demonstrated results that match or exceed those of proprietary models focused solely on

English-language tasks. The two models with the lowest dimensionality in our evaluation generate 768-dimensional vectors. The gte-multilingual-base model [34] is an open-source multilingual text encoder optimized for long-context scenarios. It supports processing long text contexts of up to 8,192 tokens while maintaining high vectorization quality. Architecturally, this model is based on a transformer with an encoder-only architecture, which reduces the model size and ensures significantly lower hardware requirements during execution compared to models based on a decoder-only architecture. This enables nearly a tenfold increase in inference speed. The second model with 768-dimensional vectors – Gecko (text-embedding-005) from Google [35] – is a proprietary solution designed to achieve high search accuracy with a compact vector representation. It implements a new two-stage process for distilling knowledge from large language models (LLMs), which uses a synthetic paired training dataset combined with the selection and relabeling of text fragments.

**Table 1.** Models, their dimensions, and MTEB scores

|                                | Embedding size | MTEB[24] |
|--------------------------------|----------------|----------|
| gemini-embedding-001           | 3072           | 1        |
| text-embedding-3-large         | 3072           | 17       |
| text-embedding-3-small         | 1536           | 33       |
| Qwen3-Embedding-0.6B           | 1024           | 4        |
| multilingual-e5-large-instruct | 1024           | 7        |
| gte-multilingual-base          | 768            | 27       |
| text-embedding-005             | 768            | 170      |

To estimate the labor costs of software development, a feedforward neural network architecture was selected, based on the recommendations and proven effectiveness of similar models in [13].

A feedforward neural network is characterized by a simple and straightforward organization of information flows without feedback, which makes feedforward networks effective tools for solving approximation, classification, and regression problems. Such networks can have two or more layers, as shown in Fig. 1. The first layer is the input layer, and the last is the output layer. The additional layers between the end layers are called hidden layers, whose neurons are called hidden neurons. These neurons act as intermediaries between the input data coming from the external environment and the network's output. It is the number of hidden layers that determines the degree of complexity of the function, which reflects the relationship between the network's input and output vectors. In particular, a single-layer

network describes first-order relationships, a two-layer network describes second-order relationships, and so on.

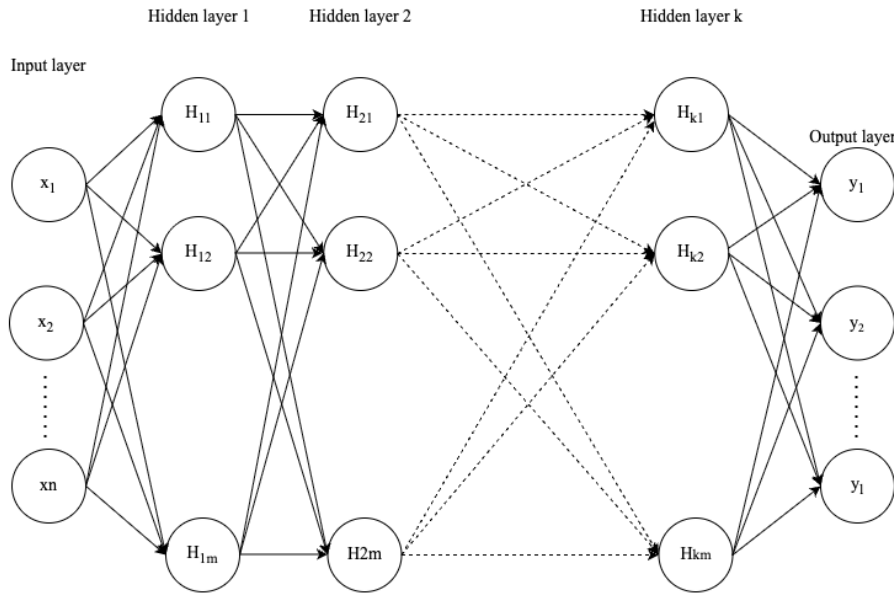


**Fig. 1.** Architecture of a feedforward neural network

Structurally, the neurons in the input layer form the elements of the input vector  $(N)$ , which is fed to the neurons of the next layer. This can be expressed as a formula:

$$N_j = \sum_{i=1}^n Iw_{ij}x_i + B_j,$$

where  $n$  – dimension of the input layer,  $Iw_{ij}$  – the weight matrix for the  $i$ -th input parameter and the  $j$ -th neuron in the hidden layer,  $B_j$  – the displacement vector of the hidden layer's  $j$ -th neuron.

This input vector is then passed through an activation function, producing an output vector that serves as the input for the next layer. It is calculated using the formula:

$$H_{ij} = f(N_j),$$

where  $i$  – hidden layer number,  $j$  – the neuron number in this layer and  $f$  – activation function.

The output signals from the neurons in this hidden layer serve as input for the next layer, and this process is repeated sequentially up to the output (final) layer. The calculation of the output layer's values can be expressed as a formula:

$$y_j = \sum_{i=1}^m Hw_{ij}H_i,$$

where  $m$  – the thickness of the last hidden layer,  $Hw_{ij}$  – the weight matrix between the  $i$ -th neuron of the hidden layer and the  $j$ -th output parameter.

Thus, the set of output signals from the neurons in the final layer determines the network's overall response to the input signal generated by the neurons in the input layer.

Model quality was evaluated using a combination of three metrics: mean absolute error (MAE), root mean square error (RMSE), and mean mean relative error (MMRE), which are among the most popular and widely used methods for evaluating models in this field [19]. MAE measures the mean absolute value of errors in the same units as the target variable, allowing for an intuitive assessment of typical deviations of predictions from actual values. It is calculated using the formula:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|.$$

RMSE amplifies the impact of large forecast errors on the model's overall accuracy by squaring the error values. It is calculated using the formula:

$$PMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}.$$

To assess relative accuracy, the MMRE metric was used, which normalizes prediction errors by dividing them by the actual values; this is particularly useful in fields where project sizes vary widely, as relative errors

provide a more meaningful assessment than absolute errors. It is calculated using the formula:

$$MMRE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i}$$

Two approaches were used to work with embedding vector generation models. The first approach involved running the models locally using the relevant Python libraries, but this is only possible for publicly available models. In the case of proprietary models developed by Google and OpenAI, vectorization is implemented via API calls to the respective services.

## 5. Research results

### 5.1. A model for estimating the labor costs of software development based on modern pre-trained zero-shot models

Let's consider the general model for estimating the efforts required for software development, shown in Fig. 2.

The input data consists of software requirements, which are documented and stored in a requirements management system; this may be a subsystem of a general project management system. The model's governing elements include domain standards, text vectorization methods, and project management methods. The model's output is a numerical representation of the labor costs required to develop the specified requirements.

Let us note the model's limitations. This model is configured for a specific project or similar projects, as the assessment of software development complexity depends heavily on the team and the domain. Therefore, the historical data used to train the model must come from this project or a very similar one. The model is designed to work with numerical estimates of the labor costs for software development. If the development complexity assessment uses other metrics, such as the "T-shirt size" estimation method, it may not work.

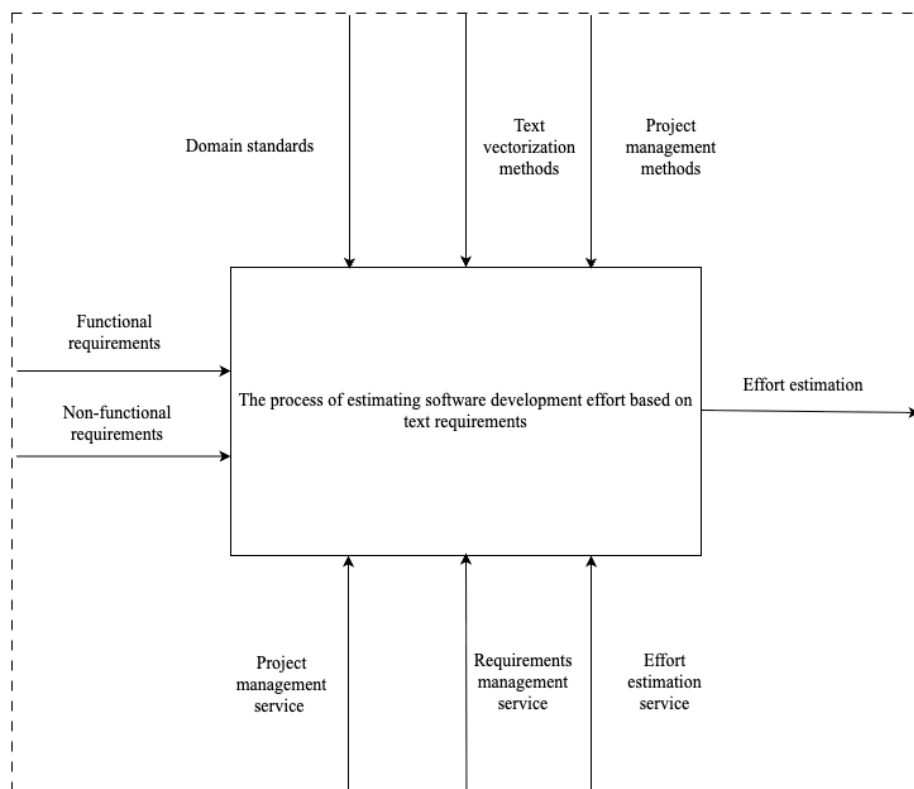


Fig. 2. General model for estimating the efforts required for software development

Let's examine the model's operation in more detail in Fig. 3. The input parameters, which represent the requirements, can be represented as a tuple  $X$ :

$$X = \{\{FR\}, \{NFR\}\},$$

where  $FR$  – functional requirements, specific functions that the software must perform;  $NFR$  – Non-functional requirements describe system characteristics that are not directly related to business functions; these may include security, productivity, and other requirements.



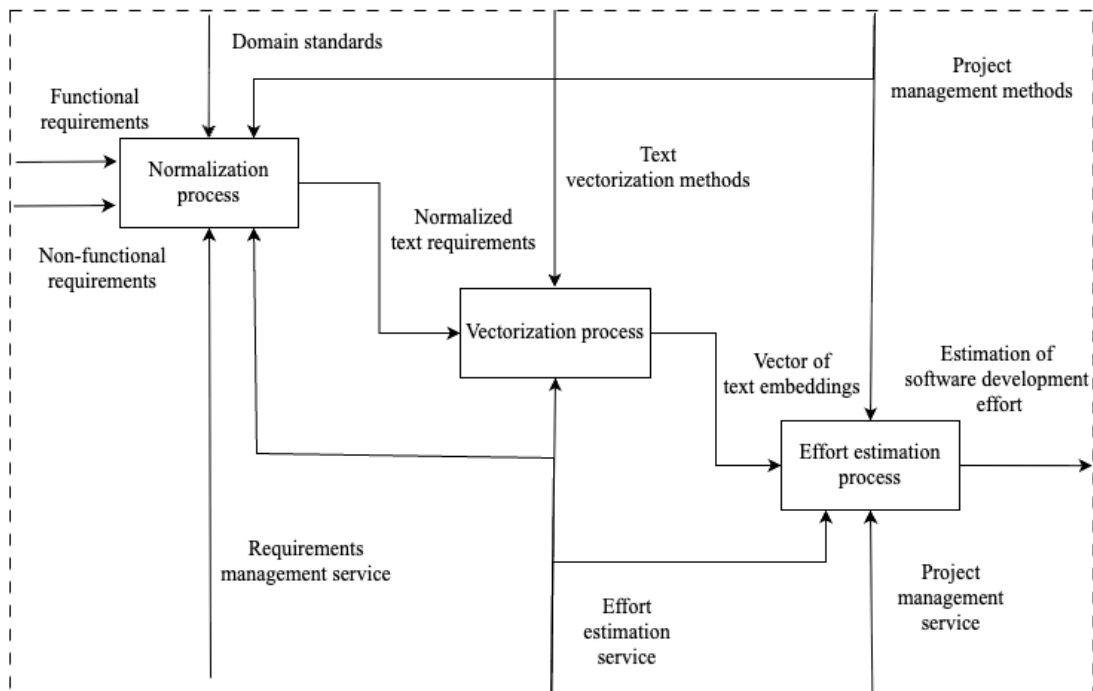


Fig. 3. Stages of estimating the efforts required for software development

In turn, the functional requirements can be presented as follows:

$$FR = \{\{TXT\}, \{MLT\}, \{FRM\}\},$$

where *TXT* – requirements presented in text form; *MLT* – visual requirements, which include mockups, interface prototypes, and so on, *FRM* – structured requirements presented in tabular form or as lists.

Non-functional requirements can be described as a tuple:

$$NFR = \{\{QAS\}, \{GRP\}\},$$

where *QAS* – quality attribute scenarios, a common approach to the structured presentation of non-functional requirements; *GRP* – The requirements are presented in graphical form, including UML diagrams, block diagrams, and more

The first step in the model is the normalization process – it is necessary to prepare the data for use with vectorization methods. There are vectorization methods that can work with different types of input data, but it is necessary to obtain a single vector from all available requirements presented in various formats. Therefore, it is at this step that we convert them to a text type and combine them into a single continuous output text.

The normalization process can be represented as a formula:

$$x' = f(X),$$

where  $x'$  – textual description of requirements;  $X$  – the previously described set of requirements;  $f$  – normalization function.

The normalization results are used as input for the vectorization process, which converts textual information into a vector of text embeddings that describes the semantic meaning of the requirements.

This can be expressed as a formula:

$$V = g(x'),$$

where  $x'$  input text description of requirements;  $V$  – output numeric vector of text embeddings;  $g$  – vectorization function.

The quality of the results from this step has the greatest impact on the model's performance. After all, the semantic elements extracted from the text embeddings are the key components of the model's operation.

Once the features have been transformed into a vector, we proceed to the prediction process, which can be represented by the following formula:

$$y = h(V),$$

where  $V$  – the result of the previous stage,  $h$  – forecasting function,  $y$  – workload assessment in story points.

For forecasting, we will use machine learning models, specifically a multi-layer feedforward neural

network. It can be represented as a tuple and expressed by the following formulas:

$$\{N_{ij}, H_{ij}, I_{ij}, V\}, \quad (1)$$

$$N_{ij} = \sum_{k=1}^n I w_{kij} H_{kj-1} + b_{ij}, \quad (2)$$

$$H_{ij} = f(N_{ij}), \quad (3)$$

$$H_{i0} = V, \quad (4)$$

where  $N_{ij}$  – input values to the  $i$ -th neuron of the  $j$ -th layer,  $H_{ij}$  – output values to the  $i$ -th neuron of the  $j$ -th layer,  $w_{ki}$  – the weight coefficient of the connection between the  $k$ -th neuron of the previous layer and the  $i$ -th neuron of the  $j$ -th layer,  $V$  – the input layer of a neural network, consisting of text embeddings,  $n$  – the number of neurons in the previous layer.

To train the model, we will use the backpropagation algorithm. A key component of this method is calculating the model's error. It can be represented as follows:

$$E = C(y, \hat{y}), \quad (5)$$

where  $E$  – is a forecasting error,  $y$  – output model value,  $\hat{y}$  – is the expected value,  $C$  – loss function.

The first step in training the model is to use the input vector as the input layer of the neural network and perform a forward pass using equations (2) – (4).

Next, the error of the output layer is calculated using the formula:

$$\delta_i^J = \frac{dC}{dH_{iJ}} f'(N_{iJ}),$$

where  $\delta$  – error,  $i$  – neuron number,  $J$  – output layer number,  $C$  – loss function,  $f'$  – the derivative of the activation function.

The error is then calculated for each layer (backpropagation) using the formula:

$$\delta_i^j = \left[ \sum_{k=1}^n \delta_k^{j+1} \times w_{kij}^{j+1} \right] f'(N_{ij}),$$

where  $j$  – layer number  $j = J - 1 \dots 1$ ,  $w$  – weight coefficients of neural connections.

Next, the weight coefficients are updated using the errors calculated at each layer of the neural network, using the formula:

$$w_i = w_{i-1} - o(\delta),$$

where  $w_i$  – updated weight coefficient value,  $w_{i-1}$  – the value of this coefficient in the previous iteration,  $o$  – an optimization function that calculates the bias based on the error calculated earlier.

These steps are performed at each iteration of the model training, allowing us to obtain the weights of the neural connections that best describe the relationship between the input and output parameters.

## 5.2. An experimental study to evaluate the accuracy of the proposed models on public datasets

We will conduct the experiment using data from the dataset [11], which consists of textual representations of requirements and includes a brief description of the user's history (title) and a detailed description (description). We will begin by normalizing the textual requirements. This process aims to reduce noise while preserving the semantic context of the requirement descriptions. In accordance with established practices [13], the cleaning stage involves removing special symbols (quotation marks, hashtags) and numerical values. Next, the entire text is converted to lowercase for unification and to facilitate tokenization. Additionally, stop words – tokens and symbols that do not carry meaningful information and typically arise from data exported from project management systems – are removed. Using a specialized stop list for this dataset helps filter out redundant or repetitive tokens, improving signal quality for subsequent vector representation and modeling. The list of special symbols and stop words to be removed during normalization is presented in Table 2.

**Table 2.** List of special symbols and stop words removed during normalization

|       |     |   |       |   |        |   |   |   |   |      |        |   |   |   |   |
|-------|-----|---|-------|---|--------|---|---|---|---|------|--------|---|---|---|---|
| 0     | 1   | 2 | 3     | 4 | 5      | 6 | 7 | 8 | 9 | 0    | !      | “ | # | % | & |
| ,     | (   | ) | *     | + | ,      | - | . | / | : | ;    | <      | = | > | ? | @ |
| [     | \   | ] | ^     | _ | '      | { | } |   | ~ | html | {html} |   |   |   |   |
| <div> | <p> |   | <pre> |   | <code> |   |   |   |   |      |        |   |   |   |   |

After text normalization, the total number of tokens was reduced by approximately 1.5 million and now stands at 2.7 million, as calculated using the text-embedding-3-large model (Fig. 4). At the same time,

the average number of tokens per unit of text was reduced from 182 to 116.

After normalization, we proceed to convert the text features into vector form so that this transformed data

can be used in the next step. This process is performed separately for each of the models mentioned earlier, and each subsequent step is carried out in parallel for all generated text embeddings. The dimension of the output vectors depends on the models used.

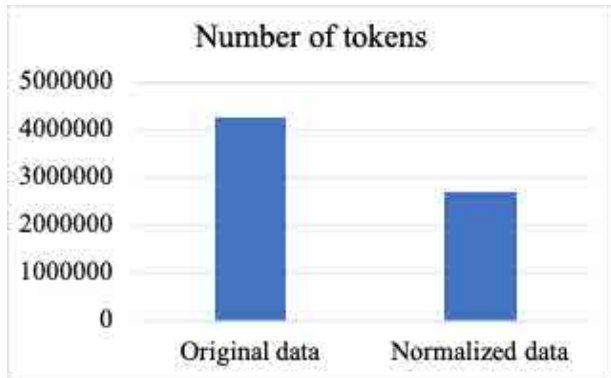


Fig. 4. Number of tokens before and after normalization

Next, we need to train the neural network responsible for forecasting. To do this, we must split the data into training and test sets, which will help evaluate the models' performance. For this, we use K-fold cross-validation. The number of folds was chosen dynamically depending on the project size to balance the ratio between the volume of training data and the reliability of validation: for projects with 1,000 or more samples, 10 folds were used; for projects with 500 to 999 samples, 5 folds; for 300–499 samples, 3 folds; and for smaller projects, 2 folds. This adaptive strategy ensures that each validation set is sufficiently representative without unduly reducing the training sample, thereby providing reliable and objective evaluations of productivity even with limited data. Results across all folds were averaged to form a comprehensive assessment of each model's predictive capabilities.

We will use a feedforward neural network for forecasting. The network consists of an input layer, the size of which corresponds to the dimension of the vectors resulting from the text transformation, followed by three dense hidden layers with the ReLU function as the activation function, and concludes with a linear output layer that generates the labor cost estimate. This architecture allows for a balance between complexity and generalization ability, ensuring effective modeling of nonlinear dependencies between text representations of user stories and the amount of labor required. Key hyperparameters were dynamically

selected according to the characteristics of the dataset: the size of the hidden layers was set to half the dimension of the vectors, with a maximum limit of 512 neurons to maintain an optimal model size. The batch size was fixed at 32, and the Adam algorithm with standard learning rate settings was used for optimization due to its adaptability and success in training similar models.

To prevent overfitting and reduce unnecessary training time, an early stopping mechanism was implemented, which monitors the mean absolute error on the validation set with a limit of 10 epochs. That is, if the model's quality does not improve over 10 iterations, we will consider that the model has achieved the best possible result in this configuration. These settings reflect a practical yet effective approach, which was validated in [18] and serves as a reliable foundation for model creation in various software development projects.

Table 3. Description of the layers of the direct expansion model

| Layer    | Dimension                   |
|----------|-----------------------------|
| Input    | N (size of text embeddings) |
| Hidden 1 | $\text{MIN}(N/2, 512)$      |
| Hidden 2 | $\text{MIN}(N/2, 512)$      |
| Hidden 3 | $\text{MIN}(N/2, 512)$      |
| Output   | 1                           |

Separate models were built for each software project in the input data, since models for estimating effort based on text data are typically project-specific and generally do not generalize well across different projects. This approach accounts for the natural variability in project characteristics, requirements, and development practices that influence the relationship between text-based user stories and the corresponding effort. By training and evaluating models for each project separately, we aim to identify project-specific patterns and improve the accuracy and relevance of predictions within the unique context of each project.

As a result of training the prediction models, the following quality metrics were obtained. The predictions based on vector representations from the gemini-embedding-001 model showed the lowest mean absolute error. On average, the error is 2.7 story points (SP). Thus, when estimating the effort required to develop a feature using a 13SP model, the actual effort will range from 10.3 to 15.7.

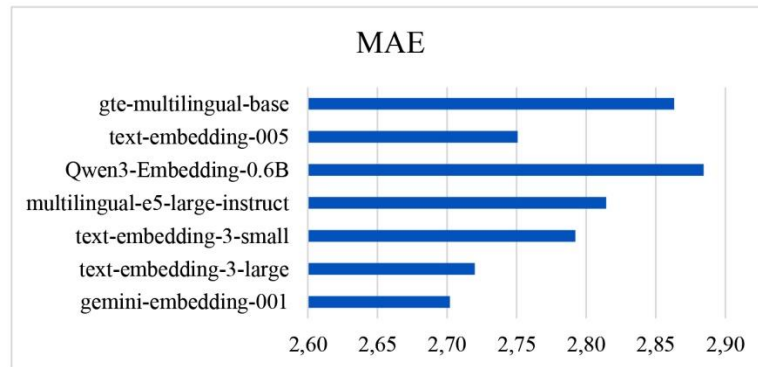


Fig. 5. Mean absolute error

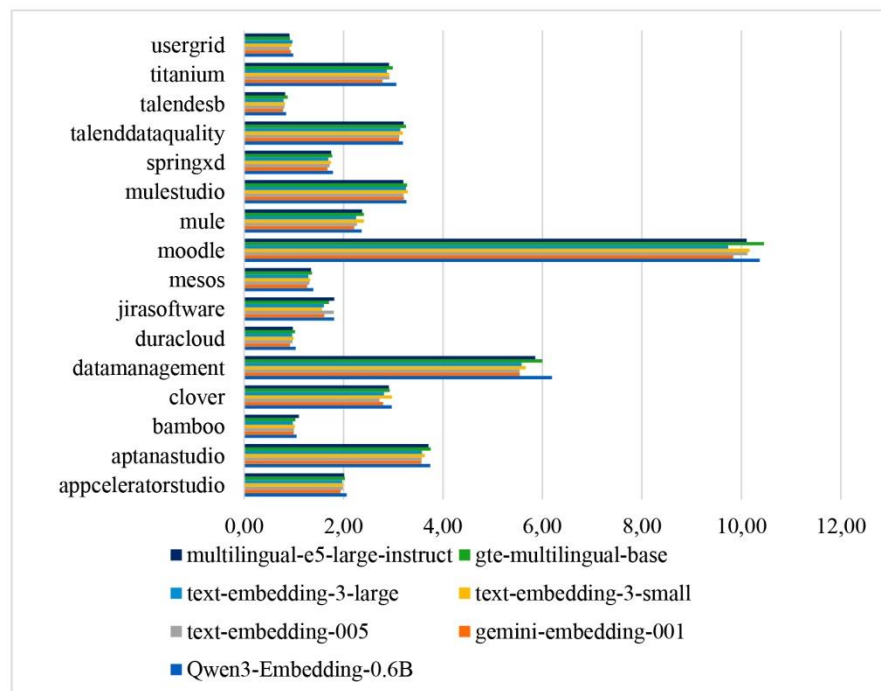


Fig. 6. Average absolute error by projects

The following evaluation is measured using the root mean square error, which is a more rigorous metric. Based on these error metrics, we can observe that gemini-embedding-001 does not perform the best (4.69), while

text-embedding-3-large takes its place with a score of 4.66. This indicates that when evaluating more complex requirements, which will result in a high SP score, the error of the second model will be lower.

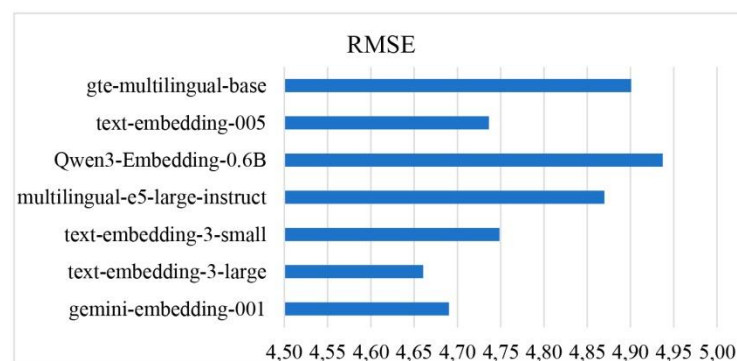


Fig. 7. Root of the root mean square deviation

Last but not least, let's examine the average relative error. Based on this metric, gemini-embedding-001 once again takes first place. It shows us that, on average, the estimation error is 0.64, or 64%. This relatively high

figure may be due to the fact that the test data contains a large number of requirements with relatively low expert estimates of effort. About 70% of the records in the data used have an estimate below 8 SP.

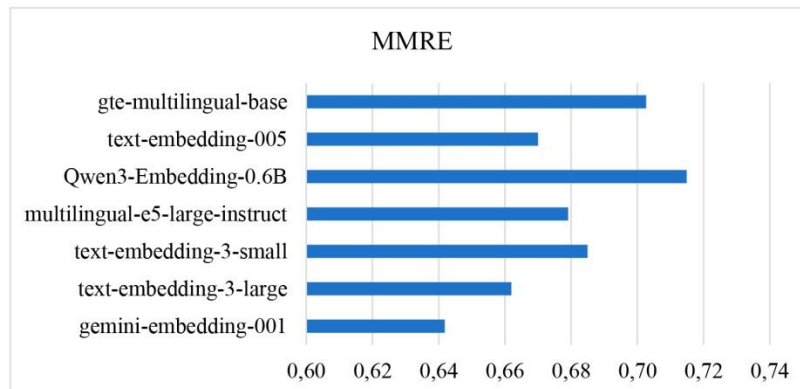


Fig. 8. Mean relative error

## 6. Discussion of the results

The experiment revealed that forecasting accuracy depends significantly on the characteristics of the dataset and the style of user story documentation. For example, the Talend ESB project showed the lowest error rates (MAE ranging from 0.79 to 0.88), while Moodle had the highest errors (MAE ranging from 9.73 to 10.37).

Based on averaged metrics across all projects for each embedding model, gemini-embedding-001 and text-embedding-3-large – the most advanced models from Google and OpenAI, respectively – performed best. These models generate the longest embedding vectors among those considered, suggesting that higher dimensionality is better able to capture the semantic context of software development requirements, which in turn directly impacts prediction accuracy.

When comparing results by MAE, the results are quite closely clustered – ranging from 2.70 (best, Gemini) to 2.88 (worst, qwen-06). Overall, across all error metrics, the qwen-06 and gte-multilingual models consistently ranked last. Given that they belong to models with medium and short embedding vector lengths, one might assume that this is the reason for their poor performance. However, the use of text-embedding-005 contradicts this. While falling within the same range of vector dimensions, it yields results similar to those of models with longer vectors. This indicates a significant influence of the training datasets, the training method itself, and the architecture of these models. Next, we should compare the obtained results with state-of-the-art models [11, 13, 25, 26]. Since most studies evaluated the quality of the built models using the MAE metric, we will use this metric as well.

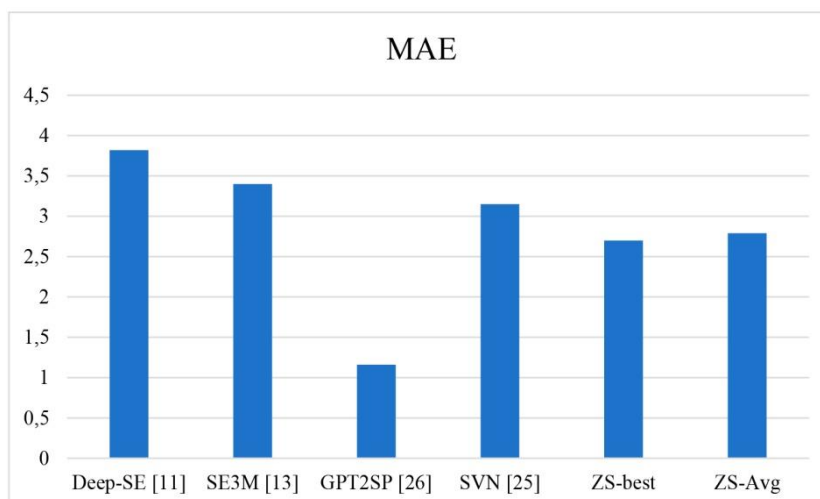


Fig. 9. MAE metrics for the developed model (ZS) and existing state-of-the-art models



The diagram shows two metrics obtained from the results of applying the developed model, namely the best metric (ZS-best) and the average metric (ZS). Overall, it can be noted that the use of modern zero-shot models in combination with feedforward neural networks yields good results compared to the metrics in existing studies [11, 13, 25], which use pre-trained models with fine-tuning on domain-specific data. In turn, GPT2SP [26] remains the leader in evaluation accuracy, and it is worth noting that such accuracy is achieved by building upon and fine-tuning the existing GPT2 model, which generally requires comparatively greater effort than using ready-made, publicly available, or proprietary zero-shot models.

## 7. Conclusions

1. A model for effort estimations based on the vectorization of software requirements has been developed. This model helps avoid the additional use of resources – in this case, experts – to evaluate user stories during planning at various stages of the software development project lifecycle. This can significantly help project managers plan time and labor resources.

2. An experimental study was conducted on the impact of using different text vectorization language models on the overall quality of labor cost prediction models based on text requirements. Considering most quality metrics, the gemini-embedding-001 model with a text embedding vector of 3072 showed the best results (MAE – 2.70, MMRE – 0.64). At the same time, the text-embedding-005 model with a significantly smaller vector (769) showed similar performance and is the best among models with small and medium-sized vectors.

3. For further development of the model for estimating effort based on software requirements, it is worth considering the possibility of adding a preprocessing step for different formats of requirement specifications. This should help identify structural elements that best reflect the complexity of implementing user stories in the generated embedding vectors. Another approach is to incorporate project-specific information, such as long-term or short-term goals, into the functionality requirements when generating text embeddings, which could improve the quality of the vectors for use in forecasting models.

## Conflict of Interest

The authors declare that they have no conflicts of interest, including financial, personal, authorship, or any other conflicts that could influence the study or the results published in this article.

## Funding

The study was conducted without financial support.

## Data Availability

Data will be provided upon reasonable request.

## Use of Artificial Intelligence

The authors confirm that they did not use artificial intelligence technologies to write this paper.

## References

1. Project Management Institute (2017), *A Guide to the Project Management Body of Knowledge (PMBOK® Guide)*, 6th edn. Newtown Square, PA: Project Management Institute.
2. Project Management Institute Sweden (2024), *Artificial Intelligence and Project Management: A Global Chapter-Led Survey 2024*, Stockholm: Project Management Institute.
3. Taboada, I., Daneshpajouh, A., Toledo, N., de Vass, T. (2023), "Artificial Intelligence Enabled Project Management: A Systematic Literature Review", *Applied Sciences*, Vol. 13(8), p. 5014. DOI: <https://doi.org/10.3390/app13085014>
4. Martovytskyi, V., Argunov, V., Ruban, I., Romanenkov, Y. (2023), "Developing a risk management approach based on reinforcement training in forming an investment portfolio", *Eastern-European Journal of Enterprise Technologies*, Vol. 2(3(122)), pp. 106–116. DOI: <https://doi.org/10.15587/1729-4061.2023.277997>
5. Dam, H. K., Tran, T., Grundy, J., Ghose, A., Kamei, Y. (2019), "Towards Effective AI-Powered Agile Project Management", *Proceedings of the 2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*, Montreal, Canada, 25-31 May. IEEE, pp. 41–44. DOI: <https://doi.org/10.48550/arXiv.1812.10578>

6. Serdechnyi, V., Barkovska, O., Kovalenko, A., Havrashenko, A., Martovytskyi, V. (2025), "Research on machine learning methods for detecting objects in difficult shooting conditions", *Radioelectronic and Computer Systems*, Vol. 2, pp. 64–77. DOI: <https://doi.org/10.32620/reks.2025.2.04>
7. Azzeh, M. (2011), "Adjusted case-based software effort estimation using bees optimization algorithm", *Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 315–324. DOI: [https://doi.org/10.1007/978-3-642-23863-5\\_32](https://doi.org/10.1007/978-3-642-23863-5_32)
8. Bardsiri, V. K., Jawawi, D. N. A., Bardsiri, A. K., Khatibi, E. (2013), "LMES: A localized multi-estimator model to estimate software development effort", *Engineering Applications of Artificial Intelligence*, Vol. 26(10), pp. 2624–2640. DOI: <https://doi.org/10.1016/j.engappai.2013.08.005>
9. Jorgensen, M., Shepperd, M. (2006), "A systematic review of software development cost estimation studies", *IEEE Transactions on Software Engineering*, Vol. 33(1), pp. 33–53. DOI: <https://doi.org/10.1109/TSE.2007.256943>
10. Alsaadi, B., Saeedi, K. (2022), "Data-driven effort estimation techniques of agile user stories: a systematic literature review", *Artificial Intelligence Review*, Vol. 55(7), pp. 5485–5516. DOI: <https://doi.org/10.1007/s10462-021-10132-x>
11. Choetkiertikul, M., Dam, H. K., Tran, T., Pham, T., Ghose, A., Menzies, T. (2018), "A deep learning model for estimating story points", *IEEE Transactions on Software Engineering*, Vol. 45(7), pp. 637–656. DOI: <https://doi.org/10.48550/arXiv.1609.00489>
12. Usman, M., Mendes, E., Weidt, F., Britto, R. (2014), "Effort estimation in agile software development: a systematic literature review", *Proceedings of the 10th International Conference on Predictive Models in Software Engineering (PROMISE '14)*, Turin, Italy, 18 September, New York: ACM, pp. 82–91. DOI: <https://doi.org/10.1145/2639490.2639503>
13. Fávero, E., Casanova, D., Pimentel, A. (2020), "SE3M: A Model for Software Effort Estimation Using Pre-trained Embedding Models", *arXiv*. DOI: <https://doi.org/10.48550/arXiv.2006.16831>
14. Moharrerri, K., Sapre, A. V., Ramanathan, J., Ramnath, R. (2016), "Cost-Effective Supervised Learning Models for Software Effort Estimation in Agile Environments", *Proceedings - International Computer Software and Applications Conference*, 2, pp. 135–140. DOI: <https://doi.org/10.1109/COMPSAC.2016.85>
15. Zhang, C., Tong, S., Mo, W., Zhou, Y., Xia, Y., Shen, B. (2016), "ESSE: an early software size estimation method based on auto-extracted requirements features", *Internetwork '16: Proceedings of the 8th Asia-Pacific Symposium on Internetwork*, Beijing, China, 18 September, New York: Association for Computing Machinery, pp. 112–115. DOI: <https://doi.org/10.1145/2993717.2993733>
16. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013), "Efficient estimation of word representations in vector space", *arXiv preprint*. DOI: <https://doi.org/10.48550/arXiv.1301.3781>
17. Floridi, L., Chiriatti, M. (2020), "GPT-3: Its Nature, Scope, Limits, and Consequences", *Minds and Machines*, 30(4), pp. 681–694. DOI: <https://doi.org/10.1007/s11023-020-09548-1>
18. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2019), "BERT: Pre-training of deep bidirectional transformers for language understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, 2–7 June, Stroudsburg, PA: Association for Computational Linguistics, pp. 4171–4186. DOI: <https://doi.org/10.18653/v1/N19-1423>
19. Jadhav, A., Kaur, M., Akter, F. (2022), "Evolution of software development effort and cost estimation techniques: five decades study using automated text mining approach", *Mathematical Problems in Engineering*, Vol. 1, 5782587. DOI: <https://doi.org/10.1155/2022/5782587>
20. Rossi, B. B., Fontoura, L. M. (2025), "AI-Based Approaches for Software Tasks Effort Estimation: A Systematic Review of Methods and Trends", *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS)*, Vol. 2, pp. 144–151. DOI: <https://doi.org/10.5220/0013218200003929>
21. Mahmood, Y., Kama, N., Azmi, A., Khan, A. S., Ali, M. (2022), "Software effort estimation accuracy prediction of machine learning techniques: A systematic performance evaluation", *Software: Practice and experience*, Vol. 52(1), pp. 39–65. DOI: <https://doi.org/10.1002/spe.3009>
22. Tenekeci, S., Ünlü, H., Dikenelli, E., Selçuk, U., Soylu, G.K., Demirörs, O. (2024), "Predicting software size and effort from code using natural language processing", in *Joint Proceedings of the 33rd International Workshop on Software Measurement and the 18th International Conference on Software Process and Product Measurement (IWSM-MENSURA 2024)*, Montréal, Canada, September, Aachen: CEUR-WS, Vol. 3852
23. Sharma, A., Chaudhary, N. (2022), "Analysis of software effort estimation based on story point and lines of code using machine learning", *International Journal of Computing and Digital Systems*, Vol. 12(1). DOI: <https://journal.uob.edu.bh/10.12785/ijcds/1201012>
24. Catak, T., Durdu, P.O., Omurca, S.I. (2024), "Enhancing agile effort estimation: An nlp approach for software requirements analysis", *2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, IEEE, pp. 1–8. DOI: <https://doi.org/10.1109/HORA61326.2024.10550870>

25. Atoum, I., Ootom, A. A. (2024), "Enhancing software effort estimation with pre-trained word embeddings: A small-dataset solution for accurate story point prediction", *Electronics*, Vol. 13(23), p. 4843. DOI: <https://doi.org/10.3390/electronics13234843>
26. Fu, M., Tantithamthavorn, C. (2022), "GPT2SP: A transformer-based agile story point estimation approach", *IEEE Transactions on Software Engineering*, Vol. 49(2), pp. 611–625. DOI: <https://doi.org/10.1109/TSE.2022.3158252>
27. Enevoldsen, K. et al. (2025), "MMTEB: Massive Multilingual Text Embedding Benchmark", *arXiv preprint*. DOI: <https://doi.org/10.48550/arXiv.2502.13595>
28. Shahinmoghadam, M., Motamedi, A. (2025), "Benchmarking pre-trained text embedding models in aligning built asset information", *Scientific Reports*, Vol. 15(1), p. 23866. DOI: <https://doi.org/10.1038/s41598-025-09052-5>
29. Lee, J., Chen, F., Dua, S., Cer, D., Shanbhogue, M., Naim, I., Duerig, T. (2025), "Gemini embedding: Generalizable embeddings from gemini", *arXiv preprint*. DOI: <https://doi.org/10.48550/arXiv.2503.07891>
30. OpenAI (2024), *OpenAI documentation*, available at: <https://platform.openai.com/docs/> (last accessed 12.05.2026)
31. Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Lin, J. (2022), "Making a MIRACL: Multilingual Information Retrieval Across a Continuum of Languages", *arXiv preprint*. DOI: <https://doi.org/10.48550/arXiv.2210.09984>
32. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Zhou, J. (2025), "Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models", *arXiv preprint*. DOI: <https://doi.org/10.48550/arXiv.2506.05176>
33. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F. (2024), "Multilingual e5 text embeddings: A technical report", *arXiv preprint*. DOI: <https://doi.org/10.48550/arXiv.2402.05672>
34. Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., Zhang, M., Li, W., Zhang, M. (2024), "mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval", *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, Miami, Florida: Association for Computational Linguistics, pp. 1198–1213. DOI: <https://doi.org/10.18653/v1/2024.emnlp-industry.103>
35. Lee, J., Dai, Z., Ren, X., Chen, B., Cer, D., Cole, J. R., Naim, I. (2024), "Gecko: Versatile text embeddings distilled from large language models", *arXiv preprint*. DOI: <https://doi.org/10.48550/arXiv.2403.20327>

Received (Надійшла) 07.04.2026

Accepted for publication (Прийнята до друку) 15.05.2026

Publication date (Дата публікації) 29.05.2026

#### Відомості про авторів / About the Authors

**Кушнір Іван Сергійович** – Харківський національний університет радіоелектроніки, аспірант кафедри електронних обчислювальних машин, Харків, Україна;

**Ivan Kushnir** – Kharkiv National University of Radio Electronics, Postgraduate Student, Department of Electronic Computers, Kharkiv, Ukraine;

e-mail: [ivan.kushnir@nure.com](mailto:ivan.kushnir@nure.com)

ORCID ID: <https://orcid.org/0009-0004-5055-0331>

**Кириї Валентина Василівна** – кандидат економічних наук, доцент, Харківський національний університет радіоелектроніки, завідувач відділу аспірантури та докторантури, Харків, Україна;

**Valentyna Kyrii** – PhD (Economic Sciences), Associate Professor, Kharkiv National University of Radio Electronics, Head of the Department of Postgraduate Studies and Doctoral Studies, Kharkiv, Ukraine;

e-mail: [valentyna.kyrii@nure.ua](mailto:valentyna.kyrii@nure.ua)

ORCID ID: <https://orcid.org/0000-0002-2537-264X>

## МОДЕЛЬ ОЦІНЮВАННЯ ТРУДОВИТРАТ НА РОЗРОБЛЕННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ НА ОСНОВІ ПОПЕРЕДНЬО НАВЧЕНИХ ZERO-SHOT-МОДЕЛЕЙ

Управління проєктами створення програмного забезпечення (ПЗ) передбачає планування, виконання та контроль робіт для забезпечення дотримання бюджету й строків. **Предметом дослідження** є оцінювання трудовитрат, що є одним

із найвагоміших завдань, оскільки дає змогу керівникам мінімізувати ризики й оптимально розподілити ресурси. За появи штучного інтелекту (ШІ) й продовження цифрової трансформації ключового значення набуває застосування методів оброблення природної мови (NLP) і великих мовних моделей (LLM). Ці інструменти здатні аналізувати текстові вимоги й користувацькі історії для прогнозування необхідних зусиль. Вони допомагають семантично структурувати складні текстові вимоги й покращувати точність оцінювання. Використання zero-shot-моделей не потребує донавчання на цільових даних, що економить ресурси й забезпечує високу адаптивність. **Мета дослідження** – розроблення моделі для оцінювання трудовитрат на основі вимог до ПЗ, поданих у різних формах. Це досягається за допомогою таких завдань: розроблення моделі оцінювання трудовитрат на створення ПЗ на основі сучасних попередньо навчених zero-shot-моделей; експериментальне дослідження процесу оцінювання точності запропонованих моделей. **Результати.** У роботі застосовано публічний набір даних і моделі ШІ, що різняться архітектурою, методами навчання й ліцензійними умовами. Моделі оцінювалися за метриками середньої абсолютної похибки, середньої відносної похибки й кореня середньоквадратичного відхилення. Побудовано модель оцінювання трудовитрат на розроблення ПЗ, яка в процесі експерименту продемонструвала, що точність прогнозування трудовитрат значною мірою залежить від характеру набору даних і стилю формулювання користувацьких історій. Порівняно з узагальненими моделями попередніх досліджень тренування проєктно-специфічних моделей значно підвищує точність прогнозів. Запропонована модель оцінювання трудовитрат на основі zero-shot-векторизації ефективно автоматизує оцінювання користувацьких історій, полегшуючи планування ресурсів на різних етапах життєвого циклу створення ПЗ. Досягнуті результати відкривають перспективи для гнучкого й масштабованого застосування NLP на основі великих мовних моделей у сфері управління проєктами розроблення програмного забезпечення.

**Ключові слова:** управління проєктами; розроблення програмного забезпечення; оцінювання трудовитрат; штучний інтелект; великі мовні моделі; векторизація тексту.

#### *Бібліографічні описи / Bibliographic descriptions*

Кушнір І. С., Кирій В. В. Модель оцінювання трудовитрат на розроблення програмного забезпечення на основі попередньо навчених zero-shot-моделей. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 196–210. DOI: <https://doi.org/10.30837/0135-1710.2026.189.196>

Kushnir, I., Kyrii, V. (2026), " Software development effort estimation model based on pre-trained zero-shot models", *Management Information System and Devices*, No. 2 (189), P. 196–210. DOI: <https://doi.org/10.30837/0135-1710.2026.189.196>

---

Маковецький С. О., Каук В. І.

## МЕТОД ПРОЄКТУВАННЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ КРАЙОВИХ ВУЗЛІВ ІНТЕРНЕТ-РЕЧЕЙ MESH-МЕРЕЖ В УМОВАХ ХИБНИХ СПРАЦЮВАНЬ

**Предметом вивчення** є моделі й методи розроблення програмного забезпечення крайових вузлів (*edge nodes*) розподілених сенсорних мереж Інтернету речей з топологією *mesh*, що функціонують у нестаціонарному шумовому середовищі. **Мета дослідження** – розробити архітектурну модель і метод проєктування програмного забезпечення крайових вузлів, які поєднують спектральне виявлення подій із адаптивним відстеженням рівня шуму й локальним прийняттям рішень без залучення хмарного оброблення. **Завдання:** формалізувати трирівневу декомпозицію програмного забезпечення вузла; запропонувати шаблон триступеневого придушення хибних спрацювань як ортогональну композицію функціональних захистів; проаналізувати обчислювальну складність і обсяг пам'яті для двох альтернативних реалізацій шаблону; отримати аналітичну модель зростання трафіку. **Методи дослідження:** архітектурне моделювання програмних систем, теорія цифрового оброблення сигналів (швидке перетворення Фур'є, медіанна й експоненціальна фільтрація), аналітичне моделювання мережевого трафіку. **Досягнуті результати.** Сформовано трирівневу архітектурну модель, що ортогонально поєднує селективну спектральну фільтрацію, темпоральну персистентну фільтрацію та адаптивне відстеження рівня шуму. Розроблено шестиступінний метод проєктування з критеріями верифікації кожного компонента. Аналітично продемонстровано, що обидва варіанти реалізації шаблону задовольняють обмеження мікроконтролерів класу ARM Cortex-M0/M0+ із тактовою частотою 64 МГц у межах розрахункових 21,5 мс на кадр і 2,7 КБ оперативної пам'яті. **Висновки.** Запропоновані модель і метод подано на рівні шаблону проєктування, що допускає повторне використання в різних апаратних і доменних інстанціях. Шаблон ідентифікує три взаємодоповнювальні захисні механізми та встановлює, що відсутність будь-якого з них породжує клас хибних спрацювань, не покритий рештою. Результати слугують методологічною основою розроблення програмного забезпечення для розподілених систем визначення руху об'єктів і периметральної безпеки.

**Ключові слова:** розподілені системи; *mesh*-мережі; архітектура програмного забезпечення; шаблон проєктування; Інтернет речей; визначення руху об'єктів; адаптивне порогове детектування; цифрове оброблення сигналів; вбудовані системи; мікроконтролери.

### 1. Вступ

Розподілені сенсорні системи на основі архітектури Інтернету речей із *mesh*-топологією дедалі частіше застосовуються для завдань, у яких визначальну роль одночасно відіграють три вимоги: висока ймовірність правильного виявлення цільових подій, низька частота хибних спрацювань і обмежений енергетичний бюджет вузла. Серед прикладних доменів – периметральна безпека критичної інфраструктури, екологічний моніторинг, логістика безпілотних транспортних засобів, виявлення несанкційних переміщень об'єктів у промислових і оборонних застосуваннях.

Незважаючи на значний прогрес апаратних платформ, зокрема 32-розрядних мікроконтролерів класу ARM Cortex-M і низькоспоживчих радіоінтерфейсів LoRa, NB-IoT, Zigbee, BLE, програмне забезпечення для таких систем досі розробляється переважно за допомогою *ad-hoc*-методів. Архітектурні рішення часто виносяться

в одну монолітну прошивку без декомпозиції на рівні відповідальностей, що ускладнює верифікацію поведінки в реальному часі, перешкоджає повторному використанню коду між апаратними платформами й також обмежує можливості незалежного тестування окремих захисних механізмів.

Дослідження останніх років розрізняють дві ортогональні стратегії боротьби з мережевими обмеженнями. Перша з них спрямована на скорочення завантаження каналу через часодомений адаптивний поріг і селективну передачу, проте робить це ціною прорідження даних і, як наслідок, ризику пропуску транзійтних (короточасних) подій. Друга стратегія забезпечує швидке прийняття рішень безпосередньо на вузлі через крайовий спектральний аналіз із фіксованим порогом, але втрачає стійкість у нестаціонарному шумовому середовищі. Поєднання переваг обох підходів у межах єдиної архітектурної моделі програмного забезпечення потребує методологічного оформлення на рівні шаблонів проєктування.



Об'єктом дослідження є процеси оброблення та передачі інформації в розподілених сенсорних мережах Інтернету речей із крайовою архітектурою, а предметом – моделі, методи й шаблони проектування програмного забезпечення крайових вузлів, що сприяють цілісності детектування за мінливих умов зовнішнього шуму.

## 2. Аналіз літературних джерел і визначення проблеми

Сучасні дослідження у сфері крайових обчислень для Інтернету речей систематизовано в роботах [1, 2], де детально розглянуто методи скорочення обсягу даних на сенсорному рівні. Серед них переважають часодоменні підходи, що використовують прогнозування або адаптивне відсікання повторюваних значень. Так, у дослідженні [3] продемонстровано підхід автономної адаптивної порогової редукції трафіку IoT-вузла, проте притаманне прорідження даних створює ризик пропуску транзйентних подій. У роботі [4] поєднано розподілене прогнозування з компресією та встановлено скорочення передачі до 96%. Автори статті [5] розширюють цю ідею на туманний (*fog*) рівень, виконуючи дворівневу агрегацію в періодичних сенсорних мережах. Окрему лінію досліджень формують роботи [6, 7], присвячені огляду методів стиснення із втратами й застосуванню глибокого навчання з підкріпленням до стиснення IoT-даних. Спільним недоліком усього цього класу методів є оптимізація для ефективності каналу зв'язку, а не цілісність детектування [8, 9].

Альтернативний клас методів становить спектральне виявлення подій безпосередньо на крайовому вузлі. У роботі [10] впроваджено STFT-аналіз для контролю якості електроенергії в IoT-системах моніторингу промислових моторів; у дослідженні [11] використано двійкові згорткові мережі для кодування часових рядів у спектральні ознаки на крайовому вузлі. У публікації [12] запропоновано систематичний огляд методів виявлення аномалій на основі штучного інтелекту в IoT, що підтверджує застосовність FFT-оброблення на ARM-сумісних мікроконтролерах. Ефективність такого оброблення зумовлена класичним результатом [13] (алгоритм Кулі – Тьюкі), який зменшує асимптотичну складність із  $O(L^2)$  до  $O(L \times \log(L))$ , що виявилось передумовою практичної реалізації спектрального оброблення на ARM-платформах [14].

В окремій групі робіт подано методи на базі TinyML. У дослідженні [15] описано неконтрольований автоенкодерний підхід до виявлення аномалій міського шуму на ресурсно-обмежених вузлах; у роботі [16] продемонстровано застосування TinyML для виявлення аномальних вібрацій у моторах. Проте таким методам властива принципова вада фіксованої референсної моделі: за повільного дрейфу шумового середовища необхідне перенавчання, що практично неможливе на мікроконтролерах класу Cortex-M0/M0+ без хмарної підтримки.

Адаптивні порогові методи з радіолокаційного кореня (CFAR), детально розглянуті в класичних роботах [17, 18], застосовують просторове усереднення сусідніх дистанційних комірок. У контексті вузлів Інтернету речей припущення про просторову однорідність порушується, оскільки вузли розташовані несинхронно, з різними локальними завадами, що робить пряме перенесення CFAR-схеми проблематичним. Натомість фундаментом стійкого непараметричного згладжування є медіанна фільтрація, запропонована в монографії [19] і формально досліджена в праці [20]. Її 50-відсоткова точка пробою забезпечує властивість, за якої до половини відліків у вікні можуть бути довільно спотворені без впливу на вихід.

Окремим напрямом досліджень є архітектурні моделі для крайових систем. У дослідженні [21] обґрунтовано доцільність попереднього локального оброблення перед хмарною передачею. У роботі [22] виокремлено інтелектуальне керування крайовими *mesh*-системами як ключову нерозв'язану проблему автономних сенсорних мереж. Огляд праць [8, 9] демонструє, що переважна більшість архітектурних робіт оперує абстракцією рівня системи (*cloud-fog-edge*), не опускаючись до рівня шаблону проектування, що описує конкретну композицію функціональних компонентів усередині прошивки крайового вузла. У студіях [23, 24] розглянуто колаборативні *cloud-edge*-схеми передбачення, проте без виокремлення повторно використовуваних шаблонів проектування програмного забезпечення. У своїй попередній роботі [25] автор аналізує енергоефективне застосування LPWAN-технологій у мережах резервного зв'язку для умов обмеженої інфраструктури, що корелює з тематикою *mesh*-надбудов поверх SX1276, обговорених у підрозд. 4.5 цієї публікації.

Аналіз літератури дає змогу виокремити прогалину в методології розроблення програмного

забезпечення крайових вузлів, що містить три взаємопов'язаних компоненти. По-перше, відсутній архітектурний шаблон, який ортогонально поєднував би адаптивний поріг для стійкості до нестационарного шуму, спектральну фільтрацію для відсікання вузькосмугових завад і темпоральну персистентність для відсікання миттєвих імпульсів. По-друге, відсутній метод проектування, що уможливував би верифікування кожного компонента захисту незалежно. По-третє, відсутня кількісна модель впливу хибних спрацювань крайового вузла на сукупний трафік багатовузлової транзитної *mesh*-мережі.

### 3. Мета й завдання дослідження

Мета дослідження – розробити архітектурну модель і метод проектування програмного забезпечення крайових вузлів *mesh*-мереж Інтернету речей, що дадуть змогу поєднувати спектральне виявлення подій з адаптивним відстеженням рівня шуму й локальним прийняттям рішень без залучення хмарного оброблення.

Для досягнення окресленої мети необхідно виконати такі завдання:

- формалізувати трирівневу декомпозицію програмного забезпечення вузла за рівнями відповідальності з явним розшаруванням компонентів спектрального оцінювання, адаптивного відстеження шуму й прийняття рішення;

- запропонувати шаблон триступеневого придушення хибних спрацювань як ортогональну композицію функціональних захистів, де кожен ступінь нейтралізує окремий клас механізмів спрацювання;

- проаналізувати обчислювальну складність і обсяг оперативної пам'яті для двох альтернативних реалізацій шаблону (на основі усереднення та медіанної фільтрації) для платформ ARM Cortex-M0/M0+;

- отримати аналітичну модель зростання трафіку в багатовузловій транзитній *mesh*-топології, що пов'яже локальну частоту хибних спрацювань вузла з обсягом сукупного навантаження мережі;

$$Q = \alpha \times FNR + \beta \times FPR \times A_{traffic}, \text{ якщо } C_{proc} \leq C_{max}, M_{RAM} \leq M_{max}, \quad (3)$$

де  $FNR$  – частка пропущених подій;  $FPR$  – частка хибних спрацювань;  $A_{traffic}$  – коефіцієнт зростання трафіку в мережі (визначено далі в розд. 5);  $C_{proc}$  – обчислювальне навантаження на мікроконтролер у тактах на кадр;  $M_{RAM}$  – обсяг оперативної пам'яті;

– продемонструвати на прикладі синтезу параметрів реалізації шаблону для типового сценарію розгортання сенсорної мережі та оцінити ефективність запропонованого підходу.

### 4. Матеріали й методи дослідження

Дослідження ґрунтується на методах архітектурного моделювання програмних систем, формальному аналізі обчислювальної складності, теорії цифрового оброблення сигналів і аналітичному моделюванні мережевого трафіку.

Розглянемо мережу з  $N_{nodes}$  сенсорних вузлів, об'єднаних у багатовузлову транзитну *mesh*-топологію із типовою середньою кількістю стрибків між вузлом-джерелом і шлюзом. Кожен вузол неперервно дискредитує сигнал свого первинного перетворювача (радар-датчик ближньої дії, акселерометр, мікрофон тощо) із частотою дискретизації  $f_s$  та обробляє його непересіченими кадрами завдовжки  $L$  відліків. Сигнал, що надходить на вхід обробника, описується співвідношенням

$$x[n] = s[n - \tau] + w_{th}[n] + w_{EMI}[n] + w_{dig}[n], \quad (1)$$

де  $s[n]$  – корисний сигнал події (спадний синусоїдальний імпульс у смузі очікуваного сигналу);  $w_{th}[n]$  – теплові гауссівські завади;  $w_{EMI}[n]$  – структурована електромагнітна завада від мережі живлення (50/60 Гц і гармоніки);  $w_{dig}[n]$  – імпульсні цифрові завади у високочастотній смузі. Потужність шуму  $P$  є нестационарною функцією часу, що дрейфує в межах  $\pm A$  дБ із періодом  $T_{cycle}$ , відтворюючи сукупний вплив температури, погодних умов і експлуатаційних чинників:

$$P(t) = P_0 \times 10^{(A/10) \times \sin(2\pi \times t / T_{cycle})}. \quad (2)$$

Завдання проектування програмного забезпечення вузла полягає в синтезі такої архітектури обробника, для якої сукупна функція якості  $Q$  є мінімальною за заданих ресурсних обмежень:

$\alpha, \beta$  – вагові коефіцієнти, що позначають доменно-специфічну ціну помилок. Для застосувань периметральної безпеки добуток  $\beta \times A_{traffic}$  має той самий порядок або перевищує  $\alpha$ , оскільки кожне

хибне спрацювання поширюється в мережі та споживає енергію всіх ретрансляторів.

#### 4.1. Трирівнева архітектурна модель програмного забезпечення крайового вузла

Запропонована модель програмного забезпечення крайового вузла побудована на принципі ортогонального розшарування за рівнями відповідальності. На відміну від типового монолітного підходу, у якому весь алгоритм оброблення реалізований єдиним блоком прошивки, модель явно виокремлює три рівні: спектральне оцінювання, адаптивне відстеження шуму й прийняття рішення.

На першому рівні – спектральне оцінювання – вхідний кадр із часової області перетворюється в частотну й селекцію інформативних спектральних бінів. Дискретне перетворення Фур'є визначається класичним співвідношенням

$$X[k] = \sum_{n=0}^{L-1} x[n] e^{-j2\pi kn/L}, \quad k = 0, 1, \dots, L-1. \quad (4)$$

Для обчислювально ефективної реалізації на 32-розрядному мікроконтролері застосовується алгоритм Кулі – Тьюкі радіус-2 із асимптотикою  $O(\log(L))$ . Замість повного спектра зберігається лише вектор бінів інтересу, що відповідає очікуваній смузі сигналу події:

$$f_i = [X[k_1], |X[k_2]|, \dots, |X[k_M]|]^T. \quad (5)$$

Така редукція з  $L/2$  унікальних бінів до  $M$  бінів інтересу одночасно зменшує обсяг пам'яті, потрібний для подальшого оброблення, фокусує чутливість детектора на смузі очікуваного сигналу та явно відсікає енергію всіх позасмугових шумових механізмів. У межах запропонованої архітектурної моделі цей етап реалізує перший ступінь захисту – селективну спектральну фільтрацію.

На другому рівні – адаптивне відстеження шуму – підтримується оцінка поточного фонового спектрального рівня, що еволюціонує в часі разом зі зміною зовнішніх умов. На відміну від просторових CFAR-схем, які усереднюють по сусідніх просторових комірках, запропонована модель здійснює усереднення темпоральне, у межах кожного біна окремо, оскільки припущення про просторову однорідність шуму між сенсорними вузлами в типовому розгортанні не виконується.

Архітектура цього рівня є каскадом двох фільтрів. Перший каскад має коротке вікно  $\gamma_d \in [3, 5]$  кадрів і

відсікає миттєві імпульсні викиди тривалістю меншою за  $\gamma_d$  кадрів – це другий ступінь захисту, темпоральна персистентна фільтрація. Другий каскад має довге вікно  $\gamma_a \in [64, 128]$  кадрів і відстежує повільний дрейф шумового рівня на масштабах десятків секунд або хвилин – це третій ступінь захисту, адаптивне відстеження рівня шуму. Каскадна структура є принциповою: вхід третього каскаду вже очищений другим каскадом від транзйентів, тому медіана або експоненціальна оцінка довгого каскаду не контамінується миттєвими піками, що виникають під час реальних подій.

На третьому рівні – прийняття рішення – формується бінарний індикатор події на основі мультиплікативного порогового правила, нормованого на поточну оцінку рівня шуму:

$$D[t] = \begin{cases} 1, & \text{якщо } |X[k][t]| > \zeta \times N[t] \\ 0, & \text{інакше,} \end{cases} \quad (6)$$

де  $\zeta \geq 1$  – мультиплікативний коефіцієнт детекторного запасу, конкретне значення якого обирається на етапі калібрування системи відповідно до співвідношення сигнал / шум очікуваної події та доменно-специфічних вимог до робочої точки на ROC-кривій. Для застосувань із низьким очікуваним SNR (порядку 3–6 дБ) типові значення  $\zeta \in [1, 2]$ , тоді як для задач периметральної безпеки з вищими SNR (15+ дБ) і критичною вимогою низької частоти хибних спрацювань калібрування дає значення  $\zeta$  у межах 5–8 (детальніше – у підрозд. 4.3). Мультиплікативна форма забезпечує сталу чутливість у дБ-шкалі, оскільки добуток  $\zeta \times N$  масштабується разом із фоновим рівнем шуму, тоді як адитивна схема потребувала б переналаштування в разі зміни рівня  $P_0$ . Для системно-рівневої події застосовується OR-агрегація по бінах:

$$E[t] = 1, \quad \text{якщо } \exists k \in K : D_k[t] = 1. \quad (7)$$

Принциповою властивістю запропонованої моделі є ортогональність трьох ступенів захисту покриття простору механізмів хибних спрацювань. Перший нейтралізує позасмугові завади (вузькосмугові 50/60 Гц і гармоніки, цифрові імпульси в смузі 0,8–2 кГц, широкосмуговий тепловий шум поза смугою події); другий ступінь нейтралізує внутрішньосмугові імпульсні транзйенти меншої тривалості за  $\gamma_d$  кадрів; третій нейтралізує повільний дрейф фонового рівня на масштабах десятків секунд і вище.

## 4.2. Метод проектування

### на основі шаблону триступеневого придушення

Запропонований метод формалізує розроблення програмного забезпечення крайового вузла як послідовність шести етапів, кожен з яких інстанціює відповідний компонент архітектурної моделі.

*Етап 1.* Доменна специфікація. Розробник задає очікувану смугу сигналу  $[f_{low}, f_{high}]$ , типову тривалість події  $T_{event}$ , очікуваний динамічний діапазон фоновому шуму  $\Delta P$  та цільові показники якості  $(FNR_{max}, FPR_{max})$ . Перелічені параметри визначаються природою прикладної задачі й слугують вхідними обмеженнями для решти етапів.

*Етап 2.* Синтез селективної спектральної фільтрації. На основі  $f_{low}, f_{high}$  і доступної частоти дискретизації  $f_s$  обираються розмір кадру  $L$ , набір бінів інтересу  $K = \{k_1, \dots, k_M\}$  і скалярна редукція спектра. Накладаються три обмеження. Перше полягає в обмеженні довжини кадру:  $L = 2^p$  (вимога алгоритму FFT radix-2). Друге обмеження – це умова дискретизації Найквіста, що пов'язує частоту дискретизації з найвищою частотою корисного сигналу:  $f_s \geq 2 \times f_{high}$ . Третє обмеження – умова спектральної роздільної здатності: частотна роздільна здатність  $\Delta f = f_s / L$  має бути достатньо малою для подання смуги  $[f_{low}, f_{high}]$  як множини щонайменше  $M_{min}$  бінів, тобто  $\Delta f \leq (f_{high} - f_{low}) / M_{min}$ . Значення  $M_{min}$  обирається доменно: типово  $M_{min} \geq 4-8$ , що забезпечує енергетичну стійкість оцінки смугової магнітуди до неточностей частотної прив'язки сигналу подій.

*Етап 3.* Синтез темпоральної персистентної фільтрації. Розмір вікна  $\gamma_d$  обирається з умови  $\gamma_d \times T_b \leq T_{event}$ , де  $T_b = L / f_s$  – тривалість кадру. Це гарантує, що персистентний фільтр пропустить корисну подію (вона тривала понад  $\gamma_d$  кадрів), але відсіче поодинокі імпульсні шумові викиди.

*Етап 4.* Синтез адаптивного відстежування рівня шуму. Розмір вікна  $\gamma_a$  (для медіанного варіанта) або коефіцієнт згладжування  $\alpha = 1 - 1/\gamma_a$  (для експоненціального) обирається з умови  $\gamma_a \times T_b \gg T_{event}$  і  $\gamma_a \times T_b \ll T_{drift}$ , де  $T_{drift}$  – характерний час зміни фоновому рівня.

*Етап 5.* Параметризація рішення. Коефіцієнт  $\zeta$  обирається на базі вимірюваного у фоновому режимі відношення  $R = X/\theta$  і необхідного запасу  $SNR_{min} = 20 \log(\zeta)$ .

*Етап 6.* Інтегральна верифікація. Кожен ступінь захисту верифікується незалежно за оцінкою його доменно-специфічної властивості: перший – за коефіцієнтом придушення позасмугової енергії; другий – відповідно до здатності відсікання одиночних кадрів-викидів; третій ступінь – за здатністю відстеження повільного дрейфу.

## 4.3. Приклад синтезу параметрів для типового сценарію

Розглянемо приклад синтезу параметрів реалізації шаблону для типового сценарію периметрального моніторингу з очікуваною смугою сигналу подій  $[1, 5]$  Гц, типовою тривалістю події  $T_{event} = 5$  с і характерним часом зміни фоновому рівня  $T_{drift} \approx 1$  год.

*Етап 2:* за  $f_s = 100$  Гц обираємо  $L = 128$  (radix-2).

Перевіряємо обмеження.

а) Найквіст:

$f_s = 100$  Гц  $\geq 2 \times f_{high} = 10$  Гц виконується із 10-кратним запасом.

б) Спектральна роздільна здатність:

$\Delta f = f_s / L = 0.78$  Гц, що дає  $M = 6$  бінів інтересу  $K = \{1, 2, 3, 4, 5, 6\}$ , які покривають смугу  $[0.78; 4.69]$  Гц і відповідають умові  $\Delta f \leq (f_{high} - f_{low}) / M_{min}$  за  $M_{min} = 6$ . Тривалість кадру  $T_b = L / f_s = 1.28$  с.

*Етап 3:*  $\gamma_d = 3 (\gamma_d \times T_b = 3.84 \text{ с} \leq T_{event} = 5 \text{ с})$ .

*Етап 4:*

$\gamma_a = 64 (\gamma_a \times T_b = 81.9 \text{ с} \gg T_{event} = 5 \text{ с} \text{ і } \gamma_a \times T_b \ll T_{drift} = 3600 \text{ с})$ .

Для експоненціального варіанта  $\alpha = 1 - 1/64 \approx 0.984$ .

*Етап 5:* для типового сценарію периметрального моніторингу обираємо  $\zeta = 6$ , що відповідає мінімальному виявленому  $SNR_{min} = 20 \log(6) \approx 15.6$  дБ. Це значення обґрунтовано результатами апаратного калібрування у фоновому режимі: відношення  $R = \bar{X}/\theta$  під час відсутності події не перевищує одиниці й має середнє значення приблизно 0.17, що залишає достатній запас для надійного спрацювання для подій зі  $SNR \geq 18$  дБ. Гейтова умова



$R_{gate} = 0.8$  запобігає проникненню енергії події в оцінку шуму.

Розрахунок ресурсів для платформи STM32G071 (ARM Cortex-M0/M0+, 64 МГц): сумарний обсяг оперативної пам'яті для медіанного варіанта становить приблизно 2,7 КБ (буфери кадрів, стан медіанних фільтрів, коефіцієнти  $\zeta_k$ ), що вдвічі менший за мінімальний обсяг ОЗП такого мікроконтролера. Час оброблення одного кадру не перевищує 21,5 мс і становить менше ніж 2% періоду кадру за  $f_s = 100$  Гц. Цей приклад підтверджує реалістичність синтезованих параметрів і ефективність запропонованого методу для практичних завдань.

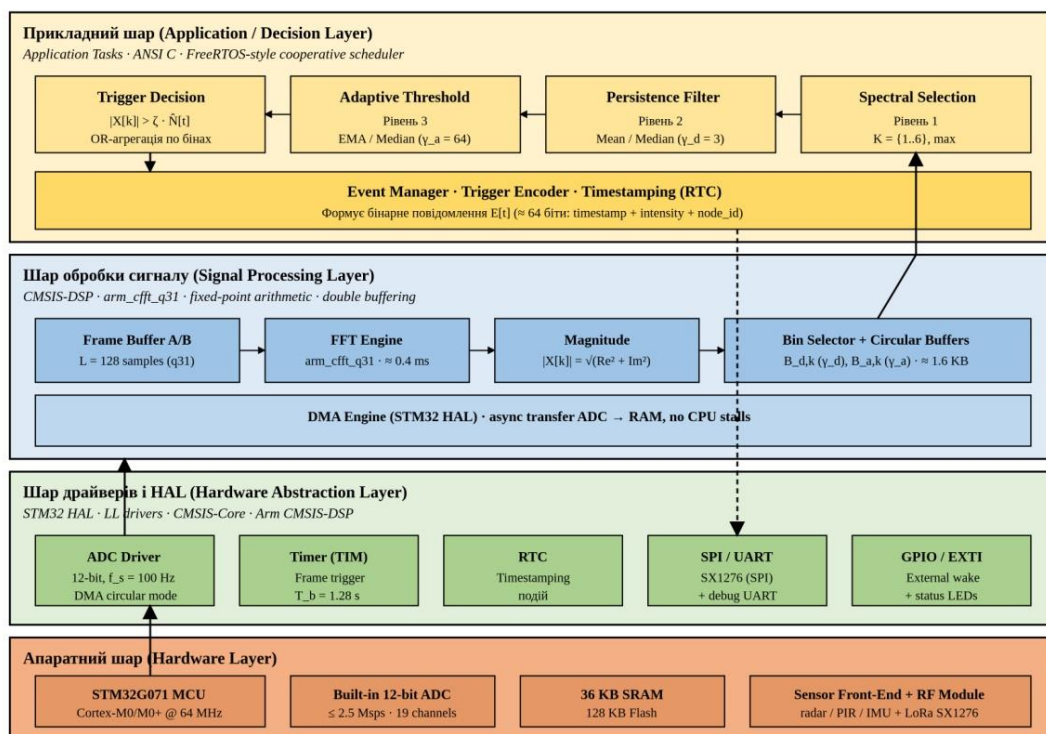
#### 4.4. Декомпозиція програмного забезпечення на функціональні шари

У підрозд. 4.1–4.3 запропоновано трирівневу архітектурну модель, що описує функціональну логіку детектування як ортогональну композицію

трьох взаємодоповнювальних захисних механізмів (рівень 1 – селективна спектральна фільтрація; рівень 2 – темпоральна персистентна фільтрація; рівень 3 – адаптивне відстеження рівня шуму). Натомість для практичної реалізації цієї логіки в прошивці крайового вузла необхідна окрема, програмно-інженерна декомпозиція, яка структурує код за принципами абстракції апаратного забезпечення, повторного використання й незалежного тестування. Така декомпозиція подана у вигляді чотиришарової структури й зображена на рис. 1. Принципова різниця між двома моделями полягає в тому, що трирівнева модель є логічною (визначає, що саме обробляється), тоді як чотиришарова – інженерною (з'ясовує, де саме в коді реалізується це оброблення). Чотиришарова декомпозиція є типовим інженерним підходом до розроблення вбудованих систем і дає змогу розділити апаратно-залежні компоненти від платформи-незалежної прикладної логіки, що зі свого боку спрощує перенесення коду між апаратними платформами й незалежне модульне тестування.

Архітектура ПЗ крайового вузла IoT mesh-мережі

(STM32G071, ARM Cortex-M0/M0+, ANSI C, STM32CubeIDE)



Стрілки: суцільні — синхронний потік даних; пунктирні — асинхронні події; контроль блоку — шар декомпозиції.

Червоний — апаратний; зелений — драйвери/HAL; блакитний — обробка сигналу; жовтий — прикладна логіка.

Рівень 1, Рівень 2, Рівень 3 — три ступені придушення хибних спрацювань (див. розд. 4.1).

Рис. 1. Архітектура програмного забезпечення крайового вузла IoT mesh-мережі: декомпозиція на чотири шари (апаратний, драйверів, оброблення сигналу, прикладний) і розташування трьох ступенів захисту від хибних спрацювань



Найнижчий апаратний шар містить мікроконтролер STM32G071 із вбудованим 12-розрядним АЦП послідовного наближення (SAR), сенсорний фронт-енд і радіочастотний модуль трансивера LoRa SX1276, під'єднаний до мікроконтролера через інтерфейс SPI. Шар драйверів і апаратної абстракції інкапсулює доступ до периферії через STM32 HAL і Low-Layer (LL) драйвери, що уніфікують програмний інтерфейс ADC, DMA, таймерів, RTC, SPI (для трансивера SX1276), UART (для діагностичного інтерфейсу) та GPIO. Шар оброблення сигналу реалізує власне обчислювальний конвеєр: подвійну буферизацію кадрів  $L = 128$  відліків, швидке перетворення Фур'є з фіксованою точкою (CMSIS-DSP `arm_cfft_q31`), обчислення магнітуд бінів, селекцію бінів інтересу й підтримку циркулярних буферів для медіанних фільтрів. Прикладний шар реалізує триступеневий шаблон придушення хибних спрацювань (рівень 1, рівень 2, рівень 3), правило прийняття рішення й формування події з мінімальною метайнформацією для подальшої передачі.

Принциповим архітектурним рішенням є виконання DMA-передач між АЦП й оперативною пам'яттю в режимі циркулярного буфера без участі ядра процесора. Це дає змогу розділити етапи набору кадру й етапи обчислення FFT у часі: у процесі обчислення FFT для одного буфера його друга половина наповнюється новими відліками від АЦП. Така подвійна буферизація є необхідною умовою детермінованого виконання реального часу й за умови коректної реалізації забезпечує стабільну тривалість кадру  $T_b$  незалежно від часу обчислення.

#### 4.5. Управління потоком даних і відповідність рівням моделі OSI

Управління потоком даних на крайовому вузлі організовано як послідовність детермінованих кроків, що ініціюються апаратними перериваннями DMA. Узагальнена блок-схема потоку продемонстрована на рис. 2 в лівій частині. Після ініціалізації клокінгу, периферії та запуску АЦП у DMA-режимі ядро процесора переходить у стан очікування переривання (WFI), у якому забезпечує мінімальне енергоспоживання. Після надходження переривання DMA *half/full-complete* активне завдання оброблення виконує FFT на неактивному буфері, послідовно застосовує три ступені захисту й у разі позитивного рішення формує бінарне повідомлення для передачі

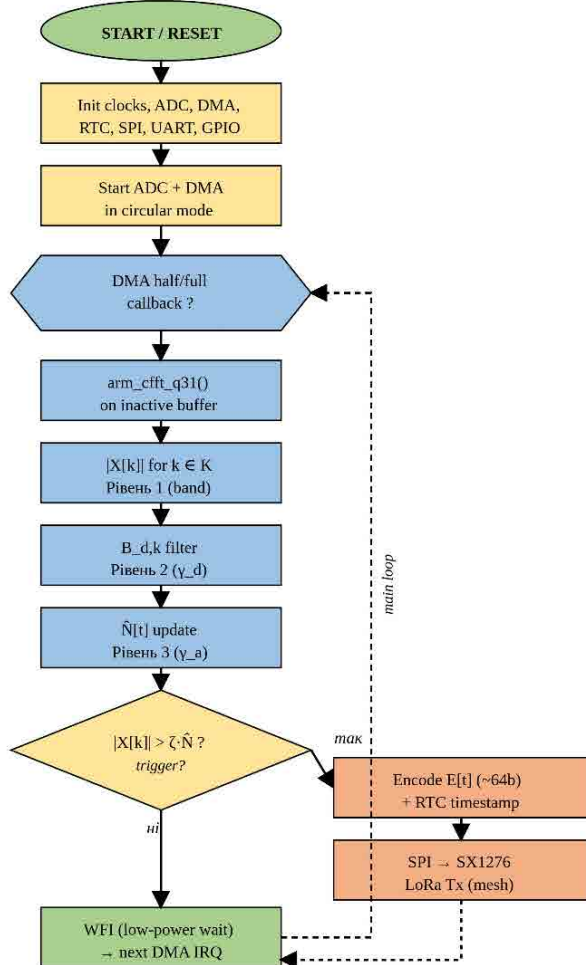
через радіотрансивер. За умови негативного рішення вузол повертається до сну, очікуючи на наступне переривання.

Запропонована архітектура природно мапиться на еталонну модель взаємодії відкритих систем (OSI), що дає змогу чітко відокремити функції різних рівнів і сприяє інтегровуваності з типовими IoT-стеками. Правий блок рис. 2 присвячений такому мапінгу. Важливо зазначити, що в цій статті під "LoRa mesh" розуміємо не стандартний LoRaWAN, який за специфікацією LoRa Alliance має *star-of-stars*-топологію, а кастомна *mesh*-надбудова побудована безпосередньо поверх фізичного рівня LoRa-трансивера SX1276 з власною реалізацією нижніх рівнів стеку. На прикладному рівні (L7) функціонує менеджер подій із кодуванням повідомлення у форматі TLV або CBOR; на рівні подання (L6) виконується серіалізація та опціональне шифрування AES-128. Сесійний рівень (L5) реалізує власну стейт-машину з підрахунком кадрів і керуванням *duty cycle* з огляду на регуляторні вимоги діапазону EU868. Транспортний (L4) і мережевий (L3) рівні реалізують, відповідно, надійну доставку й власну *mesh*-маршрутизацію (наприклад, на основі RPL-подібного протоколу) між сусідніми вузлами в багатовузловій транзитній топології. Канальний рівень (L2) реалізує власний LoRa MAC із механізмами CSMA-CA / Listen-Before-Talk, а не LoRaWAN MAC. Фізичний рівень (L1) подано модуляцією LoRa Chirp Spread Spectrum на трансивері SX1276, який під'єднано до мікроконтролера через інтерфейс SPI.

Класична модель OSI описує лише комунікаційну частину стеку, тому для повноти опису IoT-вузла на рис. 2 впроваджено додатковий рівень  $L0^*$  (*Sensing*), що охоплює сенсорний фронт-енд і вбудований АЦП мікроконтролера. Цей рівень не належить до канонічної моделі OSI, проте є фізичним фундаментом будь-якої сенсорної системи й має бути явно описаний у проєктній документації.

Важливою властивістю запропонованого мапінгу є те, що компоненти різних рівнів є слабкопов'язаними між собою через чітко визначені програмні інтерфейси (API). Це означає, що, наприклад, заміна нижніх рівнів (зокрема перехід з LoRa на NB-IoT) потребує модифікації лише компонентів  $L1-L2$ , тоді як вища логіка детектування залишається незмінною.

## а) Управління потоком даних на вузлі



## б) Відповідність рівням моделі OSI

|                    |                                                                                                                                                                                      |
|--------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| L7<br>Application  | Event Manager + Trigger Decision<br>JSON/CBOR-кодування події (timestamp, intensity, node_id), формат TLV для mesh-payload                                                           |
| L6<br>Presentation | Серіалізація + AES-128 (опц.)<br>Перетворення внутрішніх структур у байтовий потік; шифрування корисного навантаження за необхідності                                                |
| L5<br>Session      | Custom session / state machine<br>Frame counters, ACK/NACK handling, duty-cycle accounting, retransmission policy                                                                    |
| L4<br>Transport    | Reliability layer (опц. UDP/CoAP over LoRa)<br>Confirmed/unconfirmed delivery, fragmentation для повідомлень > MTU, FEC опційно                                                      |
| L3<br>Network      | Custom LoRa mesh routing (RPL-like)<br>Маршрутизація через ретранслятори до шлюзу; середнє $\bar{h}$ сканків — параметр у формулі (9)                                                |
| L2<br>Data Link    | Custom LoRa MAC + CSMA-CA / LBT<br>CRC, framing, MIC; collision avoidance — впливає на $p_{\text{getx}}$ у формулі (9)                                                               |
| L1<br>Physical     | LoRa CSS modulation, SX1276 transceiver<br>SF7-SF12, BW 125/250/500 kHz, EU868 band; Tx ~100 mA (домінує в енергобюджеті)                                                            |
| L0*<br>Sensing     | Sensor + STM32 built-in 12-bit ADC + DMA<br>Радар-датчик / акселерометр / мікрофон → ADC → кадр L = 128 відліків. Поза стандартом OSI, але є фізичною основою sensing-частини стека. |

## Зуваження про застосування моделі OSI:

Класична модель OSI описує стек комунікації, тоді як архітектура IoT-вузла потребує доповнення sensing-рівнем (L0\*) і прикладною логікою детектування. Запропонована модель ПЗ покриває L0\*–L7 у єдиній прошивці; роздільне виділення компонентів кожного рівня сприяє повторному використанню коду між апаратними платформами.

Лівий блок: блок-схема обробника від ADC-переривання до Tx-переривання. Правий блок: відповідність L0\*–L7 OSI.

Рис. 2. Управління потоком даних на крайовому вузлі (а) та відповідність функцій програмного забезпечення рівням еталонної моделі OSI (б)

## 5. Результати дослідження

5.1. Порівняльний аналіз  
двох реалізацій шаблону

Запропонована модель допускає дві альтернативні реалізації шару адаптивного відстеження: на основі

усереднення (далі – варіант А) й на основі медіанної фільтрації (далі – варіант Б). Обидва варіанти відповідають апаратним обмеженням мікроконтролерів класу ARM Cortex-M0/M0+, але різняться за вимогами до обчислювальних ресурсів. Зведене порівняння подано в табл. 1.

Таблиця 1. Порівняння двох реалізацій шаблону триступеневого придушення хибних спрацювань

| Характеристика                             | Варіант А (усереднення)                 | Варіант Б (медіана)                      |
|--------------------------------------------|-----------------------------------------|------------------------------------------|
| Агрегація за бінами                        | Максимум за 6 бінами → скаляр           | Побінне зберігання, логічне АБО (OR)     |
| Фільтр короточасних завад                  | Середнє за $\gamma_d = 3$ кадрами       | Медіана за $\gamma_d = 3 - 5$ кадрами    |
| Відстеження фоновому рівня шуму            | ЕМА із гейтуванням ( $\alpha = 0,984$ ) | Медіана за $\gamma_a = 64 - 128$ кадрами |
| Потреба у гейтуванні оновлень              | Так ( $R_{gate} = 0,8$ )                | Ні (50% точка пробую)                    |
| Збереження спектральної форми              | Ні (редукція за максимумом)             | Так (побінне рішення)                    |
| Обсяг ОЗП, байт                            | ≈ 24                                    | ≈ 1608                                   |
| Обчислювальне навантаження                 | ≈ 10 операцій / кадр                    | ≈ 9 600 порівнянь / с                    |
| Стійкість (робастність) до сильних викидів | Помірна                                 | Висока (до $\gamma_{a/2}$ викидів)       |

## 5.2. Аналіз обчислювальної складності та ресурсних обмежень

Сукупне обчислювальне навантаження на крайовий вузол містить три складники: час накопичення кадру (що виконується паралельно з обробленням завдяки подвійній буферизації), час обчислення ШПФ і час фільтраційно-вирішального етапу:

$$\tau_E = \tau_{acquire} + \tau_{FFT} + \tau_{filter} + \tau_{decision} \quad (8)$$

Для типових параметрів ( $L = 128$ ,  $M = 6$  бінів,  $\gamma_d = 3$ ,  $\gamma_a = 64$ ) і платформи ARM Cortex-M0/M0+ із тактовою частотою 64 МГц оцінено обчислювальну складність кожного складника. ШПФ із фіксованою комою (функція `arm_cfft_q31` бібліотеки CMSIS-DSP) обчислюється приблизно за 0,4 мс. Обчислення модуля спектра для 6 бінів триває менш ніж 0,1 мс. Варіант А виконується за лічені мікросекунди (10 операцій додавання та зсуву). Варіант Б із частковим сортуванням до медіанної позиції потребує приблизно  $|\gamma_d/2| \times \gamma_d$  порівнянь у короткому каскаді та  $|\gamma_a/2| \times \gamma_a$  порівнянь у довгому. Для  $\gamma_a = 64$  це становить приблизно  $64 \times 32 = 2048$  порівнянь на один бін або орієнтовно 9 600 порівнянь за секунду для 6 бінів за частоти кадрів 0,78 Гц, що охоплює менш ніж 0,02% процесорного часу.

Сумарний час оброблення одного кадру не перевищує 21,5 мс. За частоти дискретизації  $f_s = 100$  Гц тривалість кадру  $T_b = 1.28$  с, тому витрати процесорного часу становлять менше ніж 2%; за  $f_s = 1$  кГц вони дорівнюють приблизно 17%, що забезпечує достатній резерв часу для виконання інших операцій вузла. Для частот дискретизації понад 10 кГц необхідна платформа Cortex-M4 з апаратним блоком FPU.

Загальний обсяг ОЗП моделі (для варіанта Б) дорівнює сумі обсягів подвійних буферів кадру ( $2L$  відліків  $\times 4$  байти = 1 024 байти), станів медіанних фільтрів  $M \times (\gamma_d + \gamma_a) \times 4$  байти = 1608 байтів) і коефіцієнтів  $\zeta_k$  (24 байти), що загалом не перевищує 2,7 КБ. Це становить менш ніж половину обсягу ОЗП мікроконтролера STM32G071 у мінімальній конфігурації (8 КБ), що забезпечує достатній резерв пам'яті для стеку, мережевих буферів і додаткових прикладних модулів.

## 5.3. Аналітична модель зростання мережевого трафіку

У багатовузлових транзитних *mesh*-мережах хибні спрацювання локального детектора мають кумулятивний ефект на сукупний трафік мережі: кожна хибна подія потребує передачі через усі стрибки маршруту до шлюзу, до того ж в умовах перевантаження каналу можуть виникати ретрансляції внаслідок *collision backoff*. Для кількісного оцінювання додаємо коефіцієнт зростання трафіку:

$$A_{traffic} = \bar{h} \times R_{FA} \times S \times \rho_{retx}, \quad (9)$$

де  $\bar{h}$  – середня кількість стрибків маршруту;  $R_{FA}$  – частота хибних спрацювань вузла (подій/год);  $S$  – розмір корисного навантаження пакета (біт);  $\rho_{retx}$  – коефіцієнт ретрансляції ( $\geq 1$ , залежить від завантаженості каналу).

Запропонована архітектурна модель мінімізує  $A_{traffic}$  одразу за двома незалежними напрямками. По-перше, тричасткова композиція захистів зменшує  $R_{FA}$  на кілька порядків порівняно з однокомпонентними альтернативами. По-друге, локалізація прийняття рішення на крайовому вузлі редукує  $S$  до двійкового індикатора з мінімальними метаданими (мітка часу, оцінка інтенсивності) – приблизно 64 біти на подію, тоді як передача необробленого спектрального вектора з 16 бінів і 16-бітною роздільністю становить 256 бітів на кадр.

Розглянемо чисельний приклад: для мережі з  $h = 4$  стрибками, частотою кадрів 0,78 Гц і потенційним FPR на рівні 1% (наближення для системи, у якій відсутній шар адаптивного відстеження шуму) сукупний трафік хибних подій становив би  $R_{FA} = 0.01 \times 0.78 \times 3600 \approx 28$  хибних спрацювань за годину на вузол. У мережі з  $N_{nodes} = 200$  вузлів за 24 год це 134 400 хибних подій сумарно, що за  $S = 64$  біти й  $h = 4$  дає 34,4 МБіт корисного трафіку та принаймні стільки ж пропускну здатності, спожитих на ретрансляції. У межах прийнятої аналітичної моделі запропонована архітектура з трьома ортогональними механізмами захисту наближає  $R_{FA}$  до нуля. Кількісне підтвердження цього очікування для конкретних умов розгортання потребує окремої експериментальної валідації, що не є завданням цієї роботи.

## 6. Обговорення результатів

Запропоновані модель і метод подано на рівні шаблону проєктування програмного забезпечення, що допускає повторне використання в різних апаратних і доменних інстанціях. На відміну від наявних робіт, в яких зосереджено або на алгоритмічних результатах для конкретного апаратного стенда, або на загальних *cloud-fog-edge*-абстракціях, запропонований підхід явно ідентифікує три взаємодоповнювальні захисні механізми та встановлює, що відсутність будь-якого з них породжує клас хибних спрацювань, не покритий рештою.

Описані в статті архітектурна модель і метод проєктування слугують методологічною основою для розподілених систем визначення руху об'єктів, у яких крайовий вузол виявляє наявність динамічної цілі за спектральною сигнатурою її переміщення в локальному середовищі без передачі сирих сенсорних даних на хмарний рівень. Сфери застосування запропонованих результатів охоплюють такі промислові домени. Перший – системи периметральної безпеки об'єктів критичної інфраструктури (енергетичні підстанції, сховища, складські комплекси), де ціна хибних спрацювань і ціна пропусків мають співмірний порядок, а виявлення несанкційного переміщення об'єктів у захищеній зоні є основною функціональною вимогою. Другий – системи моніторингу промислового обладнання в режимі предиктивного обслуговування, де крайові вузли мають детектувати ранні ознаки відмов на тлі робочого шуму обладнання. Третій – оборонні застосування, що потребують автономної роботи без хмарної підтримки в умовах радіоелектронної протидії, зокрема виявлення руху живої сили й техніки за акустичною або сейсмічною сигнатурою. Четвертий – екологічний моніторинг (детектування браконьєрської діяльності в заповідниках, аномальних шумів у середовищі проживання охоронюваних видів), де енергетична автономність вузла є критичною.

Зв'язок із промисловістю реалізується завдяки можливості прямої інтеграції запропонованого шаблону в процес розроблення прошивки сенсорних вузлів на базі ARM-сумісних мікроконтролерів. Двофазна структура методу (синтез параметрів – інтегральна верифікація) узгоджується із промисловими процесами розроблення програмного забезпечення для систем реального часу за стандартами IEC 61508

та ISO 26262 у частині декомпозиції на верифіковані компоненти. Запропонована аналітична модель зростання трафіку дає змогу кількісно оцінювати мережеву й енергетичну вигоду локалізації прийняття рішення на крайовому вузлі ще на етапі архітектурного проєктування, що знижує ризик дорогих архітектурних перероблень на пізніх стадіях виробництва.

### Обмеження запропонованого підходу

По-перше, результати, подані в статті, ґрунтуються на аналітичній моделі, архітектурному обґрунтуванні та розрахунковому прикладі; повноцінна експериментальна валідація на стенді або польових даних виходить за межі цієї роботи та є предметом подальших досліджень. По-друге, модель припускає, що смугу сигналу події можна зафіксувати на етапі проєктування, що не завжди справедливо для застосувань зі змінними доменними характеристиками (мобільні джерела з доплерівським зсувом). По-третє, аналітична модель зростання трафіку не бере до уваги ефекти MAC-протоколу нижнього рівня (LBT, CSMA-CA), які можуть нелінійно модифікувати  $\rho_{retx}$  у режимах насичення каналу. По-четверте, формальна верифікація властивостей реального часу шаблону потребує окремого дослідження за допомогою методів модельної перевірки (TLA+, NuSMV).

## 7. Висновки

У процесі дослідження створено тривірневу архітектурну модель програмного забезпечення крайового вузла розподіленої сенсорної мережі Інтернету речей, що ортогонально композицією поєднує селективну спектральну фільтрацію, темпоральну персистентну фільтрацію й адаптивне відстежування рівня шуму. Запропоновано метод проєктування, який формалізує процес ініціалізації цієї моделі у вигляді послідовності шести етапів із чітко визначеними вхідними обмеженнями й критеріями верифікації кожного компонента.

Аналітична оцінка обчислювальної складності свідчить, що обидва варіанти реалізації шаблону відповідають обмеженням мікроконтролерів класу ARM Cortex-M0/M0+ із тактовою частотою 64 МГц у межах розрахункових 21,5 мс на кадр і 2,7 КБ оперативної пам'яті. Варіант на базі усереднення оптимальний за обсягом пам'яті (приблизно 24 байти



стану), варіант на базі медіанної фільтрації – за стійкістю до сильних викидів і збереженням спектральної форми (1,6 КБ стану). Запропонована аналітична модель зростання трафіку дає змогу кількісно оцінити мережеву й енергетичну вигоду локалізації прийняття рішення на крайовому вузлі ще на етапі архітектурного проектування.

Наукова новизна досягнутих результатів полягає в трьох позиціях. По-перше, запропоновано трирівневу архітектурну модель програмного забезпечення крайового вузла, що відрізняється явним розшаруванням компонентів виявлення, фільтрації та прийняття рішення. По-друге, створено шаблон триступеневого придушення хибних спрацювань як самостійну одиницю проектування програмного забезпечення для вузлів Інтернету речей, що допускає незалежну верифікацію кожного ступеня. По-третє, отримано аналітичну модель зростання трафіку в багатовузловій транзитній *mesh*-топології, що дає змогу кількісно зіставляти альтернативні архітектурні рішення на ранніх етапах розроблення.

Практичне значення отриманих результатів полягає в можливості системного створення програмного забезпечення для розподілених систем визначення руху об'єктів і периметральної безпеки, де ціна хибних спрацювань і ціна пропусків мають співмірний порядок. Запропоновані модель і метод можуть бути безпосередньо інтегровані в розроблення прошивки вузла Інтернету речей на базі ARM-сумісних мікроконтролерів і використані як методологічна основа подальшої верифікації та валідації за стандартами якості програмного забезпечення для систем реального часу.

Подальші дослідження автора будуть присвячені таким питанням: розширенню моделі для мобільних

вузлів-джерел (адаптивне розширення смуги бінів інтересу за оцінкою доплерівського зсуву); верифікації шаблону методами модельної перевірки на коректність часових властивостей; дослідженню *mesh*-кооперативних варіантів, у яких сусідні вузли обмінюються оцінками поточного рівня шуму для крос-валідації локальних детекторів.

### Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

### Фінансування

Дослідження проводилося без фінансової підтримки.

### Доступність даних

Рукопис не має пов'язаних даних.

### Використання засобів штучного інтелекту

Автори підтверджують, що не використовували технології ШІ для написання змістових частин цієї роботи. Допоміжні інструменти штучного інтелекту застосовувалися виключно для технічних завдань форматування тексту й перевірки граматики, що було повторно перевірено авторами вручну й не вплинуло на наукові висновки дослідження.

### References

1. Pioli, L., Dorneles, C. F., de Macedo, D. D. J., Dantas, M. A. R. (2022), "An overview of data reduction solutions at the edge of IoT systems: a systematic mapping of the literature", *Computing*, Vol. 104, No. 8, pp. 1867–1889. DOI: <https://doi.org/10.1007/s00607-022-01073-6>
2. Sadri, A. A., Rahmani, A. M., Saberikamarposhti, M., Hosseinzadeh, M. (2022), "Data reduction in fog computing and Internet of Things: a systematic literature survey", *Internet of Things*, Vol. 20, art. 100629. DOI: <https://doi.org/10.1016/j.iot.2022.100629>
3. Zhang, H., Na, J., Zhang, B. (2023), "Autonomous Internet of Things (IoT) Data Reduction Based on Adaptive Threshold", *Sensors*, Vol. 23, No. 23, art. 9427. DOI: <https://doi.org/10.3390/s23239427>
4. Hussein, A. M., Idrees, A. K., Couturier, R. (2022), "Distributed energy-efficient data reduction approach based on prediction and compression to reduce data transmission in IoT networks", *International Journal of Communication Systems*, Vol. 35, No. 9, e5282. DOI: <https://doi.org/10.1002/dac.5282>



5. Idrees, A. K., Al-Qurabat, A. K. M. (2021), "Energy-Efficient Data Transmission and Aggregation Protocol in Periodic Sensor Networks Based Fog Computing", *Journal of Network and Systems Management*, Vol. 29, No. 1, art. 4. DOI: <https://doi.org/10.1007/s10922-020-09567-4>
6. Correa, J. D. A., Pinto, A. R., Montez, C. (2022), "Lossy Data Compression for IoT Sensors: A Review", *Internet of Things*, Vol. 19, art. 100516. DOI: <https://doi.org/10.1016/j.iot.2022.100516>
7. Chen, X., Yu, Q., Dai, S., Sun, P., Tang, H., Cheng, L. (2024), "Deep Reinforcement Learning for Efficient IoT Data Compression in Smart Railroad Management", *IEEE Internet of Things Journal*, Vol. 11, No. 14, pp. 25494–25504. DOI: <https://doi.org/10.1109/JIOT.2023.3348487>
8. Hafeez, T., Xu, L., Mcardle, G. (2021), "Edge Intelligence for Data Handling and Predictive Maintenance in IIoT", *IEEE Access*, Vol. 9, pp. 49355–49371. DOI: <https://doi.org/10.1109/ACCESS.2021.3069137>
9. GabAllah, N. A., Farrag, I., Nawawy, O., Khalil, R., Sharara, H., Elbatt, T. (2021), "Towards Enabling IoT Systems with Edge Intelligence", *Proc. 2021 IEEE International Conference on Smart Internet of Things (SmartIoT)*, pp. 271–277. DOI: <https://doi.org/10.1109/SmartIoT52359.2021.00050>
10. Bhoi, S. K., Faseehuddin, M., Ravulakollu, K. K., Suthar, B., Sahoo, K. S., Pradhan, B. (2022), "Advanced Edge Computing Framework for Grid Power Quality Monitoring of Industrial Motor Drive Applications", *Proc. 2022 International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM)*, IEEE, pp. 455–459. DOI: <https://doi.org/10.1109/speedam53979.2022.9841966>
11. Xie, C., Tao, W., Zeng, Z., Dong, Y. (2023), "Binary-Convolution Data-Reduction Network for Edge-Cloud IIoT Anomaly Detection", *Electronics*, Vol. 12, No. 15, art. 3229. DOI: <https://doi.org/10.3390/electronics12153229>
12. Trilles, S., Hammad, S. S., Iskandaryan, D. (2024), "Anomaly detection based on Artificial Intelligence of Things: A Systematic Literature Mapping", *Internet of Things*, Vol. 25, art. 101063. DOI: <https://doi.org/10.1016/j.iot.2024.101063>
13. Cooley, J. W., Tukey, J. W. (1965), "An algorithm for the machine calculation of complex Fourier series", *Mathematics of Computation*, Vol. 19, No. 90, pp. 297–301. DOI: <https://doi.org/10.1090/S0025-5718-1965-0178586-1>
14. Wickert, M. A. (2015), "Using the ARM Cortex-M4 and the CMSIS-DSP library for teaching real-time DSP", *Proc. 2015 IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE)*, IEEE, pp. 283–288. DOI: <https://doi.org/10.1109/DSP-SPE.2015.7369567>
15. Hammad, S. S., Iskandaryan, D., Trilles, S. (2023), "An unsupervised TinyML approach applied to the detection of urban noise anomalies under the smart cities environment", *Internet of Things*, Vol. 23, art. 100848. DOI: <https://doi.org/10.1016/j.iot.2023.100848>
16. Arciniegas, S., Rivero, D., Piñan, J., Diaz, E., Rivas, F. (2025), "IoT device for detecting abnormal vibrations in motors using TinyML", *Discover Internet of Things*, Vol. 5, art. 142. DOI: <https://doi.org/10.1007/s43926-025-00142-4>
17. Rohling, H. (1983), "Radar CFAR Thresholding in Clutter and Multiple Target Situations", *IEEE Transactions on Aerospace and Electronic Systems*, Vol. AES-19, No. 4, pp. 608–621. DOI: <https://doi.org/10.1109/TAES.1983.309350>
18. Finn, H. M., Johnson, R. S. (1968), "Adaptive detection mode with threshold control as a function of spatially sampled clutter level estimates", *RCA Review*, Vol. 29, No. 3, pp. 414–464.
19. Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 688 p.
20. Gallagher, N. C., Wise, G. L. (1981), "A theoretical analysis of the properties of median filters", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 29, No. 6, pp. 1136–1141. DOI: <https://doi.org/10.1109/TASSP.1981.1163708>
21. Sittón-Candanedo, I., Corchado, J. M. (2019), "An Edge Computing Tutorial", *Oriental Journal of Computer Science and Technology*, Vol. 12, No. 2, pp. 34–38. DOI: <https://doi.org/10.13005/OJCST12.02.02>
22. Oikonomou, P., Karanika, A., Anagnostopoulos, C., Kolomvatsos, K. (2021), "On the Use of Intelligent Models towards Meeting the Challenges of the Edge Mesh", *ACM Computing Surveys*, Vol. 54, No. 6, pp. 1–42. DOI: <https://doi.org/10.1145/3456630>
23. Putra, M. A. P., Hermawan, A. P., Kim, D.-S., Lee, J.-M. (2023), "Data Prediction-Based Energy-Efficient Architecture for Industrial IoT", *IEEE Sensors Journal*, Vol. 23, No. 14, pp. 15856–15866. DOI: <https://doi.org/10.1109/JSEN.2023.3280485>
24. Wu, Y., Dai, H.-N., Wang, H., Xiong, Z., Guo, S. (2023), "To Transmit or Predict: An Efficient Industrial Data Transmission Scheme With Deep Learning and Cloud-Edge Collaboration", *IEEE Transactions on Industrial Informatics*, Vol. 19, No. 11, pp. 11322–11332. DOI: <https://doi.org/10.1109/TII.2023.3245673>
25. Lykov, Yu. V., Oleynikov, A. M., Lykova, H. O., Savenko, S. O., Makovetskyi, S. O. (2023), "Organization of Backup Communication Channels for Emergencies in Environments with Limited Infrastructure Using LPWAN Technology", *Radioelectronics and Communications Systems*, Vol. 66, No. 6, pp. 287–304. DOI: <https://doi.org/10.3103/S0735272723090054>

Received (Надійшла) 01.04.2026

Accepted for publication (Прийнята до друку) 12.05.2026

Publication date (Дата публікації) 29.05.2026

*Відомості про авторів / About the Authors*

**Маковецький Сергій Олександрович** – Харківський національний університет радіоелектроніки, аспірант кафедри програмної інженерії, Харків, Україна;

**Serhii Makovetskyi** – Kharkiv National University of Radio Electronics, Postgraduate Student of the Department of Software Engineering, Kharkiv, Ukraine;

e-mail: serhii.makovetskyi@nure.ua

ORCID ID: <https://orcid.org/0000-0002-9814-9465>

**Каук Віктор Іванович** – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, доцент кафедри програмної інженерії, Харків, Україна;

**Viktor Kauk** – Candidate of Technical Sciences, Associate Professor, Kharkiv National University of Radio Electronics, Associate Professor of the Department of Software Engineering Kharkiv, Ukraine;

e-mail: viktor.kauk@nure.ua

ORCID ID: <https://orcid.org/0000-0002-2780-2666>

## ARCHITECTURAL MODEL AND DESIGN METHOD OF SOFTWARE FOR EDGE NODES OF IoT MESH NETWORKS WITH THREE-STAGE FALSE-TRIGGER SUPPRESSION

**The subject** of the study is models and methods for software development of edge nodes of distributed sensor networks of the Internet of Things with mesh topology operating in non-stationary noise environments. **The aim** is to develop an architectural model and a design method for edge-node software that combine spectral event detection with adaptive noise-floor tracking and local decision-making without cloud processing. **Tasks:** to formalise a three-layer decomposition of the node software; to propose a three-stage false-trigger suppression pattern as an orthogonal composition of functional defences; to perform analysis of computational complexity and memory footprint for two alternative pattern instantiations; to derive an analytical model of traffic amplification quantifying the network benefit of localised decision-making. **Methods:** architectural modelling of software systems, formal computational complexity analysis, digital signal processing theory (Fast Fourier Transform, median and exponential filtering), analytical modelling of network traffic in multi-hop topologies. **Results.** A three-layer architectural model is formulated that orthogonally combines selective spectral filtering, temporal persistence filtering and adaptive noise-floor tracking. A six-phase design method is developed with verification criteria for each component. It is proven that both pattern instantiations satisfy the constraints of ARM Cortex-M0/M0+-class microcontrollers operating at 64 MHz within 21.5 ms per frame and 2.7 KB of RAM. The proposed analytical model of traffic amplification allows quantitative comparison of alternative architectural solutions at early development stages. **Conclusions.** The proposed model and method are presented at the level of a design pattern that allows reuse in different hardware and domain instances. The pattern identifies three complementary defence mechanisms and establishes that the absence of any of them produces a class of false triggers not covered by the rest. The results may serve as a methodological foundation for software development of distributed object-motion detection and perimeter-security systems.

**Keywords:** distributed systems; mesh networks; software architecture; design pattern; Internet of Things; object motion detection; adaptive thresholding; digital signal processing; embedded systems; microcontrollers.

*Бібліографічні описи / Bibliographic descriptions*

Маковецький С. О., Каук В. І. Метод проектування програмного забезпечення крайових вузлів інтернет-речей mesh-мереж в умовах хибних спрацювань. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 211–223. DOI: <https://doi.org/10.30837/0135-1710.2026.189.211>

Makovetskyi, S., Kauk, V. (2026), "Architectural model and design method of software for edge nodes of IoT mesh networks with three-stage false-trigger suppression", *Management Information System and Devices*, No. 2 (189), P. 211–223. DOI: <https://doi.org/10.30837/0135-1710.2026.189.211>

Просолов В. В., Халімов Г. З.

## ФОРМУВАННЯ ПРОСТОРУ ОЗНАК ДЛЯ РИЗИК-СКОРИНГУ BITCOIN НА ОСНОВІ ЛОКАЛЬНИХ І ТОПОЛОГІЧНИХ МЕТРИК

**Предметом роботи** є метод формування простору ознак для ризик-скорингу в мережі Bitcoin, що ґрунтується на поєднанні локальних характеристик вузлів, топологічних метрик їхнього околу, нелінійної нормалізації та графової трансформації ончейн-даних. **Мета дослідження** – вдосконалення методу формування простору ознак для неструктурованих ончейн-даних мережі Bitcoin способом комплексної агрегації локальних і топологічних властивостей об'єктів транзакційного графа й застосування нелінійної нормалізації, що дає змогу підвищити інформативність даних для задач ризик-скорингу. **Завдання:** обґрунтувати доцільність графового подання ончейн-даних у задачах ризик-скорингу Bitcoin; сформувати багатовимірний простір ознак на основі локальних мікроознак і глобальних топологічних макроознак; застосувати процедури попереднього оброблення, очищення та нелінійної нормалізації фінансових величин; дослідити роль локальних, структурних, темпоральних і контекстних атрибутів у підвищенні якості відокремлення ризикових об'єктів від легітимних; оцінити ефективність сформованого простору ознак як основи для подальшого використання в алгоритмах ризик-скорингу. **Методи дослідження** ґрунтуються на формалізації ончейн-даних у вигляді графа, використанні методів інженерії ознак, графових трансформацій, нелінійної нормалізації, а також на аналізі локальних і топологічних метрик транзакційної мережі Bitcoin. **Результати дослідження.** Запропоновано комплексний підхід до побудови простору ознак для ризик-скорингу Bitcoin, який передбачає агрегацію 94 локальних мікроознак і 72 глобальних топологічних макроознак, застосування нелінійної нормалізації та графової трансформації даних. Продемонстровано, що такий підхід сприяє зменшенню розрідженості ознакового простору, підвищенню інформативності вхідних даних й кращій роздільній здатності в процесі відокремлення ризикових об'єктів від легітимних. Обґрунтовано, що якісно сформований простір ознак може бути базовим компонентом для подальшої інтеграції графових і часових нейромережових механізмів у задачах фінансової аналітики. **Висновки.** Запропонований метод формування простору ознак є практично придатним інструментом для систем моніторингу ризиків у мережі Bitcoin і підтверджує доцільність поєднання локальних і топологічних метрик з метою підвищення стійкості ризик-скорингу до обфускаційних сценаріїв і складних траєкторій руху коштів.

**Ключові слова:** Bitcoin; ончейн-дані; ризик-скоринг; інженерія ознак; топологічні метрики; графові трансформації; виявлення шахрайства.

### 1. Вступ

Мережа Bitcoin, попри прозорість і доступність ончейн-даних, залишається вразливою до шахрайських і нелегальних фінансових сценаріїв, а практична потреба у швидкому й точному ризик-скорингу адрес і облікових записів нашоується на фундаментальні обмеження самих даних. Ончейн-дані є неструктурованими, високорозмірними й розрідженими, транзакційні взаємодії мають графову природу, а поведінкові патерни проявляються у часовій динаміці потоків. Додатково ситуацію ускладнює суттєвий дисбаланс між легітимними й ризиковими об'єктами. Унаслідок цього традиційні табличні підходи або втрачають структурний контекст транзакційного графа, або демонструють нестійкість і низьку предикативну здатність під час обфускації фінансових траєкторій. Отже, науково-прикладна проблема зводиться до

необхідності розроблення й обґрунтування методу формування простору ознак для ризик-скорингу Bitcoin, який поєднує локальні характеристики вузлів із топологічними метриками їхнього околу й застосовує нелінійну нормалізацію та графові трансформації, щоб зменшити розрідженість ознакового простору й підвищити якість виявлення загроз у практичних сценаріях моніторингу.

### 2. Аналіз літературних джерел і постановка проблеми дослідження

Останні дослідження з аналітики безпеки Bitcoin демонструють поступовий перехід від евристичних правил і табличних моделей до графово-орієнтованих підходів, здатних брати до уваги не лише локальні атрибути транзакцій, а й структуру їхніх взаємозв'язків. Одним із базових джерел для розвитку цього напрямку став Elliptic Data Set, у межах

якого транзакції Bitcoin подано як часовий граф із понад 200 тис. вузлів, 234 тис. спрямованих платіжних зв'язків і 166 ознак вузлів [1]. Це дало змогу формалізувати задачу виявлення іліцитних транзакцій як задачу класифікації вузлів графа й створило основу для порівняння класичних і графових алгоритмів машинного навчання.

Водночас у дослідженнях, присвячених виявленню ransomware-активності та відмиванню коштів у мережі Bitcoin, продемонстровано, що ризикові сценарії часто проявляються не через окрему фінансову величину, а через стійкі топологічні патерни руху коштів. Зокрема в роботах щодо BitcoinHeist наголошено на використанні топологічного аналізу для виявлення ransomware-транзакцій навіть за обмеженої кількості відомих прикладів [2]. Подальші дослідження Elliptic++ і підходів до фінансової форензики в Bitcoin також підтверджують доцільність поєднання транзакційного, адресного й графового подання для глибшого аналізу шахрайських сценаріїв [3].

Класичні алгоритми машинного навчання, зокрема логістична регресія, Random Forest, SVM, XGBoost і багатошарові перцептрони, здатні працювати з агрегованими табличними ознаками, однак їхня ефективність обмежується втратою структурного контексту. Такі моделі розглядають транзакції переважно як незалежні об'єкти, тоді як у мережі Bitcoin вони пов'язані між собою через модель UTXO й формують орієнтований граф руху активів. Унаслідок цього класичні підходи можуть не виявляти багатоетапні схеми обфускації, зокрема reel-chain, fan-in консолідації та щільні локальні підграфи, властиві для міксування коштів.

Перспективним напрямом подолання зазначених обмежень є застосування графових нейронних мереж. Базова архітектура GCN забезпечила можливість напівкерованої класифікації на графових структурах унаслідок агрегації ознак сусідніх вузлів [4]. Подальший розвиток цього напрямку пов'язаний із GraphSAGE, який орієнтований на індуктивне навчання й дає змогу формувати подання нових вузлів способом вибіркової агрегації інформації з їхнього локального оточення [5]. Для задач аналізу блокчейн-транзакцій це важливо, оскільки граф Bitcoin постійно розширюється, а нові транзакції мають оцінюватися без повного перенавчання моделі.

Окреме значення в розвитку графових підходів мають Graph Attention Networks, у яких

використовується механізм уваги для адаптивного зважування внеску сусідніх вузлів [6]. Проте подальші дослідження продемонстрували, що класичний GAT має обмеження, пов'язані зі статичним характером уваги. У роботі Brody, Alon та Yahav запропоновано GATv2, який реалізує більш виразний механізм динамічної уваги та краще зважає на залежність ваг ребер від стану цільового вузла [7]. Це є особливо важливим для транзакційних графів Bitcoin, де окремі слабкі або рідкісні зв'язки можуть мати неабияке значення для виявлення ризикової активності.

Суттєвим обмеженням задач ризик-скорингу Bitcoin залишається дисбаланс класів, оскільки підтверджені іліцитні транзакції становлять значно меншу частку, якщо порівнювати з легітимними. Класичний метод SMOTE дає змогу створювати синтетичні приклади міноритарного класу в табличному просторі ознак [8], однак для графових даних такого підходу недостатньо, оскільки він не бере до уваги топологічні зв'язки між об'єктами. Для подолання цього обмеження запропоновано GraphSMOTE, який генерує синтетичні вузли в просторі графових вкладень і додатково моделює ребра для збереження структурної узгодженості графа [9].

Оскільки транзакційна активність Bitcoin має не лише структурну, а й часову природу, важливим напрямом є дослідження динамічних графових моделей. Зокрема EvolveGCN демонструє можливість взяття до уваги часової еволюції графа через зміну параметрів графової згорткової мережі в часі [10]. Для задач ризик-скорингу це має значення, оскільки схеми відмивання коштів і ransomware-активність часто розгортаються як послідовність транзакційних дій, а не як одинична ізольована подія.

Додатковою вимогою до практичного використання нейромережових моделей у фінансовому моніторингу є інтерпретованість результатів. У цьому контексті метод SHAP забезпечує формальне оцінювання внеску окремих ознак у результат моделі [11]. Для задач ризик-скорингу блокчейн-транзакцій це важливо, оскільки рішення про потенційну ризиковість об'єкта має бути не лише точним, а й пояснюваним для подальшого використання в аудиті, AML/CFT-процедурах і цифровій форензиці.

Отже, аналіз літературних джерел свідчить, що сучасні дослідження поступово переходять від ізольованого аналізу транзакцій до комплексного

графового подання ончейн-даних. Водночас залишаються відкритими питання якісного формування простору ознак, зменшення його розрідженості, урахування локальних і топологічних характеристик, а також підготовки даних до подальшого використання в графових і часових нейромережових моделях. Це обґрунтовує доцільність удосконалення методу формування простору ознак для ризик-скорингу Bitcoin способом поєднання локальних мікроознак, топологічних метрик околу, нелінійної нормалізації та графової трансформації.

### 3. Мета й завдання дослідження

Мета дослідження – удосконалення методу формування простору ознак для неструктурованих ончейн-даних мережі Bitcoin способом комплексної агрегації локальних мікроознак (атрибути індивідуальних вузлів) і глобальних топологічних макроознак (структурні метрики околів) із застосуванням нелінійної нормалізації та графової трансформації, а також дослідження запропонованого методу.

#### Завдання:

- обґрунтувати доцільність графового подання ончейн-даних у задачах ризик-скорингу Bitcoin;
- сформувати багатовимірний простір ознак на основі 94 локальних мікроознак та 72 глобальних топологічних макроознак;
- застосувати процедури попереднього оброблення, очищення й нелінійної нормалізації фінансових величин;
- оцінити ефективність сформованого простору ознак як основи для подальшого використання в алгоритмах ризик-скорингу.

### 4. Матеріали й методи дослідження

Отримання ончейн-даних із розподіленого реєстру й подальше їх оброблення засобами алгоритмічної трансформації є ключовою передумовою побудови якісного простору ознак для ризик-скорингу. У задачах фінансового шахрайства здатність методів виявляти складні, багатоетапні сценарії атак визначається тим, наскільки інформативно ознаки відбивають поведінку об'єктів і наскільки коректно збережено структурний контекст їхніх взаємодій. Табличне подання, попри простоту, часто "відрізає" топологічні залежності, властиві для неевклідових мережових даних, тому більш виправданим є графове

подання, сформоване за допомогою спеціалізованого ETL-процесу.

Побудова системи збору телеметрії передбачає під'єднання до повних вузлів Bitcoin або використання інфраструктурних провайдерів, які надають доступ до інформації через API. Збір даних реалізується за допомогою розбору бінарних структур блоків із вилученням вхідних і вихідних скриптів, що формують модель UTXO. Для наближення до режиму near-real-time застосовується інтерфейс ZeroMQ, який забезпечує оперативне отримання транзакцій із mempool і дає змогу швидко реагувати на появу нових подій у мережі [14, 15].

На початковому етапі формується впорядкована в часі колекція спостережень, що відповідає дискретним часовим крокам. З огляду на масштаб транзакційного графа Bitcoin доцільним є вибіркове фільтрування та обмеження часової глибини аналізу; у роботі використано підхід із 49-часовими кроками за аналогією з Elliptic Data Set [13]. Для кожної транзакції витягуються базові дескриптори, зокрема ідентифікатор (txid), сума переказу (Value), комісія (Fee) та часовий штамп (Timestamp).

Далі стабілізується числовий простір ознак. Фінансові величини в Bitcoin мають надзвичайно широкий діапазон: від мікроплатежів у кілька сатоші до великих переказів, що різняться на багато порядків. Використання "сирих" значень у вході графових нейромереж може погіршувати збіжність оптимізації та провокувати нестійкі градієнти, тому застосовується нормалізація, що поєднує логарифмічне масштабування з min-max-перетворенням у діапазон [0; 1]:

$$x'_{ij} = \frac{\log(1 + x_{ij}) - \min(\log(1 + x_j))}{\max(\log(1 + x_j)) - \min(\log(1 + x_j))}, \quad (1)$$

де  $x_{ij}$  – початкове значення  $j$ -ї ознаки для  $i$ -го вузла;  $x_j$  – вектор усіх значень цієї ознаки в поточному часовому вікні [16].

Ключовою новизною запропонованого підходу є етап збагачення ознакового опису, у межах якого формується розширений 166-вимірний вектор ознак. Його структура побудована так, щоб поєднувати як "внутрішні" властивості об'єкта, так і контекст його взаємодій у транзакційній мережі. Перший блок містить 94 локальні ознаки, які відбивають безпосередню поведінку транзакції або вузла, зокрема характеристики обсягів переказів, конфігурацію входів і виходів у моделі UTXO, статистики комісій



і часову інтенсивність активності. Другий блок охоплює 72 структурні ознаки, що узагальнюють інформацію про найближче топологічне оточення цільового вузла на рівні 1-hop-сусідів, зокрема агреговані вхідні й вихідні потоки від сусідніх вузлів, показники центральності та локальні метрики щільності й кластеризації підграфа.

Запропонований метод формування простору ознак реалізується у вигляді послідовності кількох етапів.

1. Збір ончейн-даних. На цьому етапі отримується транзакційна інформація з повних вузлів Bitcoin або через інфраструктурні API-провайдери з огляду на часову послідовність появи транзакцій.

2. Первинне оброблення й очищення даних. Виконується видалення базових параметрів транзакцій, фільтрація некоректних або неповних записів, а також формування впорядкованої в часі колекції спостережень.

3. Екстракція локальних характеристик вузлів. Для кожної транзакції або вузла формуються локальні ознаки, що описують фінансові параметри, конфігурацію входів і виходів UTXO, комісії та часові властивості активності.

4. Нелінійна нормалізація фінансових величин. Застосовується логарифмічне масштабування та min-max-перетворення для стабілізації розподілів ознак і зменшення впливу аномально великих значень.

5. Обчислення топологічних метрик. Для вузлів транзакційного графа розраховуються структурні характеристики найближчого оточення, зокрема ступені вузлів, агреговані характеристики сусідів, центральність і локальна щільність підграфа.

6. Формування інтегрованого простору ознак. Локальні й топологічні характеристики об'єднуються в єдиний багатовимірний вектор, придатний для подальшого використання в алгоритмах ризик-скорингу.

7. Графова трансформація даних. Підготовлені ознаки й транзакційні зв'язки подаються у вигляді орієнтованого атрибутуваного графа, що зберігає структурний контекст руху коштів у мережі Bitcoin.

8. Підготовка даних до моделювання ризику. Сформований граф і матриця ознак передаються до подальших алгоритмів класифікації або ризик-скорингу для виявлення підозрілих транзакційних об'єктів.

### Багатовимірний простір ознак:

#### топологічні, темпоральні та контекстні атрибути

Формування інформативного простору ознак для ризик-скорингу в мережі Bitcoin потребує переходу

від "сирих" блокчейн-подій і транзакційної телеметрії до розрахунку поведінкових і граф-теоретичних дескрипторів, здатних відтворювати як локальну активність, так і структурний контекст взаємодій. У межах дослідження сформовано 166-вимірний простір ознак (за логікою бенчмарку Elliptic Data Set [6]), який концептуально передбачає два базові компоненти: локальні атрибути об'єкта й структурні метрики його найближчого графового оточення.

Кожний вузол  $v \in \mathcal{V}$  описується 94 локальними характеристиками, що фіксують профіль транзакційної поведінки й фінансові параметри, зокрема обсяг переказу у BTC, величину комісії, а також конфігурацію входів і виходів у моделі UTXO. Для відтворення ступеня "дроблення" коштів використовується коефіцієнт розгалуження

$$U(v) = \frac{|\mathcal{N}_{\text{unique}}(v)|}{N_{\text{in}}(v) + N_{\text{out}}(v)}, \quad (2)$$

який визначає співвідношення унікальних контрагентів до сумарної кількості входів і виходів. Значення, що істотно перевищують середні за вибіркою або за часовим вікном цього показника, є формальним індикатором багатоетапних схем розщеплення активів, типових для peel-chain-стратегій, оскільки вони вказують на підвищений ступінь розгалуження потоків.

Оскільки в графових моделях із механізмом уваги (зокрема GATv2) ваги ребер визначають ефективність передачі повідомлень, на етапі попереднього оброблення доцільно обчислювати додаткові евристичні коефіцієнти, які є "підказками" для уваги. Один із таких показників – коефіцієнт фінансової аномалії

$$H_{\text{val}}(e) = \max\left(0, \frac{\text{Value}(e) - \mu_{\text{val}}(u)}{\sigma_{\text{val}}(u)}\right), \quad (3)$$

де  $\mu_{\text{val}}(u)$  і  $\sigma_{\text{val}}(u)$  – математичне сподівання та стандартне відхилення обсягів переказів для вхідних потоків вузла  $u$ . Значення  $H_{\text{val}}$ , що значно відхиляються від середнього, сигналізують нейромережі про необхідність пріоритизації цього ребра [9, 10].

Для компенсації "топологічної сліпоти" простір ознак збагачується 72 структурними ознаками, розрахованими за методами Complex Network Analysis.

Для кожної транзакції обчислюються такі показники.

1. Ступені вузла (In/Out-Degree), що відтворюють кількість вхідних і вихідних фінансових потоків.

2. Центральність міжності (Betweenness Centrality), що вказує на функцію транзакції як "моста" в потоках ліквідності:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \sigma_{st}(v) / \sigma_{st}, \quad (4)$$

де  $C_B(v)$  – центральність міжності вузла  $v$ ;  $s$  і  $t$  – довільні вузли графа, що різняться від  $v$ ;  $\sigma_{st}$  – загальна кількість найкоротших шляхів між вузлами  $s$  та  $t$ ;  $\sigma_{st}(v)$  – кількість таких найкоротших шляхів, які проходять крізь вузол  $v$ .

3. Локальний коефіцієнт кластеризації:

$$C(v) = \frac{2|\mathcal{E}_v|}{k(v)(k(v)-1)}, \quad (5)$$

де  $C(v)$  – локальний коефіцієнт кластеризації вузла  $v$ ;  $k(v)$  – ступінь вузла  $v$ ;  $\mathcal{E}_v$  – кількість ребер між сусідами вузла  $v$ . Значення  $C(v)$ , що наближаються до 1, притаманні закритим контурам міксування коштів.

Систематизацію сформованого 166-вимірного ознакового простору (на основі бенчмарку Elliptic) подано в табл. 1.

Таблиця 1. Систематизація сформованого 166-вимірного ознакового простору

| Група ознак        | Кількість | Приклади атрибутів (ознак)                                          | Семантичне призначення                                                    |
|--------------------|-----------|---------------------------------------------------------------------|---------------------------------------------------------------------------|
| Локальні (вузли)   | 94        | Value (BTC), Fee, Timestep, UTXO counts, LockTime                   | Оцінювання фінансової ємності транзакції та її часових параметрів.        |
| Структурні (1-hop) | 72        | Сума Value сусідів, макс/мін комісії в околі, агрегований In-Degree | Виявлення аномальної поведінки за допомогою аналізу найближчого оточення. |

Аналіз табличної структури демонструє, що поєднання локальних, структурних і темпоральних компонентів перекриває ключові семантичні аспекти поведінки об'єктів у мережі Bitcoin. Зокрема перетворення абсолютних часових міток у відносні інтервали між взаємодіями дає змогу моделі аналізувати динаміку незалежно від "макроепох" блокчейну, що є важливим для виявлення атак типу ransomware, де типовим маркером є швидка консолідація дрібних сум із  $\Delta t \rightarrow 0$  [5]. Водночас обмеженням такого підходу є висока обчислювальна складність точного розрахунку деяких глобальних метрик (зокрема центральності міжності) для надвеликих графів, що зумовлює необхідність упровадження апроксимаційних алгоритмів або вибіркового обчислення на етапі ETL.

### Математична формалізація та побудова гетерогенного графа транзакцій

Після формування багатовимірного простору ознак вихідні табличні спостереження переходять у тензорний формат, придатний для обчислень на GPU та подальшого використання у фреймворках геометричного глибокого навчання, зокрема PyTorch Geometric. У межах запропонованого підходу ончейн-взаємодії Bitcoin описуються як орієнтований атрибутований граф  $G = (V, E, X)$ .

У межах дослідження (на прикладі Elliptic Data Set [1]) граф формалізується як мережа транзакцій, де вузли  $V$  – це безпосередньо транзакції, а ребра  $E$  –

потік невідпрацьованих виходів (UTXO) між ними. У пам'яті системи граф репрезентується набором тензорів.

– Матриця ознак вузлів  $X \in \mathbb{R}^{N \times F_v}$ , де  $N = |V|$  – кількість транзакцій (для Elliptic  $N \approx 203,769$ ), а  $F_v = 166$  – розмірність ознакового простору (локальні та структурні параметри).

– Матриця ознак ребер  $E \in \mathbb{R}^{M \times F_e}$ , де  $M = |E|$  – кількість зв'язків UTXO,  $F_e$  – атрибути потоку (сума, часова мітка).

– Матриця суміжності у форматі координат (COO формат)  $\text{edge\_index} \in \mathbb{Z}^{2 \times M}$ , що містить індекси джерела й приймача активів.

Для забезпечення можливості аналізу просторово-часової динаміки (STGNN) передові методи виявлення глибоких аномалій (Deep Anomaly Detection) в атрибутованих мережах потребують математичного моделювання. Граф моделюється як впорядкована послідовність знімків (Temporal Snapshots)  $\mathbb{G} = \{G^{(1)}, G^{(2)}, \dots, G^{(T)}\}$ , де кожен  $G^{(t)}$  відповідає стану мережі у фіксованому часовому кроці (timestep) [11, 12]. Це дає змогу моделі аналізувати еволюцію графових вкладень (embeddings) у часі, що критично для виявлення довгих ланцюжків відмивання коштів (peel chains).

Як було обґрунтовано раніше, частка іліцитних вузлів у розмічених даних Bitcoin становить приблизно 2%. Для компенсації цього дисбалансу

безпосередньо під час формування батчів (mini-batches) інтегровано алгоритм Graph-SMOTE [13]. Згідно з розробленою логікою тренування цей процес виконується в декілька послідовних етапів. Насамперед ідентифікується міноритарний клас, у процесі чого виділяється підмножина вузлів-аномалій (наприклад, пов'язаних із Ransomware)  $\mathcal{V}_{\min} \subset \mathcal{V}$ . Далі відбувається пошук у латентному просторі. Для кожного цільового вузла  $v_i \in \mathcal{V}_{\min}$  знаходиться  $k$  найближчих сусідів  $v_{knn} \in \mathcal{V}_{\text{syn}}$  у багатовимірному просторі вкладень. Після цього генерується вектор, унаслідок чого створюється синтетичний вузол  $x_{\text{syn}} = x_i + \delta \cdot (x_{knn} - x_i)$ , де параметр  $\delta \sim U(0,1)$  є випадковою величиною з рівномірного розподілу. На завершальному етапі проводиться топологічна реконструкція, під час якої для органічної інтеграції  $v_{\text{syn}}$  у загальний граф  $G$  застосовується генератор ребер (Edge Predictor). Він створює нові орієнтовані ребра між  $v_{\text{syn}}$  і наявними вузлами на основі ймовірнісної моделі: зокрема, якщо  $v_i$  та  $v_{knn}$  раніше мали спільних контрагентів, то новий вузол  $v_{\text{syn}}$  з високою ймовірністю отримує ребра до цих самих адрес [13]. Зрештою такий комплексний підхід дає

зможу ефективно збалансувати вибірку, водночас зберегти статистичну достовірність розподілу ступенів вершин (degree distribution) і загальну топологічну гомофілію графа.

#### Підготовка й характеристика датасету

Elliptic Data Set [13] є індустріальним стандартом де-факто для тестування графових мереж у сфері боротьби з відмиванням коштів. Він є орієнтованим графом  $\mathcal{G}_{\text{ETC}}$ , де вузли відповідають транзакціям, а орієнтовані ребра вказують напрямок руху капіталу (UTXO). Аналіз структурних характеристик демонструє масштаби мережі: граф містить  $|\mathcal{V}| = 203,769$  унікальних вузлів та  $|\mathcal{E}| = 234,355$  орієнтованих ребер. Більшість вузлів у ньому залишаються нерозміченими, що створює фундаментальну проблему екстремального дефіциту міток (label scarcity), для розв'язання якої традиційні моделі є неефективними [8].

Простір ознак  $X$  є 166-вимірним ( $F = 166$ ). Проблема екстремального дисбалансу яскраво виражена в мітках класів (Labels), що наочно подано на рис. 1.

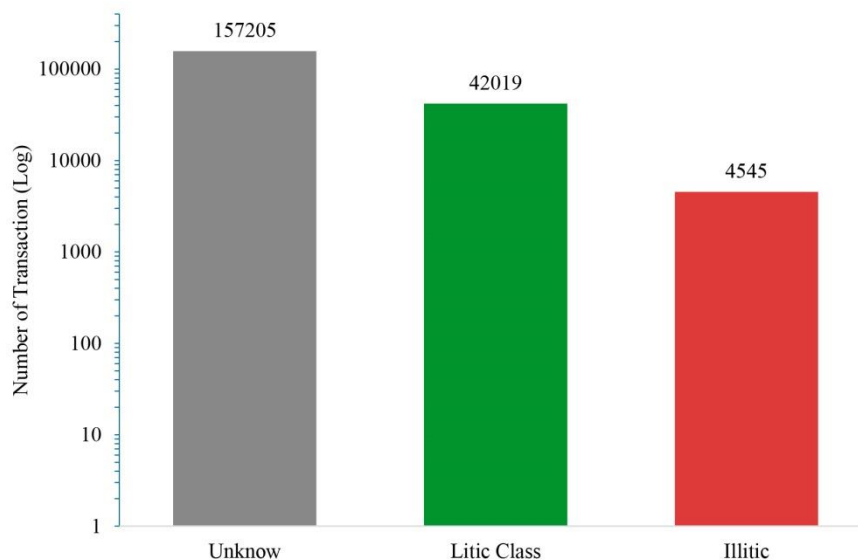


Рис. 1. Розподіл класів у транзакційному графі Elliptic Data Set

Глибокий аналіз даних з рис. 1 демонструє, що позитивний клас Illicit (шахрайські транзакції) містить лише 4 545 вузлів, що є мізерним порівняно з класом Licit (42 019 вузлів) і немаркованою категорією Unknown (157 205 вузлів). Нерівномірний розподіл між цими класами (співвідношення приблизно 1:9 серед маркованих даних) зумовлює

необхідність застосування зваженої функції втрат і методів генеративного оверсемплінгу (Graph-SMOTE). Значна кількість немаркованих вузлів свідчить про те, що методи напівкерowanego навчання (semi-supervised learning) здатні вилучити приховані структурні зв'язки для підвищення якості класифікації [6].

## 5. Результати дослідження

Для доведення ефективності запропонованого методу формування простору ознак було виконано розгорнутий порівняльний експеримент. Оцінювання проводилося в контрольованому середовищі на підготовленому просторово-часовому транзакційному датасеті Elliptic Data. У межах експерименту зіставлялися результати анізотропної архітектури GATv2, доповненої латентним оверсемплінгом Graph-SMOTE та зваженою функцією втрат (WCE), із двома базовими підходами до виявлення аномалій: класичним багатосаровим перцептроном (MLP) і графовою згортковою мережею з ізотропною агрегацією (GCN).

Порівнювані архітектури принципово відрізняються механізмом оброблення інформації, що дає змогу відокремлено оцінити внесок топологічної агрегації та механізму динамічної уваги в якість класифікації. Як табличну базову лінію використано MLP, що розглядає кожну транзакцію незалежно, з огляду лише на її 166-вимірний вектор ознак (включно зі статичними топологічними метриками), і не використовує графову структуру, задану матрицею суміжності  $A$ . Другим еталонним підходом

є GCN, де подання вузла оновлюється способом усереднення ознак сусідніх вершин із нормуванням, що визначається лише ступенем вершини  $\deg(v)$ . У такій постановці всі транзакційні ребра (UTXO-потoki) фактично вважаються рівнозначними за впливом. Третім підходом є GATv2, у якому внесок кожного сусіда  $j$  у подання цільового вузла  $i$  адаптивно зважується коефіцієнтами багатоголової уваги  $\alpha_{ij}^{(k)}$ . Це дає змогу пригнічувати структурний шум від легітимних хабів і підсилювати слабкі, але інформативні зв'язки, властиві для міноритарних аномальних патернів. Усі моделі навчалися за однакових умов для забезпечення коректності порівняння: застосовувався оптимізатор AdamW зі швидкістю навчання  $\eta=0.005$  і ваговим штрафом  $10^{-3}$ , значення Dropout становило 0.5, а розмірність прихованого простору була зафіксована як  $hidden\_dim=128$ . З огляду на екстремальний дисбаланс класів (приблизно  $\approx 1:50$ ), для всіх архітектур використовувалася зважена крос-ентропія (Weighted Cross-Entropy), що забезпечує більш коректну оптимізацію метрик якості в умовах домінування легітимного класу (див. рис. 2).

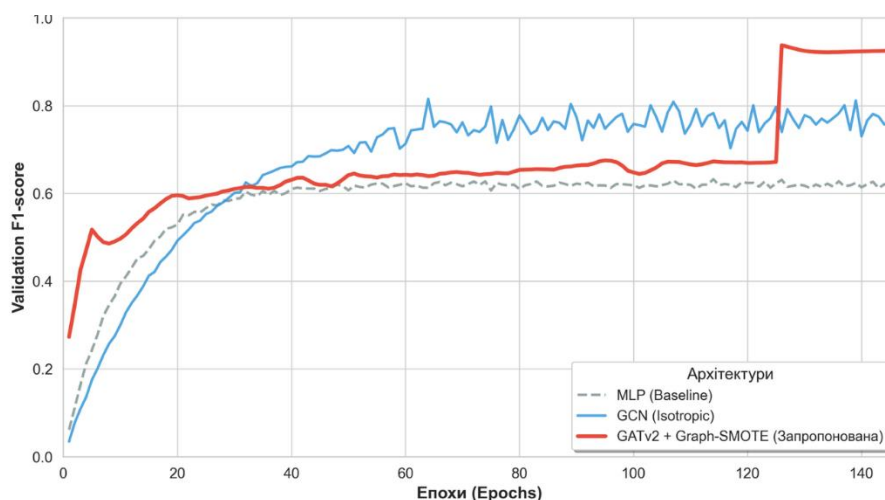


Рис. 2. Динаміка метрики Validation F1-score під час навчання моделей MLP, GCN і GATv2

Детальне дослідження кривих збіжності дає змогу виокремити кілька властивих закономірностей у процесі навчання порівнюваних архітектур. Базова модель багатосарового перцептрона (MLP) швидко покращує якість упродовж перших приблизно 20 епох, однак надалі виходить на плато з рівнем  $F_1 \approx 0,65$ . Така поведінка узгоджується з припущенням, що ізолюваний аналіз векторів ознак

без огляду на графовий контекст не дає змоги надійно відокремлювати складні шахрайські сценарії, де маскування досягається завдяки топологічним схемам і багатокроковим маршрутам руху коштів. Обмеження MLP також полягає в стабільно підвищеній частці хибнопозитивних спрацювань, оскільки модель не має доступу до структурної інформації про оточення вузла

й не може відрізнити локально схожі, але топологічно різні патерни.

Графова згорткова мережа (GCN) демонструє більш рівномірну динаміку навчання та вихід на кращий локальний оптимум із  $F_1$ -score  $\approx 0,7$ , що вказує на практичну цінність топологічної агрегації ознак. Водночас після приблизно 50-ї епохи спостерігається підвищена нестабільність кривої збіжності, яка узгоджується з ефектом надмірного згладжування: двошарова схема агрегації поступово вирівнює подання вузлів у щільних підграфах, зменшуючи роздільну здатність між легітимними й аномальними об'єктами, особливо в умовах масових транзакційних потоків і наявності "хабів" із високою активністю.

Найбільш стабільну оптимізацію демонструє гібридний підхід на основі Graph Attention Network v2 (GATv2) із багатоголовою увагою та інтегрованим Graph-SMOTE. У цій конфігурації цільові метрики зростають більш монотонно, а механізм динамічної

уваги дає змогу зменшувати вплив ребер, що несуть переважно шумовий сигнал, і цим стримувати надмірне згладжування, властиве для ізотропних схем агрегації. Завдяки поєднанню адаптивного зважування зв'язків і латентного оверсемплінгу модель досягає  $F_1$ -score 0,85 на валідаційній вибірці без ознак критичного перенавчання, а процес оптимізації завершується приблизно після 112-ї епохи внаслідок застосування ранньої зупинки.

З огляду на те, що задача виявлення аномалій у транзакційних мережах належить до класу задач із вираженим дисбалансом, використання метрики Ассурасу не є репрезентативним, оскільки вона може маскувати провали на міноритарному класі. Тому порівняння якості на незалежній Out-of-Time тестовій вибірці виконується за метриками Precision, Recall та їх гармонійним середнім F1-score для класу Illicit.

Узагальнені результати порівняльного бенчмаркінгу подано в табл. 2.

**Таблиця 2.** Узагальнені результати порівняльного бенчмаркінгу

| Архітектура моделі  | Precision (точність) | Recall (повнота) | F1-Score (гармонійне середнє) | Assurasy (загальна) |
|---------------------|----------------------|------------------|-------------------------------|---------------------|
| MLP (Baseline)      | 0.6214               | 0.5830           | 0.6016                        | 0.9420              |
| GCN (Isotropic)     | 0.7582               | 0.7815           | 0.7697                        | 0.9655              |
| GATv2 (Multi-head)  | 0.8412               | 0.9105           | 0.8744                        | 0.9780              |
| GATv2 + Graph-SMOTE | 0.8875               | 0.9022           | 0.8948                        | 0.9845              |

Аналіз інформації з табл. 2 підтверджує наукові гіпотези дослідження. Перехід від лінійної MLP до графової GCN зумовив різке зростання  $F1$ -score (з 0.6016 до 0.7697; приріст +16.8%). Це емпірично доводить, що структура зв'язків є значно інформативною для виявлення аномалій.

Модель GATv2 перевершила GCN за показником  $F1$ -score на 10.4% (0.8744 проти 0.7697). Ключовим досягненням GATv2 є максимізація повноти виявлення ( $Recall = 0.9105$ ). Це означає, що механізм уваги дає змогу системі розпізнавати понад 91% реальних кібератак.

## 6. Обговорення результатів

Досягнуті результати підтверджують практичну придатність запропонованого методу формування простору ознак для задач ризик-скорингу в мережі Bitcoin. Його перевага полягає в поєднанні локальних характеристик вузлів із топологічними метриками, що дає змогу зважати не лише на параметри окремих транзакцій, а й на структурний контекст руху

коштів. Це є особливо важливим в умовах обфускації, коли ризикові сценарії маскуються через складні графові взаємозв'язки.

Запропонований підхід може бути використаний у системах моніторингу ризиків, сервісах виявлення шахрайства й аналітичних платформах, орієнтованих на дослідження ончейн-активності. Зв'язок із промисловістю полягає в тому, що сформований простір ознак створює основу для побудови прикладних рішень у сфері криптовалютної аналітики, де критичними є стійкість до шуму, врахування топологічних патернів і підвищення якості відокремлення ризикових об'єктів від легітимних.

Отже, запропонований метод доцільно розглядати як базовий етап підготовки даних для подальшого використання в алгоритмах ризик-скорингу й графово-орієнтованих інтелектуальних системах фінансового моніторингу.

## 6. Висновки

У дослідженні обґрунтовано ефективність ризик-скорингу в мережі Bitcoin, що визначається



насамперед якістю методу формування простору ознак і здатністю методу зберігати структурний контекст транзакційного графа, оскільки багатоетапні шахрайські сценарії маскуються через топологічні патерни й обфускаційні стратегії. Сформовано багатовимірний простір ознак, який поєднує локальні мікроатрибути й топологічні макрохарактеристики околу, що сприяє зменшенню розрідженості ознакового простору й підвищенню інформативності вхідних даних для алгоритмів ризик-скорингу. Експериментальне порівняння підтвердило, що табличний підхід без явного використання графового контексту швидко досягає межі якості, тоді як залучення графової агрегації забезпечує приріст метрик, але може страждати від надмірного згладжування у щільних підграфах. Найкращі результати отримано внаслідок застосування анізотропного механізму уваги, який адаптивно зважає внесок сусідів і пригнічує структурний шум, а застосування латентного оверсемплінгу дає змогу підвищити чутливість до міноритарного класу без критичного перенавчання. Досягнуті результати підтверджують доцільність запропонованого методу як основи для побудови систем раннього виявлення ризикових об'єктів у Bitcoin і створюють підґрунтя для подальшого розвитку, зокрема внаслідок

розширення набору контекстних ознак, удосконалення часових компонентів і оптимізації обчислення глобальних графових метрик для великих мереж.

### Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

### Фінансування

Дослідження проведено без фінансової підтримки.

### Доступність даних

Рукопис не має пов'язаних даних.

### Використання засобів штучного інтелекту

Автори підтверджують, що не застосовували технології ШІ для написання статті.

## References

1. Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., Leiserson, C. E. (2019), "Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics", *Social and Information Networks*. DOI: <https://doi.org/10.48550/arXiv.1908.02591>
2. Akcora, C. G., Li, Y., Gel, Y. R., Kantarcioglu, M. (2020), "BitcoinHeist: Topological Data Analysis for Ransomware Prediction on the Bitcoin Blockchain", *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 4439–4445. DOI: <https://doi.org/10.24963/ijcai.2020/612>
3. Elmougy, Y., Liu, L. (2023), "Demystifying Fraudulent Transactions and Illicit Nodes in the Bitcoin Network for Financial Forensics". DOI: <https://doi.org/10.48550/arXiv.2306.06108>
4. Kipf, T. N., Welling, M. (2017), "Semi-Supervised Classification with Graph Convolutional Networks", *Proceedings of the International Conference on Learning Representations*. DOI: <https://doi.org/10.48550/arXiv.1609.02907>
5. Hamilton, W. L., Ying, R., Leskovec, J. (2017), "Inductive Representation Learning on Large Graphs", *Advances in Neural Information Processing Systems*, Vol. 30. DOI: <https://doi.org/10.48550/arXiv.1706.02216>
6. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y. (2018), "Graph Attention Networks", *Proceedings of the International Conference on Learning Representations*. DOI: <https://doi.org/10.48550/arXiv.1710.10903>
7. Brody, S., Alon, U., Yahav, E. (2022), "How Attentive are Graph Attention Networks?", *Proceedings of the International Conference on Learning Representations*. DOI: <https://doi.org/10.48550/arXiv.2105.14491>
8. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002), "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321–357. DOI: <https://doi.org/10.1613/jair.953>
9. Zhao, T., Zhang, X., Wang, S. (2021), "GraphSMOTE: Imbalanced Node Classification on Graphs with Graph Neural Networks", *WSDM 2021: Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 833–841. DOI: <https://doi.org/10.1145/3437963.3441720>

10. Pareja, A., Domeniconi, G., Chen, J., Ma, T., Suzumura, T., Kanezashi, H., Kaler, T., Schardl, T. B., Leiserson, C. E. (2020), "EvolveGCN: Evolving Graph Convolutional Networks for Dynamic Graphs", *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 4, pp. 5363–5370. DOI: <https://doi.org/10.1609/aaai.v34i04.5984>
11. Lundberg, S. M., Lee, S.-I. (2017), "A Unified Approach to Interpreting Model Predictions", *Advances in Neural Information Processing Systems*, Vol. 30, pp. 4765–4774. DOI: <https://doi.org/10.48550/arXiv.1705.07874>
12. Bellei, C., Robinson, T., Weber, M., Weidele, D. K. I., Domeniconi, G., Chen, J., Leiserson, C. E. (2024), "The Shape of Money Laundering: Subgraph Representation Learning on the Blockchain with the Elliptic2 Dataset", *arXiv preprint arXiv:2404.19109*. DOI: <https://doi.org/10.48550/arXiv.2404.19109>
13. Elliptic Data Set. URL: <https://www.kaggle.com/datasets/ellipticco/elliptic-data-set>
14. Taherdoost, H. (2023), "Blockchain and Machine Learning: A Critical Review on Security", *Information*, Vol. 14, No. 5. DOI: <https://doi.org/10.3390/info14050295>
15. Fedorchenko, V., Yeroshenko, O., Shmatko, O., Kolomiitsev, O., Omarov, M. (2024), "Password hashing methods and algorithms on the .Net platform", *Advanced Information Systems*, Vol. 8, No. 4, pp. 82–92. DOI: <https://doi.org/10.20998/2522-9052.2024.4.11>
16. Shaizat, M., Mussiraliyeva, S. (2025), "Enhanced identification of illicit bitcoin transactions through genetic algorithm-based feature selection", *Eastern-European Journal of Enterprise Technologies*, Vol. 4, 9 (136), pp. 34–42. DOI: <https://doi.org/10.15587/1729-4061.2025.335630>

Received (Надійшла) 31.03.2026

Accepted for publication (Прийнята до друку) 16.05.2026

Publication date (Дата публікації) 29.05.2026

#### Відомості про авторів / About the Authors

**Просолов Владислав Валерійович** – Харківський національний університет радіоелектроніки, аспірант кафедри безпеки інформаційних технологій, Харків, Україна;

**Vladyslav Prosolov** – Kharkiv National University of Radio Electronics, Postgraduate Student of the Department of Information Technology Security, Kharkiv, Ukraine;

e-mail: [vladyslav.prosolov@nure.ua](mailto:vladyslav.prosolov@nure.ua)

ORCID ID: <https://orcid.org/0009-0002-1276-4828>

**Халімов Геннадій Зайдулович** – доктор технічних наук, професор, Харківський національний університет радіоелектроніки, завідувач кафедри безпеки інформаційних технологій, Харків, Україна;

**Hennadii Khalimov** – Doctor of Technical Sciences, Professor, Kharkiv National University of Radio Electronics, Head of the Department of Information Technology Security, Kharkiv, Ukraine;

e-mail: [hennadii.khalimov@nure.ua](mailto:hennadii.khalimov@nure.ua)

ORCID: <https://orcid.org/0000-0002-2054-9186>

## FEATURE SPACE CONSTRUCTION FOR BITCOIN RISK SCORING BASED ON LOCAL AND TOPOLOGICAL METRICS

**The subject** of the study is a feature space construction method for risk scoring in the Bitcoin network based on the combination of local node characteristics, topological metrics of their neighborhood, nonlinear normalization, and graph transformation of on-chain data. **The purpose** of the study is to improve the feature space construction method for unstructured on-chain data in the Bitcoin network through the comprehensive aggregation of local and topological characteristics of transaction graph objects and the application of nonlinear normalization, which makes it possible to increase the informativeness of data for risk-scoring tasks. **Objectives:** to substantiate the feasibility of graph-based representation of on-chain data in Bitcoin risk-scoring tasks; to form a multidimensional feature space based on local micro-features and global topological macro-features; to apply preprocessing, cleaning, and nonlinear normalization procedures to financial values; to investigate the role of local, structural, temporal, and contextual attributes in improving the separation of risky objects from legitimate ones; and to evaluate the

effectiveness of the constructed feature space as a basis for further use in risk-scoring algorithms. **The research methods** are based on the formalization of on-chain data in the form of a graph, the use of feature engineering methods, graph transformations, nonlinear normalization, and the analysis of local and topological metrics of the Bitcoin transaction network. **Research results.** A comprehensive approach to constructing a feature space for Bitcoin risk scoring is proposed, which involves the aggregation of 94 local micro-features and 72 global topological macro-features, as well as the application of nonlinear normalization and graph data transformation. It is shown that such an approach makes it possible to reduce the sparsity of the feature space, increase the informativeness of the input data, and provide better discriminative capability in separating risky objects from legitimate ones. It is substantiated that a properly constructed feature space can serve as a basic component for the further integration of graph-based and temporal neural network mechanisms in financial analytics tasks. **Conclusions.** The proposed feature space construction method is a practically applicable tool for risk monitoring systems in the Bitcoin network and confirms the feasibility of combining local and topological metrics to improve the robustness of risk scoring against obfuscation scenarios and complex fund flow trajectories.

**Keywords:** Bitcoin, on-chain data, risk scoring, feature engineering, topological metrics, graph transformations, fraud detection.

#### *Бібліографічні описи / Bibliographic descriptions*

Просолов В. В., Халімов Г. З. Формування простору ознак для ризик-скорингу bitcoin на основі локальних і топологічних метрик. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 224–234. DOI: <https://doi.org/10.30837/0135-1710.2026.189.224>

Prosolov, V., Khalimov, H. (2026), "Feature space construction for bitcoin risk scoring based on local and topological metrics", *Management Information System and Devices*, No. 2 (189), P. 224–234. DOI: <https://doi.org/10.30837/0135-1710.2026.189.224>

---

Сидоренко З. М., Северінов О. В.

## МЕТОД ЗАБЕЗПЕЧЕННЯ ЦІЛІСНОСТІ ДАНИХ ІЗ ВИКОРИСТАННЯМ МОДИФІКОВАНОГО ЗАВАДОСТІЙКОГО КОДУ В СИСТЕМАХ ПРОМИСЛОВОГО ІНТЕРНЕТУ РЕЧЕЙ

**Предметом вивчення** є методи модифікації завадостійких кодів для забезпечення цілісності даних у системах промислового Інтернету речей. **Мета дослідження** – розробити метод забезпечення цілісності даних у системах промислового Інтернету речей за допомогою застосування модифікованих завадостійких кодів, а саме скорочених кодів Гоппи для виявлення та виправлення помилок, що виникають під час передавання й оброблення інформації. **Завдання:** аналіз методів модифікації завадостійких кодів, а саме кодів Гоппи, з метою зашумлення повідомлення для забезпечення цілісності даних у системах промислового Інтернету речей; дослідження методів скорочення кодів Гоппи та визначення найбільш доцільних конструкцій для забезпечення завадостійкості та цілісності даних; розроблення методу забезпечення цілісності даних із використанням скороченого завадостійкого коду в системах промислового Інтернету речей. Для виконання окреслених завдань застосовано **методи:** теоретичні методи аналізу й математичного моделювання, емпіричні методи порівняльного аналізу характеристик кодів, а також елементи комп'ютерного експерименту для розрахунку й оцінювання параметрів скорочених кодів Гоппи. **Досягнуті результати.** Проаналізовано методи модифікації кодів Гоппи, для побудови системи забезпечення цілісності даних в IoT запропоновано метод скорочення кодів, що дає змогу зберігати мінімальну кодову відстань і формувати коди із заданими параметрами. Досліджено конструкції скорочення кодів Гоппи, обґрунтовано доцільність використання скорочених кодів Гоппи на основі багаточленів, що не приводяться. Запропоновано метод забезпечення цілісності даних із застосуванням скорочених кодів Гоппи. **Висновки.** Упровадження процедур скорочення кодів Гоппи, побудованих на основі багаточленів, що не приводяться, дає змогу формувати кодові конструкції для забезпечення цілісності даних у системах промислового Інтернету речей, які одночасно мають високі властивості завадостійкості. Запропоновано метод забезпечення цілісності даних з використанням скорочених кодів Гоппи, що забезпечує гнучке налаштування параметрів коду й має перспективи гарантувати інформаційну скритність повідомлень.

**Ключові слова:** промисловий Інтернет речей; завадостійкість; цілісність; завадостійкий код; код Гоппи; модифікація коду; скорочений код.

### Вступ

Нині впровадження промислового Інтернету речей (ІоТ) супроводжується значними перешкодами. Вони пов'язані з тим, що в процесі створення цих систем промислові активи об'єднують в одну систему, під'єднану до мережі Інтернет, тобто вони стають вразливими до кібератак. Тому проблеми безпеки під час впровадження промислового Інтернету речей є основним пріоритетом у сучасних умовах. Крім того, успішна стратегія ІоТ вимагає єдиної архітектури даних, яка може обробляти величезні обсяги інформації. У таких умовах основна увага має бути зосереджена на створенні безпечного й безперебійного потоку даних.

Промисловий Інтернет речей – це система об'єднаних комп'ютерних мереж і під'єднаних до них промислових (виробничих) об'єктів із вбудованими датчиками й програмним забезпеченням для збору інформації та її обміну з можливістю віддаленого контролю й управління в автоматизованому

режимі [1, 2]. Системи промислового ІоТ визначаються значною кількістю сенсорів, контролерів і виконавчих пристроїв, що обмінюються даними через бездротові або промислові мережі. Такі канали зв'язку часто піддаються впливу електромагнітних завад, втрат пакетів, затримок або навіть цілеспрямованих атак [1, 2].

У цих умовах постає подвійне завдання – захист від спотворення інформації через ненавмисні технічні помилки й контроль цілісності даних проти зловмисних змін. Це також визначається вимогами низки стандартів для систем ІоТ, а саме IEC 61000, серії IEC 62443 та ISO/IEC 27001. У цьому разі основна увага зосереджується на необхідності забезпечення цілісності даних у системах ІоТ. Так, відповідно до стандарту IEC 62443-3-3, забезпечення цілісності даних є обов'язковим складником безпеки промислового Інтернету речей, що реалізується через вимоги до цілісності системи (System Integrity) й механізми виявлення несанкційних змін інформації та запобігання їм [3].

Одним із підходів до розв'язання завдання захисту від спотворення інформації та забезпечення цілісності даних є застосування в системах промислового ПоТ завадостійких (коригувальних) кодів, а саме кодів Гоппи [4]. Коди Гоппи – один із підкласів альтернативних кодів, які мають кращі характеристики серед лінійних блокових кодів і доведені властивості на межі Варшавова – Гілберта. Їх основна перевага перед іншими завадостійкими кодами полягає в можливості побудови значної кількості кодових слів із заданими параметрами. Це робить їх придатними для криптографічних цілей, а саме забезпечення цілісності даних у системах промислового Інтернету речей.

Перша криптографічна система на основі кодів Гоппи була запропонована Макелісом як система з відкритим ключем [5]. У цій системі секретний ключ містить породжувальну матрицю, двійковий код Гоппи  $G$ , невироджену матрицю  $S$  і матрицю перестановок  $P$ , які її маскують. Відкритим ключем є матриця кодування  $G' = S \times G \times P$  звичайного лінійного коду. Іншим варіантом використання кодів Гоппи є модифікована система Рао – Нама, в якій породжувальна матриця коду Гоппи  $G$  є секретним ключем. Системи Макеліса й Рао – Нама практично не використовувалися через велику на той час довжину ключа, що дорівнює  $k \times n$  символів породжувальної матриці коду.

Стійкість даних криптографічних систем базується на складності декодування випадкового лінійного блочного коду – однієї з фундаментальних задач у теорії завадостійкого кодування, а також відновлення структури коду. Для маскування структури коду під випадковий код використовуються різні методи: долучення матриці перестановок, які її маскують, породжувальна матриця, додавання до кодового слова випадкового вектора помилок, а також процедури модифікації коду.

У роботі [4] для забезпечення цілісності та завадостійкості даних запропоновано використовувати коди Гоппи, а ключем замість породжувальної матриці обирати примітивний багаточлен Гоппи  $G(x)$ . У цьому разі загальна кількість ключів системи визначається кількістю багаточленів примітивної довжини, що не приводяться.

Крім того, в системах захисту даних у ПоТ запропоновано застосовувати модифіковані коди Гоппи, що дасть змогу не тільки одночасно досягти високих

властивостей цілісності та завадостійкості даних, а також забезпечить конфіденційність інформації.

### Аналіз літературних джерел і визначення проблеми

Упродовж 2016–2017 років Національний інститут стандартів і технологій США (NIST) почав проєкт з відбору найкращих постквантових алгоритмів для стандартизації та заміни класичних криптографічних алгоритмів. Класична схема Макеліса [5] залишилася одним із найсильніших кандидатів і ввійшла до фіналістів NIST [6]. У зв'язку з цим з'явилась низка досліджень із застосування завадостійких кодів у криптографічних алгоритмах. Відомі кодові криптосистеми розвивали Макеліс, Рао – Нам і Нідеррайтер [7]. Здебільшого вони спрямовані на розроблення симетричних і асиметричних криптосистем на основі модифікованих схем Макеліса, Рао – Нама й Нідеррайтера.

Так, у праці [8] запропоновано механізми безпеки в гібридно-криптокодових системах, основаних на модифікованих асиметричних системах Нідеррайтера й Макеліса, для використання в методах двофакторної автентифікації на одноразових паролях. Ці модифіковані системи запропоновано будувати на еліптичних кодах замість класичних кодів Гоппи.

У роботі [9] розглянуто класичні системи на основі коду Гоппи й оцінено їх безпеку з огляду на квантові прискорення найефективніших атак, а також запропоновано схеми, основані на кодах QC-LDPC, QC-MDPC і мономіальних кодах.

Автори досліджень [10, 11] розглядають криптосистеми з відкритим ключем на основі кодових схем Макеліса й Нідеррайтера. З огляду на криптосистеми Макеліса запропоновано нову схему цифрового підпису.

У роботі [12] проаналізовано використання симетричної криптокової конструкції на основі системи Рао – Нама на алгебро-геометричних і дефектних кодах, що забезпечує можливість суттєвого зменшення обсягу ключових даних в інфраструктурі на базі мобільних і смарт-технологій.

Однак у розглянутих та інших подібних дослідженнях висвітлено питання розвитку класичних схем Макеліса, Рао – Нама й Нідеррайтера й побудови на їх основі класичних криптографічних симетричних або асиметричних схем шифрування, схем цифрового підпису та генерації ключів.



Ці системи побудовані на використанні різних завадостійких кодів, але не розглядаються для комбінованого застосування з метою захисту від спотворення інформації через ненавмисні технічні помилки й контролю цілісності даних.

Автори праці [13] проаналізували постквантові алгоритми, подані в проєкті NIST. Підтверджено, що більшість алгоритмів, побудованих на кодах, окрім класичної схеми Макеліса з кодами Гоппи, були зламані або суттєво ослаблені. Автори запропонували статичний завадостійкий код не використовувати, а здійснювати його динамічне перетворення за допомогою процедур модифікації (скорочення, додавання) коду.

#### Мета й завдання дослідження

Використання модифікованих кодів Гоппи в каналах передачі даних системи промислового Інтернету речей для захисту від помилок може забезпечити високі характеристики щодо завадостійкості та цілісності інформації, яка передається.

Аналіз літератури [14–16] дає змогу виокремити шість основних методів модифікації коду: збільшення довжини коду способом додавання інформаційних або перевірчих символів, зменшення довжини кодового блоку внаслідок викидання (вилучення) інформаційних або перевірчих символів, а також збільшення (додавання) чи зменшення (вилучення) кодових слів за незмінної довжини блоку.

Отже, постає завдання – проаналізувати методи модифікації завадостійких кодів, а саме кодів Гоппи, з метою зашумлення повідомлення для забезпечення цілісності даних у системах промислового IoT.

**Мета роботи** – підвищити ефективність забезпечення цілісності даних у системах промислового Інтернету речей за допомогою застосування модифікованих завадостійких кодів, а саме скорочених кодів Гоппи.

Крім того, необхідно розглянути особливості використання кодів Гоппи для забезпечення цілісності даних, проаналізувати методи скорочення цих кодів і визначити найбільш доцільні методи для впровадження в системах промислового IoT з метою забезпечення завадостійкості та цілісності даних.

#### Методи дослідження

Для аналізу методів модифікації завадостійких кодів Гоппи з метою сприяння цілісності даних у системах промислового Інтернету речей, дослідження особливостей їх застосування, аналізу методів скорочення кодів Гоппи та визначення найбільш доцільних методів для використання в системах промислового IoT для забезпечення завадостійкості та цілісності даних було впроваджено теоретичні методи аналізу й математичного моделювання, емпіричні методи порівняльного аналізу характеристик кодів, а також елементи комп'ютерного експерименту для розрахунку й оцінювання параметрів скорочених кодів Гоппи.

#### Аналіз методів модифікації завадостійкого коду

Розглянемо основні методи модифікації коду щодо кодів Гоппи з метою визначення доцільності їх використання для зашумлення повідомлень у системах забезпечення цілісності.

Код Гоппи може бути подовжений (Lengthening) способом долучення додаткових символів. Звичайний спосіб подовження коду полягає в послідовному виконанні двох операцій: поповнення коду внаслідок додавання вектора  $(1\ 1 \dots 1)$  (якщо він не належить коду) й розширення його за допомогою загальної перевірки на парність. Внаслідок цих операцій кількість інформаційних символів збільшиться на одиницю.

Код Гоппи може бути розширений (Extending) способом долучення додаткових перевірчих символів. Зазвичай такою модифікацією є впровадження однієї загальної перевірки на парність. Значення відповідного перевірного символу  $a(\infty)$  дорівнює

$$a(\infty) = -\sum_{i=0}^n a(\alpha),$$

де  $\alpha \in GF(q^m)$ ;  $GF(q^m)$  – Galois Field, скінченне поле Галуа;  $q$  – основа поля;  $m$  – степінь розширення поля;  $a(\alpha)$  – символи кодового слова;  $n$  – довжина коду.

Тоді вектор  $a' = (a(0), a(1), \dots, a(\alpha^{n-1}), a(\infty))$  належить розширеному коду Гоппи тоді й тільки тоді, коли [14–16]

$$\sum_{i=0}^n \frac{\alpha^i a(\alpha^i)}{G(\alpha^i)} = 0, \quad i = 0, 1, \dots, n,$$

де  $\alpha \in GF(q^m) \cap \{\infty\}$ .

Унаслідок додавання загальної перевірки на парність вага кожного кодового слова стає парною, а кодова відстань збільшується на одиницю. Так, вихідний код Гоппи  $Y(n, k, d)$  ( $n$  – довжина кодового слова,  $k$  – кількість інформаційних символів у кодовому слові,  $d$  – мінімальна відстань коду) буде модифіковано в код  $Y'(n+1, k, d+1)$ . Якщо  $Y$  має перевірку матрицю  $H$ , тоді  $Y'$  матиме перевірку матрицю

$$H' = \begin{pmatrix} 111\dots 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad H.$$

Код Гоппи може бути поповнений за допомогою додавання нових слів (Augmenting). Найпростіший спосіб поповнення коду полягає в додаванні до нього вектора  $(11\dots 1)$  (якщо він не належить коду). Це еквівалентно додаванню вектора  $(11\dots 1)$  до породжувальної матриці. Якщо  $Y$  – двійковий  $(n, k, d)$ -код Гоппи, який містить вектор  $(11\dots 1)$ , то поповнений код дорівнює  $Y' = Y \cup \{1+Y\}$ , тобто  $Y$  містить кодові слова  $Y$  та їх доповнення і є  $(n, k+1, d')$ -кодом. Мінімальна відстань поповненого коду  $d' = \min\{d, n-d_a\}$ , де  $d_a$  – найбільша вага кодових слів коду  $Y$ . Іноді код Гоппи з багаточленом Гоппи  $G(x)$  може бути поповнений до коду тієї самої довжини, багаточлен Гоппи якого є множителем  $G(x)$ .

Код Гоппи з викиданням (Expurgating) утворюється способом вилучення деяких кодових слів. Так, якщо двійковий  $(n, k, d)$ -код  $Y$  містить кодові слова парної та непарної ваги, то отримаємо код із викиданням  $Y'(n, k-1, d')$ , якщо брати кодові слова тільки парної або непарної ваги. Якщо  $d$  непарне й викидаються всі слова непарної ваги, тоді  $d' > d$ . У частотній області ця процедура

проводиться способом додаткових обмежень у систему рівнянь, що визначають код Гоппи.

Код Гоппи може бути виколотий (Puncturing) унаслідок вилучення деяких перевірчих символів. Ця процедура обернена до процедури розширення коду. Загалом у процесі виколування однієї кодової координати довжина коду зменшується на одиницю і, якщо не подбати про вибір символу, що кодується, відстань коду зменшиться на одиницю.

Розглянемо процедуру скорочення коду (Shortening). Нехай лінійний  $(n, k)$ -код має породжувальну матрицю,  $i$  стовпців якої є лінійно незалежними ( $i < k$ ). Безліч векторів довжини  $n-i$ , отриманих вилученням  $i$  компонент кодових векторів, утворюють лінійний  $(n-i, k-i)$ -код. Ця процедура називається скороченням, а код – скороченим кодом [14–16]. Породжувальна матриця скороченого коду виходить з породжувальної матриці початкового коду вилученням  $i$  рядків та  $i$  стовпців. Перевірча матриця виходить вилученням  $i$  стовпців із перевіркої матриці початкового коду. У разі скорочення коду символи, що викидаються, покладаються рівними нулю й тому не передаються, але на приймальній стороні декодер їх відновлює, так що декодування здійснюється на повній довжині коду. Якщо мінімальна відстань початкового коду дорівнює  $d$ , мінімальна відстань скороченого коду не менша ніж  $d$ .

З розглянутих методів модифікації найбільш прийнятним для побудови системи забезпечення завадостійкості та цілісності інформації, що передається, для промислового Інтернету речей є скорочення кодів. Методи скорочення не змінюють мінімальну відстань  $d$  початкового коду. Крім того, методи скорочення дають змогу будувати сімейства кодів із заданими параметрами: довжиною кодового блоку  $n$  і кількістю інформаційних символів  $k$ .

Оскільки модифікацію коду пропонується використовувати для зашумлення повідомлення, то символи скорочення можна застосовувати як секретний ключ, тому що вони не передаються, а відновлюються на приймальній стороні, і декодування здійснюється на повній довжині початкового коду. Тому вірогідно зловмисник не може однозначно відновити скорочене кодове слово до повної довжини й визначити ключовий багаточлен Гоппи  $G(x)$ .

Також скорочення кодів часто приводить до кодів із найкращими параметрами. Це видно з таблиці кращих кодів [17]. У [17] доведено, що за великих  $n$  скорочені коди досягають нижньої межі Варшамова – Гілберта. Наприклад, скорочений (55,16,19)-код Гоппи, що задається багаточленом  $G(x)$  ступеня 9, є найкращим серед відповідних кодів таблиць [17]. Скорочений (239,123,35)-код Гоппи, що задається багаточленом  $G(x)$  ступеня 17, цікавий тим, що автори [18] отримують з нього цілу низку кодів довжини 227 і мінімальною відстанню 29, які також є кращими серед відповідних відомих кодів. Отже, застосовуючи процедуру скорочення до кодів Гоппи можна будувати на їх основі кодові конструкції, які мають не тільки високі характеристики завадостійкості, але й значну кількість способів формування коду, що дає змогу будувати на їх основі системи забезпечення цілісності інформації. Постає завдання – проаналізувати методи формування скорочених кодів Гоппи.

#### Дослідження методів скорочення кодів Гоппи

Скорочені коди Гоппи будуються на елементах поля  $L$ , визначеного на  $GF(q^m) - \{\alpha_1, \alpha_2, \dots, \alpha_p\}$ , де  $\alpha_i \in GF(q^m), i = \overline{1, p}$ .

Є два варіанти побудови скороченого коду. У першому варіанті багаточленом коду Гоппи  $G(x)$  обирається багаточлен, що має некрятні корені в полі  $GF(q^m)$  і за умови скорочення викидаються нульові частоти на позиціях коренів багаточлена  $G(x)$ . Довжина скороченого коду в цьому разі дорівнюватиме  $(n-p)$ , де  $p$  – кількість коренів багаточлена коду Гоппи.

У другому варіанті як  $G(x)$  обирається багаточлен, що не приводиться, а з поля  $GF(q^m)$  викидаються деякі символи (частоти). Тоді довжина коду зменшиться на кількість символів (частот), що викидаються.

У побудові скороченого коду Гоппи на основі багаточлена  $G(x)$ , що має некрятні корені, кількість коренів  $p$ , а отже, і символів скорочення

визначатиме ступінь багаточлена Гоппи  $t$  і, відповідно, кодову відстань  $d$ . Тому в цьому разі кількість інформаційних символів дорівнюватиме

$$k' = k - p \geq n - mt - p = n - m(2p+1) - p.$$

Нехай маємо багаточлени коду Гоппи, що містять  $p_1$  і  $p_2 = p_1 + 1$  коренів, тоді:

$$k'_1 = n - m(2p_1 + 1) - p_1;$$

$$k'_2 = n - m(2p_2 + 1) - p_2 = n - m(2p_1 + 3) - p_1 - 1;$$

$$k'_1 - k'_2 = n - m(2p_1 + 1) - p_1 - n + m(2p_1 + 3) + p_1 + 1 = 2m + 1.$$

Отже, внаслідок зміни кількості символів скорочення (коренів багаточлена Гоппи) на одиницю кількість інформаційних символів змінюватиметься на  $2m+1$  біт, а також значно змінюватиметься швидкість коду  $V = k'/n'$ .

У табл. 1 подано параметри скорочених кодів Гоппи, отримані за допомогою комп'ютерного моделювання для багаточленів кодів Гоппи  $G(x)$  з різною кількістю коренів.

Таблиця 1. Параметри кодів Гоппи для багаточленів із коренями

| Кількість коренів | Розмірність поля |                                                    |             |              |
|-------------------|------------------|----------------------------------------------------|-------------|--------------|
|                   | GF(32)           | GF(64)                                             | GF(128)     | GF(256)      |
| 2                 | 30, 25, 5        | 62, 50, 5                                          | 126, 112, 5 | 254, 238, 5  |
| 3                 | 29, 14, 7        | 61, 43, 7                                          | 125, 104, 7 | 253, 229, 7  |
| 4                 | 28, 8, 9         | 60, 36, 9                                          | 124, 96, 9  | 252, 220, 9  |
| 5                 | 27, 4, 11        | 59, 29, 11                                         | 123, 88, 11 | 251, 211, 11 |
| 6                 | 26, 2, 13        | 58, 22, 13<br>58, 23, 13                           | 122, 80, 13 | 250, 202, 13 |
| 7                 | –                | 57, 15, 15<br>57, 16, 15<br>57, 17, 15             | 121, 72, 15 | 249, 193, 15 |
| 8                 | –                | 56, 8, 17<br>56, 9, 17<br>56, 10, 17<br>56, 16, 17 | 120, 64, 17 | 248, 184, 17 |

Аналіз наведених параметрів демонструє, що значна зміна інформаційної довжини коду й кодової відстані за умови скорочення на один символ не дає змогу гнучко змінювати параметри коду Гоппи.

Кількість багаточленів, обумовлених своїми коренями, дорівнює кількості комбінацій усіх неповторюваних елементів поля  $GF(q^m)$

$$N_k = C_{q^m}^p,$$

де  $p$  – кількість коренів багаточлена Гоппи  $G(x)$ .

Кількість  $N_q(t)$  багаточленів, що не приводяться, ступеня  $t$  визначається виразом

$$N_q(t) = \frac{1}{t} \sum_{r|t} \mu(t/r) q^r,$$

де  $r|t$  означає, що сума береться за всіма позитивними дільниками  $r$  числа  $t$ ;  $\mu(t/r)$  – функція Мебіуса.

На рис. 1 подано графіки залежності кількості багаточленів  $G(x)$ , що не приводяться,  $N_q(t)$ , і багаточленів з коренями  $N_k(t)$  від їх степеня для кодів над різними полями.

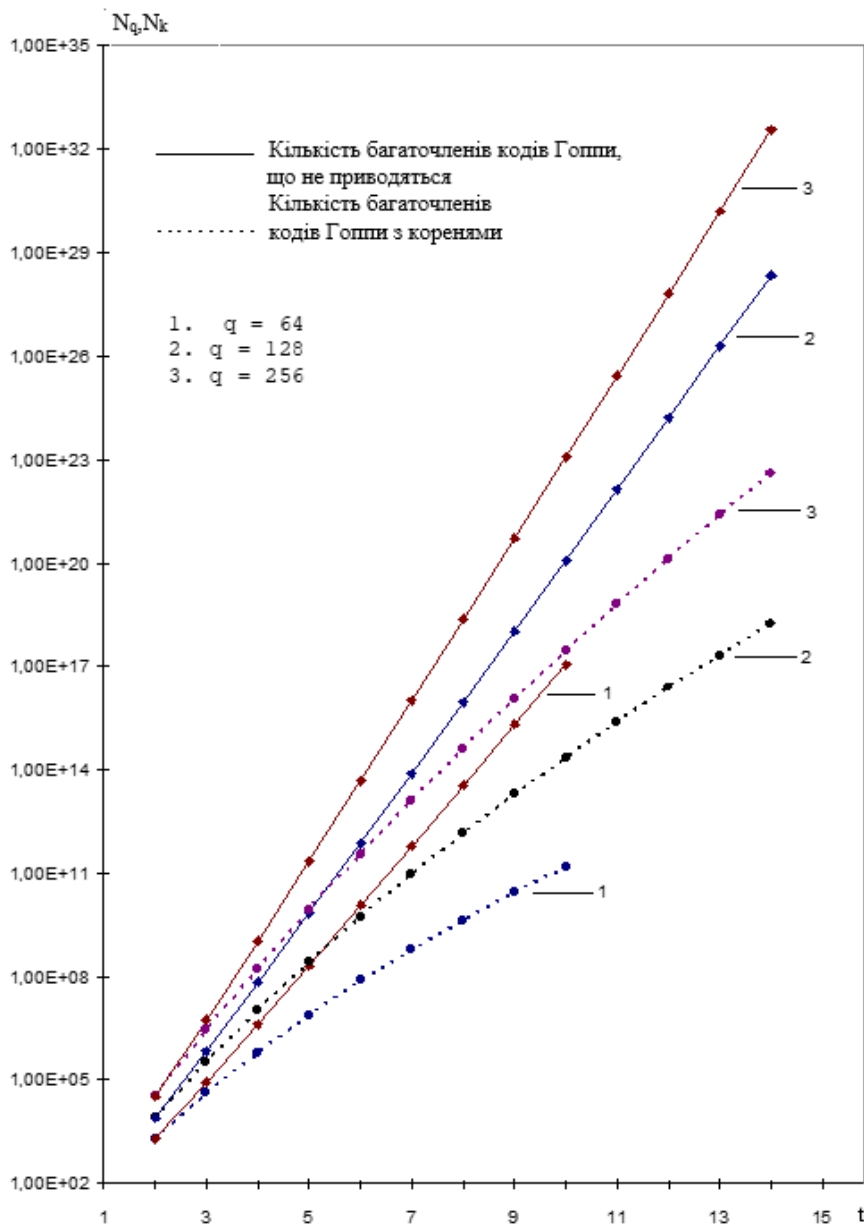


Рис. 1. Залежність кількості багаточленів кодів Гоппи від степеня багаточлена  $t$

Аналіз наведених залежностей демонструє, що за однакових ступенів кількість багаточленів, що не приводяться,  $N_q(t)$  перевершує кількість багаточленів з коренями  $N_k(t)$ . Крім того, зі збільшенням кількості символів скорочення ця

перевага зростає. Ще одним недоліком кодів Гоппи, побудованих на основі багаточленів із коренями, є те, що злоумисник під час криптоаналізу, вгадуючи символи скорочення, одночасно визначає і багаточлен Гоппи  $G(x)$ .

Отже, для побудови системи забезпечення цілісності даних у системах промислового Інтернету речей доцільно обирати скорочені коди Гоппи на основі багаточленів, що не приводяться.

Далі розглянемо методи скорочення кодів Гоппи, сформованих на основі багаточленів, що не приводяться. Для цього варіанта відомо три методи скорочення.

#### Конструкція скорочення № 1

Нехай код Гоппи  $Y(n, k, d)$ ,  $N = 2^k$  має перевірку матрицю  $H$ . Тоді матриця  $H$  з викресленими  $p$  стовпцями є перевіркою матрицею коду  $Y'(n', k', d)$  з параметрами:  $n' = n - p$ ,  $k' = k - p$ .

Якщо викреслюються  $p$  стовпців у перевірчій матриці  $H$ , що відповідають перевірчим елементам коду, то способом додаткових лінійних перетворень перевірча матриця приводиться до канонічного східчастого вигляду. Лінійні перетворення над рядками перевірчої матриці не змінюють параметрів коду й викреслення  $p$  перевірчих стовпців буде еквівалентно завданню  $p$  нулів інформаційним елементам. Такі символи не передаються. Отже, інформаційна й загальна довжини коду в цьому разі також зменшаться на  $p$  символів.

#### Конструкція скорочення № 2

Нехай код Гоппи  $Y(n, k, d)$ ,  $N = 2^k$  має перевірку матрицю  $H$  і деяку множину  $P$  із  $p$  позицій, яка покриває цілком  $h$  векторів із матриці  $H$  (тобто всі ненульові позиції цих  $h$  векторів належать множині  $P$ ). Тоді множина слів

із викресленими  $p$  позиціями множини  $P$  утворюють код  $Y'(n', k', d)$  з параметрами [17, 19]:

$$n' = n - p, \quad k' = k - p + h.$$

#### Конструкція скорочення Хелгерта і Стінаффа

Нехай код Гоппи  $Y(n, k, d)$  має породжувальну матрицю  $G(k \times n)$  такого вигляду:

$$G = \left( \begin{array}{c|c} 11\dots 1 & 00\dots 0 \\ \hline G_1 & G_2 \end{array} \right),$$

де кількість одиниць у першому рядку дорівнює  $d$ .

Тоді  $G_2$  є породжувальною матрицею альтернантного коду Гоппи  $Y'(n', k', d')$  з параметрами [17]

$$n' = n - d, \quad k' = k - 1, \quad d' \geq d/2.$$

Розглянемо можливість застосування описаних методів для забезпечення цілісності даних у системах промислового Інтернету речей.

Друга конструкція дає змогу отримати параметри коду, кращі порівняно з іншими методами скорочення. Однак для її застосування необхідно знати кількість одиниць у будь-якому рядку перевірчої матриці  $H$  коду Гоппи  $Y(n, k, d)$ , яка не менша, ніж мінімальна відстань  $d'$  дуального коду  $Y'$ . Для оцінювання дуальної відстані здебільшого використовується частковий комп'ютерний перебір слів дуального коду [20]. Однак для кодів Гоппи такі оцінки не відомі. Оскільки коди БЧХ є підкласом кодів Гоппи, скористаємося оцінками дуальної відстані для кодів БЧХ. У табл. 2 подано результати розрахунку верхньої оцінки дуальної відстані, отримані відповідно до [21], для кодів БЧХ зі швидкістю  $R = 1/2$ .

Таблиця 2. Верхня оцінка дуальної відстані для кодів БЧХ зі швидкістю  $R = 1/2$

|           |            |            |           |          |         |
|-----------|------------|------------|-----------|----------|---------|
| $n, k, d$ | 511,259,57 | 255,127,33 | 127,64,19 | 63,27,13 | 31,16,7 |
| $d'$      | 128        | 64         | 32        | 16       | 8       |

У табл. 3 запропоновано параметри скорочених двійкових кодів Гоппи  $n'$ ,  $k'$ , отримані за допомогою комп'ютерного моделювання для описаної конструкції скорочення. У цьому разі  $d_1$  позначає мінімальну кількість одиниць в одному рядку, а  $d_2$  – в об'єднанні двох рядків перевірчої матриці  $H$  коду Гоппи.

Аналіз досягнутих результатів демонструє, що цей метод не дає змогу скорочувати коди на незначну

кількість символів. Отже, у цьому разі неможливо гнучко змінювати параметри коду, а також здебільшого скорочення приводить до кодів, рівних за довжиною кодам меншого поля Галуа. А, як відомо, зі збільшенням розмірності коду, за однакових  $d$ , швидкість коду зменшується, тобто значне скорочення коду більшого поля приводить до коду з меншою кількістю інформаційних символів, ніж скорочення на кілька бітів коду меншого поля.



Таблиця 3. Параметри деяких скорочених кодів Гоппи

| $m$ | $n, k, d$    | $d_1$      | $d_2$     | $n', k', d$  | $n'_2, k'_2, d$ |
|-----|--------------|------------|-----------|--------------|-----------------|
| 7   | 127, 92, 11  | $\leq 32$  | $\leq 48$ | 95, 61, 11   | 76, 46, 11      |
|     | 127, 57, 21  | $\leq 22$  |           | 105, 37, 21  |                 |
|     | 127, 50, 23  | $\leq 16$  | $\leq 24$ | 111, 36, 23  | 103, 28, 29     |
|     | 127, 36, 29  | $\leq 14$  | $\leq 24$ | 113, 23, 29  | 103, 14, 29     |
| 8   | 255, 223, 9  | $\leq 96$  |           | 159, 128, 9  |                 |
|     | 255, 191, 17 | $\leq 64$  |           | 191, 128, 17 |                 |
|     | 255, 159, 25 | $\leq 48$  | $\leq 99$ | 207, 112, 25 | 156, 62, 25     |
|     | 255, 143, 29 | $\leq 48$  | $\leq 85$ | 207, 96, 29  | 170, 60, 29     |
| 9   | 511, 457, 13 | $\leq 184$ |           | 327, 274, 13 |                 |
|     | 511, 421, 21 | $\leq 172$ |           | 339, 250, 21 |                 |
|     | 511, 385, 29 | $\leq 156$ |           | 355, 230, 29 |                 |

Метод Хелгерта і Стінаффа дає змогу скорочувати код Гоппи мінімум на  $d_{\min}$  символів. У табл. 4 подано конструктивну кодову відстань деяких кодів Гоппи зі швидкістю  $R=1/2$ , розраховану за допомогою комп'ютерного моделювання.

Таблиця 4. Конструктивна кодова відстань кодів Гоппи зі швидкістю  $R=1/2$ 

| $n$ | 511 | 255 | 127 | 63 | 31 |
|-----|-----|-----|-----|----|----|
| $d$ | 61  | 37  | 21  | 13 | 7  |

Аналіз показників із табл. 4 демонструє, що метод Хелгерта і Стінаффа має ті самі недоліки, що й друга конструкція скорочення. Отже, найбільш ефективний захист має перший метод, що дає змогу скорочувати коди Гоппи на будь-яку кількість символів.

Отже, застосування процедур скорочення до кодів Гоппи сприяє отриманню кодів, що мають задані характеристики із завадостійкості та можуть бути використані для забезпечення цілісності даних у системах промислового Інтернету речей.

#### Метод забезпечення цілісності даних із застосуванням скорочених кодів Гоппи

На основі проведених досліджень запропоновано метод забезпечення цілісності даних у системах промислового Інтернету речей з використанням скорочених кодів Гоппи.

Метод оснований на застосуванні в каналах передачі даних систем промислового IoT кодових конструкцій скорочених кодів Гоппи, які гарантують одночасне забезпечення завадостійкості, тобто виявлення й виправлення випадкових помилок, контроль цілісності даних і мають

потенційну можливість реалізувати інформаційну скритність повідомлень.

Особливістю кодів Гоппи є наявність значної кількості способів формування коду із заданими параметрами, що визначаються багаточленом Гоппи  $G(x)$  ступеня  $t$  з коефіцієнтами з поля  $GF(qm)$ .

Як ключові дані запропоновано використовувати багаточлен Гоппи  $G(x)$ , а також кількість і позиції символів скорочення.

Параметрами цього методу обираємо:

- поле Галуа  $GF(qm)$ ;
- багаточлен Гоппи  $G(x)$ , що не приводиться;
- початковий код Гоппи  $Y(n, k, d)$ ;
- кількість символів скорочення  $p$  і множину позиції символів скорочення.

Запропонований метод дає змогу реалізувати комбінований підхід до забезпечення захисту від випадкових помилок і цілісності даних. Він поєднує методи завадостійкого кодування й елементи криптографічного захисту на основі скорочених кодових конструкцій. Використання скорочених кодів Гоппи в системах промислового IoT забезпечує високу ефективність захисту від випадкових завад, гарантує контроль цілісності даних і потенційну інформаційну скритність повідомлень.

Перевагою запропонованого методу є потенційна можливість підвищувати стійкість завдяки впровадженню динамічної зміни параметрів скороченого коду Гоппи.

#### Приклад системи забезпечення цілісності даних у системах промислового Інтернету речей

У сучасних системах промислового IoT найбільш поширеними є платформи Siemens

MindSphere, PTC ThingWorx, GE Digital Predix та IBM Maximo Application Suite. Для прикладу системи забезпечення цілісності було обрано платформу Siemens MindSphere, що забезпечує інтеграцію промислових сенсорів, контролерів, SCADA-систем і хмарних сервісів для реалізації моніторингу й автоматизованого управління виробничими процесами. Для передачі даних можуть застосовуватися різні види зв'язку, тому на каналному рівні використовуються різні протоколи, а саме Ethernet / PROFINET, Wi-Fi (IEEE 802.11ac/ax, IEEE 802.15.4 / WirelessHART / ZigBee, LoRaWAN, 5G NR Industrial IoT. У перелічених протоколах одним із методів забезпечення завадостійкості передачі є застосування CRC – циклічного завадостійкого коду в режимі виявлення помилок. Водночас залежно від призначення пакету й типів даних, що передаються, їх параметри дуже широко змінюються. Динамічний характер деяких пакетів у системах IoT не дає змоги використовувати коди Гоппи з механізмом виправлення помилок. Тому запропоновано застосовувати пакети від промислових датчиків, які передають фіксовані за довжиною дані за протоколом PROFINET IRT. Розмір корисних показників датчика становить  $k = 512$  бітів (64 Байти). Для забезпечення цілісності даних у цьому разі пропонується використати двійковий скорочений код Гоппи над полем  $GF(210)$ , що допоможе отримати максимальну довжину кодового слова  $n_{\max} = 210 = 1024$  біти. Якщо обрати можливість коду виправляти помилки, створені короткими імпульсними індустриальними завадами, то оберемо  $t = 3$ . Тоді кількість перевірчих символів дорівнюватиме  $r = mt = 10 \times 3 = 30$ . Оскільки в пакеті під перевірку зарезервовано 32 біти, то 2 біти залишаться як апаратний padding для вирівнювання рівно в 4 Байти ( $r = 32$ ). Тоді для  $k = 512$  бітів повна довжина пакету дорівнюватиме 544 біти.

### Висновки

У статті досліджено питання забезпечення цілісності даних у сучасних системах промислового Інтернету речей в умовах дії ненавмисних завад і впливів кіберзловмисників. З огляду на це запропоновано комбіноване рішення щодо застосування завадостійких кодів, які мають високі характеристики

завадостійкості й чимало способів формування коду із заданими параметрами, а саме кодів Гоппи.

У роботі встановлено, що використання модифікованих кодів Гоппи дасть змогу забезпечити не тільки завадостійкість і цілісність даних у системах IoT, але й потенційно має перспективи гарантувати інформаційну скритність повідомлень. Дослідження методів модифікації завадостійких кодів продемонструвало, що більшість із них або обмежують здатність гнучкої зміни параметрів коду, або призводять до зниження його характеристик завадостійкості, що унеможливило їх одночасне використання для забезпечення цілісності даних і захисту їх від перешкод у системах промислового Інтернету речей. На підставі проведеного аналізу визначено, що для комбінованого забезпечення цілісності та завадостійкості даних можна впроваджувати методи скорочення кодів. Ці методи потенційно здатні забезпечувати задану кодову відстань і формувати сімейства кодів Гоппи із різними параметрами.

Дослідження методів скорочення кодів Гоппи, побудованих на основі багаточленів із коренями й багаточленів, які не приводяться, підтвердили, що більш доцільним є використання кодів, сформованих на основі багаточленів, що не приводяться. Вони не тільки дають змогу формувати значну кількість кодових конструкцій, але й впливають на стійкість системи забезпечення цілісності даних унаслідок ускладнення задачі визначення структури коду.

Проведений аналіз конструкцій скорочення кодів Гоппи продемонстрував, що для побудови комбінованої системи забезпечення цілісності та завадостійкості даних найбільш ефективною є перша конструкція скорочення, яка забезпечує зміну параметрів коду на довільну кількість символів скорочення без погіршення його характеристик завадостійкості. Крім того, використання символів скорочення як ключових даних дає потенційну можливість у цьому разі забезпечити інформаційну скритність повідомлень у системах IoT.

Отже, досягнуті результати досліджень підтверджують, що застосування процедур скорочення до кодів Гоппи сприяє формуванню кодових конструкцій, які можуть забезпечити цілісність і завадостійкість даних у системах промислового Інтернету речей.

Для подальших досліджень становить інтерес розроблення методів формування кодових слів

скорочених кодів Гоппи, що забезпечують інформаційну скритність, а також застосування запропонованих рішень у сучасних системах ІоТ.

#### Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

#### Фінансування

Дослідження проводилося без фінансової підтримки.

#### Доступність даних

Рукопис не має пов'язаних даних.

#### Використання засобів штучного інтелекту

Автори підтверджують, що не застосовували технології ШІ для написання цієї роботи.

#### References

- Mishra, N., Islam, S. K. H., Zeadally, S. (2024), "A survey on security and cryptographic perspective of Industrial-Internet-of-Things", *Internet of Things*, 25, Article 101037. DOI: <https://doi.org/10.1016/j.iot.2023.101037>
- Wang, M., Sun, Y., Sun, H., Zhang, B. (2023), "Security Issues on Industrial Internet of Things: Overview and Challenges", *Computers*, 12(12), Article 256. DOI: <https://doi.org/10.3390/computers12120256>
- Makrakis, G. M., Roberson, D., Kolias, C., Cook, D. (2023), "WIPP: Towards IEC 62443-3-3 Network Compliance via Software-Defined Networking in Digital Substations", In: *2023 Resilience Week (RWS)*. IEEE, pp. 1–7. DOI: <https://doi.org/10.1109/RWS58133.2023.10284649>
- Yevheniev, A., Sydorenko, Z., Sievierinov, O. (2025), "Ensuring data integrity in industrial Internet of Things systems using error-correcting codes", *Radiotekhnika*, (221), pp. 46–50. DOI: <https://doi.org/10.30837/rt.2025.2.221.06>
- Singh, H. (2019), "Code based cryptography: Classic McEliece"? *arXiv preprint arXiv:1907.12754*. DOI: <https://doi.org/10.48550/arXiv.1907.12754>
- García-Morchón, O., Marojevic, V., Kumar, S. (2024), "Exploring Post-Quantum Cryptography: Review and Directions for the Transition Process", *Technologies*, 12(12), Article 241. DOI: <https://doi.org/10.3390/technologies12120241>
- Halim, S.K., Sugeng, K.A. (2024), "Application of Goppa Code in Niederreiter Cryptosystem", *AIP Conference Proceedings*, 3163(1), Article 030002. DOI: <https://doi.org/10.1063/5.0213630>
- Yevseiev, S., Kots, H., Minukhin, S., Korol, O., Kholodkova, A. (2017), "The development of the method of multifactor authentication based on hybrid crypto-code constructions on defective codes", *Eastern-European Journal of Enterprise Technologies*, Vol. 5(9 (89)), pp. 19–35. DOI: <https://doi.org/10.15587/1729-4061.2017.109879>
- Baldi, M., Santini, P., Cancellieri, G. (2017), "Post-quantum cryptography based on codes: State of the art and open challenges", in *2017 AEIT International Annual Conference*. Piscataway, NJ: IEEE, pp. 1–6. DOI: <https://doi.org/10.23919/AEIT.2017.8240549>
- Kuznetsov, A., Pushkar'ov, A., Kiyan, N., Kuznetsova, T. (2018), "Code-based electronic digital signature", in *2018 IEEE 9th International Conference on Dependable Systems, Services and Technologies (DESSERT)*. Kyiv, Ukraine: IEEE, pp. 331–336. DOI: <https://doi.org/10.1109/DESSERT.2018.8409154>
- Gorbenko, Y., Kiian, A., Pushkar'ov, A., Korneiko, O., Smirnov, S., Kuznetsova, T. (2019), "Code-based hybrid cryptosystem: Comparative studies and analysis of efficiency", *International Journal of Computing*, 18(4), pp. 372–380. DOI: <https://doi.org/10.47839/ijc.18.4.1608>
- Melenti, Y., Korol, O., Shulha, V., Milevskiy, S., Sievierinov, O., Voitko, O., Rzayev, K., Husarova, I., Kravchenko, S., Pashayeva, S. (2025), "Development of post-quantum cryptosystems based on the Rao-Nam scheme", *Eastern-European Journal of Enterprise Technologies*, Vol. 19(133), pp. 35–48. DOI: <https://doi.org/10.15587/1729-4061.2025.323195>
- Balamurugan, C., Singh, K., Ganesan, G., Rajarajan, M. (2021), "Post-quantum and code-based cryptography-some prospective research directions", *Cryptography*, Vol. 5(4), pp. 38. DOI: <https://doi.org/10.3390/cryptography5040038>
- Admaty, N., Litsyn, S., Keren, O. (2012), "Puncturing, expurgating and expanding the q-ary BCH based robust codes", In: *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel (IEEEI 2012)*. IEEE. DOI: <https://doi.org/10.1109/IEEEI.2012.6376995>

15. Chen, X., Cheng, K., Li, X., Mao, S. (2023), "Random Shortening of Linear Codes and Applications". In: *Computing and Combinatorics: 29th International Conference, COCOON 2023, Tianjin, China, August 4–6, 2023, Proceedings*. Lecture Notes in Computer Science, vol. 14455. Cham: Springer, pp. 184–197. DOI: [https://doi.org/10.1007/978-3-031-49193-1\\_14](https://doi.org/10.1007/978-3-031-49193-1_14)
16. Balamurugan, C., Singh, K., Ganesan, G., Rajarajan, M. (2021), "Post-Quantum and Code-Based Cryptography - Some Prospective Research Directions", *Cryptography*, Vol. 5(4), Article 38. DOI: <https://doi.org/10.3390/cryptography5040038>
17. Fedorchenko, V., Yeroshenko, O., Shmatko, O., Kolomiitsev, O., Omarov, M. (2024), "Password hashing methods and algorithms on the .Net platform", *Advanced Information Systems*, Vol. 8, No. 4, pp. 82–92. DOI: <https://doi.org/10.20998/2522-9052.2024.4.11>
18. Bezzateev, S. V., Shekhunova, N. A. (1995), "Subclass of binary Goppa codes with minimal distance equal to the design distance", *IEEE Transactions on Information Theory*, Vol. 41(2), pp. 554–555. DOI: <https://doi.org/10.1109/18.370170>
19. Zinoviev V. A., Litsyn S. N. (1984), "On Shortening of Codes". *Problems of Information Transmission*, Vol. 20, No. 1, pp. 1–7.
20. Krasikov, I., Litsyn, S. (2001), "On the Distance Distributions of BCH Codes and Their Duals". *Designs, Codes and Cryptography*, Vol. 23(2), pp. 223–232. DOI: <https://doi.org/10.1023/A:1011220817609>
21. Ding, C., Li, C. (2024), "BCH cyclic codes", *Discrete Mathematics*, Vol. 347(6), Article 113918. DOI: <https://doi.org/10.1016/j.disc.2024.113918>

Received (Надійшла) 06.04.2026

Accepted for publication (Прийнята до друку) 17.05.2026

Publication date (Дата публікації) 29.05.2026

#### *Відомості про авторів / About the Authors*

**Сидоренко Зоя Михайлівна** – Харківський національний університет радіоелектроніки, аспірантка кафедри безпеки інформаційних технологій; м. Харків, Україна;

**Zoia Sydorenko** – Kharkiv National University of Radio Electronics, Postgraduate Student of the Department of Information Technology Security; Kharkiv, Ukraine;

e-mail: [zoia.sydorenko@nure.ua](mailto:zoia.sydorenko@nure.ua)

ORCID ID: <https://orcid.org/0000-0002-0104-6807>

**Северінов Олександр Васильович** – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, професор кафедри безпеки інформаційних технологій; м. Харків, Україна;

**Oleksandr Sievierinov** – PhD (Technical Sciences), Associate Professor, Kharkiv National University of Radio Electronics, Professor of the Department of Information Technology Security, Kharkiv, Ukraine;

e-mail: [oleksandr.sievierinov@nure.ua](mailto:oleksandr.sievierinov@nure.ua)

ORCID ID: <http://orcid.org/0000-0002-6327-6405>

## METHOD FOR ENSURING DATA INTEGRITY USING A MODIFIED ERROR-CORRECTING CODE IN INDUSTRIAL INTERNET OF THINGS SYSTEMS

The subject of the study is methods for modifying error-correcting codes to ensure data integrity in Industrial Internet of Things (IIoT) systems. The purpose of the research is to develop a method for ensuring data integrity in Industrial Internet of Things systems through the application of modified error-correcting codes, namely shortened Goppa codes, for the detection and correction of errors arising during information transmission and processing. Objectives: to analyze methods for modifying error-correcting codes, particularly Goppa codes, in order to introduce message obfuscation for ensuring data integrity in Industrial Internet of Things systems; to investigate methods for shortening Goppa codes and determine the most appropriate constructions for providing error resistance and data integrity; to develop a method for ensuring data integrity using a shortened error-correcting code in Industrial Internet of Things systems. To accomplish the outlined objectives, the following methods were applied: theoretical methods of analysis and mathematical modeling, empirical methods of comparative analysis

of code characteristics, as well as elements of computer experimentation for calculating and evaluating the parameters of shortened Goppa codes. Results achieved. Methods for modifying Goppa codes were analyzed. For constructing a data integrity assurance system in IIoT, a code-shortening method was proposed that preserves the minimum code distance and enables the formation of codes with specified parameters. Constructions for shortening Goppa codes were investigated, and the feasibility of using shortened Goppa codes based on irreducible polynomials was substantiated. A method for ensuring data integrity using shortened Goppa codes was proposed. Conclusions. The implementation of shortening procedures for Goppa codes constructed on the basis of irreducible polynomials makes it possible to form code constructions for ensuring data integrity in Industrial Internet of Things systems while simultaneously maintaining high error-resistance properties. A method for ensuring data integrity using shortened Goppa codes has been proposed, providing flexible adjustment of code parameters and offering prospects for guaranteeing the information concealment of messages.

**Keywords:** Industrial Internet of Things; error resistance; integrity; error-correcting code; Goppa code; code modification; shortened code.

#### *Бібліографічні описи / Bibliographic descriptions*

Сидоренко З. М., Сєверінов О. В. Метод забезпечення цілісності даних із використанням модифікованого завадостійкого коду в системах промислового інтернету речей. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 235–246. DOI: <https://doi.org/10.30837/0135-1710.2026.189.235>

Sydorenko, Z., Sievierinov, O. (2026), "Method for ensuring data integrity using a modified error-correcting code in industrial internet of things systems", *Management Information System and Devices*, No. 2 (189), P. 235–246. DOI: <https://doi.org/10.30837/0135-1710.2026.189.235>

---



Ситнік О. О.

## ПІДХІД ДО ДИСПЕТЧЕРИЗАЦІЇ БІЗНЕС-ПРОЦЕСІВ В УМОВАХ ФУНКЦІОНУВАННЯ ВИРОБНИЧИХ ОБ'ЄКТІВ НА ОСНОВІ АНТРОПОЦЕНТРИЧНОСТІ

**Предметом дослідження** є підхід до антропоцентричної диспетчеризації бізнес-процесів в умовах функціонування виробничих об'єктів, який ґрунтується на узгодженні розподілу завдань між виконавцями з огляду на комбінації їхніх хронобіологічних, когнітивних та психологічних якостей. **Мета роботи** – підвищення ефективності бізнес-процесів функціонування виробничих об'єктів (ВО) за допомогою розроблення, формального описання й апробування антропоцентричного методу диспетчеризації (АМД) як загального підходу до класу задач планування у ВО з людьми-виконавцями, а також демонстрація його застосовності на прикладі діяльності реального виробничого підприємства авіаційної галузі й автоматизованих систем управління (АСУ) навчальним процесом у закладах вищої освіти. **Завдання дослідження:** формалізація задачі антропоцентричної диспетчеризації як такої, що є інваріантною щодо типу виробничого об'єкта; розроблення хронотип-зберігального оператора схрещування CPC; реалізація алгоритму антропоцентричного оператора мутації AWM. Крім цього, у роботі розглянуто механізм адаптивного коригування ваг SRA, проаналізовано й порівняно АМД з базовими конфігураціями ГА. Застосовано такі **методичні засоби дослідження:** еволюційний алгоритм з антропоцентрично орієнтованими генетичними операторами, OLS-регресія задоволеності всіх учасників навчального процесу, абляційний аналіз і порівняльне тестування на 30 незалежних запусках прототипу програмного забезпечення. **Практичні результати** довели, що АМД скорочує час досягнення  $F(x) > 0.75$  на 34.0% (38.4 проти 58.2 хв), підвищує фінальне значення цільової функції до 0.84 (зі збільшенням на 18.3% щодо *Baseline* GA) і знижує дисперсію результатів до  $\sigma = 0.029$ . Норма  $\|w - w^*\|_2$  скорочується на 92.8% за три семестри. **Висновки.** У роботі вперше формально описано інваріанти збереження циркадної узгодженості та когнітивної якості безпосередньо в структурі генетичних операторів, а не через штрафні доданки цільової функції. Крім того, доведено загальність підходу для широкого класу виробничих задач диспетчеризації з людиною-виконавцем.

**Ключові слова:** антропоцентрична диспетчеризація; бізнес-процеси виробничого об'єкта; еволюційний алгоритм; генетичні оператори; хронотип-зберігальний кросовер; адаптивне навчання; регресія задоволеності; автоматизовані системи управління навчальним процесом.

### Вступ

Ефективність виробничих процесів на сучасних підприємствах і в установах здебільшого залежить не лише від технологічного забезпечення, а й від раціонального розподілу завдань між людьми-виконавцями. Варто зауважити, що якість когнітивно навантажених операцій, зокрема проектування, конструювання, технічна підготовка виробництва, навчання фахівців, підпорядкована біологічним ритмам людини: хронотипу, добовим коливанням когнітивної продуктивності, психологічному навантаженню [1–3]. Ігнорування цих факторів в умовах диспетчеризації бізнес-процесів призводить до зниження якості результатів, надмірної втоми виконавців і зростання кількості помилок [3, 4].

Диспетчеризація бізнес-процесів (*Business Process Dispatching*) у контексті автоматизованих систем управління (АСУ) виробничими об'єктами

традиційно розглядається як оптимізаційна задача розподілу завдань у часі з огляду на ресурсні обмеження: наявності виконавців, обладнання, аудиторій чи виробничих ліній. Проте сучасний рівень розвитку людини орієнтованих виробничих систем вимагає пильної уваги до нового класу "антропоцентричних" обмежень, пов'язаних із біологічними, психологічними й когнітивними якостями людини-виконавця [5–7].

Антропоцентрична диспетчеризація – це підхід до планування бізнес-процесів, за умови якого розподіл завдань між виконавцями здійснюється з огляду на всі обмеження: технічні (відсутність конфліктів ресурсів), хронобіологічні (відповідність часових слотів хронотипам виконавців), психологічні (рівномірність навантаження, запобігання перевантаженню) та когнітивні (сумісність суміжних задач за типом розумової діяльності). Перелічені чотири категорії обмежень інтегруються в єдину

багатокритеріальну цільову функцію, що оптимізується за допомогою еволюційного алгоритму зі спеціалізованими операторами.

Задача антропоцентричної диспетчеризації є інваріантною щодо типу виробничого об'єкта: вона виникає в будь-якій системі, де є виконавці з різними хронотипами й когнітивними профілями, яким призначаються завдання різної когнітивної складності в часових слотах. Підтвердженням цьому є те, що задачі диспетчеризації в промисловому виробництві [8, 9], де команда розробників отримує завдання різної інженерно-конструкторської складності, та в академічній сфері [10, 11], де студенти й викладачі мають навчальні заняття різного навантаження, описуються ідентичним формальним апаратом.

Метою роботи є підвищення ефективності бізнес-процесів функціонування виробничих об'єктів (ВО) за допомогою розроблення, формального опису й апробації антропоцентричного методу диспетчеризації (АМД) як загального підходу до класу задач планування у ВО з людьми-виконавцями, а також демонстрація його застосовності на прикладі діяльності реального виробничого підприємства авіаційної галузі та АСУ навчальним процесом у закладах вищої освіти.

### Огляд суміжних підходів

У публікації [10] описано результати застосування *Constraint-preserving* операторів GA для задач синтезу розкладу. *Burke* та *Petrovic* виокремлюють клас операторів, що зберігають технічні жорсткі обмеження після схрещування, але водночас не розглядають антропоцентричних якостей виконавців. *Haitan* і *Nazarov* [11] застосовують гібридний підхід на основі мурашиного алгоритму та GA для автоматизованого складання розкладу у ЗВО, однак без огляду на хронотипи. *Mokhtari et al.* [8] розробляють модель змішаного цілочисельного програмування для задачі *university course timetabling*, зважаючи на преференції учасників. *Oude Vrielink et al.* [9] систематизують практики складання розкладу у ЗВО та ідентифікують розрив між дослідницькими підходами й практикою. Загальні методи *constraint-preserving* планування розглядаються також у контексті маршрутизації та розкладу [12, 13] і в метагевристичних підходах [10, 14]. Отже, жоден із цих підходів не передбачає впровадження хронотипових інваріантів безпосередньо в структуру генетичних операторів.

Зважена мутація із спрямованим вибором точки вже багаторазово застосовувалася в задачах планування з нерівномірним розподілом якості результатів [15]. *Eiben* і *Smith* у своїх роботах описують *temperature-based softmax*-вибір у мутаційних операторах [16]. Адаптацію параметрів алгоритму на основі зворотного зв'язку розглянуто в контексті інтерактивних еволюційних систем [12]. Вплив хронотипу виконавців на результативність виробничого процесу підтверджено хронобіологічними дослідженнями [17, 18]. Аналіз публікацій дав змогу виявити відсутність досліджень з адаптацією ваг компонентів цільової функції диспетчеризації за допомогою регресії задоволеності учасників виробництва або навчання.

## 1. Загальний опис антропоцентричного методу диспетчеризації

Для математичного опису антропоцентричного методу диспетчеризації необхідно впровадити позначки для основних параметрів. Нехай задано множину виконавців задач  $E = \{e_1, e_2, \dots, e_n\}$ , що мають хронотипові профілі *chronotypes* ( $e_i$ ). Позначимо набір задач  $T = \{t_1, t_2, \dots, t_m\}$  з когнітивними профілями для кожної з них. Також задамо скінченну множину часових слотів  $S = \{(d, s) \mid d \in \{1..D\}, s \in \{1..H\}\}$ , де  $d$  – день із навантаженням і  $s$  – порядковий номер слота в межах дня. Для цих параметрів знайдемо розклад  $x: T \rightarrow E \times S$ , що максимізує багатокритеріальну цільову функцію:

$$F(x) = w_1 \cdot F_{tech}(x) + w_2 \cdot F_{circ}(x) + w_3 \cdot F_{psych}(x) + w_4 \cdot F_{cogn}(x),$$

де  $F_{tech}(x)$  – компонент технічних обмежень і відсутності конфліктів ресурсів. У поданій формулі  $F_{circ}(x)$  – це компонент циркадної узгодженості та відповідності часових слотів хронотипам виконавців;  $F_{psych}(x)$  – компонент психологічного комфорту й рівномірності навантаження кожного учасника виробничого чи навчального процесу;  $F_{cogn}(x)$  – компонент когнітивної ефективності та сумісності суміжних задач. Позначки  $w_1 \dots w_4$  є ваговими коефіцієнтами, які адаптуються механізмом *Satisfaction Regression Adaptation* (SRA).

Отже, метод АМД реалізує оптимізацію зазначеної функції за допомогою трьох спеціалізованих алгоритмічних компонентів: хронотип-зберігального оператора схрещування СРС, антропоцентричного оператора мутації АWM і механізму адаптивного навчання SRA. Детальний математичний опис кожного з цих компонентів наведено в розд. 3. Основною властивістю методу є антропоцентричні обмеження, вбудовані безпосередньо в структуру генетичних операторів, а не виражені штрафними доданками цільової функції, що запобігає їх систематичному порушенню в процесі рекомбінації [10, 11].

## 2. Застосування АМД у виробничій системі на прикладі приватної фірми "Спейс"

"Спейс" є приватною фірмою (ПФ) авіаційного профілю, що спеціалізується на виробництві й ремонті аеродромної техніки, зокрема самохідних пасажирських трапів (СПТ-154 та СПТ-114). Загальна кількість персоналу налічує 52 співробітники, з яких у виробництві залучені конструктори, технологи й виробничі фахівці [19, 20]. У межах проєкту зі створення нової моделі СПТ (проєкт розроблення комплексу причалювального обладнання, КПО) формується команда виконавців

із фахівців підрозділів НДДКР (науково-дослідницька й дослідно-конструкторська робота) і ТПВ (технологічна підготовка виробництва).

### 2.1. Формалізація задачі диспетчеризації для ПФ "Спейс"

Розглянемо задачу антропоцентричної диспетчеризації виробничих завдань у межах проєкту розроблення КПО. Формально нехай  $E = \{e_1, \dots, e_8\}$  – команда виконавців із хронотиповими профілями з табл. 1;  $T = \{t_1, \dots, t_{12}\}$  – множина виробничих завдань;  $T_A \subset T$  – завдання типу А (висока концентрація);  $T_B = T \setminus T_A$  – завдання типу В;  $D = 5$  робочих днів,  $H = 4$  часові слоти по 2 год.

Обмеження: (I) кожне  $t_j$  призначається рівно одному  $e_i$  у рівно один слот ( $s, d$ ); (II) кожен  $e_i$  не перевищує одне завдання у слоті; (III)  $t_j \in T_A$  призначається у слот ( $d, s$ ) з  $\max F_{circ}(e_i, s)$  для відповідного виконавця  $e_i$ . Нехай команда виконавців налічує  $n=8$  фахівців (табл. 1), яким необхідно призначити  $m=12$  виробничих завдань у робочих слотах тижневого горизонту планування. Хронотипи виконавців визначені на основі стандартизованих анкет MEQ [18] та відображені у табл. 1.

Таблиця 1. Склад команди виконавців проєкту КПО ПФ "Спейс" та їх хронотипові профілі

| № | Роль                              | Хронотип (MEQ) | F_circ-пік             |
|---|-----------------------------------|----------------|------------------------|
| 1 | Провідний конструктор (КПО)       | Ранковий (67)  | Слот 1–2 (8:00–12:00)  |
| 2 | Конструктор I кат. (шасі)         | Проміжний (54) | Слот 2–3 (10:00–14:00) |
| 3 | Конструктор I кат. (привод)       | Вечірній (38)  | Слот 3–4 (14:00–18:00) |
| 4 | Конструктор II кат. (гідрравліка) | Проміжний (50) | Слот 2–3 (10:00–14:00) |
| 5 | Провідний технолог ТПВ            | Ранковий (71)  | Слот 1–2 (8:00–12:00)  |
| 6 | Технолог ТПВ (виготовлення)       | Проміжний (48) | Слот 2–3 (10:00–14:00) |
| 7 | Інженер-випробувальник            | Вечірній (35)  | Слот 4 (16:00–18:00)   |
| 8 | Технік-нормувальник               | Ранковий (62)  | Слот 1 (8:00–10:00)    |

Виробничі завдання проєкту КПО розрізняють за когнітивним профілем. Завдання типу А, що передбачають високу концентрацію, як-от ескізне проєктування й розроблення конструкторської документації, мають пріоритет у слотах з максимальним  $F_{circ}$ -значенням виконавця. Задачі типу Б передбачають посередній рівень концентрації. Це, наприклад, розроблення специфікацій і нормування, які є гнучкими щодо часового слота й складності навантаження. Натомість завдання типу В рутинні, як-от внесення змін до

документації та звіти. Їх доцільно призначати в слоти з мінімальним значенням  $F_{circ}$  виконавця. За такого розподілу зберігаються пікові слоти для задач типу А, що є найбільш раціональним рішенням.

### 2.2. Застосування операторів АМД у процесі диспетчеризації виробничих завдань

За умови диспетчеризації хромосома  $x$  кодує розклад виробничих завдань:  $x[d][s][r] = (t, e)$ . У наведеній формулі  $d$  – це день тижня;  $s$  – часовий

слот;  $r$  – виробничий ресурс, наприклад робоче місце або верстат;  $t$  – ідентифікатор завдання;  $e$  – виконавець. Ініціалізація популяції методом рандомізованої жадібної побудови розміщує завдання типу  $A$  в слоти з найвищим  $F_{circ}$ -значенням для відповідного виконавця.

Оператор CPC у процесі схрещування двох розкладів самостійно обирає точку розрізу, пропорційну різниці циркадних внесків між батьківськими особинами за днями. Тобто дні, де батьківські розклади суттєво різняться за  $F_{circ}$ , є більш інформативними точками рекомбінації. Це дає змогу нащадку успадковувати денні блоки з кращою хронотиповою структурою. Оператор AWM спрямовує мутаційний бюджет на часові слоти з найнижчим антропоцентричним внеском  $q(d,s) = w_2 \cdot a(d,s) + w_3 \cdot (1 - l(d,s)) + w_4 \cdot c(d,s)$ ,

де  $a(d,s)$  – циркадний внесок слота. У формулі позначка  $l(d,s)$  є рівнем навантаження, а  $c(d,s)$  – когнітивною сумісністю сусідніх завдань [2, 3].

### 2.3. Результати аналізу сценарного прикладу виробничої діяльності ПФ "Спейс"

Для ілюстрації роботи АМД у межах дослідження виконано модельний розрахунок на основі інформації про персонал ПФ "Спейс" [19]. Весь персонал підприємства налічує вісім виконавців. Також є 12 виробничих завдань, 5-денний горизонт планування, чотири часові слоти. Хронотипові профілі визначені анкетним методом [18]. Порівняно два варіанти розкладу: стандартний (рівномірний розподіл завдань без огляду на хронотипи) й оптимізований АМД (табл. 2).

Таблиця 2. Порівняння варіантів диспетчеризації для команди проєкту КПО ПФ "Спейс"

| Критерій                             | Стандартний розклад | АМД-оптимізований |
|--------------------------------------|---------------------|-------------------|
| $F_{circ}$ (циркадна узгодженість)   | 0.51                | 0.79              |
| $F_{psych}$ (психологічний комфорт)  | 0.58                | 0.76              |
| $F_{cogn}$ (когнітивна ефективність) | 0.60                | 0.81              |
| $F(x)$ загальне                      | 0.57                | 0.79              |
| Завдань типу А в пікових слотах      | 42%                 | 83%               |
| Сумішень різнотипних завдань на день | 6                   | 1                 |

АМД-оптимізований розклад забезпечив призначення 83% завдань типу А, як-от ескізне проєктування й конструкторська документація, у часові слоти, що відповідають піку когнітивної активності конкретного виконавця. Для провідного конструктора й провідного технолога, які мають ранкові хронотипи, критичні завдання зосереджені в першій половині дня; для інженера-випробувача з вечірнім хронотипом – у другій. Кількість психологічно несумісних сумішень різнотипних завдань у межах одного дня скорочено з 6 до 1. Такий приклад демонструє прикладну ефективність АМД для реальних виробничих команд і практичний ефект від застосування цього методу у фірмі.

### 3. Спільна природа задачі диспетчерського управління для виробничих об'єктів і ЗВО

Порівняння задачі диспетчеризації для ПФ "Спейс" і для АСУ навчальним процесом у ЗВО [20] виявляє структурну ідентичність за різного предметного

контексту. В обох випадках: (1) є виконавці з різними хронотипами й когнітивними профілями; (2) є завдання з різною когнітивною складністю, що вимагають розміщення в часових слотах; (3) жорсткі технічні обмеження (виконавець не може бути у двох місцях одночасно) поєднуються з м'якими антропоцентричними; (4) якість розкладу безпосередньо впливає на результативність виконавця.

Розбіжності між двома застосуваннями є непринциповими для формальної постановки: в обох випадках хромосома  $x$  кодує відображення завдань на пари виконавця й часового слота, цільова функція  $F(x)$  має ідентичну структуру. Оператори CPC, AWM і механізм SRA не залежать від типу виробничого об'єкта. Це підтверджує загальність запропонованого підходу та його застосовність до широкого класу задач диспетчеризації бізнес-процесів виробничих систем з командою виконавців. Відповідність між елементами задачі диспетчеризації та їх виконавцями для випадків виробництва і навчального процесу подано у табл. 3.

Таблиця 3. Відображення елементів задачі диспетчеризації між виробничою системою та ЗВО

| Елемент задачі        | ПФ "Спейс"                                    | АСУ ЗВО                                            |
|-----------------------|-----------------------------------------------|----------------------------------------------------|
| Виконавці $E$         | Інженери й конструктори                       | Викладачі                                          |
| Реципієнти            | Виробничі ресурси або підрозділи              | Академічні групи студентів                         |
| Завдання $T$          | Виробничі або проєктні операції               | Навчальні заняття                                  |
| Часові слоти $S$      | Робочі зміни або годинні блоки                | Навчальні пари                                     |
| $F_{circ}$            | Відповідність зміни хронотипу інженера        | Відповідність пари хронотипу студентів і викладача |
| $F_{psych}$           | Рівномірність денного навантаження            | Рівномірність навчального навантаження             |
| $F_{cogn}$            | Сумісність суміжних задач за типом діяльності | Сумісність суміжних дисциплін за складністю        |
| SRA-зворотний зв'язок | Оцінки задоволеності виконавців               | Оцінки задоволеності студентів і викладачів        |

#### 4. Формальний опис методу АМД для АСУ навчальним процесом у ЗВО

##### 4.1. Кодування хромосоми та ініціалізація

Основна різниця між методом АМД і класичним генетичним алгоритмом полягає в перенесенні антропоцентричних обмежень з рівня штрафних доданків цільової функції на рівень структури самих операторів. І в разі такого підходу CPC-оператор гарантує збереження хронотипової узгодженості в процесі схрещування, а АWM-оператор зосереджує мутаційний бюджет на проблемних слотах. Цільова функція для задачі розкладу ЗВО має такий вигляд:  $F(x) = w_1 \cdot F_{tech}(x) + w_2 \cdot F_{circ}(x) + w_3 \cdot F_{psych}(x) + w_4 \cdot F_{cogn}(x)$  (компоненти подано в розд. 1). Детальний опис математичних моделей компонентів наведено в роботі [5].

Тобто хромосома  $x$  кодує розклад на один тиждень у вигляді тривимірного масиву:  $x[d][s][g] = (e, r)$ ,  $d \in \{1..6\}$ ,  $s \in \{1..8\}$ ,  $g \in \{1..|G|\}$ , де  $d$  – день тижня;  $s$  – порядковий номер пари;  $g$  – індекс групи;  $e$  – ідентифікатор викладача;  $r$  – ідентифікатор аудиторії. Кожна особина доповнюється вектором кешованих метрик  $m(x)$ , який оновлюється внаслідок обчислення фітнесу. Хронотипові профілі учасників визначаються на основі стандартизованих методик [18], які широко застосовуються в хронобіологічних дослідженнях [17]. Найкращим прикладом таких методик є опитувальник MEQ ( $a$ ). Алгоритм ініціалізації популяції за цими даними методом рандомізованої жадібної побудови у вигляді псевдокоду представлено на рис. 1.

```

ALGORITHM Init-Population(|P|, plan, chronotypes)
INPUT:  |P| – розмір популяції
        plan – доручення {(g, e, d_id, hours)}
        chronotypes – хронотип кожного учасника
OUTPUT: P – початкова популяція
P <- empty
FOR i <- 1 TO |P| DO
  x <- empty_schedule()
  assignments <- shuffle(plan)
  FOR EACH (g, e, d_id, h) IN assignments DO
    slots <- free_slots(x, g, e)
    slots <- sort_by_circadian(slots, chronotypes[g], chronotypes[e])
    slot <- slots[random_top_k(slots, k=3)] // e-greedy
    x[slot.d][slot.s][g] <- (e, assign_room(slot, d_id))
  END FOR
  m(x) <- compute_metrics(x)
  P <- P + {x}
END FOR
RETURN P
  
```

Рис. 1. Псевдокод ініціалізації популяції методом рандомізованої жадібної побудови

##### 4.2. Хронотип-зберігальний оператор схрещування CPC

У процесі роботи алгоритму оператор CPC (*Chronotype-Preserving Crossover*) обирає точки

розрізу на рівні днів пропорційно абсолютній різниці циркадного внеску між батьківськими особинами:  $P(\text{cut at day } d) \propto |F_{circ}(x_A, d) - F_{circ}(x_B, d)|$ . У цьому разі дні із значною різницею є більш інформативними



точками розрізу, тобто нащадок успадковує кращий з двох варіантів для кожного дня. Важливість збереження хронотипової узгодженості

підтверджена в дослідженнях [1, 3]. Алгоритм оператора CPC у вигляді псевдокоду відображено на рис. 2.

```
ALGORITHM CPC(x_A, x_B, w, chronotypes)
INPUT:  x_A, x_B – батьківські хромосоми
        w = (w1, w2, w3, w4) – поточні вагові коефіцієнти
        chronotypes – хронотипові профілі
OUTPUT: x_C, x_D – дочірні хромосоми
// Крок 1: обчислити різницю циркадного внеску за днями
FOR d <- 1 TO 6 DO
    D[d] <- |F_circ(x_A, d) - F_circ(x_B, d)|
END FOR
p_cut <- softmax(D)
// Крок 2: обрати точку розрізу
d_cut <- sample(p_cut)
// Крок 3: побудувати нащадків
x_C <- x_A[1..d_cut] || x_B[d_cut+1..6]
x_D <- x_B[1..d_cut] || x_A[d_cut+1..6]
// Крок 4: відновлення технічних обмежень
x_C <- repair(x_C)
x_D <- repair(x_D)
m(x_C) <- compute_metrics(x_C)
m(x_D) <- compute_metrics(x_D)
RETURN x_C, x_D
```

Рис. 2. Псевдокод алгоритму оператора CPC

#### 4.3. Антропоцентричний оператор мутації AWM

У виконанні алгоритму, поданого вище, оператор AWM (*Anthropocentric Weighted Mutation*) спрямовує мутацію до слотів із найнижчим антропоцентричним внеском, використовуючи *softmax*-зважений розподіл.  $P(\text{select}(d, s)) \propto \exp(-\beta \cdot q(d, s))$ , де

$a(d, s)$  – нормований внесок слота у  $F_{\text{circ}}$ ;  $l(d, s)$  – психологічне навантаження;  $c(d, s)$  – когнітивна сумісність із сусідніми часовими слотами [2, 3]. За умови  $\beta \rightarrow 0$  оператор AWM вироджується в рівномірну мутацію. Алгоритм антропоцентричного оператора мутації AWM у вигляді псевдокоду наведено на рис. 3.

```
ALGORITHM AWM(x, w, beta=2.0)
INPUT:  x хромосома
        w поточні вагові коефіцієнти
        beta температурний параметр
OUTPUT: x' мутована хромосома
// Крок 1: антропоцентрична властивість кожного слота
active <- {(d, s, g) : x[d][s][g].e != 0}
FOR EACH (d, s, g) IN active DO
    a <- F_circ_slot(x, d, s, g)
    l <- L_slot(x, d, s, g)
    c <- cogn_compat(x, d, s, g)
    q[d, s, g] <- w2*a + w3*(1-l) + w4*c
END FOR
// Крок 2: зважений вибір двох слотів
logits <- -beta * q[.]
p_select <- softmax(logits)
slot_1 <- sample(p_select)
slot_2 <- sample(p_select \ {slot_1})
// Крок 3: перевірка та обмін
IF compatible_swap(x, slot_1, slot_2) THEN
    x' <- swap(x, slot_1, slot_2)
    x' <- repair(x')
    m(x') <- compute_metrics(x')
ELSE
    x' <- x
END IF
RETURN x'
```

Рис. 3. Псевдокод алгоритму антропоцентричного оператора мутації AWM

#### 4.4. Механізм адаптивного навчання SRA

Після виконання оператора AWM механізм SRA (*Satisfaction Regression Adaptation*) коригує вагові коефіцієнти після завершення кожного семестру на основі агрегованих оцінок задоволеності студентів і викладачів. Адаптивне коригування параметрів через зворотний зв'язок розглядалося в контексті застосування нечіткої логіки в освітніх системах дистанційного навчання [6, 7]. Оцінки збираються

за шкалами  $Q_{tech}, Q_{circ}, Q_{psych}, Q_{cogn} \in [5, 15]$ . OLS-регресія

$\hat{\beta} = (X^T X)^{-1} X^T Q$  виявляє вагу кожного компонента в задоволеності учасників навчального процесу, після чого вагові коефіцієнти оновлюються з ЕМА-згладжуванням ( $\alpha=0.4$ ) та обмеженням  $w_j \geq 0.05$  [5]. Алгоритм механізму адаптивного навчання SRA наведено на рис. 4.

```

ALGORITHM SRA(w_old, F_components, Q_survey, alpha=0.4)
INPUT:  w_old - поточні ваги (w1..w4)
        F_components - вектори (F_tech, F_circ, F_psych, F_cogn) для N учасників
        Q_survey - матриця оцінок [N x 4]
        alpha - EMA learning rate
OUTPUT: w_new - оновлені ваги
FOR i <- 1 TO N DO
    Q_bar[i] <- mean(Q_survey[i, .])
END FOR
X <- [1 | F_components]           // матриця ознак [N x 5]
b_hat <- OLS(X, Q_bar)             // b_hat = (X'X)^{-1} X' Q_bar
b1..b4 <- b_hat[1..4]
b_pos <- max(b_hat_j, 0) for j=1..4
IF sum(b_pos) > 0 THEN
    w_hat <- b_pos / sum(b_pos)
ELSE
    w_hat <- w_old
END IF
FOR j <- 1 TO 4 DO
    w_new[j] <- alpha * w_hat[j] + (1-alpha) * w_old[j]
    w_new[j] <- max(w_new[j], 0.05)
END FOR
w_new <- w_new / sum(w_new)
RETURN w_new
  
```

Рис. 4. Псевдокод алгоритму механізму адаптивного навчання SRA

#### 4.5. Повна схема алгоритму АМД

Метод АМД об'єднує всі три компоненти. Зовнішній цикл (між семестрами) оновлює вагові коефіцієнти через SRA; внутрішній цикл (всередині семестру) виконує еволюційну оптимізацію з операторами CPC і AWM. Повна схема антропоцентричного методу диспетчеризації показана на рис. 5.

Параметри алгоритму:  $|P| = 200$ ,  $p_{cross} = 0.85$ ,  $p_{mut} = 0.15$ ,  $k_t = 5$ ,  $\beta = 2.0$ , максимум 500 поколінь, критерій стагнації: 50 поколінь з  $\Delta F < 0.001$ ,  $\alpha = 0.4$  (SRA learning rate), мінімальна вага  $w_j \geq 0.05$ . Вибір цих значень обґрунтовано в підрозд. 5.3.

```

ALGORITHM AMD(plan, chronotypes, K_semesters)
INPUT:  plan навчальний план і доручення
        chronotypes хронотипові профілі учасників
        K_semesters кількість семестрів
OUTPUT: x* оптимальний розклад, w* адаптовані ваги
w <- (0.25, 0.25, 0.25, 0.25)
P <- Init-Population(|P|=200, plan, chronotypes)
FOR k <- 1 TO K_semesters DO
    gen <- 0; best_F <- 0; stagnation <- 0
    WHILE gen < 500 AND stagnation < 50 DO
        gen <- gen + 1
        parents <- tournament_select(P, k_t=5, |P|/2)
        offspring <- empty
        FOR EACH (x_A, x_B) IN parents DO
            IF random() < 0.85 THEN
                x_C, x_D <- CPC(x_A, x_B, w, chronotypes)
            ELSE
                x_C, x_D <- x_A, x_B
            END IF
            offspring <- offspring + {x_C, x_D}
        END FOR
        FOR EACH x IN offspring DO
            IF random() < 0.15 THEN x <- AWM(x, w, beta=2.0) END IF
        END FOR
        all <- P + offspring
        F_vals <- {F(x,w) : x IN all}
        P <- top_k(all, F_vals, k=|P|)
        curr_best <- max(F_vals)
        IF |curr_best - best_F| < 0.001 THEN
            stagnation <- stagnation + 1
        ELSE
            stagnation <- 0; best_F <- curr_best
        END IF
    END WHILE
    x*_k <- argmax F(x,w) over P
    Q_survey <- collect_satisfaction(x*_k)
    F_comp <- compute_components(x*_k)
    w <- SRA(w, F_comp, Q_survey, alpha=0.4)
    P <- Init-Population(|P|=200, plan_{k+1}, chronotypes)
END FOR
RETURN x*_{K_semesters}, w

```

Рис. 5. Повна схема антропоцентричного методу диспетчеризації

### 5. Експериментальна перевірка ефективності для задачі диспетчеризації занять у ЗВО

Для практичного оцінювання ефективності алгоритму проведено експеримент, який має такі вхідні показники: 400 студентів, 18 академічних груп, 60 викладачів, 42 дисципліни. Створено матрицю когнітивної сумісності  $S$  та хронотипові

профілі, зафіксовані відповідно до [5]. Кожну конфігурацію запущено 30 разів для статистичної надійності тесту.

#### 5.1. Порівнювані конфігурації

Конфігурації для порівняльного аналізу перевірки ефективності для задачі диспетчеризації занять у ЗВО наведено у табл. 4.

Таблиця 4. Конфігурації для порівняльного аналізу

| Конфігурація | Кросовер    | Мутація             | Ваги / адаптація                         |
|--------------|-------------|---------------------|------------------------------------------|
| Baseline GA  | Двоточковий | Рівномірна swap     | Рівномірні, фіксовані                    |
| CPC-GA       | CPC         | Рівномірна swap     | Рівномірні, фіксовані                    |
| AWM-GA       | Двоточковий | AWM ( $\beta=2.0$ ) | Рівномірні, фіксовані                    |
| АМД (повний) | CPC         | AWM ( $\beta=2.0$ ) | Рівномірні $\rightarrow$ SRA (три цикли) |

### 5.2. Результати збіжності

Отримані значення показників збіжності у результаті 30 тестових запусків кожної з порівнюваних конфігурацій подано у табл. 5.

Експериментальна перевірка ефективності АМД продемонструвала, що застосування цього методу дало змогу скоротити час досягнення  $F(x) > 0.75$  на 34.0% (тобто 38.4 проти 58.2 хв). Абляційний аналіз засвідчує, що CPC робить більший внесок

(приріст на 7.0%  $F$  за умови 500 поколінь), ніж АWM (у разі якого фіксувалося збільшення на 4.2%). Однак їх комбінація дає синергетичний ефект (зі зростанням на 18.3%). У наборі даних результату дисперсія  $\sigma=0.029$  у АМД проти  $\sigma=0.043$  у *Baseline* вказує на підвищену стабільність, оскільки антропоцентрично орієнтовані оператори звужують область ефективного пошуку значення функції [1, 4].

Таблиця 5. Показники збіжності (середнє  $\pm \sigma$  по 30 запусків)

| Конфігурація | F за умови 100 пок. | F за умови 500 пок. | t(F > 0.65), хв | t(F > 0.75), хв | $\sigma$ фін. |
|--------------|---------------------|---------------------|-----------------|-----------------|---------------|
| Baseline GA  | 0.53 $\pm$ 0.04     | 0.71 $\pm$ 0.04     | 21.4 $\pm$ 2.1  | 58.2 $\pm$ 4.8  | 0.043         |
| CPC-GA       | 0.59 $\pm$ 0.04     | 0.76 $\pm$ 0.04     | 15.7 $\pm$ 1.8  | 41.3 $\pm$ 4.1  | 0.037         |
| AWM-GA       | 0.57 $\pm$ 0.04     | 0.74 $\pm$ 0.04     | 17.2 $\pm$ 1.9  | 45.6 $\pm$ 4.3  | 0.039         |
| АМД          | 0.64 $\pm$ 0.03     | 0.84 $\pm$ 0.03     | 14.1 $\pm$ 1.6  | 38.4 $\pm$ 3.7  | 0.029         |

### 5.3. Вибір гіперпараметрів

Гіперпараметр мутації  $\beta$  відіграє вагоме значення в балансуванні результату методу АМД. Чутливість до  $\beta$  під час експерименту перевірялася в діапазоні  $\beta \in \{0.5, 1.0, 2.0, 3.0, 5.0\}$ . Значення  $\beta=2.0$  забезпечує оптимальний баланс, оскільки за умови  $\beta=0.5$  АWM еквівалентний рівномірній мутації ( $F=0.72$ ), натомість, якщо  $\beta=5.0$ , виникає надмірна спрямованість ( $F=0.77$ ,  $\sigma=0.052$ ). Значення  $\beta=2.0$  дає  $F=0.84$  з  $\sigma=0.029$ . На відміну від непараметризованих підходів [14, 16], фіксований набір гіперпараметрів забезпечує відтворюваність експерименту й полегшує порівняння конфігурацій.

Параметр  $\alpha$  SRA перевірявся в діапазоні  $\{0.2, 0.4, 0.6, 0.8\}$ . У разі значення параметра  $\alpha=0.2$  система занадто повільно реагує на зміни пріоритетів. За умови задання  $\alpha=0.8$  ваги нестабільно коливаються внаслідок шуму в опитуванні ( $\sigma_w > 0.04$  між циклами). Посереднє значення  $\alpha=0.4$  забезпечує стійку збіжність до  $w^*$  за три цикли, якщо  $\sigma_{noise} \leq 0.3$ .

Це робить результат експерименту більш точним, порівняно з іншими значеннями цих параметрів.

### 5.4. Динаміка адаптації ваг через SRA

Динаміка адаптації вагових коефіцієнтів у процесі балансування результату антропоцентричного методу диспетчеризації відображена у таблиці 6.

За умови різних значень параметрів у процесі експерименту норма адаптації вагових коефіцієнтів  $\|w - w^*\|_2$  скорочується з 0.152 до 0.011 (зі зменшенням на 92.8%), тоді як кореляція між  $w$  і  $w^*$  зростає з 0.61 до 0.89. Домінантним напрямом процесу адаптації є підвищення  $w_3$  (*psych*) і  $w_2$  (*circ*) внаслідок зниження  $w_1$  (*tech*), коли учасники процесу стабільно оцінюють психологічний комфорт і хронотипову відповідність як більш значущі, ніж формальну відсутність конфліктів. Це повністю узгоджується з інформацією про вплив часу проведення занять на успішність здобувачів освіти, що належать до певного хронотипу [4, 7].

Таблиця 6. Динаміка адаптації вагових коефіцієнтів (еталон  $w^* = (0.15; 0.30; 0.35; 0.20)$ )

| Цикл SRA          | $w_1$ (tech) | $w_2$ (circ) | $w_3$ (psych) | $w_4$ (cogn) | $\ w - w^*\ _2$ |
|-------------------|--------------|--------------|---------------|--------------|-----------------|
| Еталон $w^*$      | 0.150        | 0.300        | 0.350         | 0.200        | —               |
| $w^0$ (початок)   | 0.250        | 0.250        | 0.250         | 0.250        | 0.152           |
| $w^1$ (1 семестр) | 0.192        | 0.278        | 0.311         | 0.219        | 0.089           |
| $w^2$ (2 семестр) | 0.167        | 0.292        | 0.336         | 0.205        | 0.039           |
| $w^3$ (3 семестр) | 0.156        | 0.298        | 0.347         | 0.199        | 0.011           |

### Обговорення результатів

Наукова новизна запропонованого підходу полягає в трьох взаємопов'язаних аспектах. По-перше, перенесено антропоцентричні обмеження з рівня штрафних доданків цільової функції на рівень структури генетичних операторів через формальний опис інваріантів збереження циркадної узгодженості та когнітивної якості. По-друге, формалізовано метод адаптації вагових коефіцієнтів на основі регресійного аналізу зворотного зв'язку з гарантованою збіжністю за скінченну кількість циклів. По-третє, продемонстровано інваріантність підходу щодо типу виробничого об'єкта – від авіаційного машинобудування (ПФ "Спейс") до академічного середовища (АСУ ЗВО).

Ключова відмінність від підходів, запропонованих у працях [10, 11, 8], полягає в такому: у *Baseline GA* антропоцентрична властивість впливає лише на відбір через фітнес, але кросовер і мутація систематично порушують досягнутий рівень  $F_{circ}$  у процесі рекомбінації. CPC запобігає деградації, зберігаючи денну структуру як базову одиницю рекомбінації. Зв'язок між хронотиповою відповідністю розкладу й результативністю підтверджено емпіричними дослідженнями [1], що також узгоджується з підходом *constraint-preserving* до планування в задачах маршрутизації [13].

**Обчислювальна вартість.** CPC вимагає  $O(|G| \cdot 6)$  додаткових обчислень на схрещування, AWM –  $O(|G| \cdot 8 \cdot 6)$  обчислень  $q(d, s, g)$ . Загальне збільшення часу одного покоління:  $11.8\% \pm 0.9\%$ . Але, оскільки загальна кількість поколінь скорочується на  $\sim 34\%$ , підсумкова обчислювальна вартість для досягнення  $F(x) > 0.75$  є меншою, ніж у *Baseline GA*.

**Обмеження.** SRA потребує щонайменше  $N = 50$  учасників опитування для надійної OLS-регресії. Якщо  $N < 50$ , рекомендовано ридж-регресію. Лінійна модель SRA адекватна для поступової адаптації ваг, але може не відобразити нелінійні залежності через різкі зміни контингенту. У цьому разі доцільне примусове скидання ваг до рівномірних із подальшою повторною адаптацією.

### Висновки

У роботі запропоновано антропоцентричний підхід до диспетчеризації бізнес-процесів за умови

функціонування виробничих об'єктів, що формалізує задачу розподілу завдань між виконавцями з огляду на їх хронобіологічні, когнітивні та психологічні якості. Підхід є інваріантним щодо типу об'єкта; він може бути використаний на прикладах авіаційного виробництва (ПФ "Спейс") й управління навчальним процесом у ЗВО.

Розроблений у процесі дослідження антропоцентричний метод диспетчеризації (АМД) оснований на перенесенні антропоцентричних обмежень із рівня штрафних доданків на рівень структури генетичних операторів. Під час виконання алгоритму хронотип-зберігальний оператор CPC обирає точки розрізу даних пропорційно різниці циркадної якості між батьківськими особинами. Тоді як оператор AWM реалізує *softmax*-зважений вибір часових слотів для мутації, концентруючи мутаційний бюджет на проблемних слотах із найнижчим антропоцентричним внеском. Отже, метод АМД значно підвищує ефективність процесу диспетчеризації, який може бути використаний як у виробничих процесах, так і в закладах вищої освіти.

Крім цього, необхідно зауважити, що в алгоритмі застосовується механізм адаптивного навчання *Satisfaction Regression Adaptation* (SRA). Він коригує вагові коефіцієнти цільової функції через OLS-регресію агрегованих оцінок задоволеності всіх учасників виробничого чи навчального процесу з ЕМА-згладжуванням ( $\alpha = 0.4$ ) та обмеженням  $w_j \geq 0.05$ . Це усуває необхідність ручного налаштування параметрів і забезпечує автоматичне розгортання в складі автоматизованої системи управління.

Порівняльний аналіз результату експерименту з 30 незалежних запусків підтвердив, що АМД скорочує час досягнення  $F(x) > 0.75$  на 34.0% (38.4 проти 58.2 хв). Також запропонований метод підвищує фінальне  $F$  до 0.84 (зі збільшенням на 18.3% щодо *Baseline GA*) і знижує дисперсію ( $\sigma = 0.029$  проти 0.043). SRA за три цикли скорочує  $\|w - w^*\|_2$  з 0.152 до 0.011 (зі зменшенням на 92.8%), підвищуючи кореляцію з 0.61 до 0.89.

Перспективними результатами дослідження є заміна лінійної регресії SRA на навчання з підкріпленням для нелінійних залежностей задоволеності та застосування CPC/AWM до задачі динамічного перепланування впродовж виробничого циклу. Це уможливило розроблення паралельної



острівної версії програмного забезпечення для великих виробничих систем із розподіленою АСУ.

#### Доступність даних

Рукопис не має пов'язаних даних.

#### Конфлікт інтересів

Автор заявляє, що не має конфлікту інтересів щодо цього дослідження.

#### Використання засобів штучного інтелекту

Автор підтверджує, що під час написання цієї статті не було застосовано технологій штучного інтелекту.

#### Фінансування

Дослідження виконано без фінансової підтримки.

#### References

1. Maučec, K., Štukovnik, V. (2024), "The Relationship between Chronotype and Academic Achievement among Slovene University Students: The Mediating Role of Trait Self-Control and Sleep Quality", *Center for Educational Policy Studies Journal*. DOI: <https://doi.org/10.26529/cepsj.1790>
2. Jankowski, K. S., Díaz-Morales, J. F., Vollmer, C. (2023), "Chronotype, Time of Day, and Performance on Intelligence Tests in the School Setting", *Journal of Intelligence*, Vol. 11, No. 1, p. 13. DOI: <https://doi.org/10.3390/jintelligence11010013>
3. Al-Rfooh, O. F., Khater, W. (2023), "The impact of chronotype on physical health, psychological health, and job performance among health care providers in acute care settings", *International Journal of Healthcare Management*, Vol. 16, No. 4, pp. 581–589. DOI: <https://doi.org/10.1080/20479700.2023.2177665>
4. Souza, S. S., Brito, K. S., Duarte, L. L. (2024), "Sleep quality, chronotype and academic performance of university students in the health area", *Biological Rhythm Research*. DOI: <https://doi.org/10.1080/09291016.2024.2437643>
5. Sytnik, O. O. (2025), "Anthropocentric model of class scheduling in higher education institutions of aerospace profile", *Aerospace Technic and Technology*, No. 6 (215), pp. 1–12. DOI: <https://doi.org/10.32620/akt.2025.6.07>
6. Chen, S.-C., Chen, H.-M., Chen, H.-K., Li, C.-L. (2024), "Multi-Objective Optimization in Industry 5.0: Human-Centric AI Integration for Sustainable and Intelligent Manufacturing", *Processes*, Vol. 12, No. 12, p. 2723. DOI: <https://doi.org/10.3390/pr12122723>
7. Ситнік, О., Вдовіченко, О. (2025), "Математичні структури та інструменти антропоцентричної диспетчеризації в закладах вищої освіти", *Відкриті інформаційні та комп'ютерно-інтегровані технології*, № 105, С. 212–226. DOI: <https://doi.org/10.32620/oikit.2025.105.17>
8. Mokhtari, M., Vaziri Sarashk, M., Asadpour, M., Saeidi, N., Boyer, O. (2021), "Developing a Model for the University Course Timetabling Problem: A Case Study", *Complexity*, Article e9940866. DOI: <https://doi.org/10.1155/2021/9940866>
9. Oude Vrielink, R. A., Jansen, E. A., Hans, E. W., van Hillegersberg, J. (2019), "Practices in timetabling in higher education institutions: a systematic review", *Annals of Operations Research*, Vol. 275, pp. 145–160. DOI: <https://doi.org/10.1007/s10479-017-2688-8>
10. Burke, E. K., Petrovic, S. (2002), "Recent research directions in automated timetabling", *European Journal of Operational Research*, Vol. 140, No. 2, pp. 266–280. DOI: [https://doi.org/10.1016/S0377-2217\(02\)00069-3](https://doi.org/10.1016/S0377-2217(02)00069-3)
11. Haitan, O., Nazarov, O. (2020), "Hybrid approach to solving the automated timetabling problem in higher educational institution", *Системи Управління Навігації та Зв'язку*, Vol. 2, pp. 60–69. DOI: <https://doi.org/10.26906/SUNZ.2020.2.060>
12. Косолап, А. І., Дубовик, Т. М. (2021), "Оптимізація розкладу занять в університеті", *Радіоелектроніка, інформатика, управління*, № 3, С. 175–183. DOI: <https://doi.org/10.15588/1607-3274-2021-3-15>
13. Weiszer, M., Burke, E. K., Chen, J. (2022), "Search graph structure and its implications for multi-graph constrained routing and scheduling problems", *Scientific Reports*, Vol. 12, Art. 14860. DOI: <https://doi.org/10.1038/s41598-022-18026-w>
14. Wang, S., Drake, J. H., Melder, D. et al. (2025), "A selection hyper-heuristic approach to the single airport slot allocation problem", *Computers & Operations Research*, Vol. 184, Art. 107223. DOI: <https://doi.org/10.1016/j.cor.2025.107223>
15. Mahlous, A. R., Mahlous, H. (2023), "Student timetabling genetic algorithm accounting for student preferences", *PeerJ Computer Science*, Vol. 9, Article e1200. DOI: <https://doi.org/10.7717/peerj-cs.1200>
16. Eiben, A. E., Smith, J. E. (2015), *Introduction to Evolutionary Computing*, 2nd ed., Springer, Berlin, 302 p. DOI: <https://doi.org/10.1007/978-3-662-44874-8>
17. Roenneberg, T., Wirz-Justice, A., Mellow, M. (2003), "Life between clocks: Daily temporal patterns of human chronotypes", *Journal of Biological Rhythms*, Vol. 18, No. 1, pp. 80–90. DOI: <https://doi.org/10.1177/0748730402239679>
18. Horne, J. A., Ostberg, O. (1976), "A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms", *International Journal of Chronobiology*, Vol. 4, No. 2, pp. 97–110.

19. Феоктистова, О. І. (2017), Модель та методи компетентісно-орієнтованого експертного оцінювання команди виконавців високотехнологічного проекту: дисертація канд. техн. наук: 05.13.22, Національний аерокосмічний університет «ХАІ», Харків, 215 с.
20. Носиков, О., Ситнік, О. (2025), "Формування єдиного інформаційного простору в закладах вищої освіти: виклики, рішення та перспективи розвитку", *Відкриті інформаційні та комп'ютерно-інтегровані технології*, № 104, С. 200–213. DOI: <https://doi.org/10.32620/oikit.2025.104.13>

Received (Надійшла) 27.01.2026

Accepted for publication (Прийнята до друку) 30.03.2026

Publication date (Дата публікації) 29.05.2026

*Відомості про авторів / About the Authors*

**Ситнік Олег Олександрович** – Національний аерокосмічний університет "Харківський авіаційний інститут", аспірант кафедри інженерії програмного забезпечення, Харків, Україна;

**Oleh Sytnik** – National Aerospace University "Kharkiv Aviation Institute", Postgraduate Student of the Department of Software Engineering, Kharkiv, Ukraine;

e-mail: [o.sytnik@khai.edu](mailto:o.sytnik@khai.edu)

ORCID ID: <http://orcid.org/0009-0009-4504-3489>

## APPROACH TO BUSINESS PROCESS DISPATCHING IN THE OPERATION OF INDUSTRIAL FACILITIES BASED ON ANTHROPOCENTRICITY

The article presents an approach to anthropocentric dispatching of business processes in the operation of production facilities, aligning task distribution among executors by accounting for a combination of their chronobiological, cognitive, and psychological characteristics. The goal of this work is to improve the efficiency of business processes in the operation of production facilities (PF) through the development, formal description, and validation of the anthropocentric dispatching method (ADM) as a general approach to the class of scheduling problems in PFs with human executors, as well as to demonstrate its applicability using a real aviation industry enterprise and an automated control system for the educational process of a higher education institution (HEI). The objectives of this study are the formalization of the anthropocentric dispatching problem as one that is invariant with respect to the type of production facility, the development of the chronotype-preserving crossover operator CPC, and the implementation of the anthropocentric weight mutation operator AWM algorithm. The paper also examines the adaptive weight adjustment mechanism (SRA) and compares ADM with baseline GA configurations. The methodological tools employed are an evolutionary algorithm with anthropocentrically oriented genetic operators, OLS regression of satisfaction across all participants in the educational process, ablation analysis, and comparative testing over 30 independent runs of a software prototype. The practical results obtained show that ADM reduces the time to reach  $F(x) > 0.75$  by 34.0% (38.4 vs. 58.2 min), increases the final value of the objective function to 0.84 (an 18.3% improvement over Baseline GA), and reduces result variance to  $\sigma = 0.029$ . The norm  $\|w - w^*\|_2$  decreases by 92.8% over a period of 3 semesters. Conclusions: the scientific novelty of this study lies in the formal description of invariants that preserve circadian coherence and cognitive quality within the structure of the genetic operators, rather than through penalty terms in the objective function. The generality of the approach for a broad class of production dispatching problems with a human executor has also been demonstrated.

**Keywords:** anthropocentric dispatching; business process management; industrial facility AMS; evolutionary algorithm; genetic operators; chronotype-preserving crossover; adaptive learning; educational process management system.

*Бібліографічні описи / Bibliographic descriptions*

Ситнік О. О. Підхід до диспетчеризації бізнес-процесів в умовах функціонування виробничих об'єктів на основі антропоцентричності. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 247–258. DOI: <https://doi.org/10.30837/0135-1710.2026.189.247>

Sytnik, O. (2026), "Approach to business process dispatching in the operation of industrial facilities based on anthropocentricity", *Management Information System and Devices*, No. 2 (189), P. 247–258. DOI: <https://doi.org/10.30837/0135-1710.2026.189.247>

Далека В. Х., Бомбандьоров Ю. П., Метешкін К. О., Андрусевич А. О., Чумаченко І. В.

## ІНФОРМАЦІЙНЕ ЗАБЕЗПЕЧЕННЯ ЕНЕРГЕТИЧНОЇ ЄМНОСТІ ЕЛЕКТРОТРАНСПОРТУ В ПРОЄКТАХ І ПРОГРАМАХ СТАЛОГО РОЗВИТКУ УКРАЇНИ

**Предметом вивчення** є моделі й методики визначення застосовності типоряду тягових акумуляторних батарей електротранспорту в мультипроектному середовищі для забезпечення автономного ходу тролейбусів у містах України. **Мета дослідження** – розробити інформаційне забезпечення енергетичної ємності електротранспорту в проєктах і програмах сталого розвитку й обґрунтувати достатність типоряду тягових акумуляторних батарей (ТАБ) для задоволення потреби міст в електротранспорті та сталого розвитку України в цьому напрямі. Зважаючи на окреслену мету, необхідно виконати такі **завдання**: оцінити енергетичну ємність тягових акумуляторних батарей; зібрати інформацію по містах України про характеристики тролейбусних маршрутів і обґрунтувати доцільність використання тролейбусів з автономним ходом з огляду на містобудівельні та енергетичні обмеження; дослідити й розробити інформаційне забезпечення на основі моделі визначення енергетичної ємності та методики визначення застосовності типоряду ТАБ у мультипроектному середовищі для забезпечення автономного ходу тролейбусів у містах України. **Методи**: системний, аналітичний, комп'ютерне моделювання, програмний продукт EASY WAY. **Досягнуті результати**. Запропоновано інформаційне забезпечення на основі моделі встановлення енергетичної ємності та методику визначення застосовності типоряду тягових акумуляторних батарей. Продемонстровано, що типоряд ТАБ здатний забезпечити автономний хід тролейбусів майже для всіх міст України, якщо його довжина не перевищуватиме 15 км. **Висновки**. За результатами дослідження визначено, що наявний типоряд ТАБ може забезпечувати автономний хід на відстань не меншу ніж 30 км, крім окремих міст України, що пов'язано з наявністю витрат енергоносія. Надалі необхідно вдосконалити інформаційне забезпечення енергетичної ємності електротранспорту в проєктах і програмах сталого розвитку й дослідити способи уточнення витрат енергоносія в містах України, де наразі неможливо застосувати жодну із ТАБ, які виготовляються для довжин автономного ходу понад 15 км.

**Ключові слова**: інформаційне забезпечення; проєкт; транспортний засіб; енергетична ємність; енергоносіє; сталий розвиток.

### Вступ

Збільшення кількості транспортних засобів з електричною тяговою установкою в мультипроектному середовищі вдосконалення електротранспорту має на меті покращення екологічного стану міст і сталого розвитку України. Нині Україна посідає одне із провідних місць Європи, де здійснюються перевезення пасажирів міським електричним транспортом, до складу якого належить тролейбусний транспорт. Прийнятий 2023 року закон України [1] передбачає розроблення проєктів і програм заміни автобусів, що мають дизельні силові установки в українських містах, на транспортні засоби з електричними тяговими установками. Для реалізації вимоги цього закону необхідно насамперед з'ясувати, які витрати енергоносія має електробус чи тролейбус з автономним ходом під час роботи на маршруті для обґрунтування необхідної для цього енергетичної ємності тягової акумуляторної батареї, і розробити інформаційну технологію на основі моделі

визначення енергетичної ємності для обґрунтування різних типів ТАБ.

### Аналіз останніх досліджень і публікацій

Надання чинності закону України [1] сприяло збільшенню інтересу до визначення енергоефективності транспортних засобу категорії МЗ класу І, до яких належать автобуси й тролейбуси.

Так, дослідження фахівців Національного транспортного університету [2–6] присвячені методу обґрунтування енергетичної ємності тягової акумуляторної батареї для проєктів забезпечення автономного ходу тролейбусів. У студіях [5, 6] висвітлено доцільність використання транспортних засобів з різними силовими установками. Крім того, вчені Луцького політехнічного університету виконали випробування щодо витрат енергоносія для різних умов руху маршрутом [7, 8]. На особливу увагу заслуговують роботи фахівців Харківського національного університету міського господарства імені О. М. Бекетова, які дають теоретичну базу

для проектів визначення динамічних властивостей транспортних засобів з електричною тяговою установкою, що також містять і витрати енергоносія на їх забезпечення [9–11]. Програми за напрямом оцінювання витрат енергоносіїв та їх поповнення дослідили фахівці інших країн [12–15]. Доцільність використання тролейбусів із функцією автономного ходу (АХ) підтверджено виконаними проектами й досвідом їх експлуатації в низці міст України (Київ, Харків, Кременчук, Одеса тощо), а також у багатьох інших країнах. Удосконалюється й наукова база програм енергоефективності. Публікації за результатами досліджень можна поділити на ті, в яких теоретично обґрунтовано проекти витрат ресурсів [9–11] і висвітлено наукові, технічні та організаційні питання розроблення проектів ресурсозбереження, підвищення якості програм транспортного обслуговування пасажирів і охорони довкілля [1, 9–11].

Особливо проблемними в експлуатації тролейбусів із функцією АХ є питання зарядження / дозарядження ТАБ, що здійснюються різними методами. Зокрема в статті [10] запропоновано застосування на рухомому складі сонячних панелей, у студіях [11, 13–16] ідеться про вдосконалення графіків зарядження батарей в депо і на кінцевих станціях, а в публікації [12] описано використання балансу електроенергії протягом визначеного часу.

#### Мета й завдання дослідження

Мета дослідження – розробити інформаційне забезпечення енергетичної ємності електротранспорту в проектах і програмах сталого розвитку, обґрунтувати

достатність типоряду ТАБ, які виготовляються в Україні ТОВ "ПОЛІТЕХНОСЕРВІС", для забезпечення потреби міст у тролейбусах, що мають функцію автономного ходу, і підтримки сталого розвитку країни в цьому напрямі.

Для досягнення окресленої мети необхідно виконати такі завдання: оцінити енергетичну ємність ТАБ; зібрати інформацію по українських містах про характеристики тролейбусних маршрутів і обґрунтувати доцільність використання тролейбусів з АХ з огляду на містобудівельні та енергетичні обмеження; дослідити й розробити інформаційну технологію на основі моделі визначення енергетичної ємності й методик визначення застосовності типоряду ТАБ у мультипроектному середовищі для забезпечення АХ тролейбусів у містах України.

У цій статті функція АХ тролейбуса передбачає рух транспортного засобу тільки із живленням від ТАБ або інших джерел енергії на рухомому складі.

#### Виклад основного матеріалу

Нині в Україні виготовляють ТАБ, які забезпечують автономний хід тролейбуса на ділянках міського маршруту, де відсутня контактна мережа.

Крім акумуляторних батарей, для АХ тролейбусів можуть використовуватись іоністори (суперконденсатори), дизель-генераторні установки, паливні елементи й газові турбіни. Але найбільш перспективним напрямом вважаються програми застосування ТАБ.

Типоряд ТАБ, виробництво якого впроваджено в Україні, має технічні властивості, подані в табл. 1.

Таблиця 1. Технічні властивості типоряду ТАБ

| Тип і виконання ТАБ | Маса, кг | Енергетична ємність, кВт*год | Максимальна напруга ТАБ за 100% заряду, В | Мінімальна напруга ТАБ для 40% зарядженої батареї на опорі 6,1 Ом (після нормального розряду) | Мінімальна напруга ТАБ для 30% зарядженої батареї на опорі 5,7 Ом (критичний розряд) |
|---------------------|----------|------------------------------|-------------------------------------------|-----------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| ТБЕТЗ.01            | 615      | 32,4                         | 540                                       | 405                                                                                           | 375                                                                                  |
| ТБЕТЗ.02            | 511      | 65,7                         | 589                                       | 469                                                                                           | 378                                                                                  |
| ТБЕТЗ.03            | 511      | 65,7                         | 574                                       | 469                                                                                           | 406                                                                                  |
| ТБЕТЗ.04            | 381      | 45                           | 590                                       | 482                                                                                           | 418                                                                                  |
| ТБЕТЗ.05            | 540      | 57                           | 590                                       | 482                                                                                           | 418                                                                                  |
| ТБЕТЗ.06            | 671      | 90                           | 590                                       | 482                                                                                           | 418                                                                                  |
| ТБЕТЗ.07            | 590      | 78,0                         | 590                                       | 417                                                                                           | 388,8                                                                                |

Для оцінювання енергетичної ємності ТАБ було зібрано інформацію про довжину міських тролейбусних маршрутів. Передбачено, що інформаційне забезпечення енергетичної ємності електротранспорту буде використовувати програмний продукт EASY WAY, який працює з відомостями про рух автобусів,

тролейбусів і трамвайних вагонів на маршрутах заданого міста, з характеристиками маршрутів, технічними властивостями типоряду ТАБ та показниками питомих витрат енергоносія тролейбусним транспортом. Параметри довжини маршрутів тролейбусного транспорту подано в табл. 2.

Навність такої інформації дало змогу визначити довжину ділянки АХ, яку можна застосувати в разі необхідності для подовження маршруту. Типові довжини ділянок обрано за умови, що рух АХ не має перевищувати 40% від загальної довжини одного оборотного рейсу руху із живленням від контактної мережі.

Вибір співвідношення 40% АХ і 60% руху із живленням від контактної мережі пов'язаний із властивістю тягової акумуляторної батареї бути в режимах розрядження й зарядження. Як правило, потужність зарядного пристрою ТАБ завжди менша за потужність, задіяну під час розрядження тягової акумуляторної батареї, що забезпечує встановлений ресурс її роботи. Тому час зарядження ТАБ завжди

перевищує час її розрядження, а за умови однакової швидкості руху на маршруті ділянка АХ має бути завжди менша, ніж ділянка руху, де вона отримує заряд.

У табл. 2 подано довжини ділянок АХ, які можуть бути запропоновані для подовження маршрутів без використання капітальних вкладень на будівництво системи енергоживлення таких ділянок.

Обґрунтування необхідної енергетичної ємності ТАБ для забезпечення АХ на ділянках різних довжин (див. табл. 3) може ґрунтуватися на статистичних відомостях про витрати енергоносія для кожного міста України. Ця інформація береться з показань лічильників, установлених на тягових підстанціях, що живлять контактну мережу трамвайного й тролейбусного транспорту.

Таблиця 2. Довжини маршрутів тролейбусного транспорту

| Місто            | Довжина маршрутів, км |       |       |       |       |       |       |       |       |       |
|------------------|-----------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Вінниця          | 14,1                  | 7,67  | 12,97 | 13,42 | 8,93  | 4,77  | 7,35  | 7,64  | 9,44  | 7,64  |
|                  | 10,69                 | 7,13  | 10,65 | 9,9   | 9,24  | 9,52  | 12,92 | 12,92 | 9,04  | 16,73 |
|                  | 13,47                 | 16,09 |       |       |       |       |       |       |       |       |
| Дніпро           | 7,61                  | 21,58 | 10,44 | 7,25  | 7,65  | 16,63 | 7,62  | 9,88  | 13,39 | 9,11  |
|                  | 5,4                   | 3,68  | 11,58 | 8,67  | 11,07 | 11,7  | 10,95 | 10,03 | 12,47 |       |
| Житомир          | 10,96                 | 11,95 | 10,31 | 12,57 | 11,82 | 7,77  | 7,9   | 6,61  | 12,76 | 15,2  |
|                  | 9,97                  | 10,03 |       |       |       |       |       |       |       |       |
| Запоріжжя        | 14,18                 | 14,24 | 8,85  | 16,91 | 5,88  |       |       |       |       |       |
| Київ             | 4,89                  | 7,17  | 17,28 | 14,03 | 14,24 | 14,82 | 9,34  | 8,95  | 13,1  | 17,07 |
|                  | 7,22                  | 5,57  | 11,64 | 6,96  | 12,25 | 19,43 | 9,24  | 7,93  | 9,41  | 7,83  |
|                  | 13,25                 | 9,02  | 14,79 | 24,25 | 21,96 | 6,17  | 14,05 | 9,85  | 19,24 | 10,56 |
|                  | 11,84                 | 7,32  | 6,69  | 9,25  | 9,1   | 4,81  | 14,57 | 18,12 | 8,29  | 6,51  |
|                  | 21,52                 | 11,93 |       |       |       |       |       |       |       |       |
| Львів            | 10,47                 | 8,6   | 8,38  | 6,67  | 7,45  | 6,44  | 11,71 | 6,63  | 5,14  | 9,54  |
| Одеса            | 5,78                  | 8,91  | 19,76 | 8,23  | 10,98 | 13,03 | 11,83 |       |       |       |
| Івано-Франківськ | 8,48                  | 10,93 | 10,47 | 7,16  | 7,61  | 9,78  | 9,37  |       |       |       |
| Кременчук        | 8,18                  | 9,83  | 11,44 | 15,53 | 11,28 | 12,23 | 11,2  | 12,4  | 13,77 | 20,4  |
|                  | 13,25                 | 15,05 | 17,31 | 13,82 | 15,85 |       |       |       |       |       |
| Кривий Ріг       | 11,27                 | 7,21  | 11,05 | 27,87 | 14,47 | 7,43  | 20,34 | 11,47 | 16,15 | 9,21  |
|                  | 27,04                 | 16,04 | 13,25 | 18,48 | 14,36 | 15,29 | 6,29  | 18,57 | 16,01 | 15,67 |
|                  | 29,13                 | 22,63 |       |       |       |       |       |       |       |       |
| Луцьк            | 15,41                 | 13,3  | 14,51 | 11,45 | 11,99 | 11,21 | 12,18 | 11,95 | 11,06 |       |
| Миколаїв         | 20,23                 | 15,19 | 8,14  | 8,1   | 12,4  | 14,26 | 14,18 | 27,87 | 15,68 | 27,05 |
|                  | 12,87                 | 11,76 |       |       |       |       |       |       |       |       |
| Полтава          | 10,56                 | 10,85 | 10,63 | 14,65 | 15,46 | 15,21 | 10,35 | 15,2  | 9,2   |       |
| Рівне            | 6,72                  | 9,11  | 5,77  | 6,23  | 6,36  | 9,2   | 8,46  | 5,22  | 5,35  | 10,45 |
|                  | 13,92                 | 14,48 |       |       |       |       |       |       |       |       |
| Суми             | 8,39                  | 15,1  | 8     | 10,7  | 11,21 | 8,98  | 5,69  | 12,17 | 11,27 | 13,84 |
|                  | 10,14                 | 6,55  | 6,77  | 10,42 | 9,92  | 8,11  | 5,81  |       |       |       |
| Тернопіль        | 8,16                  | 6,31  | 12,32 | 12,95 | 13,25 | 10,85 | 6,13  | 10,87 |       |       |
| Херсон           | 3,77                  | 8,02  | 6,07  | 6,19  | 10,23 |       |       |       |       |       |
| Хмельницький     | 8,98                  | 12,73 | 12,47 | 11,89 | 6,98  | 13,71 | 13,95 | 7,32  | 10,17 | 9,95  |
|                  | 8,5                   | 10,76 | 10,55 | 9,65  | 13,17 | 7,43  | 4,97  | 7,5   | 7,93  | 10,47 |
|                  | 9,63                  | 5,1   | 9,07  | 8,35  |       |       |       |       |       |       |
| Черкаси          | 12,47                 | 12,41 | 8,83  | 9,03  | 11,17 | 7,48  | 15,44 | 18,7  | 12,52 | 13,97 |
|                  | 12,09                 | 12    | 11,04 | 12,49 | 10,22 | 13,75 | 10,99 | 12,78 | 16,1  | 12,49 |
|                  | 11,93                 |       |       |       |       |       |       |       |       |       |
| Чернівці         | 5,84                  | 6,51  | 9,04  | 5,13  | 14,63 | 7,77  | 6,47  | 5,44  |       |       |
| Чернігів         | 8,38                  | 10,62 | 9,59  | 19,28 | 14,43 | 11,61 | 9,94  |       |       |       |
| Кропивницький    | 13,61                 | 12,68 | 10,96 | 17,44 | 18,75 | 11,43 | 7,8   | 12,39 |       |       |



Таблиця 3. Довжина ділянок АХ, які пропонуються

| Довжина ділянки АХ, км | Довжина ділянки живлення від контактної мережі, км | Кількість маршрутів, де можна реалізувати АХ | Відсоток маршрутів, де можна реалізувати АХ від загальної кількості маршрутів, % |
|------------------------|----------------------------------------------------|----------------------------------------------|----------------------------------------------------------------------------------|
| 25                     | понад 30 км                                        | 1                                            | 0,34                                                                             |
| 20                     | понад 25 км                                        | 6                                            | 2,06                                                                             |
| 15                     | понад 18 км                                        | 11                                           | 3,78                                                                             |
| 10                     | понад 12,5                                         | 79                                           | 27,15                                                                            |
| 5                      | понад 5 км                                         | 194                                          | 66,67                                                                            |

**Примітка.** Довжину ділянки АХ подано для одного оборотного рейсу, а ділянки із живленням від контактної мережі – для 0,5 оборотного рейсу.

Показання лічильників спожитої електроенергії, встановлені на тягових підстанціях, містять не тільки витрати енергоносія на виконання транспортної роботи, а й втрати енергоносія:

- на підстанціях в процесі перетворення змінної високої напруги 12 кВ у напругу постійного струму номінальною напругою 550 В;

- під час передавання до струмоприймачів тролейбуса в контактній та кабельній мережі.

Щодо втрати енергії в контактній і кабельній мережі, то їх величина визначається перерізом контактного дроту й величиною струму споживання. Для подальших розрахунків можна прийняти значення втрат енергоносія на рівні 10%.

Витрати енергоносія на рух тролейбуса від депо до заданого розкладом маршруту й у зворотному напрямку після завершення роботи на маршруті, а також витрати на виконання маневрових робіт під час перебування транспортного засобу в депо дотепер не досліджувалися. Для оцінювання цього складника витрат енергоносія приймемо, що їх доля в загальних витратах може становити приблизно 5%.

Тоді, з огляду на зазначені вище припущення для інформації, поданої в табл. 4, можна застосувати коригувальний коефіцієнт 0,85, щоб отримати показники витрат енергоносія безпосередньо на здійснення транспортної роботи.

Табл. 4 містить скореговану інформацію щодо витрат енергоносія на перевезення пасажирів, а також:

- середнє значення питомих витрат, яке має бути застосовано для встановлення робочого діапазону розряду ТАБ;

- стандартне відхилення значення питомих витрат, використане для визначення максимальної витрати енергоносія;

- максимальне значення питомих витрат, що має бути застосовано для встановлення критичного діапазону розряду ТАБ.

Максимальне значення питомих витрат енергоносія є максимальним значенням двобічного довірчого інтервалу, визначеного для довірчої ймовірності 95%, яке обчислено за методикою ДСТУ ISO 2602 [21].

Загальна модель розрахунку для встановлення необхідної енергетичної ємності ТАБ запропонована в роботі [4], але може бути записана в спрощеному вигляді, оскільки споживання електричного устаткування транспортного засобу на власні потреби вже міститься в статистичних показниках, поданих у табл. 4, а саме:

$$C_{ТАБ} = \max \left\{ \begin{matrix} e_{cp} \cdot L / k_{pd} \\ e_{max} \cdot L / k_{kd} \end{matrix} \right., \quad (1)$$

де  $k_{pd}$  – коефіцієнт, що регламентує робочий діапазон розряду ТАБ (0,6);  $k_{kd}$  – коефіцієнт, що регламентує критичний діапазон розряду ТАБ (0,8);  $e_{cp}$  – середнє значення питомих витрат енергоносія;  $e_{max}$  – можливе максимальне значення питомих витрат енергоносія;  $L$  – довжина ділянки маршруту для руху автономним ходом (далі АХ).

Отримане за виразом (1) значення необхідної енергетичної ємності ТАБ використано для вибору її з типоряду, який виготовляє ТОВ "ПОЛІТЕХНОСЕРВІС". У такий спосіб було обчислено кількість випадків можливого застосування ТАБ із визначеною енергетичною ємністю для різних довжин ділянок АХ щодо загальної кількості міст, які розглядаються в цьому дослідженні.

У табл. 5 на базі показників, узятих з моделі визначення енергетичної ємності для обґрунтування різних типів тягових акумуляторних батарей, подано частоту застосування ТАБ (у відсотках) по містах України залежно від можливої довжини ділянки АХ.

**Таблиця 4.** Статистичні показники питомих витрат енергоносія тролейбусним транспортом

| Місто            | Витрати енергоносія, кВт*год/км |         |         | Середнє значення, кВт*год/км | Стандартне значення, кВт*год/км | Максимальне значення, кВт*год/км |
|------------------|---------------------------------|---------|---------|------------------------------|---------------------------------|----------------------------------|
|                  | 2022 р.                         | 2023 р. | 2024 р. |                              |                                 |                                  |
| Вінниця          | 1,43                            | 1,62    | 1,49    | 1,51                         | 0,10                            | 1,57                             |
| Дніпро           | 1,38                            | 1,56    | 2,03    | 1,66                         | 0,34                            | 1,86                             |
| Житомир          | 3,18                            | 4,15    | 3,46    | 3,60                         | 0,50                            | 3,90                             |
| Запоріжжя        | 1,57                            | 1,93    | 1,52    | 1,67                         | 0,22                            | 1,81                             |
| Київ             | 1,09                            | 1,76    | 1,47    | 1,44                         | 0,34                            | 1,64                             |
| Львів            | 3,46                            | 4,10    | 1,45    | 3,00                         | 1,38                            | 3,84                             |
| Одеса            | 3,42                            | 4,22    | 4,66    | 4,10                         | 0,63                            | 4,48                             |
| Івано-Франківськ | 1,52                            | 1,61    | 1,72    | 1,62                         | 0,10                            | 1,68                             |
| Кременчук        | 3,92                            | 5,12    | 4,72    | 4,59                         | 0,61                            | 4,96                             |
| Кривий Ріг       | 1,87                            | 3,04    | 3,58    | 2,83                         | 0,88                            | 3,36                             |
| Луцьк            | 2,66                            | 3,16    | 3,11    | 2,98                         | 0,27                            | 3,14                             |
| Миколаїв         | 1,76                            | 3,47    | 2,56    | 2,60                         | 0,86                            | 3,12                             |
| Полтава          | 3,36                            | 2,68    | 2,86    | 2,97                         | 0,35                            | 3,18                             |
| Рівне            | 2,75                            | 3,17    | 3,02    | 2,98                         | 0,21                            | 3,11                             |
| Суми             | 2,08                            | 2,17    | 1,97    | 2,07                         | 0,10                            | 2,13                             |
| Тернопіль        | 1,29                            | 1,53    | 1,54    | 1,46                         | 0,14                            | 1,54                             |
| Херсон           |                                 | 1,34    | 1,00    | 1,17                         | 0,24                            | 1,32                             |
| Хмельницький     | 1,77                            | 1,80    | 1,73    | 1,77                         | 0,03                            | 1,79                             |
| Черкаси          | 0,85                            | 1,34    | 2,43    | 1,54                         | 0,81                            | 2,03                             |
| Чернівці         |                                 | 2,09    | 0,02    | 1,06                         | 1,46                            | 1,94                             |
| Чернігів         | 2,88                            | 3,19    | 3,36    | 3,14                         | 0,24                            | 3,29                             |
| Кропивницький    | 4,95                            | 4,96    | 5,01    | 4,97                         | 0,03                            | 4,99                             |

**Таблиця 5.** Можлива частота застосування типів ТАБ, %

| Виконання ТАБ та її енергетична ємність | Довжина АХ, км |        |       |       |       |       |
|-----------------------------------------|----------------|--------|-------|-------|-------|-------|
|                                         | 5              | 10     | 15    | 20    | 25    | 30    |
| ТБЕТЗ.01, 32.4 кВт*год                  | 86,36          | 50,00  | 4,55  | –     | –     | –     |
| ТБЕТЗ.03, 65.7 кВт*год                  | –              | 9,09   | 4,55  | 9,09  | 22,73 | 4,55  |
| ТБЕТЗ.04, 45 кВт*год                    | 13,64          | 4,55   | 40,91 | 4,55  | –     | –     |
| ТБЕТЗ.05, 57 кВт*год                    | –              | 22,73  | 4,55  | 27,77 | 4,55  | –     |
| ТБЕТЗ.06, 90 кВт*год                    | –              | 4,55   | 13,64 | 4,55  | 4,55  | 18,18 |
| ТБЕТЗ.071, 78, кВт*год                  | 0,00           | 9,09   | 27,27 | 4,55  | 18,18 | 22,73 |
| Разом                                   | 100,00         | 100,00 | 95,47 | 50,51 | 50,01 | 45,46 |

У табл. 6 подано міста України, для яких неможливо застосувати ТАБ із типоряду, що виготовляє ТОВ "ПОЛІТЕХНОСЕРВІС", якщо замовлена містом бажана довжина АХ буде перевищувати 10 км.

Обмеження застосовності типоряду ТАБ для АХ завдовжки понад 10 км для міст України (див. табл. 6) насамперед пов'язано із вищими, ніж у інших містах, питомими витратами енергоносія.

На нашу думку, причиною більших питомих витрат енергоносія можуть бути:

- недосконала методика розподілу їх витрат між тролейбусним і трамвайним транспортом у містах Житомирі, Львові, Одесі, Миколаєві, Кривому Розі;
- похибки обліку витрат енергоносія або наявність додаткових його споживачів, які не пов'язані із забезпеченням руху тролейбусів (Кременчук, Кривий Ріг, Луцьк, Полтава, Рівне, Суми, Чернігів).

Таблиця 6. Міста України, де неможливо забезпечити АХ завдовжки понад 10 км

| Місто         | Витрати енергоносія, кВт*год/км за роки |      |      | Середнє значення витрати енергоносія, кВт*год/км | Довжина АХ |    |    |    |    |    |
|---------------|-----------------------------------------|------|------|--------------------------------------------------|------------|----|----|----|----|----|
|               | 2022                                    | 2023 | 2024 |                                                  | 5          | 10 | 15 | 20 | 25 | 30 |
| Житомир       | 3,18                                    | 4,15 | 3,46 | 3,60                                             | +          | +  | +  | –  | –  | –  |
| Львів         | 3,46                                    | 4,10 | 1,45 | 3,00                                             | +          | +  | +  | –  | –  | –  |
| Одеса         | 3,42                                    | 4,22 | 4,66 | 4,10                                             | +          | +  | –  | –  | –  | –  |
| Кременчук     | 3,92                                    | 5,12 | 4,72 | 4,59                                             | +          | +  | –  | –  | –  | –  |
| Кривий Ріг    | 1,87                                    | 3,04 | 3,58 | 2,83                                             | +          | +  | +  | –  | –  | –  |
| Луцьк         | 2,66                                    | 3,16 | 3,11 | 2,98                                             | +          | +  | +  | –  | –  | –  |
| Миколаїв      | 1,76                                    | 3,47 | 2,56 | 2,60                                             | +          | +  | +  | +  | –  | –  |
| Полтава       | 3,36                                    | 2,68 | 2,86 | 2,97                                             | +          | +  | +  | –  | –  | –  |
| Рівне         | 2,75                                    | 3,17 | 3,02 | 2,98                                             | +          | +  | +  | –  | –  | –  |
| Суми          | 2,08                                    | 2,17 | 1,97 | 2,07                                             | +          | +  | +  | +  | +  | –  |
| Чернігів      | 2,88                                    | 3,19 | 3,36 | 3,14                                             | +          | +  | +  | –  | –  | –  |
| Кропивницький | 4,95                                    | 4,96 | 5,01 | 4,97                                             | +          | +  | –  | –  | –  | –  |

Примітка. "+" – можливо забезпечити АХ, "-" неможливо забезпечити АХ наявним типорядом ТАБ.

### Обговорення результатів дослідження

Модель, методика та інформаційне забезпечення на їх основі можуть бути застосовані з метою обґрунтування різних типів ТАБ для забезпечення автономного ходу тролейбусів і перевезення пасажирів електробусами. Результати цього дослідження наразі є науковою базою вдосконалення ТАБ, що виготовляє ТОВ "ПОЛІТЕХНОСЕРВІС".

### Висновки

У процесі дослідження визначено, що описаний типоряд ТАБ здатний забезпечувати АХ на відстані не меншій ніж 30 км, крім поодиноких українських міст. Це пов'язано з наявністю витрат енергоносія, які в інших містах країни є вищими.

З метою уточнення типоряду ТАБ надалі необхідно вдосконалити інформаційне забезпечення на основі моделі визначення енергетичної ємності різних типів тягових акумуляторних батарей, а також дослідити способи уточнення витрат енергоносія в містах України, де наразі неможливо застосувати жодну з ТАБ, які нині виготовляються, для довжин АХ понад 15 км.

Матеріали цієї статті можуть бути впроваджені для розроблення інформаційного забезпечення

експлуатації автономного електротранспорту в безпілотній системі.

### Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

### Фінансування

Випробування проводиться в межах дослідницького проєкту 0126U003139 НФД України з теми "Наукові засади формування та управління людським потенціалом у мультипроєктному середовищі для забезпечення сталого розвитку програм відновлення України".

### Доступність даних

Рукопис не має пов'язаних даних.

### Використання засобів штучного інтелекту

Автори підтверджують, що не застосовували технології ШІ для написання цієї роботи.

### References

1. Про деякі питання використання транспортних засобів, оснащених електричними двигунами, та внесення змін до деяких законів України щодо подолання паливної залежності і розвитку електрозарядної інфраструктури та електричних транспортних засобів [Електронний ресурс]: Закон України від 24.02.2023 №2956-IX: станом на 01 серп.2025р. – URL: <https://zakon.rada.gov.ua/laws/show/2956-20#Text>

2. Андрусенко, С. І., Будниченко, В. Б., Подпісов, В. С. (2022), "Методика оцінки споживання енергії електробусом та параметрів тягової акумуляторної батареї в умовах експлуатації", *Автомобіль і електроніка. Сучасні технології*, № 22, с. 64–71. DOI: <https://doi.org/10.30977/VEIT.2022.22.08>
3. Андрусенко, С. І., Будниченко, В. Б., Подпісов, В. С. (2022), "Розробка методики визначення експлуатаційних витрат автобусів та тролейбусів з різними видами силових установок", *Науково-виробничий журнал «Автошляховик України» (Автомобільний транспорт)*, № 2 (270), с. 15–25. DOI: 10.33868/0365-8392-2022-2-270-15-25
4. Андрусенко, С. І., Будниченко, В. Б., Подпісов, В. С. (2021), "Математична модель енергетичної ємності тягової акумуляторної батареї", *Вісник Національного транспортного університету. Серія «Технічні науки». Науково-технічний збірник*, Вип. 3 (50), с. 3–10. DOI: <https://doi.org/10.33744/2308-6645-2021-3-50-003-010>
5. Андрусенко, С., Дембіцький, В., Будниченко, І., Дикий, В. (2024), "Дослідження доцільності використання електробусів на автобусних та тролейбусних маршрутах у містах", *Сучасні технології в машинобудуванні та транспорті*, № 1(22), с. 76–88. DOI: <https://doi.org/10.36910/automash.v1i22.1348>
6. Андрусенко, С. І., Будниченко, В. Б., Подпісов, В. С. (2021), "Оптимізація параметрів тягової акумуляторної батареї у тролейбусах із частковим автономним ходом", *Науково-виробничий журнал «Автошляховик України» (Автомобільний транспорт)*, № 3 (267), с. 15–21. DOI: <https://doi.org/10.33868/0365-8392-2021-3-267-15-21>
7. Дембіцький, В., Андрусенко, С., Будниченко, І., (2023), "Оцінка показників долі рекуперації енергії тролейбусами в умовах міського руху", *Сучасні технології в машинобудуванні на транспорті. Луцький національний технічний університет*, том 2, № 21. DOI: <https://doi.org/10.36910/automash.v2i21.1202>
8. Dembitskyi, V., Sitovskyi, O., Pavliuk, V. (2019), "Influence of a system "vehicle – driver – road – environment" on the energy efficiency of the vehicles with electric drive", *Proceedings of ICCPT 2019, May 28-29, 2019. – Tern. : TNTU, Scientific Publishing House "SciView"*, pp. 162–173.
9. Сорока, К. О. (2022), *Тягові розрахунки на основі аналізу процесів перетворення енергії : монографія / К.О.Сорока; Харк. Нац. Ун-т міськ. Госп-ва ім. О.М. Бекетова.-Харків :ХНУМГ ім О.М.Бекетова, 243 с.*
10. Кульбашна, Н., Лукашова, Н., Далека, В., Скуріхін, В.. (2025), "Шляхи підвищення енергоефективності електробусів за рахунок сучасних фотоелектричних технологій", *Комунальне господарство міст : наук.-техн. зб. Харків : ХНУГХ ім. О. М. Бекетова*, том 1. № 189, С. 52–60. DOI: <https://doi.org/10.33042/2522-1809-2025-1-189-52-60>
11. Kulbashna, N., Daleka, V., Lukashova, N., Furtat, S, (2025), "Energy Aspects of Forming Electric Bus Charging Schedules", *Lighting Engineering & Power Engineering*, Vol. 64, No. 2, pp. 153–162. DOI: <https://doi.org/10.33042/2079-424X.2025.64.2.03>
12. Dirks, N., Schiffer, M., Walther, G. (2022), "On the integration of battery electric buses into urban bus networks", *Transportation Research Part C: Emerging Technologies*, No. 139, 103628. DOI: <https://doi.org/10.1016/j.trc.2022.103628>
13. Heendeniya, C. B., Nespoli, L., Belliardi, M., Medici, V. (2023), "Exploring trade-offs in public bus electrification under stochastic conditions", *Energy Informatics*, No. 6 (1), 35, pp. 1–20. DOI: <https://doi.org/10.1186/s42162-023-00279-y>
14. Zhang, L., Zeng, Z., Gao, K. (2022), "A bi-level optimization framework for charging station design problem considering heterogeneous charging modes", *Journal of Intelligent and Connected Vehicles*, No. 5/1, pp. 8–16. DOI: <https://doi.org/10.1108/JICV-07-2021-0009>
15. Bie, Y., Ji, J., Wang, X., Qu, X. (2021), "Optimization of electric bus scheduling considering stochastic volatilities in trip travel time and energy consumption", *Computer-aided civil and infrastructure engineering*, No. 36 (12), pp. 1530–1548. DOI: <https://doi.org/10.1111/mice.12684>
16. Estrada, M., Mensión, J., Salicrú, M., Badía, H. (2022), "Charging operations in battery electric bus systems considering fleet size variability along the service", *Transportation Research Part C: Emerging Technologies*, 138, 103609. DOI: <https://doi.org/10.1016/j.trc.2022.103609>

Received (Надійшла) 27.03.2026

Accepted for publication (Прийнята до друку) 08.04.2026

Publication date (Дата публікації) 29.05.2026

#### Відомості про авторів / About the Authors

**Далека Василь Хомич** – доктор технічних наук, професор, Харківський національний університет міського господарства імені О. М. Бекетова, професор кафедри електричного транспорту, Харків, Україна;

**Vasyl Daleka** – Doctor of Technical Sciences, Professor, O. M. Beketov National University of Urban Economy, Professor of the Department of Electric Transport, Kharkiv, Ukraine;

e-mail: [dalekavf@ukr.net](mailto:dalekavf@ukr.net)

ORCID ID: <https://orcid.org/0000-0002-3074-5500>

**Бомбандьоров Юрій Петрович** – радник директора, Товариство відкритого типу "ПОЛІТЕХНОСЕРВІС", Бровари, Україна;

**Yuri Bombandorov** – Advisor to the Director, Open Society "POLYTECHNOSERVICE", Brovary, Ukraine;

e-mail: [bombandorov@gmail.com](mailto:bombandorov@gmail.com)

ORCID ID: <https://orcid.org/0009-0001-9848-1644>

**Метешкін Костянтин Олександрович** – доктор технічних наук, професор, Харківський національний університет міського господарства імені О. М. Бекетова, професор кафедри земельного адміністрування та геоінформаційних систем, Харків, Україна;

**Kostyantyn Meteshkin** – Doctor of Technical Sciences, Professor, O. M. Beketov National University of Urban Economy, Professor of the Department of Land Administration and Geoinformation Systems, Kharkiv, Ukraine;

e-mail: meteshkin@gmail.com

ORCID ID: <https://orcid.org/0000-0002-1170-2062>

**Андрусевич Анатолій Олександрович** – доктор технічних наук, професор, Харківський національний університет радіоелектроніки, професор кафедри комп'ютерно-інтегрованих технологій, автоматизації, робототехніки та безпекової інженерії, Харків, Україна;

**Anatoly Andrusevich** – Doctor of Technical Sciences, Professor, Kharkiv National University of Radioelectronics, Professor at the Department of Computer-Integrated Technologies, Automation, Robotics and Safety Engineering, Kharkiv, Ukraine;

e-mail: [anatolii.andrusevych@nure.ua](mailto:anatolii.andrusevych@nure.ua)

ORCID ID: <https://orcid.org/0000-0002-3142-635X>

**Чумаченко Ігор Володимирович** – доктор технічних наук, професор, Харківський національний університет міського господарства імені О. М. Бекетова, завідувач кафедри управління проектами в міському господарстві та будівництві, Харків, Україна;

**Igor Chumachenko** – Doctor of Technical Sciences, Professor, O. M. Beketov National University of Urban Economy in Kharkiv, Head of Project Management in Urban Management and Construction Department, Kharkiv, Ukraine;

e-mail: [ivchumachenko@gmail.com](mailto:ivchumachenko@gmail.com)

ORCID ID: <http://orcid.org/0000-0003-2312-2011>

## INFORMATION SUPPORT OF THE ENERGY CAPACITY OF ELECTRIC TRANSPORT IN PROJECTS AND PROGRAMS OF SUSTAINABLE DEVELOPMENT OF UKRAINE

The **subject matter** of the study is the models and methods for determining the applicability of the type series of traction batteries for electric transport in a multi-project environment to ensure autonomous operation of trolleybuses in the cities of Ukraine. **The purpose of the study** is to develop information support for the energy capacity of electric transport in sustainable development projects and programs and to substantiate the sufficiency of the type series of traction batteries (TAB) to ensure the needs of cities and sustainable development of Ukraine in electric transport. The article requires the following **tasks** to be performed: to assess the energy capacity of traction batteries, to collect information on the cities of Ukraine on the characteristics of trolleybus routes and to substantiate the feasibility of using trolleybuses with autonomous operation, taking into account urban planning and energy restrictions. To investigate and develop information support based on a model for determining the energy capacity and a methodology for determining the applicability of the TAB type series in a multi-project environment to ensure autonomous operation of trolleybuses in the cities of Ukraine. **Methods:** system, analytical, computer modeling, EASY WAY software product. Achieved results: Information support based on the model for determining the energy capacity and the method for determining the applicability of the traction battery type series are proposed. It is shown that the TAB type series is sufficient to ensure autonomous travel of trolleybuses for almost all cities of Ukraine, if its length does not exceed 15 km. **Conclusions:** According to the results of this study, it was determined that the existing TAB type series is capable of providing autonomous travel for a distance of at least 30 km, except for some cities of Ukraine, which is associated with the presence of energy carrier costs. In the future, it is proposed to improve the information support for the energy capacity of electric transport in sustainable development projects and programs and to conduct research to clarify energy carrier costs in cities of Ukraine where it is currently impossible to apply any of the TABs that are currently manufactured for autonomous travel lengths of more than 15 km.

**Keywords:** information support; project; vehicle; energy capacity; energy carrier; sustainable development.

### *Бібліографічні описи / Bibliographic descriptions*

Далека В. Х., Бомбандьоров Ю. П., Метешкін К. О., Андрусевич А. О., Чумаченко І. В. Інформаційне забезпечення енергетичної ємності електротранспорту в проектах і програмах сталого розвитку України. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 259–266. DOI: <https://doi.org/10.30837/0135-1710.2026.189.259>

Daleka, V., Bombandorov, Y., Meteshkin, K., Andrusevich, A., Chumachenko, I. (2026), "Information support of the energy capacity of electric transport in projects and programs of sustainable development of Ukraine", *Management Information System and Devices*, No. 2 (189), P. 259–266. DOI: <https://doi.org/10.30837/0135-1710.2026.189.259>



Dmytro Dymertsov

## COMPUTER SIMULATION OF GAS-DYNAMIC PROCESSES: A ONE-DIMENSIONAL APPROACH

**The relevance of this work** derives from the need for effective information technologies and software tools for modeling gas-dynamic processes in technical systems. Under modern conditions, the use of full-scale multidimensional models is associated with significant computational costs and complexity of software implementation, which limits their application in practical problems. In this context, one-dimensional models gain particular importance as a basis for developing optimized numerical solution algorithms that provide a reasonable balance between accuracy, computational efficiency, and implementation complexity. The **objective** of the study is to develop a one-dimensional mathematical model and a software package for numerical simulation of thermogasdynamic processes in flow elements of technical systems, as well as to verify and analyze the accuracy of the developed model. The **aim** of the study is to develop, implement, and verify a unified method for investigating gas-dynamic processes based on one-dimensional mathematical modeling using modern numerical analysis techniques. Special attention is paid to the development of efficient computational algorithms that take into account key physical processes (gas flow, heat and mass transfer, energy transformations, turbulence, and hydraulic losses), as well as ensure stability, convergence, and computational efficiency in software implementation. The **object** of the study is the processes of numerical modeling of gas-dynamic phenomena in flow elements of technical systems (pipelines, channels, heat exchangers, piston units), which are characterized by unsteady behavior, wave effects, and intensive heat and mass transfer processes. The **subject** of the study is methods based on a generalized one-dimensional mathematical model for calculating gas-dynamic processes in flow elements of technical systems. The results of the study showed that methods for improving numerical stability and reducing computational errors have been investigated and implemented, which ensured stable operation of algorithms in modeling unsteady gas-dynamic processes. The developed model is implemented as a software package designed for computational experiments and further analysis of simulation results. The **performed** verification, based on comparison with known analytical, numerical, and experimental data, confirmed the correctness, accuracy, and efficiency of the proposed algorithms. The **conclusion** confirms the feasibility of using one-dimensional computational models as an effective tool for computer modeling of complex physical processes. The proposed approach can be applied in the development of modern information technologies and in further scientific research in the field of gas dynamics and thermophysics, while the relative error does not exceed 5%.

**Keywords:** modeling; numerical methods; mathematical modeling; software implementation; computational systems; systems analysis; technologies; design.

### 1. Introduction

A current trend in computer science is the development of algorithmically efficient software tools for modeling gas-dynamic processes. The use of modern numerical methods and computational approaches makes it possible to create software systems capable of accurately reproducing complex non-stationary phenomena in technical systems. Mathematical modeling is a fundamental tool of computer science that formalizes complex physical processes into computational models suitable for implementation in a software environment. In today's context, the development of effective numerical methods and algorithms for modeling processes in technical and energy systems [1] is of particular relevance, where high calculation accuracy must be ensured with limited computational resources. This necessitates the creation of optimization models

focused on performance, scalability, and integration into application software packages.

In this work, as part of a project to develop a software package for calculating gas-dynamic processes in gas-jet acoustic generators, a unified computational model of gas-dynamic processes is proposed, focused on studying the filling and emptying of closed containers in technical systems. Particular attention is paid to the algorithmization of the model, the selection of discretization methods, and the implementation of efficient computational procedures. The developed model was verified by comparing the results of numerical simulations with known experimental and calculated data, which confirms its correctness and suitability for use as part of engineering analysis software tools. Further development of this approach opens up opportunities for creating intelligent decision-support systems, parameter optimization,

and the automation of the analysis of complex energy processes.

The aim of the study is to develop a generalized mathematical and computational model of thermogasdynamic processes in gas-jet cooling-heating acoustic generators with subsequent software implementation. The main focus is on constructing an efficient numerical solution algorithm that provides a comprehensive description of the interaction of wave, thermal, and hydrodynamic phenomena in a technical system.

The ultimate goal is to create a software-oriented complex for conducting numerical experiments, analyze operating modes, and optimize the performance parameters of gas-jet acoustic generators, which will improve the energy efficiency of the systems, enable adaptive adjustment of operating modes, and expand their application potential in modern information-controlled technical systems. The use of the developed software complex within the scope of this research will allow for the creation of a database of operational characteristics and results of numerical experiments for further analysis and systematization of the object's operating parameters.

---

## 2. Analysis of Recent Publications

In modern computer science research, mathematical modeling is a key tool for analyzing complex dynamic systems, particularly physical and gas-dynamic processes, as it allows for the formalization of equations into computational models and their implementation as software algorithms within software packages [2]. Computational models are based on numerical methods for solving nonlinear differential equations, including approaches such as the finite difference method, the finite volume method, and other algorithmic schemes that enable the construction of approximate solutions [3, 4].

Modern approaches in computer modeling include multidimensional (2D, 3D) computational models, which provide a detailed reproduction of the spatiotemporal dynamics of processes, but are characterized by high computational complexity and significant demands on resources and optimization of implementation algorithms [5]. In turn, one-dimensional models are viewed as a method for reducing the complexity of the problem, allowing for reduced computational costs, simplified algorithm structures, and efficient software implementation while maintaining

acceptable accuracy for analyzing general trends in system behavior [6]. Such approaches are widely used in the development of software for modeling technical systems, ensuring a balance between computational productivity and accuracy of results.

Within the framework of modern information technologies, hybrid and adaptive computational models are actively being developed, combining various numerical methods to improve algorithm stability and resource efficiency. Such approaches allow for the dynamic adaptation of the discretization of the computational domain depending on the complexity of the process, which is particularly important for optimizing the operation of software simulators of complex physical systems [4, 5].

A key role in such approaches is played by the algorithmic implementation of models, the selection of numerical methods, and the optimization of computational procedures. This includes the use of finite difference methods, finite volume methods, solvers for hyperbolic systems of equations, as well as their integration into software environments using modern programming languages and libraries (Python, C++, Lazarus, etc.). The result is the creation of flexible software packages that provide numerical modeling, visualization, and analysis of the dynamics of complex processes.

An analysis of contemporary scientific sources shows that computer modeling methods are a fundamental tool of information technology in scientific and engineering problems, as they combine high computational accuracy, algorithmic efficiency, and the ability to automate process analysis across various applied areas. Based on such approaches, it is possible to develop formalized mathematical and computational models of complex physical processes, which provides the foundation for the development of effective software systems for simulation and optimization.

---

## 3. Materials and Methods for Modeling Technical Systems

Methods for modeling technical systems involve representing one-dimensional unsteady gas flow as a formalized system of computational gas dynamics equations (see Equations 1–3) [6–8], which enables their algorithmization and subsequent software implementation within the framework of numerical methods for computer modeling. This approach allows

---

for the creation of computational models suitable for integration into software packages for simulating dynamic processes and forms the basis for performing

numerical analysis, optimizing system parameters, and supporting decision-making in information-controlled technical systems [9, 10].

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho \omega)}{\partial x} = 0 \quad (1)$$

$$\frac{\partial(\rho \omega)}{\partial t} + \frac{\partial(p + \rho \omega^2)}{\partial x} + \frac{\xi}{2d} \rho \omega^2 = 0 \quad (2)$$

$$\frac{\partial y}{\partial t} \left[ \rho \left( u + \frac{\omega^2}{2} \right) \right] + \frac{\partial y}{\partial x} \left[ \rho \omega \left( u + \frac{p}{\rho} + \frac{\omega^2}{2} \right) \right] - \left[ \frac{4}{d} q + \frac{\partial}{\partial x} \left( \alpha \frac{\partial T}{\partial x} \right) \right] = 0 \quad (3)$$

where  $\rho$  is the density of the working fluid;  $\omega$  is the flow rate;  $p$  is the pressure;  $\alpha$  is the thermal conductivity coefficient;  $\xi$  is the gas resistance coefficient;  $T$  is the temperature;  $d$  is the tube diameter.

The model for calculating technical processes is based on the use of linear approximations, which allows the system to be formalized as a set of elements with concentrated parameters, provided that its characteristic dimensions are significantly smaller than the wavelength. Within this approach, the computational description of processes boils down to the numerical solution of the equations of unsteady gas flow in vessels of constant cross-section (see Equations 4–6) with given boundary conditions, which are determined by the operational logic of the technical system [10]. Implementing this approach as a computational model allows for the reproduction of the main patterns of process dynamics while significantly reducing computational complexity and optimizing the use of software system resources.

$$\frac{\partial p}{\partial x} = \frac{\partial(\rho \omega)}{\partial t} + 2\alpha(\rho \omega) \quad (4)$$

$$-\frac{\partial p}{\partial t} = C_u^2 \frac{\partial(\rho \omega)}{\partial x} \quad (5)$$

$$C_u^2 = \frac{dp}{d\rho} = \frac{P_0}{\rho_0} \quad (6)$$

where  $\rho$  is the density of the working fluid;  $\omega$  is the flow velocity;  $p$  is the pressure.

Of particular interest in the context of computer modeling is the analysis of processes in high-pressure technical systems [10], which requires accounting for nonlinear effects in mathematical and computational models. Research on the influence of nonlinear components shows that as system parameters increase, the contribution of nonlinearity to the state equation becomes dominant, leading to significant deviations in the numerical representation of physical characteristics

compared to the contribution of nonlinear terms in the equations of motion and continuity.

To solve such problems in software packages, an iterative (successive approximation) method is used, which is implemented within the system of equations via a reduced computational subsystem (see Equations 7–8). This algorithmic approach ensures a step-by-step refinement of the numerical solution and allows for the integration of the influence of key physical parameters into the process of computer modeling of gas system dynamics [9, 10].

$$\frac{1}{\rho} \frac{\partial \rho}{\partial x} + \frac{\partial \omega}{\partial t} + \omega \frac{\partial \omega}{\partial x} + \xi \frac{\omega^2}{2d} = 0 \quad (7)$$

$$\frac{\partial \rho}{\partial t} + \rho \frac{\partial \omega}{\partial x} + \omega \frac{\partial \rho}{\partial x} = 0 \quad (8)$$

where  $\rho$  is the density of the working fluid;  $\omega$  is the flow velocity;  $p$  is the pressure;  $\xi$  is the gas resistance coefficient.

#### 4. Structure and implementation of the composite one-dimensional model

Based on the methods discussed, a mathematical model is developed that implements the ideal gas equation of state (see Equation 9) and provides a formalized description of the key parameters of the working medium as it moves and interacts with the system's characteristics. Within the scope of computer modeling, this model is used as a basis for developing computational algorithms for the analysis and simulation of dynamic processes in technical systems.

$$p = \rho RT \quad (9)$$

where  $p$  – is the pressure;  $\rho$  – is the density;  $T$  – is the temperature;  $R$  – is the gas constant.

The implementation of this model involves a number of assumptions, according to which the working medium is treated as a system of material points,

and the internal energy is determined solely by temperature. This formalized representation allows for a significant reduction in the complexity of the mathematical description of processes while maintaining acceptable accuracy over a wide range of material parameters, which is important for the subsequent software implementation of computational models.

To numerically describe the dynamics of the processes, a one-dimensional system of gas dynamics differential equations is used (see Equations 1–3), while thermal processes are modeled using a one-dimensional unsteady heat conduction equation (see Equation 10). Within the framework of computer modeling, the technical system is considered as a thin-walled cylinder with one end closed, which allows the spatial problem to be reduced to a one-dimensional formulation and simplifies its numerical implementation in the software environment.

$$\frac{\partial T}{\partial t} = a \frac{\partial^2 T_c}{\partial x^2} + \frac{1}{\delta C_c \rho_c} \left[ \alpha_g (T_g - T_c) + \alpha_o (T_c - T_o) \right], \quad (10)$$

where  $\alpha$  is the heat transfer coefficient;  $\delta$  is the tube wall thickness;  $C_c$  is the heat capacity of the tube wall;  $\alpha_g$  – thermal conductivity coefficient;  $\rho_c$  – density of the tube wall;  $T_g$  – gas temperature;  $T_c$  – tube wall temperature;  $T_o$  – ambient temperature.

## 5. Methods for the numerical solution of gas dynamics systems

Computational algorithms that take into account the discretization structure and the specific characteristics of each model element are used for the numerical implementation of the system of equations. The state parameters of the working medium in the pipe are determined using the finite difference method, implemented as a discrete computational grid with a time – density at the new time step

$$\rho_{i+1/2}^j = \rho_{i+1/2}^{j-1} - \frac{\Delta t}{h} \left[ (\rho \omega)_{i+1}^j - (\rho \omega)_i^{j-1} \right]; \quad (11)$$

– velocity at the new time step

$$\omega_{i+1/2}^j = \left( (\rho \omega)_{i+1/2}^{j-1} - \frac{\Delta t}{h} \left[ (p + \rho \omega^2)_{i+1}^j - (p + \rho \omega^2)_i^{j-1} \right] + \frac{\Delta t}{2d} (\xi \rho \omega^2)_{i+1/2}^{j-1} \right) / \rho_{i+1/2}^j; \quad (12)$$

– pressure on the new time layer

$$p_{i+1/2}^j = (k-1) \left\{ \left[ \frac{p}{k-1} + \frac{\rho \omega^2}{2} \right]_{i+1/2}^{j-1} - \frac{\Delta t}{h} \left[ \left[ \omega \left( \frac{p}{k-1} + \frac{\rho \omega^2}{2} \right) \right]_{i+1}^j - \left[ \omega \left( \frac{p}{k-1} + \frac{\rho \omega^2}{2} \right) \right]_i^{j-1} \right] \right\} - \left[ \frac{\rho \omega^2}{2} \right]_{i+1/2}^j + \frac{4\alpha}{d} \Delta t (T_{i+1/2}^{j-1} - T_i^{j-1}) \quad (13)$$

step  $t$  and a spatial step  $h$ , as well as an additional offset grid with a step size of  $h/2$  (see Figs. 1, 2). Based on this discretization, an explicit central-difference scheme is constructed, which provides first-order accuracy in time and second-order accuracy in the spatial variable  $h$ , which is a typical approach in the numerical modeling of dynamic processes.

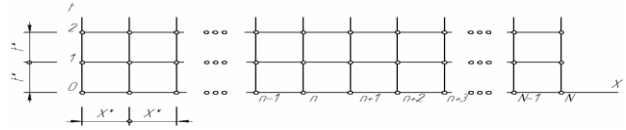


Fig. 1. Flat grid for process calculation:  
 $\Delta t$  – time step;  $x$  – spatial step

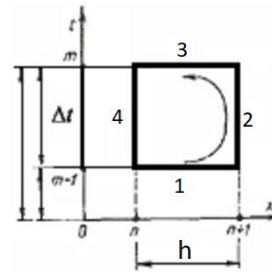


Fig. 2. Computational cell of a planar grid:  
where  $\Delta t$  is the time step;  $h$  is the spatial step

When transitioning to a new time step, numerical discontinuities and jumps in model parameters may occur at the nodes of a staggered computational grid. To eliminate such instabilities and correctly determine gas-dynamic quantities at intermediate points  $h/2$ , an algorithm for solving the arbitrary discontinuity breakdown problem is applied [11, 12]. The use of this approach within the software implementation ensures the numerical stability of the computational process and improves the accuracy of reproducing gas flow dynamics in the simulation software package:

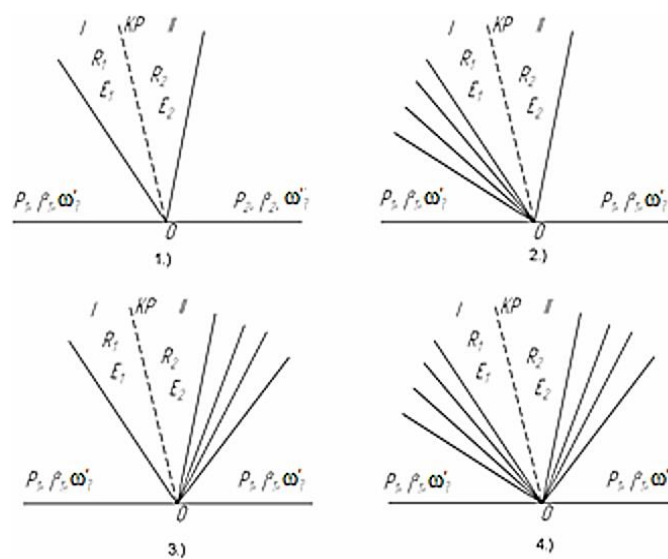
Gas parameter jumps may occur between nodes in the computational grid within a pipe, leading to calculation errors and potential failure of the finite difference scheme.

To eliminate numerical instabilities at the nodes of the computational grid, an approach based on solving the problem of the decay of an arbitrary discontinuity is employed. Within this framework, the model considers two semi-infinite regions of the working medium, in each of which the gas parameters are constant at the initial time but may differ from one another. This formalization allows for a correct description of the evolution of the distribution of physical properties after the occurrence of

a discontinuity and ensures the stability and consistency of the numerical algorithm during the computer simulation process.

### The problem of an arbitrary discontinuity

When two gas media come into contact, the nature of their interaction is determined by the values of the main parameters, which can result in the formation of various flow patterns (see Fig. 3). On both sides of the initial separation point (point 0), the gases are in states with constant parameters: on the left –  $p_1 \rho_1 \omega_1$ , on the right –  $p_2 \rho_2 \omega_2$ .



- 1 – shock waves propagate to the left and right;
- 2 – a rarefaction wave propagates to the left, a shock wave to the right;
- 3 – a shock wave propagates to the left, a discharge wave to the right;
- 4 – a discharge wave propagates to the left and right

**Fig. 3.** Diagram of the location of the discontinuity: where

I – is the space on the left; II – is the space on the right; KP – is the contact gap;  
 $\rho$  is the density of the working fluid;  $\omega$  is the flow velocity;  $p$  is the pressure

At the boundary of the flow, which may manifest as a shock wave or a rarefaction wave, a change in flow parameters occurs. In this region, a redistribution of physical quantities takes place; in particular, the values of gas pressure and velocity change to new, consistent values that ensure compliance with the laws of conservation of mass, momentum, and energy [13, 14].

The calculation is performed in the following sequence:

$$a = \sqrt{k \frac{p_i + p_{i+1}}{2} \frac{\rho_i + \rho_{i+1}}{2}} \quad (14)$$

$$P_k = \frac{p_i + p_{i+1}}{2} - a \frac{\omega_i + \omega_{i+1}}{2} \quad (15)$$

$$U_k = \frac{\omega_i + \omega_{i+1}}{2} - \frac{p_i - p_{i+1}}{2a} \quad (16)$$

where  $a$  – is the mass velocity of sound;  $U_k$  is the contact discontinuity velocity;  $P_k$  is the contact discontinuity pressure.

The wave velocity propagating to the left and right of the discontinuity (see Equations 17, 18):

$$D_I = \omega_i - a/\rho_i \quad (17)$$

$$D_{II} = \omega_{i+1} + a/\rho_{i+1} \quad (18)$$



Depending on the sign of the speeds  $D_I$  and  $D_{II}$ , the following options are possible:

$$D_I > 0, D_{II} > 0, \text{ then } U = \omega_i, P = p_i, R = \rho_i \quad (19)$$

$$D_I < 0, D_{II} < 0, \text{ then } U = \omega_{i+1}, P = p_{i+1}, R = \rho_{i+1} \quad (20)$$

$$D_I < 0, D_{II} > 0, U_k > 0, \text{ then } U = U_k, R = \rho_i \frac{(k+1)P_k + (k-1)p_i}{((k-1)P_k + (k+1)p_i)} \quad (21)$$

$$D_I < 0, D_{II} > 0, U_k < 0, \text{ then } U = U_k, R = \rho_i \frac{(k+1)P_k + (k-1)p_{i+1}}{(k-1)P_k + (k+1)p_{i+1}} \quad (22)$$

## 6. Analysis and verification of modeling algorithms

The developed unified one-dimensional computational model was implemented using the Lazarus Free Pascal programming language [15] as a software package

for the numerical calculation of the operating parameters of gas-jet acoustic generators (see Fig. 4). The chosen language ensures high computational speed, flexibility in implementing numerical algorithms, and the ability to create an intuitive graphical user interface.

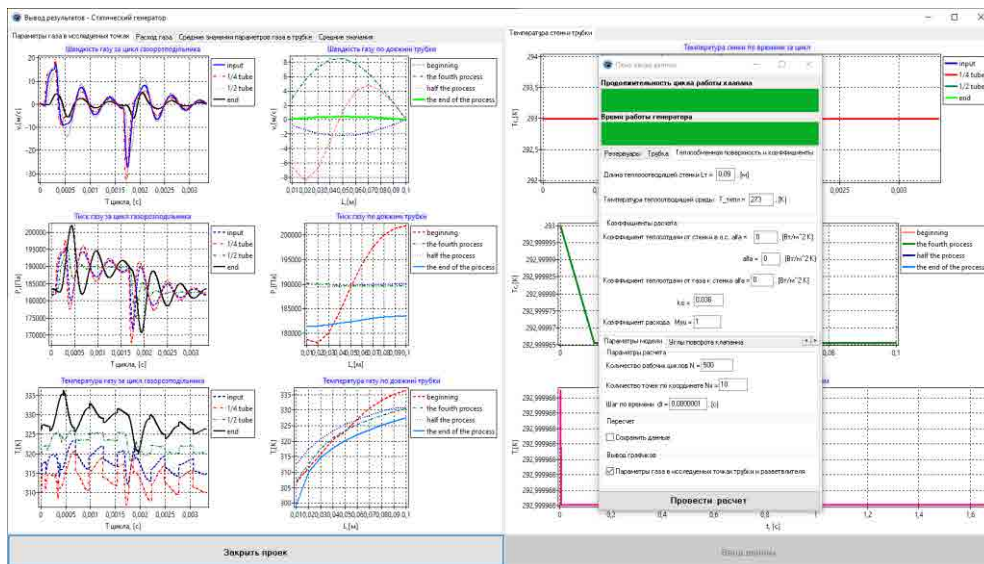


Fig. 4. Software package for calculation

The developed software enables numerical simulation of the thermo-gasdynamic processes occurring in the system's working elements, specifically the study of wave processes associated with the propagation of shock waves and rarefaction waves, as well as thermal processes involving heat transfer between the working medium and the walls of the receiving tube. This approach enables the integration of the computational model into software packages for the analysis, optimization, and verification of gas-jet acoustic generators.

A key element of such a system is the data storage subsystem, implemented using modern database technologies. The use of a specialized data warehouse allows for the accumulation of information on geometric parameters, operating modes, performance

characteristics, and numerical simulation results for a set of objects of the same type. Given the structure of the subject domain, it is advisable to use a relational data model, which ensures integrity, consistency, and efficient query execution. At the same time, for tasks involving complex relationships between objects, network or hybrid approaches to data organization may be used.

Integrating the computational model with the database enables the automated generation of input parameters for modeling based on accumulated operational data, as well as the storage of calculation results for further analysis. Periodic updates of equipment status data and the re-execution of numerical calculations allow for the implementation of mechanisms to predict changes in technical characteristics over time.

Based on this approach, it becomes possible to model equipment aging processes, assess parameter degradation, and predict remaining service life. This, in turn, creates the prerequisites for building decision support systems for maintenance and repair, including the optimization of preventive maintenance schedules. Thus, the computational model, combined with an information subsystem for data storage and processing, transforms into a tool for modeling the life cycle of technical objects.

To confirm the reliability of the results, the model was verified by comparing it with known literature data [16, 17] prior to conducting numerical experiments. A comparison of the simulation results with experimental and literature data showed that the relative error does not exceed 5% (see Fig. 5). This indicates a sufficient level of model adequacy for describing the physical processes under study in flow elements of technical systems.

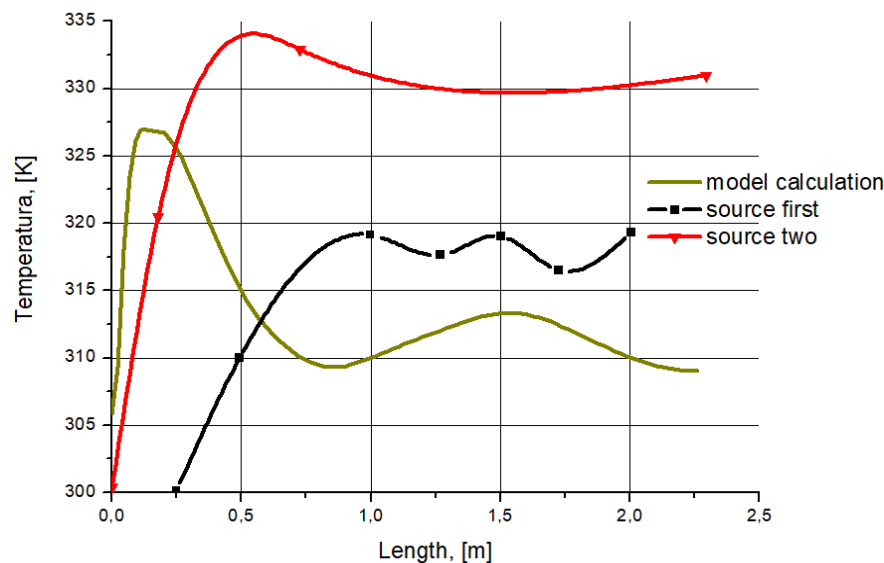


Fig. 5. Testing the model using existing results

Further development of the presented computational model should be directed not only toward improving its physical accuracy but also toward expanding the system's information component. In particular, the creation of an integrated software and information complex using distributed computing technologies, cloud services, and digital twin elements is promising, as it will enable real-time monitoring, modeling, and forecasting of operations.

## 7. Conclusions

As a result of the study, a unified one-dimensional computational model of gas-dynamic processes was developed, which accounts for key physical mechanisms, including gas flow, energy conversion, heat and mass transfer, turbulent effects, and local hydraulic losses, and enables the software implementation of numerical algorithms for engineering modeling. The model is based on the fundamental laws of conservation of mass, momentum, and energy, ensuring its physical validity and suitability for application to a wide class of gas-dynamic problems. During development, the structure of the

computational model was defined, the selection of initial and boundary conditions was justified, and an approach to discretizing the computational domain was implemented, ensuring numerical stability, convergence, and computational efficiency.

The proposed model has been implemented as a software package that enables engineering calculations, analysis of operating modes, and parametric optimization of gas-dynamic systems. Verification of the model's adequacy through comparison with experimental and literature data confirmed a high degree of agreement between the results (the error does not exceed 5%), which indicates the correctness of the assumptions made and the practical applicability of the developed approach.

Further analysis showed that the effective use of the developed model should be considered not only in the context of numerical engineering modeling but also as part of information technology for supporting the life cycle of technical systems. In this case, the model is integrated into an information and analytical system that ensures the collection, storage, and processing of data on technical objects with individual parameters.

A key element of such information technology is a data management subsystem, implemented on the basis of databases or data warehouses. The choice of storage architecture (relational, hierarchical, or networked) is determined by the structure of the subject domain and the nature of the relationships between object parameters. The use of a centralized data warehouse enables the systematic updating of object parameters and their subsequent use in computational procedures.

Integrating the database with the computational model enables the implementation of a mechanism for recalculating operational characteristics based on current input data, which creates a foundation for predicting changes in the technical condition of equipment. This approach enables the modeling of aging processes, the assessment of parameter degradation, and the planning of maintenance interventions.

The results obtained confirm the feasibility of using one-dimensional computational models as a component of modern information technologies for engineering analysis and the management of the life cycle of technical systems. The developed approach can be used to create integrated software and information systems

for decision support in the operation and optimization of gas-dynamic installations.

---

### Conflict of Interest

The author declares that there is no conflict of interest regarding this study, including financial, personal, authorship, or other conflicts that could influence the study and its results presented in this article.

---

### Funding

The work was performed within the framework of the institution's economic-contractual project No. 0126U001791.

---

### Data Availability

The manuscript has no associated data.

---

### Use of Artificial Intelligence

The author confirms that no artificial intelligence technologies were used in the creation of this work.

## References

1. Потапов, В., Димерцов, Д. (2025), "Перспективи застосування газоструменевих акустичних генераторів у якості вторинних джерел енергії", *Refrigeration Engineering and Technology*, Vol.61, No 2, pp. 128–137, DOI: <https://doi.org/10.15673/ret.v61i2.3178>
  2. Вислоух, С., Волошко, О., Тимчик, Г., Філіппова, М. (2021), "Комп'ютерне моделювання процесів та систем. Чисельні методи", Київ: КПІ ім. Ігоря Сікорського.
  3. Dymertsov, D. (2025), "Modeling of Gas-Dynamic Processes of Wave Low-Temperature Heat Generators Dynamic Gas Distribution", In: Tonkonogyi, V., Ivanov, V., Trojanowska, J., Oborskyi, G. (eds) *Advanced Manufacturing Processes VI. Interpartner 2024. Lecture Notes in Mechanical Engineering*. Springer, Cham. DOI: [https://doi.org/10.1007/978-3-031-82746-4\\_68](https://doi.org/10.1007/978-3-031-82746-4_68)
  4. Zeidan, D., Hidalgo, A. (2025), "Computational Fluid Dynamics: Novel Numerical and Computational Approaches". Cham: Springer Nature. DOI: <https://doi.org/10.1007/978-981-97-8152-2>
  5. Nadamani, M. N., Shadloo, M. S., Dbouk, T. (2025), "A Review Theoretical and Computational Fluid Dynamics Modeling of Coupled Heat and Mass Transfer", *Energies*, Vol. 18, No. 24. Article 6418, DOI: <https://doi.org/10.3390/en18246418>
  6. Anderson, J. D. (2003), "Modern Compressible Flow with Historical Perspective", New York: McGraw-Hill, McGraw-Hill Series in Aeronautical and Aerospace Engineering.
  7. Thompson, P. A. (1988), "Compressible Fluid Dynamics". New York: McGraw-Hill, p. 665.
  8. Chaudhry, M. H. (2014), "Pipes and Flow Systems", *Applied Hydraulic Transients*. New York: Springer. DOI: <https://doi.org/10.1007/978-1-4614-8538-4>
  9. Hirsch, C. (2007), "Numerical Computation of Internal and External Flows", Oxford: Butterworth-Heinemann.
  10. Toro, E. F. (2009), "Riemann Solvers and Numerical Methods for Fluid Dynamics", Dordrecht: Springer. DOI: <https://doi.org/10.1007/b79761>
  11. Chaudhry, M. H. (1987), "Fluid Transients in Systems", Englewood Cliffs: Prentice Hall.
  12. Zucrow, M. J., Hoffman, J. D. (1976), "Gas Dynamics", New York: Wiley.
  13. Hirsch, C. (2007), "Numerical Computation of Internal and External Flows", Butterworth-Heinemann.
  14. Anderson, J. D. (2003), "Modern Compressible Flow: With Historical Perspective", McGraw-Hill.
  15. Dymertsov, D. (2026), "Development of software for modeling gas-dynamic processes in wave cold generators", *Simon Kuznets Kharkiv National University of Economics*, №0126U001791, <https://nrat.ukrintei.ua/en/searchdoc/0126U001791/#>
  16. Bauer, C., Hauser, M., Haidn, O. (2015), "Investigation of stabilization effects in Hartmann-Sprenger-Tubes", In *Proceedings of the 30th International Symposium on Space Technology and Science*, Kobe-Hyogo, Japan, 4–9 July.
-

17. Li, B., Hu, G., Zhou, Z. (2007), "Numerical simulation of flow in Hartmann resonance tube and flow in ultrasonic gas atomizer", Applied Mathematics and Mechanics (English Edition), Vol. 28.

Received (Надійшла) 05.03.2026

Accepted for publication (Прийнята до друку) 30.04.2026

Publication date (Дата публікації) 29.05.2026

#### Відомості про авторів / About the Authors

Димерцов Дмитро Олександрович – Харківський національний економічний університет ім. С. Кузнеця, старший викладач кафедри здорового способу життя, технологій і безпеки життєдіяльності, Харків, Україна;

Dmytro Dymertsov – Simon Kuznets Kharkiv National University of Economics, Senior Lecturer of the Department of Healthy Lifestyle, Technologies and Life Safety, Kharkiv, Ukraine;

e-mail: dmytro.dymertsov@hneu.net

ORCID ID: <https://orcid.org/0009-0000-4894-3495>

## КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ СКЛАДНИХ ПРОЦЕСІВ НА ОСНОВІ ОДНОВИМІРНОГО ПІДХОДУ

**Актуальність роботи** зумовлена необхідністю в ефективних інформаційних технологіях і програмних засобах для моделювання газодинамічних процесів у технічних системах. Використання повномасштабних багатовимірних моделей обмежено значними обчислювальними витратами та складністю програмної реалізації, що знижує їх практичну ефективність. У цьому контексті одновимірні моделі є основою для побудови оптимізаційних обчислювальних алгоритмів, які забезпечують раціональний баланс між точністю, швидкістю та складністю реалізації. Додаткову актуальність визначає перехід до інтегрованих інформаційно-аналітичних систем для класів технічних об'єктів зі змінними параметрами, що потребує застосування баз даних і сховищ інформації для зберігання та оновлення експлуатаційних характеристик. **Об'єктом дослідження** є процеси чисельного моделювання газодинамічних явищ у проточних елементах технічних систем, а **предметом** – методи на основі узагальненої одновимірної математичної моделі розрахунку газодинамічних процесів у проточних елементах технічних систем. **Мета дослідження** – розроблення узагальненої математичної та обчислювальної моделі термогазодинамічних процесів у газоструменевих холодильно-нагрівальних акустичних генераторах із подальшою програмною реалізацією. **Завдання:** розроблення одновимірної математичної моделі та програмного комплексу для чисельного моделювання термогазодинамічних процесів у проточних елементах технічних систем, їх верифікація та аналіз точності; формування інформаційної технології підтримки життєвого циклу технічних об'єктів на основі інтеграції обчислювальної моделі з базами даних для зберігання, оброблення й оновлення параметрів із подальшим використанням для прогнозування технічного стану й аналізу експлуатаційних характеристик обладнання. Особливу увагу приділено формуванню ефективних обчислювальних алгоритмів, що беруть до уваги ключові фізичні процеси (рух газового середовища, тепло- й масообмін, енергетичні перетворення, турбулентність і гідравлічні втрати), а також забезпечують стійкість, збіжність і обчислювальну ефективність під час реалізації в програмному середовищі. **Результати дослідження** продемонстрували, що застосування методів підвищення чисельної стійкості та зменшення похибок обчислень забезпечує стабільну роботу програмних алгоритмів у моделюванні нестационарних газодинамічних процесів. Запропонована модель реалізована у вигляді програмного комплексу, як складник інформаційної технології чисельного моделювання й аналізу даних. Проведена верифікація на основі порівняння з аналітичними, чисельними й експериментальними показниками підтвердила коректність і високу точність запропонованих алгоритмів (похибка не перевищує 5%). **Висновки.** Підтверджено доцільність використання одновимірних обчислювальних моделей як базового елемента сучасних інформаційних технологій комп'ютерного моделювання складних фізичних процесів і створення програмно-інформаційних систем підтримки інженерного аналізу.

**Ключові слова:** моделювання; чисельні методи; математичне моделювання; програмна реалізація; обчислювальні системи; системний аналіз; технології; проектування; бази даних.

#### Бібліографічні описи / Bibliographic descriptions

Димерцов Д. О. Комп'ютерне моделювання складних процесів на основі одновимірної підходу. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 267–275. DOI: <https://doi.org/10.30837/0135-1710.2026.189.267>

Dymertsov, D. (2026), "Computer simulation of gas-dynamic processes: a one-dimensional approach", *Management Information System and Devices*, No. 2 (189), P. 267–275. DOI: <https://doi.org/10.30837/0135-1710.2026.189.267>

Євтушенко О. С., Заковоротний О. Ю.

## АПРОКСИМАЦІЯ КРИТИЧНОЇ ШВИДКОСТІ ПОЇЗДА МЕТОДАМИ МАШИННОГО НАВЧАННЯ

**Предметом дослідження** є ефективність методів машинного навчання для апроксимації критичної швидкості в умовах нелінійних залежностей та обмеженого обсягу даних. **Мета статті** – порівняльне оцінювання трьох методів навчання з учителем щодо їх здатності апроксимувати критичну швидкість залізничного рухомого складу в умовах сильно нелінійних залежностей параметрів і обмеженого обсягу даних способом проведення низки експериментів і аналізу їх результатів. З огляду на окреслену мету необхідно виконати такі **завдання**: сформувати датасети малого й великого обсягу на основі математичної моделі криволінійного руху поїзда; реалізувати й навчити моделі багатопарового перцептрона, регресії на основі гауссівських процесів і регресії на опорних векторах у середовищі MATLAB; порівняти й оцінити точність апроксимації за метриками RMSE й  $R^2$  на обох датасетах; визначити модель, що забезпечує найкращий компроміс між точністю та масштабованістю; обґрунтувати вибір оптимального методу для впровадження в системах підтримки прийняття рішень. **Методи**. Для генерації датасетів застосовано математичну модель криволінійного руху поїзда. Якість апроксимації оцінювалась за метриками RMSE й  $R^2$ . Усі експерименти реалізовано в середовищі MATLAB. **Досягнуті результати**: моделі багатопарового перцептрона й регресії на основі гауссівських процесів продемонстрували високу точність прогнозування як на малому, так і на великому обсязі даних, до того ж перцептрон продемонстрував кращу масштабованість в умовах зростання обсягу навчальних даних, ніж регресія гауссівських процесів. **Висновки**. На основі порівняльного аналізу доцільно використовувати MLP як ключову модель для визначення критичної швидкості в систему підтримки прийняття рішень машиніста й діагностики залізничного транспорту. Упровадження запропонованого підходу дасть змогу підвищити безпеку руху поїздів в умовах реальної експлуатації, скоротити витрати на технічне обслуговування й забезпечити основу для розроблення адаптивних бортових систем контролю швидкості.

**Ключові слова**: критична швидкість; коливання рухомого складу; багатопаровий перцептрон; регресії на основі гауссівських процесів; регресія на опорних векторах; апроксимація; машинне навчання; система підтримки прийняття рішень машиніста; криволінійний рух.

### 1. Вступ

Дослідження коливальних процесів і нестійкості руху залізничних транспортних засобів, обладнаних традиційними колісними парами з жорсткою посадкою, набуває дедалі більшої актуальності, що зумовлено необхідністю компенсувати погіршення технічного стану рухомого складу за допомогою комплексної оптимізації його руху.

Найменша швидкість, за якої виникають стійкі поперечні коливання системи "колісна пара – візок", називається критичною швидкістю. Визначення критичної швидкості є важливим завданням у розробленні та проектуванні залізничних транспортних засобів. Якщо транспортний засіб рухається зі швидкостями, що перевищують критичну, динамічні процеси, спричинені збуреннями, можуть не затухати, а навпаки, призводити до стійких коливань. Хоча такий рух може бути стійким у математичному сенсі, його прийнято називати нестійким рухом. Оскільки

інтенсивні коливання можуть спричиняти значні бічні сили між колесом і рейкою та, як наслідок, підвищувати ризик зміщення колії або навіть сходу рухомого складу, такої поведінки необхідно уникати за нормальних умов експлуатації [1, 2]. Інакше кажучи, критична швидкість є верхньою межею допустимої швидкості руху залізничного транспортного засобу.

Традиційні аналітичні та чисельні методи визначення критичної швидкості потребують детальних математичних моделей рухомого складу з великою кількістю параметрів і є обчислювально витратними. Вони, як правило, не беруть до уваги ефекти деградації (зношування профілю коліс, гістерезис підвіски) і стохастичних збурень (стики рейок, нерівності колії), що призводить до надто оптимістичних оцінок запасу стійкості й робить їх непридатними для застосування в системах реального часу [3, 4]. У зв'язку з цим виникає потреба в розробленні альтернативних підходів, здатних забезпечити швидко й достатньо точну апроксимацію



критичної швидкості на основі обмеженого набору вимірюваних параметрів.

З огляду на постійне погіршення стану залізничної інфраструктури й рухомого складу, що призводить до динамічного зниження безпечної швидкості руху поїздів у режимі реального часу, здатність оперативно й надійно оцінювати критичну швидкість стала не тільки перевагою, а нагальною операційною необхідністю для запобігання нестійкості, спричиненій коливаннями, та зниження ризику сходу рухомого складу.

Практичне значення дослідження полягає у впровадженні отриманих результатів у бортові системи підтримки прийняття рішень машиніста, що сприятиме підвищенню безпеки залізничного руху, зниженню ризику сходу рухомого складу й оптимізації витрат на технічне обслуговування інфраструктури й рухомого складу в умовах реальної експлуатації.

## 2. Аналіз сучасних наукових публікацій. Визначення проблеми дослідження

Питанню критичної швидкості присвячено значну кількість досліджень. Перевищення критичної швидкості визначається як одна з потенційних причин сходу поїздів з колії [3]. Вивчено вплив колії та її основи на значення критичної швидкості [6–8], а також можливість установлення цього параметра за допомогою аналізу наслідків, які спричиняє рух на критичній швидкості для колії [7, 9]. Зазвичай розглядаються два обчислювальні підходи до оцінювання критичної швидкості: метод відстеження траєкторії, що забезпечує автоматизований розрахунок, і метод "грубого перебору". Проте через залежність від періодичності розв'язків метод відстеження траєкторії обмежений винятково суворо-періодичними режимами руху. Відповідно, метод "грубого перебору" є більш придатним для оцінювання критичної швидкості вздовж складних (неперіодичних) траєкторій [4].

Обчислення критичної швидкості є нетривіальним завданням. Це значення залежить від значної кількості змінних, постійних параметрів і характеристик руху поїзда, чимало з яких впливають на критичну швидкість вкрай нелінійно. З огляду на складність динаміки залізничних транспортних засобів безпосереднє обчислення критичної швидкості на основі наявної математичної моделі

динаміки криволінійного руху [10] є практично неможливим. Класичні методи оцінювання критичної швидкості ґрунтуються на лінеаризованому аналізі стійкості або часовому моделюванні з поступовим збільшенням швидкості. Однак окреслені підходи передбачають ідеалізовану геометрію колії, не беруть до уваги ефекти деградації (наприклад, зношування профілю коліс, гістерезису підвіски) й потребують високоточних моделей, що є обчислювально непридатними для реалізації в режимі реального часу. Крім того, вони часто не зважають на стохастичні збурення (наприклад, стики рейок, нерівності колії), що призводить до надто оптимістичних оцінок запасу стійкості.

Методи машинного навчання є перспективним інструментом для розв'язання подібних задач регресії в галузі залізничної динаміки, оскільки вони здатні виявляти складні нелінійні залежності безпосередньо з даних, не потребуючи явного задання фізичних рівнянь руху [11–13]. Зокрема нейронні мережі й методи на основі ядрових функцій уже застосовувалися для прогнозування динамічних навантажень у системі "колесо – рейка" [14, 15]. Проте порівняльний аналіз ефективності різних алгоритмів машинного навчання саме для задачі апроксимації критичної швидкості в умовах обмеженого обсягу даних залишається недостатньо дослідженим, що й зумовлює актуальність цієї роботи.

## 3. Мета й завдання дослідження

**Метою статті** є порівняльне оцінювання трьох методів навчання з учителем: багат шарового перцептрона, регресії на основі гауссівських процесів і регресії на опорних векторах – щодо їх здатності апроксимувати критичну швидкість залізничного рухомого складу в умовах сильно нелінійних залежностей параметрів і обмеженого обсягу даних способом проведення низки експериментів і аналізу їх результатів.

Для досягнення поставленої мети необхідно виконати низку завдань.

1. На основі наявної математичної моделі криволінійного руху поїзда сформувані два датасети – малого обсягу й великого обсягу, що відображають нелінійну залежність критичної швидкості від конусності колісної пари  $\gamma$ , номінального радіуса кочення  $r_0$  та кутової жорсткості  $C$ .

2. Реалізувати й навчити три різні моделі машинного навчання в середовищі MATLAB і забезпечити відтворюваність експериментів.

3. Кількісно оцінити якість апроксимації кожної моделі за метриками середньоквадратичної похибки (RMSE) та коефіцієнта детермінації ( $R^2$ ) окремо для малого й великого датасетів.

4. Дослідити масштабованість кожного методу способом порівняння точності на малому й великому датасетах.

5. На основі результатів порівняльного аналізу обґрунтувати вибір оптимального методу й визначити перспективи його впровадження як ключового модуля апроксимації критичної швидкості в системі підтримки прийняття рішень машиніста й у складі повної моделі криволінійного руху поїзда.

Найбільш ефективний метод надалі буде впроваджено для розроблення модуля оцінювання критичної швидкості в складі системи підтримки прийняття рішень машиніста і як компонента моделі криволінійного руху поїзда.

#### 4. Матеріали й методи дослідження

Розглянемо поїзд, що рухається криволінійною ділянкою колії постійного радіуса кривизни  $R = 500$ . У межах наявної моделі криволінійного руху поїзда для критичної швидкості  $V_c$  застосовуються такі залежності: від конусності колісної пари  $\gamma$ , від номінального радіуса кочення  $r_0$ , від кутової жорсткості зв'язків візка з кузовом  $C$ . Ці залежності є нелінійними й можуть бути умовно подані таким чином:

$$V_c \propto r_0 C / \sqrt{\gamma}.$$

З наведеного співвідношення випливає, що критична швидкість прямо пропорційна номінальному радіусу кочення  $r_0$  і кутовій жорсткості  $C$  й обернено пропорційна кореню квадратному з конусності колісної пари  $\gamma$ . На основі цих залежностей апроксимація критичної швидкості є можливою.

Розглянемо інструменти, придатні для апроксимації критичної швидкості.

**Багатошаровий перцептрон (Multilayer Perceptron – MLP)** – клас нейронних мереж прямого поширення, що містять щонайменше три шари: вхідний, один або кілька прихованих шарів

і вихідний. За винятком вхідного шару, всі нейрони використовують нелінійну функцію активації.

**Регресія на основі гауссівських процесів (Gaussian Process Regression – GPR)** – потужний і гнучкий непараметричний метод регресії, що застосовується в машинному навчанні й статистиці. Він особливо корисний для задач з неперервними показниками, де зв'язок між вхідними змінними та вихідним значенням є невідомим або надзвичайно складним. GPR – це баєсівський підхід, який кількісно оцінює невизначеність у прогнозах, що робить його цінним для задач оптимізації та прогнозування часових рядів. Він ґрунтується на концепції гауссівського процесу – набору випадкових змінних, будь-яка скінченна підмножина яких має спільний гауссівський розподіл.

**Регресія на опорних векторах (Support Vector Regression – SVR)** – різновид методів опорних векторів (SVM), що використовується для задач регресії. Її мета полягає в знаходженні функції, яка найкраще прогнозує неперервне вихідне значення для заданого входу. SVR може застосовувати як лінійні, так і нелінійні ядра. Лінійне ядро обчислює простий скалярний добуток двох вхідних векторів, нелінійне ядро (наприклад, радіальна базисна функція – RBF) дає змогу виявляти більш складні закономірності в даних. Вибір ядра залежить від характеристик даних і складності задачі.

Для оцінювання продуктивності обраних методів у процесі прогнозування критичної швидкості використовувалися такі стандартні метрики:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i^{pred} - y_i^{true})^2},$$

де  $y_i^{true}$  – справжнє (еталонне) значення критичної швидкості з датасету;  $y_i^{pred}$  – прогнозоване значення критичної швидкості;  $N$  – кількість зразків у тестовій вибірці.

Коефіцієнт детермінації (*Coefficient of Determination*)  $R^2$ :

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i^{true} - y_i^{pred})^2}{\sum_{i=1}^N (y_i^{true} - \bar{y}^{true})^2},$$

де  $\bar{y}^{true}$  – це середнє значення істинних значень.

Для аналізу ефективності обраних моделей машинного навчання сформовано два датасети –

малого й великого обсягу, що містять варіації критичної швидкості, які відповідають різним комбінаціям параметрів ( $\gamma, r_0, C$ ).

Для формування датасетів використовувалась попередньо побудована математична модель криволінійного руху поїзда. Набори даних були сформовані способом прямого перебору. Датасет малого обсягу подано 57 наборами значень параметрів  $\gamma, r_0, C$  як вхідні дані й відповідне їм значення  $V_c$  як вихідні дані. Датасет великого обсягу містить 212 наборів значень вхідних і вихідних даних. Фрагмент даних зі створених датасетів подано в табл. 1. Ці датасети застосовувалися для навчання моделей, реалізованих у MATLAB.

Таблиця 1. Фрагмент датасету

| $r_0$ | $\gamma$ | $C$  | $V_c$ |
|-------|----------|------|-------|
| 0,1   | 0,05     | 1403 | 58    |
| 0,125 | 0,05     | 1403 | 63    |
| ...   | ...      | ...  | ...   |
| 0,8   | 0,05     | 1403 | 133,6 |

Графіки точності прогнозування, отримані під час експериментів, зображено на рис. 1–4. Істинні показники швидкості подано на графіках у вигляді пунктирної лінії, прогнози швидкості – у вигляді точок.

На рис. 1–2 продемонстровано точність прогнозування моделей, навчених на невеликому наборі даних. На малому наборі даних модель MLP прогнозує критичну швидкість із показником точності  $R^2 = 0.9876$ .

За таких саме умов за незначної кількості даних модель GPR здатна прогнозувати критичну швидкість із прийнятним показником точності  $R^2 = 0.9987$ . Видно, що в межах тестового набору даних прогнозовані значення критичної швидкості збігаються з реальними показниками (рис. 1).

Модель SVR за умови незначного обсягу даних може прогнозувати критичну швидкість з показником точності  $R^2 = 0.9721$ . Такий показник є недостатнім, що призводить до стабільних розбіжностей між прогнозами швидкості та реальними показниками (рис. 2).

Отже, за результатами експерименту можна помітити, що як MLP, так і GPR успішно виконують завдання прогнозування критичної швидкості, до того ж точність, якої досягає GPR, вища

за точність MLP. Натомість з усіх трьох моделей SVR демонструє найгірші результати у виконанні поставленого завдання в умовах незначного обсягу даних для навчання.

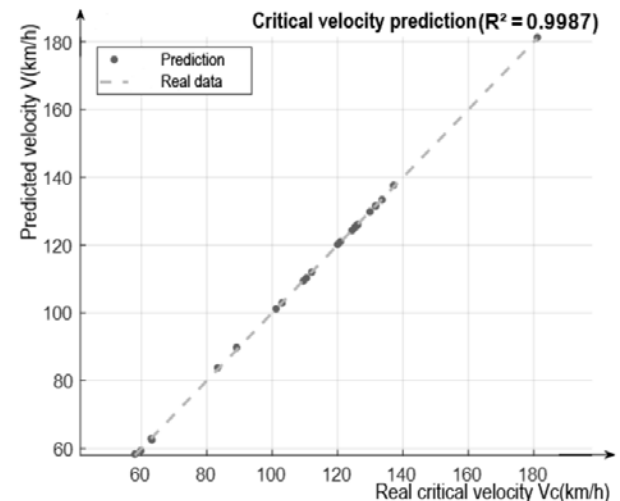


Рис. 1. Прогнозування критичної швидкості за допомогою моделі GPR, навченої на невеликому наборі даних

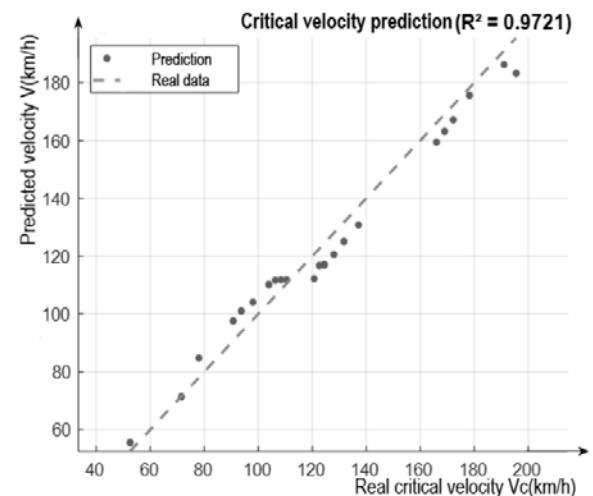


Рис. 2. Прогнозування критичної швидкості за допомогою моделі SVR, навченої на невеликому наборі даних

На рис. 3–4 подано результати для моделей, навчених на великому наборі даних. На цьому наборі даних модель MLP виконує прогнозування критичної швидкості з достатньо високим показником точності  $R^2 = 0.9997$ . Такий показник є майже ідеальним збігом прогнозів із реальними показниками, що можна побачити на графіку (рис. 3).

За саме таких умов модель GPR здатна прогнозувати критичну швидкість із показником точності  $R^2 = 0.9989$ .

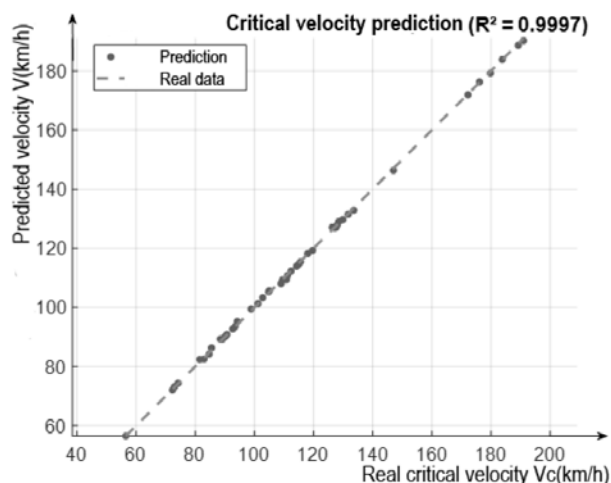


Рис. 3. Прогнозування критичної швидкості за допомогою моделі MLP, навченої на великому наборі даних

Модель SVR у разі значного обсягу даних може прогнозувати критичну швидкість з показником точності  $R^2 = 0.9783$ , що виражається у видимій різниці між прогнозованою та реальною критичними швидкостями (рис. 4).

На великому датасеті MLP і GPR залишаються ефективними, проте точність GPR знижується зі збільшенням обсягу даних, тоді як точність MLP зростає, що має кращу масштабованість. SVR продовжує демонструвати незадовільні результати як

на малому, так і на великому датасетах. Відповідно, з огляду на перспективи подальшого розширення даних і параметрів модель MLP є найбільш доцільним вибором для прогнозування критичної швидкості.

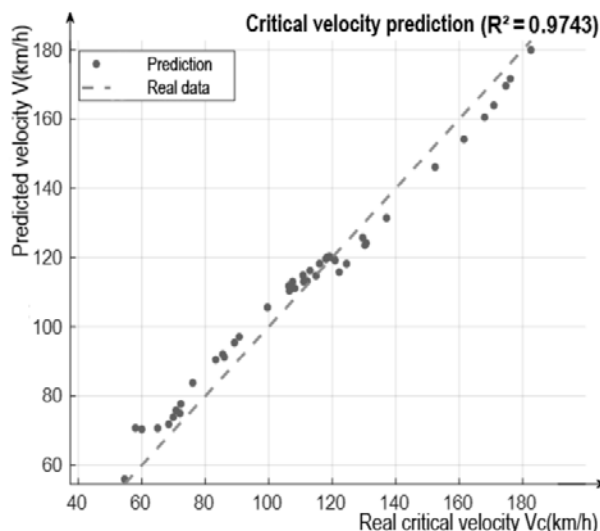


Рис. 4. Прогнозування критичної швидкості за допомогою моделі SVR, навченої на великому наборі даних

Для додаткового підтвердження ефективності MLP було порівняно залежності критичної швидкості від конусності колісної пари  $\gamma$  (за незмінних інших параметрів) (рис. 5).

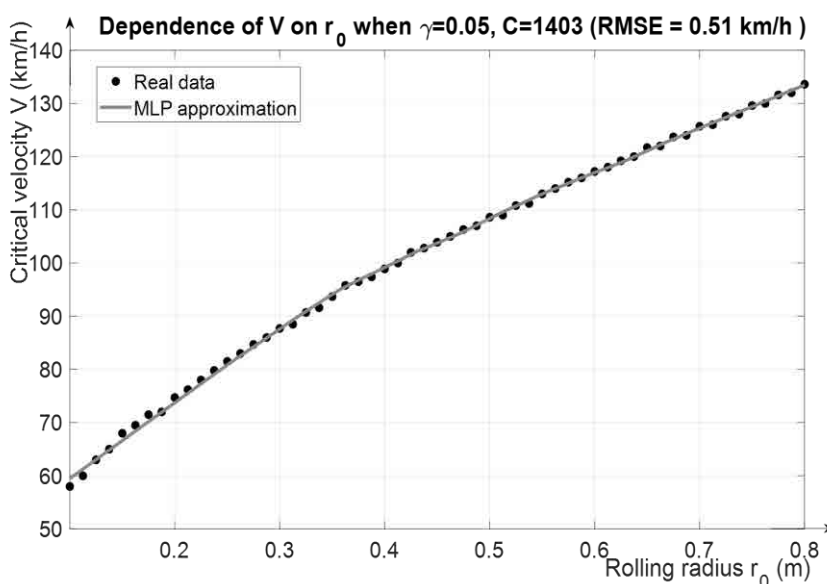


Рис. 5. Порівняння критичної швидкості та конусності колісної пари за незмінних інших параметрів

Як зображено на рис. 5, MLP апроксимує критичну швидкість із похибкою  $RMSE = 0.51$  км/год. Така похибка є незначною з погляду задачі

прогнозування швидкості руху поїзда, що підтверджує дуже високий рівень точності прогнозів зазначеної моделі.

## 5. Результати дослідження

Порівняльний аналіз трьох моделей машинного навчання: багат шарового перцептрона, регресії гауссівських процесів і регресії опорних векторів – виявляє різні закономірності в роботі відповідно до розміру набору даних і залежностей базових параметрів (табл. 2).

Таблиця 2. Результати порівняльного аналізу

| Модель | Малий набір даних | Великий набір даних |
|--------|-------------------|---------------------|
| MLP    | $R^2 = 0.9876$    | $R^2 = 0.9997$      |
| GPR    | $R^2 = 0.9987$    | $R^2 = 0.9989$      |
| SVR    | $R^2 = 0.9721$    | $R^2 = 0.9783$      |

Усі моделі були навчені з використанням обмеженої кількості комбінацій параметрів руху поїзда. Як продемонстровано на рис. 1–3, на невеликому наборі даних як MLP, так і GPR успішно відобразили нелінійну залежність, що визначає критичну швидкість унаслідок досягнення високої точності щодо еталонних значень, отриманих з високоточної моделі криволінійного руху [8]. Зауважимо, що GPR продемонструвала дещо вищу точність, що можна пояснити її баєсівським характером і здатністю виводити плавні функціональні форми з розріджених даних. Натомість SVR (рис. 3) не змогла правильно апроксимувати критичну швидкість: її прогнози продемонстрували значне відхилення, особливо на крайніх значеннях діапазону конічності, що вказує на погану адаптивність до базової динаміки.

У процесі переходу до великих наборів даних поведінка моделей помітно змінилася (рис. 4–6). Ефективність MLP значно покращилася: значення  $R^2$  зросло з 0.9876 до 0.9997, що підтверджує її високу здатність до навчання й масштабованість у разі збільшення обсягу даних. Це відповідає відомим перевагам глибоких мереж із прямим поширенням у процесі апроксимації високорозмірних нелінійних відображень за умови достатнього навчання. Водночас відповідно до показника  $R^2$  точність GPR майже не змінилася. Це є свідченням недостатньої здатності до масштабування. Зі свого боку SVR залишився стабільно ненадійним для обох розмірів наборів даних:  $R^2 = 0.9721$  на малому датасеті та  $R^2 = 0.9783$  на великому. Додатковий експеримент з порівнянням критичної швидкості та конусності

колісної пари за незмінних інших параметрів підтверджує високу точність і надійність апроксимації швидкості за допомогою MLP та його перспективність для подальшого використання.

Результати демонструють, що MLP пропонує найкращий компроміс між точністю, надійністю та масштабованістю; GPR підходить для прототипування на невеликих наборах даних, але стає неефективним і менш точним із зростанням обсягу даних; SVR не підходить для цього конкретного завдання регресії в тестованих конфігураціях.

## 6. Висновки

Отже, у межах дослідження проведено порівняльний аналіз трьох методів навчання з учителем – багат шаровий перцептрон (MLP), гауссівська регресія процесів (GPR) і регресія на опорних векторах (SVR) – щодо здатності апроксимувати критичну швидкість залізничного рухомого складу в умовах сильно нелінійних залежностей параметрів і обмеженого обсягу даних. Для визначення продуктивності моделей обрано відповідні метрики, виконано порівняльну оцінку їх ефективності й точності. Порівняльний аналіз трьох моделей машинного навчання: багат шарового перцептрона (MLP), гауссівської регресії процесів (GPR) і регресії на опорних векторах (SVR) – виявив різні властивості продуктивності залежно від розміру датасету й характеру залежностей параметрів.

Результати показують, що і MLP, і GPR досягають високої точності прогнозування як на малих, так і на великих датасетах. У цьому разі MLP демонструє кращу масштабованість і покращення точності зі збільшенням обсягу датасету, тоді як ефективність GPR за тих самих умов знижується. Натомість для SVR властива недостатня точність прогнозування на всіх розмірах датасету.

Додатковий експеримент із виокремленням залежності критичної швидкості від конусності колісної пари  $\gamma$  за фіксованих  $r_0$  та  $C$  підтверджує не лише загальну апроксимаційну здатність, але й збереження локальної чутливості – критичну вимогу для систем безпеки реального часу, наприклад системи підтримки прийняття рішень машиніста.

Відповідно, модель MLP обрано для подальшого впровадження як ключовий компонент модуля оцінювання критичної швидкості – складника



системи підтримки прийняття рішень машиніста й частини моделі криволінійного руху поїзда – і рекомендовано для виконання аналогічних задач в інших сферах діяльності.

Подальше дослідження обраної моделі буде присвячено модифікації її алгоритму й структури з огляду на фізичні параметри системи руху поїзда для підвищення якості досягнутих результатів.

#### Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

#### References

1. Alves Costa, P., Colaço, A., Calçada, R., Cardoso, A. (2014), "Critical velocity of railway tracks. Detailed and simplified approaches", *Transportation Geotechnics*, No. 2, pp. 30–32. DOI: <https://doi.org/10.1016/J.TRGeo.2014.09.003>
2. Iwnicki, S., Spiryagin, M., Cole, C., McSweeney, T. (eds.) (2019), "Handbook of Railway Vehicle Dynamics", 2nd edn. Boca Raton: CRC Press. DOI: <https://doi.org/10.1201/9780429469398>
3. Dusza, M. (2020), "Rail vehicle model motion analysis on curved track with vertical irregularity", *World Journal of Transport and Environment*, No. 130, pp. 75–84. DOI: <https://doi.org/10.5604/01.3001.0014.5456>
4. Wu, Y., Fu, H., Bian, X. (2024), "Critical speed and dynamic response of high-velocity railway in ballastless track", *Journal of Physics: Conference Series*, No 2647. DOI: <https://doi.org/10.1088/1742-6596/2647/20/202001>
5. Polach, O., Kaiser, I. (2012), "Comparison of methods analyzing bifurcation and hunting of complex rail vehicle models", *Journal of Computational and Nonlinear Dynamics*, No. 7(4), 041005. DOI: <https://doi.org/10.1115/1.4006825>
6. Castanheira-Pinto, A., Fernández-Ruiz, J., Colaço, A., Alves Costa, P., Connolly, D.P. (2022), "A simplified approach for predicting the non-linear critical speed of railway tracks", *Transportation Geotechnics*, No. 37, 100869. DOI: <https://doi.org/10.1016/j.trgeo.2022.100869>
7. Alves Costa, P., Soares, P., Colaço, A., Lopes, P., Connolly, D. (2020), "Railway critical velocity assessment: A simple experimental-analytical approach", *Soil Dynamics and Earthquake Engineering*, No. 134, 106156. DOI: <https://doi.org/10.1016/j.soildyn.2020.106156>
8. Estaire, J., Crespo-Chacón, I. (2024), "On the use of dispersion curves to determine the critical velocity of railway tracks. Application to case studies", *Transportation Geotechnics*, No. 46, 101247. DOI: <https://doi.org/10.1016/j.trgeo.2024.101225>
9. Mezher, S., Connolly, D., Woodward, P., Laghrouche, O., Pombo, J., Alves Costa, P. (2015), "Railway critical velocity - Analytical prediction and analysis", *Transportation Geotechnics*, No. 6, P. 84–96. DOI: <https://doi.org/10.1016/j.trgeo.2015.09.002>
10. Yevtushenko, O. S., Zakovorotnyi, O. Yu. (2025), "Mathematical model of rolling stock vibrations on a curved track section for the on-board decision support system", *Systems of Control, Navigation and Communication*, No. 3, P. 33–36. DOI: <https://doi.org/10.26906/SUNZ.2025.3.033>
11. Rasmussen, C. E., Williams, C. K. I. (2006), "Gaussian Processes for Machine Learning", Cambridge, MA: MIT Press.
12. Goodfellow, I., Bengio, Y., Courville, A. (2016), "Deep Learning", Cambridge, MA: MIT Press.
13. Smola, A. J., Schölkopf, B. (2004), "A tutorial on support vector regression", *Statistics and Computing*, No. 14(3), pp. 199–222
14. Liu, H., Song, L., Xu, L., Yu, Z. (2025), "Identification of wheel-rail forces on high-speed railways based on physical model and hybrid recursive neural networks", *Engineering Structures*, No. 338, pp. 120547. DOI: <https://doi.org/10.1016/j.engstruct.2025.120547>
15. Ye, Y., Shi, D., Krause, P., Hecht, M. (2019), "A data-driven method for estimating wheel flat length", *Vehicle System Dynamics*, No. 58, pp. 1329–1347. DOI: <https://doi.org/10.1080/00423114.2019.1620956>

#### Фінансування

Дослідження проводилося без фінансової підтримки.

#### Доступність даних

Дані будуть надані за обґрунтованим запитом.

#### Використання засобів штучного інтелекту

Автори застосовували технології ШІ (*Qwen-3.5*, *DeepL*) для перевірки граматики й орфографії. Після використання цих інструментів автори перевірили й відредагували контент і несуть повну відповідальність за зміст публікації. Автори підтверджують, що не застосовували технології штучного інтелекту для досягнення результатів, поданих у роботі.

Received (Надійшла) 01.04.2026

Accepted for publication (Прийнята до друку) 14.04.2026

Publication date (Дата публікації) 29.05.2026

*Відомості про авторів / About the Authors*

**Євтушенко Олександр Сергійович** – Національний технічний університет "Харківський політехнічний інститут", аспірант кафедри "Комп'ютерна інженерія та програмування", Харків, Україна;

**Oleksandr Yevtushenko** – National Technical University "Kharkiv Polytechnic Institute", Postgraduate Student, Department of Computer Engineering and Programming, Kharkiv, Ukraine;

e-mail: Oleksandr.Yevtushenko@cs.khpi.edu.ua

ORCID ID: <https://orcid.org/0009-0000-7798-8545>

**Заковоротний Олександр Юрійович** – доктор технічних наук, професор, Національний технічний університет "Харківський політехнічний інститут", завідувач кафедри "Комп'ютерна інженерія та програмування", Харків, Україна;

**Oleksandr Zakovorotnyi** – Doctor of Technical Sciences, Professor, National Technical University "Kharkiv Polytechnic Institute", Head of the Department of Computer Engineering and Programming, Kharkiv, Ukraine;

e-mail: Oleksandr.Zakovorotnyi@khpi.edu.ua

ORCID ID: <https://orcid.org/0000-0003-4415-838X>

## ESTIMATION OF A TRAIN'S CRITICAL SPEED USING MACHINE LEARNING METHODS

**The subject** of this study is the effectiveness of machine learning methods for approximating critical speed under conditions of nonlinear dependencies and limited data volume. **The goal** of this article is to conduct a comparative evaluation of three teacher-guided learning methods in terms of their ability to estimate the critical speed of railway rolling stock under conditions of highly nonlinear parameter dependencies and limited data, through a series of experiments and an analysis of their results. **Given the stated objective, the following tasks must be performed:** create small- and large-scale datasets based on a mathematical model of a train's curvilinear motion; implement and train models of a multilayer perceptron, Gaussian process regression, and support vector machine regression in the MATLAB environment; conduct a comparative assessment of approximation accuracy using the RMSE and  $R^2$  metrics on both datasets; identify the model that provides the best trade-off between accuracy and scalability; justify the selection of the optimal method for implementation in the DSS. **Methods.** A mathematical model of a train's curvilinear motion was used to generate the datasets. The quality of the approximation was evaluated using the RMSE and  $R^2$  metrics. All experiments were implemented in MATLAB. **The following results were obtained:** the multilayer perceptron and Gaussian process regression models demonstrated high prediction accuracy for both small and large datasets, with the perceptron showing better scalability as the training dataset size increased than Gaussian process regression. **Conclusions.** Based on the comparative analysis, it is advisable to use the MLP as the primary model for determining the critical speed in the train driver's decision support system and railway transport diagnostics. Implementing the proposed approach will improve train safety under actual operating conditions, reduce maintenance costs, and provide a foundation for the development of adaptive on-board speed control systems

**Keywords:** critical speed; rolling stock vibrations; multilayer perceptron; Gaussian process-based regression; support vector regression; approximation; machine learning; train driver decision support system; curved-track movement.

## *Бібліографічні описи / Bibliographic descriptions*

Євтушенко О. С., Заковоротний О. Ю. Апроксимація критичної швидкості поїзда методами машинного навчання. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 276–283. DOI: <https://doi.org/10.30837/0135-1710.2026.189.276>

Yevtushenko, O., Zakovorotnyi, O. (2026), "Estimation of a train's critical speed using machine learning methods", *Management Information System and Devices*, No. 2 (189), P. 276–283. DOI: <https://doi.org/10.30837/0135-1710.2026.189.276>

Anna Nikolaichuk, Oleg Kobylin, Ilya Kobylin, Oleksandra Putyatina

## A COMPREHENSIVE EVALUATION OF TRANSFORMER MODELS FOR SENTENCE-LEVEL SEMANTIC SIMILARITY IN ENGLISH AND UKRAINIAN

The subject of this study is transformer-based models for semantic textual similarity and approaches to their evaluation in monolingual and cross-lingual settings for English and Ukrainian. **The aim of the work** is to evaluate and compare the effectiveness of transformer models on English, Ukrainian, and English-Ukrainian sentence pairs, using different approaches to assessing their quality and practical applicability. To achieve this aim, the following tasks are addressed: comparing models on monolingual and cross-lingual datasets; analyzing results using Pearson and Spearman correlation coefficients; assessing practical applicability by classifying predictions based on error thresholds; and investigating the impact of language on model accuracy. **Methods.** The study employs transformer models GTE, LaBSE, MiniLM, and MPNet, evaluated on the STS-B dataset, as well as its Ukrainian and English-Ukrainian versions. Cosine similarity is used to compute sentence similarity, while evaluation is performed using Pearson and Spearman correlation coefficients together with classification of predictions based on an error threshold. **Results.** The findings show that the GTE model demonstrates the best overall performance across combined metrics, while MiniLM provides an optimal balance between accuracy and computational efficiency. They also indicate that high correlation scores do not always correspond to a high percentage of correct predictions, revealing limitations of traditional evaluation approaches. In addition, a systematic tendency of models to overestimate similarity in the lower range is observed. It is further found that cross-lingual pairs may achieve higher accuracy under threshold-based evaluation despite lower correlation values, indicating differences in model behavior depending on the evaluation metric. **Conclusions.** The study demonstrates the effectiveness of a combined evaluation approach for semantic similarity models, which enables a more comprehensive assessment of their real-world performance. The results confirm the importance of considering language-specific characteristics and practical task requirements when selecting models and evaluation methods, and highlight the role of comprehensive evaluation strategies.

**Keywords:** semantic similarity; transformer models; model evaluation; Pearson correlation coefficient; Spearman rank correlation coefficient; English language; Ukrainian language; cross-lingual similarity; practical model performance.

### 1. Introduction

One of the key challenges in the field of natural language processing (NLP) is the accurate assessment of semantic similarity between two texts, commonly formalized as the semantic textual similarity (STS) task. Effective solutions to this problem provide the foundation for a wide range of applied NLP tasks.

Substantial progress in semantic modeling has been achieved through methods for learning vector representations (embeddings) of linguistic units, most notably Word2Vec [13] and GloVe [15]. However, these methods share an inherent limitation: they primarily operate at the level of individual words and therefore fail to adequately capture context-dependent variations in meaning. This was followed by the development of transformer-based models, particularly BERT (Bidirectional Encoder Representations from Transformers), which generate significantly higher-quality contextualized representations [7]. This progress,

in turn, facilitated the emergence of models specifically designed for text comparison tasks and optimized for more accurate semantic similarity assessment.

The fundamental role of the STS task within NLP is reflected in its frequent use as a component of more complex applied systems. In particular, STS is widely used for automatic essay grading, plagiarism detection, question answering, and filtering text pairs based on semantic similarity. In this context, the accuracy and robustness of STS models directly determine the effectiveness of the corresponding applications.

At the same time, the performance of semantic similarity models may vary considerably depending on the task, dataset properties, and linguistic context. This variability highlights the importance of systematic analysis of existing models and evaluation methodologies, as well as a comprehensive investigation of the factors that influence their performance.

## 2. Analysis of sources and definition of the problem

Recent studies increasingly demonstrate that general-purpose transformer models do not exhibit consistently stable performance across different application scenarios. Their effectiveness depends on the domain, dataset composition, and the chosen evaluation metric [21, 20]. This indicates that even high-quality vector representations, when considered in isolation, do not guarantee stable performance in STS tasks. Therefore, comparative studies across different types of data are necessary to assess the practical suitability of such models.

Separately, the literature addresses the problem of evaluating the quality of model predictions in semantic similarity tasks [1, 11, 16]. Traditionally, particularly within the framework of SemEval-2017 Task 1 [4], Pearson's correlation coefficient is used to compare model predictions with human annotations. This metric enables the assessment of linear correspondence between predicted and ground-truth scores. However, its application has certain limitations, as identical coefficient values may correspond to different error distributions and do not always adequately reflect the practical effectiveness of a model.

As an alternative, Spearman's rank correlation coefficient [24] is often used, as it better captures monotonic and nonlinear relationships between scores. This metric has also been successfully applied in the evaluation of models on STS tasks [18]. However, it is not universal either: high rank consistency does not necessarily ensure effectiveness of the system in applied scenarios where it is critical to make threshold-based decisions, ranking decisions, or selection of text pairs.

For this reason, other evaluation metrics are being investigated, including normalized cumulative gain (nCG), normalized discounted cumulative gain (nDCG), precision, F1-score, and combined approaches [19]. However, existing results indicate that there is no universal metric for evaluating STS models, and the choice of a specific evaluation method should be determined by the properties of the task at hand. A similar conclusion is supported by findings in other domains where system performance monitoring, early error detection, and optimization with respect to speed and accuracy are essential [9, 10, 23].

Another under-researched problem is cross-linguistic semantic similarity, particularly for languages with different grammatical and syntactic structures,

such as Ukrainian and English. Most existing studies focus on monolingual datasets or high-resource language pairs. Although recent approaches to constructing multilingual sentence embeddings have demonstrated promising results in zero-shot transfer settings [3], their performance remains limited for typologically distant languages, where a decrease in semantic alignment quality is still observed.

In this context, the Ukrainian language faces additional challenges due to the lack of standardized datasets, which complicates accurate benchmarking and fine-tuning of the models for natural language processing tasks, particularly STS [12, 14]. Although cross-lingual benchmarks have been proposed, such as those in SemEval-2017 Task 1 [4], most comparative studies continue to focus on high-resource languages, thereby limiting the evaluation of models for Ukrainian and constraining cross-lingual comparisons between English and Ukrainian.

Thus, two interrelated research gaps can be identified. First, there is no universal approach to evaluating the quality of semantic textual similarity (STS) models that adequately reflects their performance across different application scenarios. Second, the behavior of modern transformer-based models in cross-lingual settings, particularly for English–Ukrainian sentence pairs, remains underexplored. These gaps jointly motivate a comparative analysis of STS models on both monolingual and cross-lingual datasets using multiple evaluation approaches.

## 3. Research objectives and tasks

Given that the performance of STS models is strongly language-dependent, this study compares several models on English, Ukrainian, and English–Ukrainian sentence pairs. This approach enables the assessment of model stability as well as sensitivity to typological differences in grammar and syntax.

Since different correlation-based metrics capture different aspects of model behavior, the evaluation is conducted using both Pearson and Spearman correlation coefficients. This allows for identifying potential differences in performance assessment depending on the evaluation metric and provides a more comprehensive view of the agreement between model predictions and human annotations.

However, correlation analysis alone is insufficient for assessing the practical applicability of the models.

Therefore, it is complemented by classifying sentence pairs based on the magnitude of the error between predicted and ground-truth similarity scores. This makes it possible to analyze how the models handle sentence pairs with varying levels of semantic similarity and to identify cases in which they exhibit large deviations or, conversely, achieve high accuracy.

Accordingly, the purpose of this work is to evaluate and compare transformer-based models on monolingual and cross-lingual datasets of English and Ukrainian sentences, taking into account multiple approaches to semantic similarity assessment.

**Table 1.** *Parameters of the models under comparison*

| Model                                          | Embedding dimension | Supported languages | Base model |
|------------------------------------------------|---------------------|---------------------|------------|
| gte-multilingual-base (GTE)                    | 768                 | 75                  | GTE        |
| LaBSE                                          | 768                 | 110                 | BERT       |
| paraphrase-multilingual-MiniLM-L12-v2 (MiniLM) | 384                 | 50                  | BERT       |
| paraphrase-multilingual-mpnet-base-v2 (MPNet)  | 768                 | 50                  | XLNet      |

The gte-multilingual-base (GTE) model is based on an encoder-only transformer architecture, supports 75 languages, and processes input sequences of up to 8192 tokens. A key distinguishing feature of GTE is its ability to generate both dense and sparse embeddings, which improves performance in retrieval and re-ranking tasks. As a result, GTE achieves state-of-the-art performance on text benchmarks while maintaining high computational efficiency [25].

The LaBSE model (Language-Agnostic BERT Sentence Embedding) is a multilingual extension of BERT that produces language-independent embeddings and ensures strong cross-lingual alignment [8]. Consequently, it demonstrates strong performance in cross-lingual retrieval tasks and supports over 100 languages. However, this design focus may lead to reduced performance on monolingual semantic similarity tasks.

The paraphrase-multilingual-MiniLM-L12-v2 (MiniLM) model is also based on a multilingual BERT-derived architecture and supports 50 languages. Through knowledge distillation, the model remains compact and computationally efficient while maintaining competitive performance relative to BERT-base [22]. However, the reduced dimensionality of its embeddings may limit accuracy on more complex semantic tasks.

The paraphrase-multilingual-mpnet-base-v2 (MPNet) model is based on XLM-RoBERTa and supports 50 languages. It combines masked and permutation-based language modeling strategies, enabling more effective

#### 4. Materials and research methods

##### *Selected models and their characteristics*

To compare performance on sentence-level semantic similarity tasks, several transformer-based models were selected. These models differ in architectural characteristics, support a wide range of languages (including English and Ukrainian), and have demonstrated strong performance in prior studies. Table 1 presents the embedding dimensionality of each model, the number of supported languages, and their base architectures.

modeling of token dependencies [6]. This results in improved performance compared to BERT and XLNet across a wide range of tasks. Its main limitation is relatively high computational cost due to the higher embedding dimensionality.

All selected models were fine-tuned using the Sentence-Transformers framework and adapted for semantic similarity tasks.

##### *Semantic textual similarity benchmark datasets*

To compare model performance, the STS-B (Semantic Textual Similarity Benchmark) dataset [4] is selected as the evaluation benchmark. It contains sentence pairs from various domains, including news headlines, image captions, and natural language inference tasks. Each pair is annotated with human-assigned similarity scores ranging from 0 (completely dissimilar sentences) to 5 (semantically equivalent sentences).

In the experiments, a normalized version of the English dataset is used, in which similarity scores are rescaled to the interval [0, 1]. The dataset consists of 5749 sentence pairs, each associated with a corresponding similarity score.

A Ukrainian version of the dataset is also created [2], where the original data are translated using state-of-the-art neural machine translation systems to ensure semantic consistency. Although human translation may preserve certain linguistic nuances more accurately in some cases, machine translation is chosen due to



its accessibility, efficiency, and high quality for general-domain texts. A comparison of human and machine translation results is left for future work investigating the impact of translation quality on semantic similarity assessment.

#### Evaluation metrics

Cosine similarity is used to compute the similarity between vector representations of sentences, and the results are evaluated using the Pearson and Spearman correlation coefficients, which measure the correlation between predicted similarity scores and ground-truth labels.

Cosine similarity measures the similarity between two vectors in a high-dimensional space by computing the cosine of the angle between them:

$$\cos(\theta) = \frac{AB}{\|A\| \cdot \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}, \quad (1)$$

where  $A$  and  $B$  are vector representations of two sentences, with  $\|A\|$  and  $\|B\|$  denoting their L2-norms.

To enable comparison between cosine similarity values, which lie in the interval  $[-1, 1]$ , and ground-truth labels, min-max normalization is applied to rescale the values to the interval  $[0, 1]$ :

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \quad (2)$$

where  $\min(x)$  and  $\max(x)$  are the minimum and maximum cosine similarity values. Thus, the minimum value is mapped to 0, the maximum to 1, and all intermediate values are linearly scaled within this range.

The Pearson correlation coefficient ( $r$ ) measures the linear relationship between predicted and ground-truth values and is defined as:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}, \quad (3)$$

**Table 2.** Pearson and Spearman correlation coefficients for each model

| Model  | EN-EN pairs |        | UK-UK pairs |        | EN-UK pairs |        | Averages    |             |
|--------|-------------|--------|-------------|--------|-------------|--------|-------------|-------------|
|        | $r$         | $\rho$ | $r$         | $\rho$ | $r$         | $\rho$ | $r$         | $\rho$      |
| GTE    | 87.4        | 87.4   | 85.1        | 84.1   | 80.6        | 78.1   | <b>84.4</b> | <b>83.2</b> |
| LaBSE  | 74.9        | 73.6   | 74.4        | 73.1   | 71.7        | 69.9   | 73.7        | 72.2        |
| MiniLM | 85.1        | 83.9   | 81.6        | 80.1   | 78.1        | 74.6   | 81.6        | 79.5        |
| MPNet  | 86.3        | 85.9   | 83.1        | 82.1   | 80.6        | 77.8   | <b>83.3</b> | <b>81.9</b> |

Based on the average correlation values computed across different datasets, the GTE and MPNet models

where  $x_i$  and  $y_i$  are the predicted and true similarity scores, with  $\bar{x}$  and  $\bar{y}$  denoting their mean values.

To address the limitations of Pearson correlation in capturing non-linear relationships, Spearman's rank correlation coefficient ( $\rho$ ) is also used, as it evaluates the monotonic relationship between ranked values:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \quad (4)$$

where  $d_i$  is the difference between the ranks of the predicted and true values, and  $n$  is the number of sentence pairs.

Additionally, an error-based evaluation is performed by classifying sentence pairs as correctly or incorrectly predicted based on a predefined threshold (0.3 in this study). Correct predictions correspond to small deviations between predicted and ground-truth scores, whereas incorrect predictions correspond to large deviations. This classification enables the identification of patterns in which models most frequently achieve high or low accuracy and helps reveal systematic biases, such as consistent overestimation or underestimation of semantic similarity.

## 5. Research results

### Correlation-based performance comparison

At the first stage of the study, model performance was evaluated using correlation-based metrics to assess how well the models align with benchmark sentence-level semantic similarity scores. Pearson and Spearman correlation coefficients were computed for different language pairs: EN-EN (monolingual English), UK-UK (monolingual Ukrainian), and EN-UK (cross-lingual English-Ukrainian). The results are presented in Table 2, with values expressed as percentages ( $r \times 100$  and  $\rho \times 100$ ).

demonstrated the best performance, achieving scores exceeding 80% for both coefficients. This indicates their

ability to generate embeddings that effectively capture both semantic content and syntactic properties in both monolingual and cross-lingual sentence pairs.

The MiniLM model ranked third, lagging by up to 3% in Pearson correlation and up to 4% in Spearman correlation. At the same time, due to its compact architecture, it provides an optimal balance between accuracy and computational cost in practical applications.

The LaBSE model demonstrated the lowest performance, with a gap of 12-13% for both coefficients compared to the best-performing models. This reflects the so-called "curse of multilinguality", in which the inclusion of a large number of languages leads to reduced performance under fixed model capacity [5]. In other words, broader language coverage is achieved at the cost of less specialized optimization for individual languages. At the same time, due to its language-agnostic embeddings, LaBSE exhibits relatively stable performance across different language pairs (varying within 1–7%). If stability across both monolingual and cross-lingual settings is prioritized, LaBSE may be preferable, as other models exhibit substantially greater variability, ranging from 3% to 12% depending on the language combination.

Overall, Spearman correlation was lower than Pearson correlation, indicating stronger linear agreement between predictions and ground-truth labels, while rank order is not fully preserved. Nevertheless, the overall conclusions regarding model performance remain consistent, suggesting that both metrics lead to similar conclusions depending on the evaluation context.

A comparison of model performance across monolingual and cross-lingual settings shows that monolingual pairs yield higher scores for both correlation metrics and exhibit smaller differences between them. This result is expected, as cross-lingual tasks are more challenging due to grammatical and semantic differences, as well as cultural context, which require more specialized model training. At the same time, general-purpose models still achieve strong results in such settings, although the performance gap remains noticeable.

Based on these evaluations, an initial ranking of the models was established, enabling the selection of an optimal model depending on priorities such as

accuracy, computational efficiency, or stability across languages. However, correlation-based results do not always guarantee similar performance in real-world applications. Therefore, to obtain a more detailed assessment of practical applicability, an additional analysis is conducted by classifying semantic similarity predictions for sentence pairs.

#### *Detailed analysis of semantic similarity predictions for sentence pairs. The problem of score scaling*

To further investigate model behavior, predictions are compared with human-assigned labels. At the initial stage, min-max scaling is applied for data preprocessing, with the minimum and maximum values defined according to the theoretical range of cosine similarity. Figure 1 presents the distributions of ground-truth scores and model predictions obtained from vector representations of the three models with the highest correlation performance (GTE, MPNet, and MiniLM) for English and Ukrainian sentence pairs.

The results indicate a substantial difference between the distributions of ground-truth scores and model predictions. For all three models, predicted values are shifted toward higher similarity scores, with insufficient differentiation in the lower and mid-range values. Most predictions are concentrated within a narrow interval from 0.7 to 1.0, while low scores (0–0.4) are nearly absent. In contrast, ground-truth labels are more evenly distributed across the entire range, from low to high similarity.

This behavior is consistent across all models and language pairs and complicates both overall performance evaluation and the analysis of language effects, as the narrow distribution of predictions limits the ability to identify subtle differences in model accuracy.

Based on these observations, an adjusted scaling approach is proposed in which the actual minimum and maximum prediction values for a given model and dataset are used instead of the theoretical bounds of cosine similarity. This transformation distributes predicted scores more evenly across the entire scale, providing a clearer representation of differences between sentence pairs with low, medium, and high semantic similarity.

Figure 2 illustrates the results after applying the adjusted scaling.

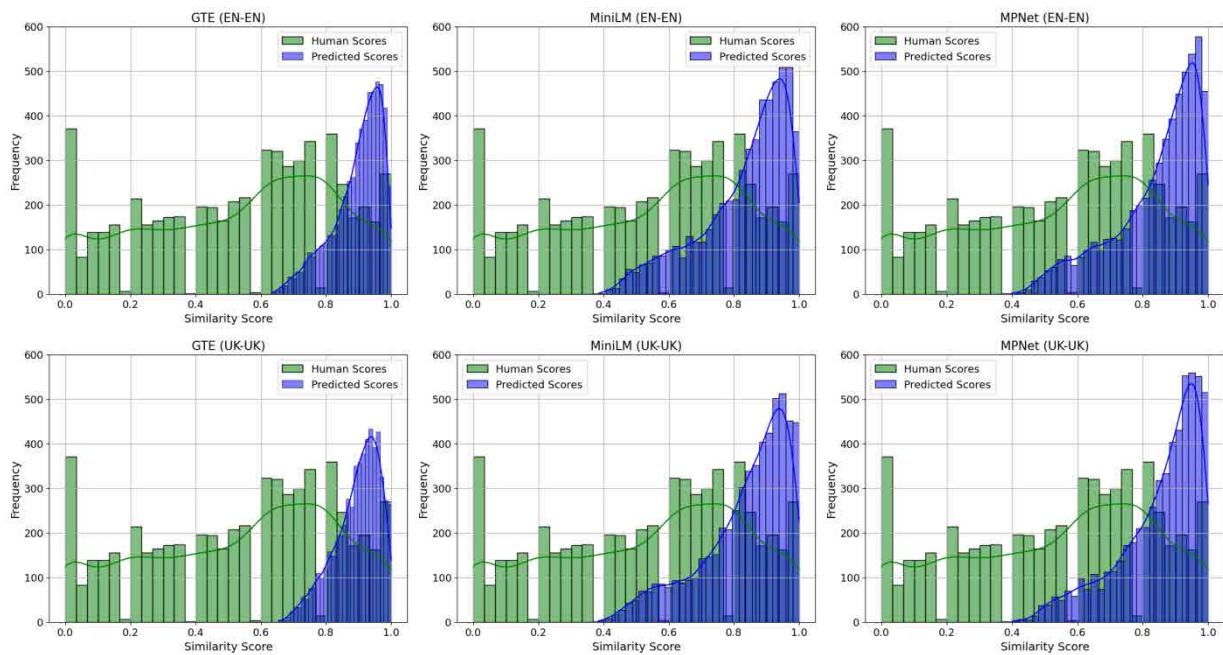


**Fig. 1.** Comparison of the distributions of predicted and gold semantic similarity scores for monolingual sentence pairs

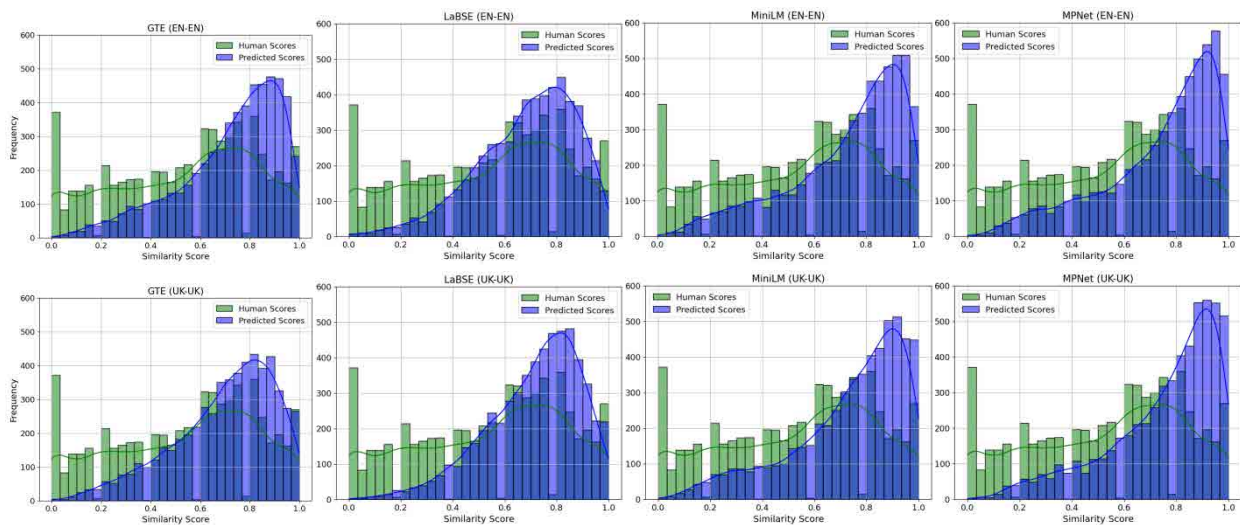


**Fig. 2.** Comparison of the distributions of predicted and gold semantic similarity scores for monolingual sentence pairs after scaling correction

The plots demonstrate improved alignment between model predictions and ground-truth labels, particularly in the lower and middle ranges, where substantially more predictions are observed than with the initial scaling. This transformation does not alter the relative ordering of sentence pairs by similarity, but only modifies the mapping of predictions onto the target scale. Thus, the adjusted scaling serves as a calibration mechanism that improves the interpretability of the results without affecting the actual accuracy of the models.

Overall, the analysis of distributions highlights the need to adapt scaling to the observed range of prediction values, which is an important step in assessing the practical applicability of the models.

#### *Comparison of model suitability for practical use based on classification of semantic similarity predictions*

Figure 3 presents the results of the classification of model predictions for all evaluated models. In the plots, green corresponds to correctly predicted



similarity scores, i.e., predictions that fall within a threshold of 0.3 from the corresponding ground-truth labels. Red indicates predictions with substantial deviations, including both overestimation and underestimation, where the model assigns higher or

lower similarity scores than those provided by human annotators. The dashed diagonal line represents the ideal case in which predicted scores exactly match the ground-truth labels. For high-quality models, most predictions are expected to lie close to this line.

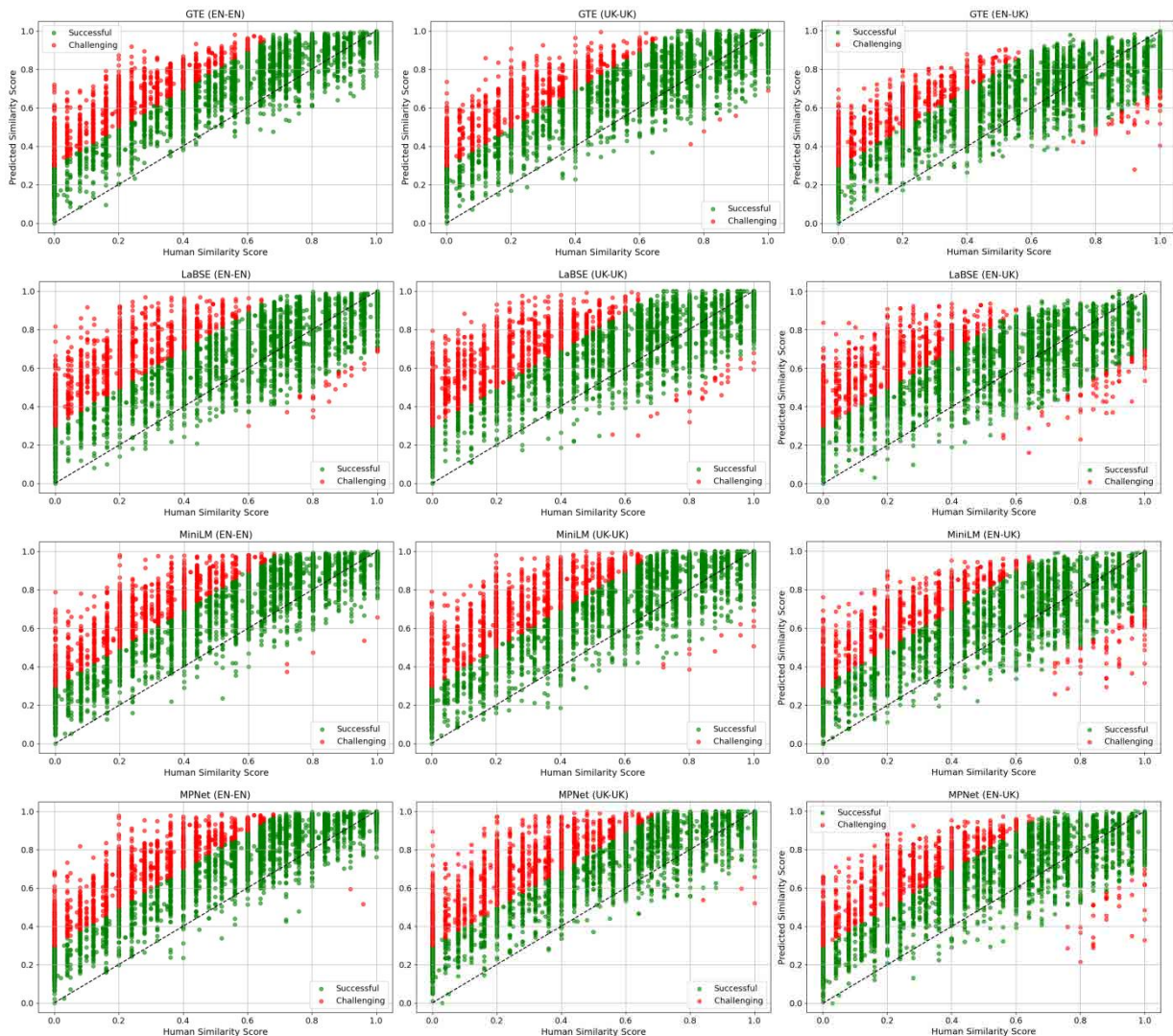


**Fig. 3.** Comparison of predicted and gold semantic similarity scores for sentence pairs

To complement the graphical analysis, Table 3 reports the percentage of correctly predicted similarity

scores for each model, separately for monolingual (EN-EN and UK-UK) and cross-lingual (EN-UK) sentence pairs.

**Table 3.** Percentage of correctly predicted sentence pair similarity for each model

| Model  | EN-EN pairs | UK-UK pairs | EN-UK pairs | Averages |
|--------|-------------|-------------|-------------|----------|
| GTE    | 78.4        | 80.1        | 82.3        | 80.3     |
| LaBSE  | 76.6        | 72.7        | 77.6        | 75.6     |
| MiniLM | 77.7        | 75.3        | 80.4        | 77.8     |
| MPNet  | 75.9        | 71.4        | 78.3        | 75.2     |

#### *Performance on monolingual sentence pairs*

As shown in Table 3, the ranking of models based on the percentage of correct predictions differs slightly from the ranking obtained using correlation-based metrics. The GTE and MiniLM models maintain high performance, while LaBSE and MPNet exhibit lower average values. Notably, despite achieving high correlation coefficients, MPNet shows the lowest percentage of correct predictions. This indicates that high correlation does not necessarily imply better agreement between predicted similarity scores and ground-truth values: a model may preserve the correct ordering of sentence pairs while making larger errors in the actual numerical predictions.

The analysis of the distribution plots (Figure 2) and the comparison of predictions with ground-truth values (Figure 3) show that, for all models, the largest number of correct predictions occurs in the medium and high similarity ranges, as reflected by peaks in the histograms within the interval 0.6–0.9. At the same time, the highest number of incorrect predictions is observed for sentence pairs with low human-assigned similarity scores. In this range, incorrect predictions are predominantly located above the diagonal, indicating systematic overestimation of similarity. Thus, the main difficulty for the models lies in their limited ability to distinguish between dissimilar sentence pairs.

The GTE model demonstrates the best agreement with human annotations. Although its prediction distribution is shifted toward higher values, it exhibits a relatively smooth shape and partially overlaps with the ground-truth distribution in the range 0.4–0.8. A substantial proportion of predictions lies close to the diagonal, and overestimation in the low-similarity range is less pronounced than in the other models. Cases of underestimation for highly similar sentence pairs are rare for Ukrainian and even rarer for English. Therefore, GTE not only preserves the relative ranking of sentence pairs, as reflected by its high correlation values, but also achieves the closest alignment between predicted scores and human annotations.

The MiniLM model exhibits a similar distribution pattern; however, its predictions are more concentrated in the upper part of the scale (0.8–1.0). In the medium similarity range, they align well with the ground-truth values, whereas in the high-similarity range, underestimation occurs more frequently, particularly for Ukrainian sentence pairs. In the low-similarity range, a tendency toward overestimation is also observed, although it is less pronounced than in the other models. Overall, these results confirm the competitiveness of MiniLM, especially given its compact architecture.

The LaBSE model is characterized by a less smooth distribution, with peak values in the range 0.7–0.9 and a near absence of low scores. Consequently, the largest number of errors occurs for sentence pairs with low similarity, where overestimation is predominant. In addition, cases of underestimation in the high-similarity range are more frequent than in the previously discussed models. Although these deviations are generally not large, as indicated by the relatively close proximity of predictions to the diagonal, the model overall provides stable but less accurate predictions compared to GTE and MiniLM in this evaluation.

The distribution of MPNet predictions is among the most compressed across all models, with a strong concentration in the narrow interval 0.8–1.0. The most pronounced pattern is the overestimation of similarity for sentence pairs with low ground-truth scores, where predicted similarity scores are substantially higher. At the same time, in the medium and high similarity ranges, the model demonstrates relatively good agreement with human annotations, and cases of underestimation are comparatively rare. Thus, while MPNet effectively preserves the relative ranking of sentence pairs (as reflected in its high correlation scores), systematic overestimation in the low-similarity range results in a lower percentage of correct predictions.

#### *Effect of language on prediction quality*

Separately, it is important to consider the effect of language on model prediction performance. According to the results presented in Table 3, English sentence pairs generally exhibit a higher percentage of correct predictions compared to Ukrainian pairs. The largest differences, ranging from 5–6%, are observed for models with lower average performance, such as LaBSE and MPNet, whereas for GTE and MiniLM the difference between languages is minimal (2–3%). This is due to the more complex morphology and greater syntactic variability of Ukrainian, which make accurate prediction of exact similarity scores more challenging and amplify the limitations of models with lower overall performance.

Interestingly, high correlation values for English sentence pairs do not always correspond to higher accuracy in terms of exact similarity scores. For example, in the case of GTE, the relative ranking of sentence pairs is better preserved for English, whereas the percentage of predictions within the specified threshold is slightly higher for Ukrainian.

Overall, the results indicate that models make more errors on Ukrainian sentence pairs. In particular,



underestimation is more common for highly similar pairs, whereas overestimation is less language-dependent, as it is also prevalent for English pairs. These findings highlight the need to account for language-specific characteristics when evaluating the practical applicability of semantic similarity models.

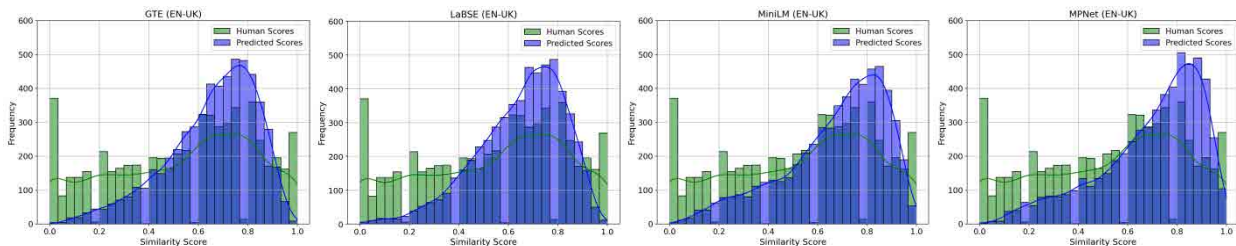


**Fig. 4.** Comparison of the distributions of predicted and gold semantic similarity scores for cross-lingual sentence pairs

According to the results in Table 3, all models achieve a higher percentage of correct predictions for cross-lingual pairs than for monolingual ones. GTE once again achieves the highest values, with an average increase of approximately 4% compared to English and Ukrainian pairs. A similar pattern is observed for MiniLM, with an improvement of about 5%. LaBSE and MPNet also show slightly better results on cross-lingual data, though the increase is more modest, at around 4%.

The plots in Figure 3 indicate that the overall pattern of predictions for cross-lingual pairs is similar to that observed for monolingual data. Most correct predictions occur in the medium and high similarity ranges, while the most challenging cases correspond to sentence pairs with low ground-truth scores. However, underestimation is more frequent for highly similar pairs. At the same time, performance in the low-similarity range improves: overestimation occurs less frequently, and such predictions lie closer to the diagonal.

The histograms in Figure 4 further confirm the higher percentage of correct predictions. Although predicted scores remain skewed toward higher values, most predictions are concentrated near the peak of the ground-truth distribution (0.6–0.9). This suggests that the improved performance on cross-lingual pairs does not imply the absence of distortions; rather, it indicates that models handle sentence pairs with medium and high semantic similarity more effectively, reflecting patterns observed in human annotations.

Thus, contrary to the expectation that comparing sentences across different languages is more difficult

### *Performance on cross-language sentence pairs*

To evaluate how model behavior changes under more challenging conditions, predictions for English–Ukrainian sentence pairs were analyzed. Figure 4 presents the distribution of predictions for these cross-lingual pairs, complementing the comparison of predicted and ground-truth scores shown in Figure 3.

due to the need to construct a shared semantic representation for English and Ukrainian, the classification results indicate higher accuracy. At the same time, this finding contrasts with the correlation analysis, where model performance on cross-lingual pairs is lower than on monolingual data. In other words, for English–Ukrainian pairs, models are less effective at preserving the overall ranking of similarity, while their predicted scores more often fall within the acceptable threshold. This discrepancy highlights the importance of analyzing not only aggregate statistical metrics but also the nature of prediction errors.

## **6. Discussion of results**

The results demonstrate that the evaluation of semantic similarity models cannot be limited to correlation-based metrics alone. Although Pearson and Spearman coefficients provide a general measure of agreement between model predictions and human annotations, they do not always reflect the practical applicability of models in real-world systems, where decisions are often made based on predefined thresholds.

Prediction classification thus serves as an important complement to correlation analysis: even with high correlation values, a model may be less suitable for practical use due to systematic overestimation or underestimation of similarity. This effect is particularly evident in the low-similarity range, where model behavior is less predictable because similarity scores are frequently overestimated for both monolingual and cross-lingual pairs. Therefore, evaluation based solely

on correlation-based metrics fails to capture important aspects of model behavior in practical scenarios.

Given these findings, model selection should be guided by the specific application context. A clear example of this is GTE, which achieves the best results in terms of both correlation metrics and the percentage of correct predictions and is therefore the most versatile model, suitable for systems requiring high overall accuracy and reliability.

A comparable pattern is observed for the MiniLM model, which demonstrates slightly lower accuracy while maintaining a consistently high level of practical applicability. This makes it a strong candidate as a trade-off between performance and computational efficiency. It is particularly well suited for large-scale production systems, where processing speed, latency, and scalability are critical.

The results for MPNet illustrate why evaluation based solely on correlation metrics is insufficient. Although this model achieves one of the highest correlation scores, the percentage of correct predictions is lower. At the same time, this characteristic makes MPNet well suited for tasks where relative ranking is the primary objective, such as information retrieval or re-ranking, but less suitable for systems that rely on strict thresholds, such as plagiarism detection.

Another important consideration is model stability across languages. LaBSE, despite lower average performance, demonstrates relatively stable behavior across different languages and their combinations. This property is particularly important for systems in which consistent performance across languages is more critical than achieving maximum accuracy for a single language, such as cross-lingual retrieval or multilingual applications.

The analysis of cross-lingual pairs reveals a distinct pattern of model behavior: according to correlation metrics, such pairs appear more challenging, whereas threshold-based classification shows improved results. In practical systems, it is often more important to ensure sufficiently accurate separation between similar and dissimilar sentence pairs than to perfectly preserve the global ranking of all examples. Thus, in cross-lingual scenarios, models may perform comparably, or in some cases even more effectively, when evaluated in terms of threshold-based accuracy.

## 7. Conclusions

This study demonstrates the continued relevance of the semantic textual similarity (STS) task for modern

natural language processing systems, particularly in multilingual settings, where no universal evaluation approach currently exists. Previous research shows that evaluating transformer-based models requires a comprehensive framework that goes beyond correlation-based metrics alone. A combined evaluation approach is therefore employed, incorporating both correlation-based metrics and classification-based measures of prediction accuracy. Such an approach provides a more complete assessment of model performance across different linguistic scenarios, capturing not only the agreement between predicted and human-assigned scores but also the practical applicability of models in real-world tasks.

Building on this evaluation framework, the comparison of four models on English, Ukrainian, and English-Ukrainian sentence pairs revealed significant differences in performance. The GTE model demonstrated the highest overall effectiveness, achieving balanced results across the combined evaluation framework. MiniLM proved to be a competitive alternative, offering an optimal balance between accuracy and computational efficiency. In contrast, MPNet and LaBSE showed that high correlation scores or cross-lingual stability alone do not necessarily guarantee superior performance in applied scenarios.

A more detailed analysis of predictions revealed important patterns in model behavior. In particular, a systematic bias toward higher similarity scores was observed, and correlation-based results did not always align with classification-based evaluations. These findings highlight the importance of a combined evaluation approach that integrates multiple metrics to provide a more reliable assessment of model performance.

These findings have both theoretical and practical significance. They contribute to a deeper understanding of transformer model behavior under different linguistic conditions and emphasize the need to adapt evaluation methodologies for cross-lingual scenarios. From a practical perspective, they enable the informed application of STS models across a range of tasks, such as information retrieval, classification, automated text evaluation, and other applied scenarios.

Overall, this study highlights the continued relevance of the problem addressed and demonstrates the effectiveness of comprehensive evaluation approaches for transformer models. Such approaches enable the selection of models that can simultaneously achieve high accuracy in multilingual environments and meet the requirements of specific application scenarios.

Finally, a promising direction for future research is the development of new evaluation metrics for semantic similarity models that go beyond correlation coefficients and account for practical applicability. In particular, combining correlation-based and ranking-based metrics, along with a more detailed analysis of error distributions, may provide deeper insights, especially considering the observed tendency of models to overestimate similarity in the low-similarity range. Another important direction is the adaptation of models to domain-specific applications, such as legal or medical texts, where additional fine-tuning can improve accuracy, robustness, and the ability to handle highly specialized data. These directions can contribute to the development of practical recommendations for model selection, evaluation strategies, and deployment in real-world systems.

The results of this research were obtained under the international research project INITIATE under the grant No. 101136775-HORIZON-WIDERA-2023-ACCESS-03.

#### Conflict of Interest

The authors declare that they have no conflicts of interest, including financial, personal, copyright,

or any other conflicts that could influence the research or the results published in this article.

#### Funding

The study was conducted without financial support.

#### Data Availability

The manuscript contains data in the form of electronic supplementary material.

#### Use of Artificial Intelligence

During the preparation of this article, the GPT-5 mini model was used for technical proofreading of the text, including checking spelling, punctuation, and style. All changes suggested by this tool were critically reviewed, edited, and verified by the authors. The main content, scientific ideas, and conclusions belong exclusively to the authors and have not been influenced by any third parties.

#### References

1. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W. (2013), "SEM 2013 shared task: Semantic Textual Similarity", *Joint Conference on Lexical and Computational Semantics*, Vol. 1, pp. 32–43.
2. anikol12 (2026), "STSB-UK", *Huggingface.co*. Available at: <https://huggingface.co/datasets/anikol12/STSB-UK> (Accessed 1 Apr. 2026).
3. Artetxe, M., Schwenk, H. (2019), "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond", *Transactions of the Association for Computational Linguistics*, Vol. 7, pp. 597–610. DOI: [https://doi.org/10.1162/tacl\\_a\\_00288](https://doi.org/10.1162/tacl_a_00288)
4. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L. (2017), "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation", *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*. DOI: <https://doi.org/10.18653/v1/S17-2001>
5. Chang, T.A., Arnett, C., Tu, Z., Bergen, B. (2024), "When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 4074–4096. DOI: <https://doi.org/10.18653/v1/2024.emnlp-main.236>
6. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020), "Unsupervised Cross-lingual Representation Learning at Scale", *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. DOI: <https://doi.org/10.18653/v1/2020.acl-main.747>
7. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, Vol. 1, pp. 4171–4186. DOI: <https://doi.org/10.18653/v1/n19-1423>
8. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W. (2022), "Language-agnostic BERT Sentence Embedding", *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. DOI: <https://doi.org/10.18653/v1/2022.acl-long.62>
9. Gorokhovatskyi, V., Tvoroshenko, I., Kobylin, O., Vlasenko, N. (2023), "Search for Visual Objects by Request in the Form of a Cluster Representation for the Structural Image Description", *Advances in Electrical and Electronic Engineering*, Vol. 21, No. 1. DOI: <https://doi.org/10.15598/aeec.v21i1.4661>

10. Kobylin, I., Nikolaichuk, A. (2024), "Monitoring and Diagnosing Faults in Online Mode Using Time Series Data", *Information Processing Systems*, No. 3(178), pp. 27–32. DOI: <https://doi.org/10.30748/soi.2024.178.03>
11. Kour, G., Ackerman, S., Farchi, E., Raz, O., Carmeli, B., Tavor, A.A. (2022), "Measuring the Measuring Tools: An Automatic Evaluation of Semantic Metrics for Text Corpora", *Proceedings of the Workshop on General Evaluation of NLP Models*. DOI: <https://doi.org/10.18653/v1/2022.gem-1.35>
12. Maksymenko, D., Turuta, O. (2024), "Interpretable Conversation Routing via the Latent Embeddings Approach", *Computation*, Vol. 12, No. 12, p. 237. DOI: <https://doi.org/10.3390/computation12120237>
13. Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013), "Efficient Estimation of Word Representations in Vector Space", *arXiv*. DOI: <https://doi.org/10.48550/arXiv.1301.3781>
14. Panchenko, D., Maksymenko, D., Turuta, O., Luzan, M., Tytarenko, S. (2022), "Ukrainian News Corpus as Text Classification Benchmark", *Communications in Computer and Information Science*, pp. 550–559. DOI: [https://doi.org/10.1007/978-3-031-14841-5\\_37](https://doi.org/10.1007/978-3-031-14841-5_37)
15. Pennington, J., Socher, R., Manning, C. (2014), "Glove: Global Vectors for Word Representation", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543. DOI: <https://doi.org/10.3115/v1/d14-1162>
16. Poliak, A. (2020), "A Survey on Recognizing Textual Entailment as an NLP Evaluation", *Proceedings of the Workshop on Evaluation for NLP*. DOI: <https://doi.org/10.18653/v1/2020.eval4nlp-1.10>
17. Reimers, N., Gurevych, I. (2019), "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. DOI: <https://doi.org/10.18653/v1/d19-1410>
18. Reimers, N., Gurevych, I. (2020), "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. DOI: <https://doi.org/10.18653/v1/2020.emnlp-main.365>
19. Reimers, N., Beyer, P., Gurevych, I. (2016), "Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity", *Proceedings of the International Conference on Computational Linguistics*, pp. 87–96.
20. Mehri, S., Eric, M., Hakkani-Tur, D. (2020), "DialogGLUE: A Natural Language Understanding Benchmark for Task-Oriented Dialogue", *arXiv*. DOI: <https://doi.org/10.48550/arXiv.2009.13570>
21. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S. (2018), "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding", *Proceedings of the Workshop on Evaluation of NLP Systems*. DOI: <https://doi.org/10.18653/v1/W18-5446>
22. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M. (2020), "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers", *arXiv*. DOI: <https://doi.org/10.48550/arXiv.2002.10957>
23. Yakovleva, O., Kovtunencko, A., Liubchenko, V., Honcharenko, V., Kobylin, O. (2023), "Face Detection for Video Surveillance-Based Security System", *Proceedings of the International Conference on Computational Linguistics and Intelligent Systems*, Vol. III, pp. 69–86.
24. Zesch, T. (2010), "Study of Semantic Relatedness of Words Using Collaboratively Constructed Semantic Resources", *Dissertation*, p. 130. Available at: <https://d-nb.info/1001286782> (Accessed 1 Apr. 2026).
25. Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., Zhang, M., Li, W., Zhang, M. (2024), "mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1393–1412. DOI: <https://doi.org/10.18653/v1/2024.emnlp-industry.103>

Received (Надійшла) 06.03.2026

Accepted for publication (Прийнята до друку) 09.04.2026

Publication date (Дата публікації) 29.05.2026

#### Відомості про авторів / About the Authors

**Ніколайчук Анна Ігорівна** – Харківський національний університет радіоелектроніки, здобувач другого рівня вищої освіти, Харків, Україна;

**Anna Nikolaichuk** – Kharkiv National University of Radio Electronics, Master's student, Kharkiv, Ukraine;

e-mail: [anna.nikolaichuk@nure.ua](mailto:anna.nikolaichuk@nure.ua)

ORCID ID: <https://orcid.org/0009-0001-8643-332X>

**Кобилін Олег Анатолійович** – кандидат технічних наук, доцент, Харківський національний університет радіоелектроніки, завідувач кафедри інформатики, Харків, Україна;

**Oleg Kobylin** – PhD (Technical Sciences), Associate Professor, Kharkiv National University of Radio Electronics, Head of the Department of Informatics, Kharkiv, Ukraine;

e-mail: [oleg.kobylin@nure.ua](mailto:oleg.kobylin@nure.ua)

ORCID ID: <https://orcid.org/0000-0003-0834-0475>

**Кобилін Ілля Олегович** – кандидат технічних наук, Харківський національний університет радіоелектроніки, старший викладач кафедри інформатики, Харків, Україна;

**Ilya Kobylin** – PhD (Technical Sciences), Kharkiv National University of Radio Electronics, Senior Professor of the Department of Informatics, Kharkiv, Ukraine;

e-mail: [ilya.kobylin@nure.ua](mailto:ilya.kobylin@nure.ua)

ORCID ID: <https://orcid.org/0000-0002-4552-9616>

**Путятіна Олександра Євгенівна** – кандидат технічних наук, Харківський національний університет радіоелектроніки, старший викладач кафедри інформатики, Харків, Україна;

**Oleksandra Putyatina** – PhD (Technical Sciences), Kharkiv National University of Radio Electronics, Senior Professor of the Department of Informatics, Kharkiv, Ukraine;

e-mail: [oleksandra.putyatina@nure.ua](mailto:oleksandra.putyatina@nure.ua)

ORCID ID: <https://orcid.org/0000-0003-4853-7125>

## КОМПЛЕКСНЕ ОЦІНЮВАННЯ ТРАНСФОРМЕРНИХ МОДЕЛЕЙ ДЛЯ ЗАДАЧІ СЕМАНТИЧНОЇ ПОДІБНОСТІ РЕЧЕНЬ ДЛЯ АНГЛІЙСЬКОЇ ТА УКРАЇНСЬКОЇ МОВ

**Предметом дослідження** є трансформерні моделі семантичної подібності речень і підходи до їх оцінювання в одномовних і міжмовних сценаріях для англійської та української мов. **Мета роботи** – оцінити й порівняти ефективність трансформерних моделей на англійськомовних, українськомовних і англо-українських парах речень з огляду на різні підходи до оцінювання їх якості та практичної придатності. Зважаючи на окреслену мету, необхідно було виконати такі **завдання**: порівняти моделі на одномовних і міжмовних наборах даних; проаналізувати результати за коефіцієнтами Пірсона та Спірмена; оцінити практичну придатність моделей способом класифікації прогнозів за величиною похибки; дослідити вплив мовного чинника на точність моделей. **Методи.** У дослідженні використано трансформерні моделі GTE, LaBSE, MiniLM та MPNet і набір даних STS-B, його українськомовну й англо-українську версії. Для обчислення подібності застосовано косинусну міру, а результати оцінено за допомогою коефіцієнтів кореляції Пірсона та Спірмена й класифікації прогнозів за порогом похибки. **Результати дослідження.** Визначено, що модель GTE демонструє найкращу загальну ефективність за сукупністю метрик, а MiniLM забезпечує оптимальний баланс між точністю та обчислювальними витратами. З'ясовано, що високі значення кореляції не завжди відповідають високій частці коректних прогнозів, що вказує на обмеженість традиційних підходів оцінювання. Встановлено систематичну тенденцію моделей до переоцінювання подібності в низькому діапазоні, а також виявлено, що міжмовні пари можуть демонструвати вищу точність за пороговими оцінками, незважаючи на нижчі значення кореляції, що свідчить про різну поведінку моделей залежно від типу метрики. **Висновки.** Обґрунтовано доцільність використання комбінованого підходу для оцінювання моделей семантичної подібності, який дає змогу більш повно відтворити їх реальну ефективність. Досягнуті результати підтверджують необхідність врахування мовної специфіки та практичних вимог задач у виборі моделей і підходів до їх оцінювання та наголошують на важливості переходу до комплексних підходів оцінювання.

**Ключові слова:** семантична подібність; трансформерні моделі; оцінювання моделей; коефіцієнт кореляції Пірсона; коефіцієнт кореляції рангу Спірмена; англійська мова; українська мова; міжмовна подібність; практична ефективність моделей.

### Бібліографічні описи / Bibliographic descriptions

Ніколайчук А. І., Кобилін О. А., Кобилін І. О., Путятіна О. Є. Комплексне оцінювання трансформерних моделей для задачі семантичної подібності речень для англійської та української мов. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 284–296. DOI: <https://doi.org/10.30837/0135-1710.2026.189.284>

Nikolaichuk, A., Kobylin, O., Kobylin, I., Putyatina, O. (2026), "A comprehensive evaluation of transformer models for sentence-level semantic similarity in English and Ukrainian", *Management Information System and Devices*, No. 2 (189), P. 284–296. DOI: <https://doi.org/10.30837/0135-1710.2026.189.284>



Островецький С. В.

## ЗАСТОСУВАННЯ ШТУЧНОГО ІНТЕЛЕКТУ В ТРАНСПОРТНІЙ ЛОГІСТИЦІ: КОМПЛЕКСНИЙ ОГЛЯД МОДЕЛЕЙ, МЕТОДІВ І ПРАКТИЧНИХ РІШЕНЬ

**Предмет дослідження** охоплює теоретичні та прикладні аспекти впровадження технологій штучного інтелекту (ШІ) в процеси транспортної логістики, зокрема математичне моделювання, алгоритмічну оптимізацію та автоматизацію рішень в умовах переходу до Індустрії 5.0. **Мета роботи** полягає в комплексному аналізі сучасних моделей, алгоритмів і практичних рішень на базі ШІ для підвищення ефективності функціонування логістичних мереж. Особливо наголошено на вивченні синергії детермінованих методів, машинного навчання й генеративних моделей. **Завдання:** систематизувати підходи до оптимізації; порівняти ефективність статистичних методів та ML; проаналізувати вплив евристичних алгоритмів на маршрутизацію; розробити структурно-логічну схему взаємодії компонентів цифрового двійника; оцінити потенціал GenAI. **Методи:** системний аналіз, порівняльний аналіз (ARIMA, ETS, LSTM), концептуальне й математичне моделювання (для формалізації архітектури логістичної екосистеми). **Результати.** Систематизовано показники ефективності: доведено, що використання LSTM-мереж знижує похибку прогнозування попиту (MAPE) до 2,3% порівняно з 3,7–4,1% у традиційних методів. Визначено, що застосування CNN для транспортного планування підвищує вчасність доставки до 98%, забезпечуючи економію витрат на рівні 6%. Обґрунтовано доцільність біонічних методів: мурашині алгоритми скорочують витрати палива на 35,25% і час доставки на 30,12%, а генетичні знижують операційні витрати на 4,5%. Розроблено структурно-логічну схему екосистеми "Логістика 5.0", що описує наскрізний потік даних від IoT-сенсорів крізь ML-ядро до генеративного інтерфейсу. Запропоновано математичну формалізацію цієї екосистеми у вигляді упорядкованого кортежу множин і функціональних відображень, що дає змогу розмежувати зони відповідальності алгоритмів прогнозування, оптимізації та генерації рішень. **Висновки.** Обґрунтовано перехід до адаптивних "розумних" мереж. Підтверджено, що інтеграція генеративного ШІ сприяє гуманізації технологій, спрощуючи взаємодію "людина – машина". Доведено економічну доцільність упровадження комплексних ШІ-систем і перспективи використання цифрових двійників для мінімізації ризиків.

**Ключові слова:** логістика; штучний інтелект; оптимізація; нейромережі; прогнозування; маршрутизація; алгоритми; автоматизація.

### Вступ

Стрімка цифровізація глобальної економіки [1] й експоненційне зростання обсягів електронної комерції вимагають фундаментального перегляду підходів до управління транспортно-логістичними системами. В умовах високої невизначеності ринку, нестабільності ланцюгів постачання й зростання вимог клієнтів до швидкості доставки традиційні лінійні моделі управління втрачають свою ефективність [2]. Сучасна логістика трансформується в складну екосистему, де критичним фактором успіху стає здатність обробляти великі масиви даних у реальному часі й приймати на їх основі адаптивні рішення.

Забезпечення стійкості та гнучкості логістичних мереж неможливе без упровадження новітніх інформаційних технологій. У цьому контексті штучний інтелект (ШІ) є не просто допоміжним інструментом автоматизації, а ключовим драйвером переходу до концепції "розумної логістики" (*Smart Logistics*) [3] та "Логістики 5.0" [4]. Застосування

методів машинного навчання, нейронних мереж, генетичних алгоритмів і комп'ютерного зору уможливорює розв'язання задач багатокритеріальної оптимізації, що раніше вважалися нерозв'язними через високу обчислювальну складність.

Актуальність теми дослідження зумовлена необхідністю систематизації та аналізу наявних моделей і методів штучного інтелекту для підвищення ефективності транспортної логістики. Упровадження інтелектуальних систем дає змогу мінімізувати операційні витрати, оптимізувати маршрутизацію, підвищити точність прогнозування попиту й забезпечити прозорість ланцюгів постачання. Однак, попри значний прогрес у цій сфері, існує потреба в комплексному огляді практичних рішень і оцінюванні їх реального впливу на ключові показники ефективності логістичних процесів.

### Аналіз останніх досліджень і публікацій

Проблема застосування штучного інтелекту в логістиці є предметом активних наукових

дискусій [5], які еволюціонують від класичних задач оптимізації до комплексних адаптивних систем. Сучасні дослідження можна класифікувати за трьома основними напрямками: математичне моделювання мереж, алгоритмічна маршрутизація та застосування новітніх методів машинного навчання.

### **Математичне моделювання та багатокритеріальна оптимізація**

Сучасні системи оптимізації в транспортній логістиці базуються на складних математичних моделях, що дають змогу одночасно брати до уваги множинні критерії ефективності (витрати, час доставки, коефіцієнт готовності мережі та живучість системи) [6]. Практичне застосування цих принципів демонструється через моделі оптимізації багаторівневих мереж. Експериментальні дослідження оптимізації маршрутів перевезень у системі замкненої логістики виявили поліноміальну часову складність окресленого підходу [7]. Отримана середня похибка  $\bar{\varepsilon} = 0,024$  підтверджує практичну застосовність методу для реальних логістичних завдань, оскільки така мінімальна розбіжність сприяє ефективному масштабуванню математичної моделі на великі логістичні мережі без значних втрат у точності планування [7]. Для формалізованого опису топології мереж використовується структурне подання, що забезпечує систематичний аналіз різних конфігурацій логістичних мереж [7].

### **Алгоритмічні підходи до маршрутизації**

Ключовим завданням транспортної логістики залишається пошук оптимальних маршрутів. Порівняльний аналіз алгоритмів пошуку найкоротшого шляху виявив суттєві розбіжності в їх ефективності [8]. Дослідження п'яти основних алгоритмів (Greedy, Best First Search, Dijkstra, A\* та Floyd-Warshall) на прикладі мережі доставки товарних вантажів в Індонезії продемонструвало, що алгоритм A\* має найкращі результати за співвідношенням швидкості й точності знаходження валідного маршруту [8]. У розглянутому дослідженні доведено, що, хоча цей алгоритм не завжди є найбільш швидким за абсолютним часом виконання, він гарантує оптимальність знайденого рішення, на відміну від швидших, але менш надійних "жадібних" алгоритмів [8].

Окрему низку досліджень присвячено біонічним методам. Мурашині алгоритми є інноваційним

підходом до логістичного управління [9], що симулюють поведінку мурашок для оптимізації планування маршрутів. Дослідження підтверджують, що імітація поведінки біологічних систем дає змогу суттєво скоротити довжину маршрутів, зекономити витрати палива та зменшити загальний час доставки.

Генетичні алгоритми (ГА) застосовуються в складних системах, зокрема в морській логістиці [10], де система підтримки прийняття рішень містить модулі евристики, оптимізації та симуляції. Наведені результати експериментів для системи з параметрами популяції 20 індивідуумів і 200 поколінь сприяли суттєвому покращенню показників ефективності. Дослідження оптимізації логістики демонструють, що застосування ГА для великомасштабних завдань (наприклад, управління морським флотом) здатне забезпечити значне зниження загальних операційних витрат і уможливити багатомільйонну економію для глобальних корпорацій [11].

### **Машинне навчання й новітні парадигми**

Значна увага в сучасній літературі приділяється інтеграції машинного навчання (ML) у ланцюги постачання. Систематичний огляд [12] визначає ключові напрями застосування ML, виокремлюючи нейронні мережі як домінуючий інструмент. Зокрема в роботі [13] проаналізовано застосування нейромереж для "справедливого планування турів" (Fair Tour Planning), що дає змогу балансувати навантаження між водіями.

Для комплексного розв'язання логістичних задач у сучасних наукових працях виокремлюється низка ключових методів. Серед класичних статистичних підходів можна виокремити ARIMA (інтегровану модель авторегресії – ковзного середнього), яка ефективна для стаціонарних часових рядів. З-поміж методів глибокого навчання (*Deep Learning*) аналізуються LSTM (рекурентні нейронні мережі довгої короткочасної пам'яті), що здатні вловлювати нелінійні довгострокові залежності в прогнозуванні, та CNN (згорткові нейронні мережі), які ефективно працюють з просторовими й топологічними даними. Для задач комп'ютерного зору (наприклад, на складах) досліджується комбінація алгоритму швидкої детекції об'єктів YOLOv4 і системи багатокритеріального трекінгу DeepSort [14].

Окремим вектором є розвиток генеративного штучного інтелекту (GenAI). GenAI відкриває

фундаментально нові можливості з метою покращення операційної діяльності завдяки здатності генерувати синтетичні дані для моделювання ризикових сценаріїв, використання семантичного пошуку для глибокого оцінювання постачальників, підтримки предиктивного обслуговування обладнання й оптимізації взаємодії людини й роботів на складах, хоча й несе певні технічні обмеження [15]. Це корелює з концепцією "Логістика 5.0", яка розглядає синергію людини й штучного інтелекту в контексті сталого розвитку.

Водночас упровадження інтелектуальних систем породжує нові виклики. Дослідники [16] наголошують на необхідності діагностики потенціалу "руйнування вартості" за умови використання ІІІ як бізнес-інструменту. Критично важливим є також питання прогнозування ризиків ланцюга постачання [17], де існує компроміс між продуктивністю алгоритмів та їх інтерпретованістю, що вимагає зваженого вибору між моделями типу "чорна скринька" та інтерпретованими деревами рішень.

Екологічні аспекти впровадження ІІІ, розглянуті в роботі [18], підтверджують, що інтелектуальна логістика має значний потенціал для зменшення вуглецевого сліду, проте потребує розв'язання проблема енергоспоживання самих обчислювальних систем.

### Постановка завдання

Попри значну кількість наукових праць, присвячених окремим аспектам оптимізації перевезень, аналіз останніх досліджень виявив низку нерозв'язаних питань. Більшість наявних підходів зосереджена або на класичних методах дослідження операцій, або на ізольованому застосуванні окремих алгоритмів машинного навчання без їх системного порівняння. Зокрема недостатньо вивченим залишається синергетичний ефект від поєднання традиційних евристичних методів із сучасними технологіями глибокого навчання та комп'ютерного зору в єдиному контурі управління логістикою. Також потребує більш детального дослідження потенціал інтеграції генеративного ІІІ для автоматизації прийняття рішень у складних ланцюгах постачання.

Мета роботи полягає в комплексному аналізі сучасних математичних моделей, алгоритмів і практичних рішень на базі штучного інтелекту

для підвищення ефективності функціонування логістичних мереж і ланцюгів постачання.

Для досягнення поставленої мети необхідно виконати такі завдання:

- систематизувати підходи до логістичної оптимізації та виявити ключові розбіжності між детермінованими й адаптивними моделями;
- провести порівняльний аналіз ефективності традиційних методів управління в логістиці та алгоритмів машинного навчання;
- оцінити перспективи й обмеження використання сучасних методів ІІІ в логістичних процесах;
- розробити концептуальну модель екосистеми "Логістика 5.0", яка описує наскрізну взаємодію цифрових двійників, алгоритмів прогнозування, оптимізації та генеративного штучного інтелекту в єдиному контурі управління.

### Матеріали й методи дослідження

Методологічну основу роботи становить системний підхід до аналізу й узагальнення науково-практичних результатів упровадження технологій штучного інтелекту в транспортну логістику. Дослідження базується на комплексному вивченні наукових праць, галузевих звітів і практичних кейсів, що дало змогу систематизувати наявні рішення й оцінити їх ефективність.

Власне дослідження реалізовано у форматі систематичного огляду й метааналізу емпіричних показників. Процес збору інформації охоплював пошук у міжнародних наукометричних базах за визначеними ключовими словами для публікацій останніх років.

З початкового масиву статей способом застосування жорстких критеріїв долучення (наявність кількісних метрик ефективності алгоритмів, детальний опис архітектури моделей) та вилучення (відсутність емпіричної бази або валідації на реальних логістичних даних) було відібрано ключові праці для глибокого порівняльного аналізу. Це дало змогу сформувати власну доказову базу для математичної агрегації розрізнених показників ефективності різних алгоритмів в єдину порівняльну матрицю впливу.

Для досягнення поставленої мети впроваджено такі методи теоретичного дослідження:

- бібліометричний аналіз – для відбору й систематизації джерельної бази, що охоплює сучасні

моделі машинного навчання, алгоритми оптимізації та системи комп'ютерного зору;

- *порівняльний аналіз* – для зіставлення показників ефективності (точність прогнозування, час виконання, економічна вигода) традиційних статистичних методів і алгоритмів на базі нейронних мереж (LSTM, CNN), описаних у фаховій літературі;

- *метод класифікації та групування* – для структурування підходів до логістичної оптимізації за такими напрямками: прогнозування попиту, маршрутизація, управління складом і управління ризиками;

- *синтез і узагальнення* – для формування висновків щодо доцільності переходу від традиційних детермінованих моделей управління до адаптивних інтелектуальних систем.

Окрім аналітичних методів, у роботі застосовано елементи математичного апарату теорії множин і концептуального моделювання. Це дало змогу здійснити математичну формалізацію запропонованої екосистеми "Логістика 5.0" (у вигляді впорядкованого кортежу множин і функціональних відображень) для розмежування зон відповідальності алгоритмів багатокритеріальної оптимізації та нейромережових моделей. Самі ж методи нейромережевого моделювання та евристичної оптимізації в цьому дослідженні є об'єктом метааналізу, результати якого математично агрегуються для об'єктивного порівняння їх ефективності.

Зважаючи на різноманітність логістичних завдань, у дослідженні розглянуто архітектурні принципи інноваційних методів машинного навчання й штучного інтелекту, водночас уникнено надмірної деталізації класичних статистичних підходів і базових алгоритмів пошуку, які слугують переважно детермінованим рівнем (baseline) для порівняння.

Зокрема у сфері прогнозування попиту наголошено на архітектурі рекурентних нейромереж LSTM. Вона розроблена спеціально для розв'язання проблеми зникаючого градієнта й використовує систему контрольних "вентилів" (вхідних, вихідних і забування), що дає змогу моделі вибірково зберігати або відкидати інформацію. Це забезпечує ефективну ідентифікацію нелінійних довгострокових залежностей і прихованих сезонних патернів, з якими не справляються стандартні лінійні моделі.

Для задач просторової оптимізації та транспортного планування розглядаються згорткові нейронні мережі. Принцип їх роботи полягає

у використанні згорткових фільтрів для багатошарової екстракції ознак із топологічних даних. CNN здатні "розуміти" просторову ієрархію транспортної мережі, перетворюючи географічні координати й матриці відстаней у структуровані карти ознак (feature maps).

Щодо алгоритмічних підходів до маршрутизації, то класичні методи пошуку найкоротшого шляху (зокрема алгоритм Дейкстри чи A\*) розглядаються лише як статична основа. Натомість основну увагу приділено принципам роботи біонічних методів – мурашиних (ACO) і генетичних алгоритмів (GA). Їх архітектура базується на стохастичній імітації еволюційних процесів і ройового інтелекту. Такий підхід дає змогу уникати збіжності до локального оптимуму й ефективно розв'язувати NP-складні задачі динамічної маршрутизації в умовах постійної зміни зовнішніх факторів.

Для автоматизації складських процесів розглядається синергія методів комп'ютерного зору: YOLOv4 і DeepSort. Принцип роботи YOLO полягає в однопрохідному розділенні зображення на сітку й прогнозуванні обмежувальних рамок з імовірностями класів. Інтеграція з алгоритмом DeepSort доповнює цю архітектуру рекурсивним фільтром Калмана, що допомагає математично прогнозувати кінематичну траєкторію об'єктів і стабільно відстежувати їх навіть під час візуального перекриття в умовах щільного складського трафіку.

Генеративний штучний інтелект розглядається як інструмент високорівневої оркестрації. На основі архітектури трансформерів і великих мовних моделей GenAI здатний генерувати синтетичні дані для моделювання стрес-сценаріїв у ланцюгах постачання, а також здійснювати глибокий семантичний аналіз неструктурованої текстової інформації для динамічного оцінювання надійності логістичних вузлів.

Інформаційною базою дослідження слугували відомості про результати застосування евристичних алгоритмів (генетичних, мурашиних) і моделей глибокого навчання (YOLOv4, DeepSort) у реальних логістичних операціях, отримані з опублікованих емпіричних досліджень і галузевих експериментів [9–11, 14].

Оцінка ефективності була основана на порівнянні кількісних показників, наведених у проаналізованих джерелах. Як наслідок, показники було розмежовано на дві групи: критерії якості



математичних моделей (зокрема середня абсолютна відсоткова похибка – MAPE) та критерії якості їх практичного застосування (скорочення витрат палива й підвищення вчасності доставки). Для комплексного порівняння в роботі проаналізовано такі методи: LSTM – рекурентні нейромережі, здатні ідентифікувати нелінійні залежності в довгих часових рядах; ARIMA – класичні статистичні моделі, що базуються на припущенні про стаціонарність ряду; CNN, які дають змогу здійснювати багатопланову екстракцію просторових ознак із топологічних даних; системи комп'ютерного зору, де комбінуються алгоритми детекції об'єктів (YOLOv4) та їх рекурсивного відстеження (DeepSort).

### Результати досліджень та їх обговорення

Комплексний аналіз науково-практичних джерел дає змогу констатувати, що інтеграція штучного інтелекту в логістику трансформувє галузь із лінійної моделі "постачання – збут" у динамічну екосистему адаптивного управління. Отримані результати доцільно систематизувати за ключовими логістичними завданнями, що дає змогу об'єктивно оцінити ефективність традиційних та інтелектуальних методів.

#### Прогнозування й планування попиту

Упровадження технологій машинного навчання кардинально змінює підходи до прогнозування в транспортній логістиці. Особливу увагу привертають результати порівняльного аналізу методів прогнозування попиту. Дослідження застосування рекурентних нейронних мереж демонструє їх значні переваги порівняно з традиційними стохастичними методами. Результати експериментів наведені в роботі *V. Pasupuleti* та співавторів показують, що LSTM-мережі досягають середньої абсолютної відсоткової похибки (MAPE) лише 2,3%, що суттєво перевершує показники класичних моделей: ARIMA з 3,7% MAPE та ETS з 4,1% MAPE [19]. Такі результати зафіксували автори під час прогнозування попиту для 12 млн комбінацій "товар – локація" (SKU) за шестимісячний період. Статистична надійність цих результатів підтверджується високим значенням F-статистики 1245,61 за умови  $p < 0,001$ . Природа такої переваги нейромережових підходів над традиційною статистикою криється в архітектурі моделей.

Стохастичні моделі типу ARIMA мають тенденцію згладжувати раптові коливання попиту, сприймаючи їх як статистичний шум, що призводить до систематичної недооцінки потреб у періоди пікових навантажень. Натомість механізм "пам'яті" LSTM дає змогу адаптивно реагувати на складні патерни взаємодії безлічі зовнішніх факторів, мінімізуючи ефект відкладеного дефіциту.

Вибір архітектури рекурентних нейронних мереж, зокрема LSTM, для прогнозування попиту зумовлений їх здатністю до розв'язання проблеми зникаючого градієнта, що є критичним у роботі з довгими часовими рядами в логістиці. На відміну від класичних моделей ARIMA, які базуються на припущенні про стаціонарність ряду, LSTM ефективно ідентифікує нелінійні залежності та сезонні коливання з мінімальною MAPE на рівні 2,3%.

Інфраструктурна підтримка цих технологій для прогнозування попиту реалізується за допомогою платформ TensorFlow, PyTorch і Scikit-learn [20, 21].

Значний потенціал у цій сфері демонструє генеративний штучний інтелект. Його застосування підвищує точність прогнозів, що дає змогу скоротити втрати продажів (lost sales) до 65% і зменшити надлишкові запаси на 20–50% [15]. Наприклад, у фармацевтичній галузі завдяки цьому точність прогнозів зросла на 30%, а відходи простроченої продукції зменшилися на 20% [22]. Висока ефективність GenAI у зниженні втрат продажів пояснюється його здатністю виходити за межі аналізу суто внутрішніх історичних даних. Інструменти генеративного ШІ аналізують глобальні контекстуальні масиви в режимі реального часу, виявляючи приховані кореляції між попитом і зовнішніми подіями, що допомагає превентивно коригувати політику управління запасами.

#### Оптимізація маршрутизації та транспортне планування

Для завдань транспортного планування доцільність використання згорткових нейронних мереж була обґрунтована в дослідженнях *V. Pasupuleti* та співавторів [19]. Вони довели, що CNN уможливають здійснення багатопланової екстракції ознак із топологічних даних для оптимізації розподілу ресурсів і виявлення просторових зв'язків у ланцюгах постачання [19]. На думку авторів, інтеграція алгоритмів глибокого



навчання сприяє досягненню 95% точності в прогнозуванні термінів виконання замовлень, що безпосередньо корелює з підвищенням загальної вчасності доставки [19].

На стратегічному рівні транспортного планування згорткові нейронні мережі демонструють високу ефективність порівняно з класичними алгоритмами. Застосування моделі CNN, натренованої на масивному датасеті з 800 тис. записів (зокрема зображення карт маршрутів і географічні атрибути), дало змогу підвищити частку вчасних доставок з історичного середнього показника 94% до 98%. Цей результат має високу статистичну значущість ( $t(800) = 23.54$ ,  $p < 0.001$ ) і забезпечує середню економію витрат у розмірі 6% на одне відправлення [19]. Для оперативного пошуку найкоротшого шляху алгоритм A\* забезпечує оптимальність рішення із середнім часом виконання 7056,67 мс, перевершуючи більш швидкі "жадібні" алгоритми за показниками надійності знайденого маршруту [8]. Суттєві результати отримано також за умови впровадження біонічних евристичних алгоритмів. Використання мурашиних алгоритмів (ACO) для динамічного планування парку транспортних засобів сприяло скороченню витрати палива на 35,25% і часу доставки на 30,12% [9]. Ці показники було агреговано з результатів симуляційного моделювання, проведеного в дослідженні [9] для парку вантажних автомобілів. Генетичні алгоритми (ГА) продемонстрували ефективність у морській логістиці, де стратегічна оптимізація маршрутів флоту сприяла зниженню загальних операційних витрат на 4,5%, що еквівалентно значній економії ресурсів [10, 11]. Сучасні системи динамічного планування й управління автопарком інтегрують телематичні дані (наприклад, системи Galileo) для оптимізації маршрутів у реальному часі з огляду на погодні умови й трафік [20, 24].

#### **Управління складськими операціями та комп'ютерний зір**

Системи комп'ютерного зору революціонізують складські операції. Практична система трекінгу [14] довела ефективність комбінації алгоритмів YOLOv4 (для детекції об'єктів) і DeepSort (для їх відстеження). Ця архітектура використовує CNN для розділення зображення на сітку й рекурсивну фільтрацію Калмана для прогнозування стану цілі. Опубліковані

результати, основані на тестуванні систем комп'ютерного зору в умовах щільного складського трафіку, демонструють поріг толерантності  $\pm 20$  пікселів для горизонтального центрування, що дає змогу стабільно відстежувати вантажі та об'єкти навіть за умови їх взаємного перекриття. Технічний спектр рішень для автономної роботи на складах також передбачає застосування LiDAR і SLAM для навігації роботизованої техніки, а також гіперспектрального зображення (Hyperspectral Imaging) для автоматизованого контролю якості [23].

У сфері "останньої милі" підтверджено ефективність автономних транспортних засобів і БпЛА (дронів), що сприяє безперебійній доставці в режимі 24/7 та знімає обмеження, пов'язані з людським фактором [24].

#### **Комплексне оцінювання ефективності та управління ризиками**

Для комплексного оцінювання логістичних мереж використовуються математичні моделі на основі адитивно-мультиплікативних функцій корисності [7], описані формулою

$$P(s) = \sum_{i=1}^m \alpha_i \mu_i(s) + \sum_{i=1}^m \sum_{j=i}^m \alpha_{ij} \mu_i(s) \mu_j(s) + \dots, \quad (1)$$

де  $P(s)$  – загальна функція корисності (ефективності) логістичної мережі;  $\alpha_i$  – вагові коефіцієнти значущості окремих критеріїв;  $\mu_i(s)$ ,  $\mu_j(s)$  – функції корисності за  $i$ -м та  $j$ -м критерієм (наприклад, за часом доставки, витратами тощо);  $\alpha_{ij}$  – коефіцієнти, що беруть до уваги взаємовплив різних критеріїв.

Цей підхід допомагає математично обґрунтувати взаємовплив множинних параметрів ефективності.

Крім того, ефективним виявився гібридний метод ML з MCDM-агрегацією для оцінювання надійності постачальників [25]. Його висока результативність зумовлена тим, що алгоритми машинного навчання дають змогу динамічно коригувати вагу критеріїв і ефективно управляти в реальному часі багатьма параметрами, що конфліктують. Такий підхід забезпечує значно вищу надійність, гнучкість і об'єктивність оцінювання порівняно з традиційними методами багатокритеріального вибору [25].

У сфері прогнозування й управління ризиками [26] виявлено суттєвий компроміс

між точністю та інтерпретованістю алгоритмів. Як свідчать результати порівняльного аналізу, проведеного в дослідженні [17], метод опорних векторів (SVM) забезпечує високі показники точності ( $AP = 0,851$ ,  $F1 = 0,791$ ,  $Recall = 0,973$ ,  $MCC = 0,775$ ,  $Acc = 0,943$ ), але визначається низькою інтерпретованістю ("чорна скринька"). Водночас дерева рішень, маючи дещо нижчі показники точності ( $AP = 0,698$ ,  $F1 = 0,821$ ,  $Recall = 0,823$ ,  $MCC = 0,792$ ,  $Acc = 0,950$ ), забезпечують кращий баланс і є пріоритетними для задач, де необхідне чітке пояснення прийнятих рішень. Управління ризиками й оброблення великих масивів даних для розв'язання цих завдань ефективно забезпечують спеціалізовані платформи, як-от *Apache Mahout*. Для зниження аварійності в транспортних мережах упроваджуються телематичні моделі й технологія V2X (*Vehicle-to-Everything*), яка сприяє обміну інформацією між учасниками руху та інфраструктурою в реальному часі [26].

Поданий у статті процес порівняння методів базується на систематизації та математичній агрегації розрізнених метрик ефективності, які були опубліковані іншими дослідниками за результатами незалежних тестувань, які вони виконали на реальних масивах логістичних даних. Відсоткові показники ефективності, наведені у зведеній табл. 1, є результатом виокремлення й усереднення цих емпіричних показників.

Оскільки кожен алгоритм тестували автори першоджерел у різних умовах (детальні параметри вибірок, обсяги датасетів, як-от 800 тис. записів для CNN чи 12 млн комбінацій для LSTM, та умови симуляцій докладно розкрито у відповідних підрозділах вище), для коректного зіставлення вони були зведені до єдиної бази порівняння. Аналіз передбачав виокремлення ключових критеріїв (KPI) для кожної логістичної задачі (похибка прогнозування, вчасність доставки, економія витрат і середня точність) й агрегації результатів у єдину матрицю впливу. Зведені результати цього літературного метааналізу подано в табл. 1.

Таблиця 1. Ефективність застосування різних класів інтелектуальних алгоритмів у логістиці

| Метод                  | Кількісний результат (значення KPI)            | Особливості                                                | Переваги                                                            | Недоліки                                                                                                                                       |
|------------------------|------------------------------------------------|------------------------------------------------------------|---------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>LSTM</b>            | Похибка прогнозування (MAPE) = <b>2,3%</b>     | Рекурентна архітектура для оброблення довгих часових рядів | Висока точність прогнозування попиту порівняно з ARIMA (3,7%)       | Висока складність налаштування гіперпараметрів і потреба у великих історичних даних для навчання; висока енергоємність обчислювальних процесів |
| <b>CNN</b>             | Рівень вчасності доставки = <b>98%</b>         | Екстракція ознак із топологічних даних і зображень карт    | Зниження витрат на 6% унаслідок оптимізації планування              | Потребує великих масивів тренувальних даних                                                                                                    |
| <b>ACO</b> (мурашиний) | Зниження витрат палива: <b>35,25%</b>          | Симуляція біонічної поведінки агентів для пошуку шляхів    | Ефективне динамічне планування маршрутів у реальному часі           | Значна обчислювальна складність для масштабних мереж                                                                                           |
| <b>ГА</b> (генетичний) | Зниження операційних витрат: <b>4,5%</b>       | Популяційний пошук на основі механізмів еволюції           | Значна економія в стратегічній логістиці                            | Імовірність збіжності до локального оптимуму                                                                                                   |
| <b>SVM</b>             | Середня точність $AP = $ <b>0,85</b>           | Метод опорних векторів для класифікації ризиків            | Висока точність прогнозування збоїв у ланцюгах постачання           | Низька інтерпретованість рішень ("чорна скринька") порівняно з деревами рішень                                                                 |
| <b>YOLO + DeepSort</b> | Толерантність відстеження: $\pm $ <b>20 px</b> | Комбінація детекції об'єктів і рекурсивної фільтрації      | Стабільне відстеження вантажів навіть за умови взаємного перекриття | Для повноцінної автономної навігації потребує інтеграції з додатковими сенсорними системами (наприклад, LiDAR)                                 |

Наведені в табл. 1 результати засвідчують, що кожен із розглянутих інтелектуальних методів демонструє високу ефективність переважно в розв'язанні вузькоспеціалізованих, локальних задач. Зокрема моделі LSTM оптимізують виключно прогнозування попиту, біонічні алгоритми (ACO, ГА)

орієнтовані на маршрутизацію, а системи комп'ютерного зору (YOLO, DeepSort) – на складський моніторинг. Проте сучасні ланцюги постачання є складними взаємозалежними системами, де ізольована оптимізація одного етапу не здатна забезпечити наскрізну гнучкість усієї мережі. Саме ця

фрагментарність рішень зумовлює об'єктивну необхідність їх системного об'єднання. Подальший аналіз дав змогу розробити комплексну модель, що ілюструє синергетичний ефект від інтеграції цих розрізнених технологій (коли результати комп'ютерного зору й ML-прогнозування автоматично стають вхідними показниками для алгоритмів динамічної маршрутизації) у єдиному контурі управління.

### Синергетичний ефект технологій

Дослідження продемонструвало, що максимальний економічний ефект досягається не внаслідок ізольованого впровадження окремих

алгоритмів, а завдяки створенню комплексних екосистем. Поєднання комп'ютерного зору (YOLOv4, DeepSort) для автоматизації складів, генеративного ШІ для управління ланцюгами постачання й ML-алгоритмів для прогнозування створює синергетичний ефект, що є основою концепції "Логістика 5.0". Ринок генеративного ШІ в логістиці демонструє потенціал зростання у 38 разів протягом наступного десятиліття, що підтверджує стратегічну важливість цього напрямку.

Концептуальну модель інтеграції, яка демонструє взаємодію потоків даних та інструментів оброблення в єдиному контурі управління, зображено на рис. 1.

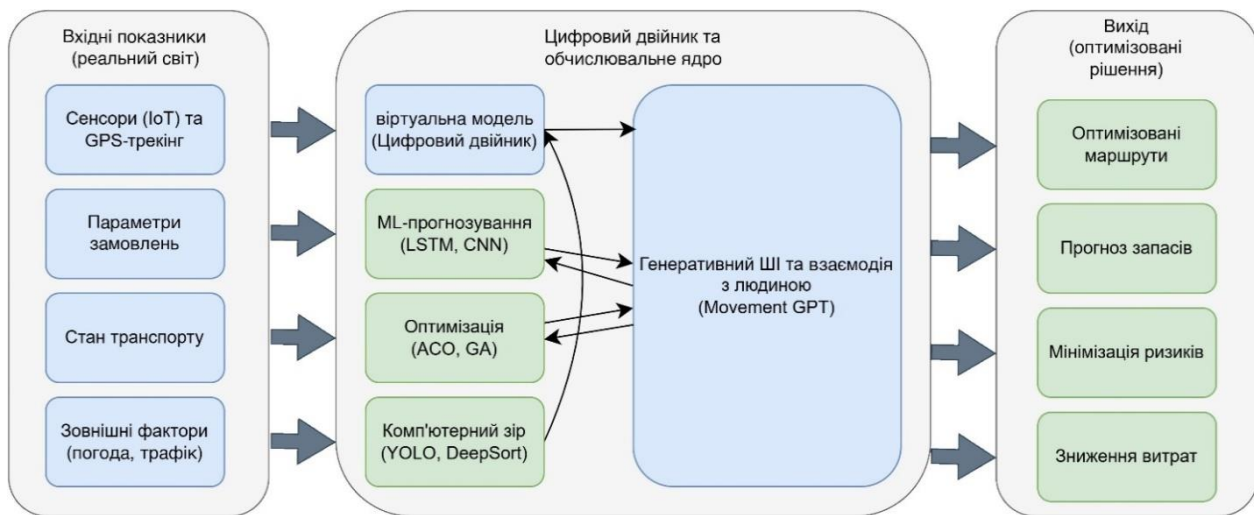


Рис. 1. Структурно-логічна схема екосистеми "Логістика 5.0" на базі синергії ШІ та цифрових двійників

### Математична формалізація моделі екосистеми "Логістика 5.0"

Для більш глибокого розуміння природи синергетичних зв'язків, зображених на рис. 1, доцільно формалізувати запропоновану екосистему  $S_{L5.0}$  як упорядкований кортеж множин і функціональних відображень. Це дає змогу перейти від концептуального опису до алгоритмічної реалізації моделі управління.

Систему можна подати в такому вигляді:

$$S_{L5.0} = \langle D_{in}, M_{DT}, \Psi_{AI}, R_{out} \rangle, \quad (2)$$

де  $D_{in}$  – вектор вхідних показників;  $M_{DT}$  – простір станів цифрового двійника;  $\Psi_{AI}$  – композиція інтелектуальних операторів;  $R_{out}$  – множина управлінських рішень.

Вектор вхідних показників ( $D_{in}$ ) формується з гетерогенних джерел інформації:

$$D_{in} = \{d_{IoT}, d_{GPS}, d_{order}, d_{env}\}, \quad (3)$$

де  $d_{IoT}$  – потокові показники із сенсорів (температура, вібрація, завантаження);  $d_{GPS}$  – геопросторові координати;  $d_{order}$  – параметри замовлень (обсяг, дедлайни);  $d_{env}$  – зовнішні стохастичні фактори (погода, затори), що обробляються модулем комп'ютерного зору (YOLOv4, DeepSort).

Оператор цифрового двійника ( $M_{DT}$ ) виконує функцію відтворення реального стану системи у віртуальний простір станів  $V$ :

$$M_{DT} : D_{in} \rightarrow V(t), \quad (4)$$

де  $V(t)$  – матриця стану системи в момент часу  $t$ , що містить положення всіх активів і рівень запасів.

Композиція інтелектуальних операторів ( $\Psi_{AI}$ ) є ядром системи й описується як суперпозиція функцій прогнозування, оптимізації та генерації:

$$\Psi_{AI}(V) = G_{GenAI}(\phi_{pred}(V), \xi_{opt}(V), Q_{human}), \quad (5)$$

де  $\phi_{pred}(V)$  – функція прогнозування (реалізується через LSTM/CNN), яка повертає вектор імовірних станів  $V(t+1)$ ;  $\xi_{opt}(V)$  – функція евристичної оптимізації (реалізується через ACO/GA), яка знаходить локальний екстремум цільової функції витрат;  $Q_{human}$  – запити або обмеження, впроваджені оператором природною мовою;  $G_{GenAI}$  – генеративна функція (Movement GPT), яка агрегує результати  $\phi$  та  $\xi$ , інтерпретує їх у контексті  $Q$  і формує остаточне рішення.

Множина вихідних рішень ( $R_{out}$ ) є результатом перетворення

$$R_{out} \rightarrow \{r_{route}^*, r_{inv}^*, r_{risk}^*\}, \quad (6)$$

де  $r_{route}^*$  – оптимальний маршрут;  $r_{inv}^*$  – скоригований план запасів;  $r_{risk}^*$  – предиктивне попередження про ризик.

Така формалізація допомагає чітко розмежувати зони відповідальності кожного алгоритму: ML-моделі відповідають за зменшення ентропії вхідних показників, евристичні алгоритми – за пошук в просторі рішень, а генеративний ШІ – за семантичну інтеграцію та адаптацію результатів під бізнес-контекст.

Саме наявність компонента  $G_{GenAI}$ , що діє як "оркестратор" між жорсткою логікою оптимізації та неструктурованими показниками реального світу, є ключовою відмінністю запропонованої моделі від традиційних автоматизованих систем управління (TMS).

### Висновки й перспективи подальшого розвитку

Проведене комплексне дослідження доводить, що інтеграція технологій штучного інтелекту є фундаментальною зміною парадигми управління транспортною логістикою. З огляду на окреслене завдання досягнуто ключових результатів.

1. Систематизовано підходи до логістичної оптимізації: обґрунтовано незворотний перехід від детермінованих лінійних моделей управління до адаптивних "розумних" мереж, здатних до самоорганізації.

2. Проведено порівняльний аналіз статистичних методів і алгоритмів ML: доведено статистично значущу перевагу нейромереж, зокрема LSTM-архітектури, яка знижує похибку прогнозування попиту (MAPE) до 2,3% проти 3,7–4,1% у класичних моделей ARIMA й ETS. Застосування CNN для стратегічного планування підвищує вчасність доставки до 98%.

3. Визначено вплив евристичних алгоритмів на маршрутизацію: встановлено, що мурашині алгоритми скорочують витрати палива на 35,25% і час доставки на 30,12%, а генетичні алгоритми знижують загальні операційні витрати на 4,5%.

4. Розроблено концептуальну модель: сформовано структурно-логічну схему й математичну формалізацію екосистеми "Логістика 5.0", де комп'ютерний зір (YOLOv4) та оптимізаційні алгоритми інтегровані в єдиний контур управління цифровим двійником.

5. Оцінено перспективи використання GenAI: підтверджено, що інтеграція генеративного ШІ спрощує взаємодію "людина – машина", забезпечує гуманізацію технологій і мінімізує системні ризики в ланцюгах постачання.

### Перспективи подальшого розвитку

Найбільш перспективним вектором розвитку є конвергенція генеративного ШІ з технологіями цифрових двійників. Це дасть змогу моделювати й оптимізувати логістичні процеси у віртуальному середовищі перед їх фізичною реалізацією, мінімізуючи ризики. Майбутні тренди також передбачають еволюцію бізнес-моделей до форматів Supply Chain as a Service (SCaaS) та Logistics as a Service (LaasS).

### Виклики й ризики

Крім переваг, аналіз виявив формування нових системних ризиків. Зростання автономності інтелектуальних систем підвищує їх вразливість до кібератак і технічних збоїв. Також існує проблема енергоємності великих мовних моделей і необхідність забезпечення інтерпретованості

("прозорості") алгоритмічних рішень для дотримання етичних норм і соціальної справедливості.

Отже, подальший науковий пошук має зосередитися на розробленні гібридних моделей, які поєднують високу точність нейромереж із прозорістю логічних правил, а також на створенні надійних протоколів кібербезпеки для захисту критичної логістичної інфраструктури.

#### Конфлікт інтересів

Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який

міг би вплинути на дослідження, а також на результати, опубліковані в цій статті.

#### Фінансування

Дослідження проводилося без фінансової підтримки.

#### Доступність даних

Рукопис не має пов'язаних даних.

#### Використання засобів штучного інтелекту

Автори підтверджують, що не застосовували технології ШІ для написання статті.

#### References

1. UNCTAD. (2024), Digital Economy Report 2024: Shaping an environmentally sustainable and inclusive digital future. United Nations. URL: <https://unctad.org/publication/digital-economy-report-2024>
2. Medvediev, Ye., Popova, Yu., Kovalenko, M. (2023), "Innovative artificial intelligence technologies in social enterprise logistics management", *Economy and Society*, No. 56. DOI: <https://doi.org/10.32782/2524-0072/2023-56-53>
3. Woschank, M., Rauch, E., Zsifkovits, H. (2020), "A Review of Further Directions for Artificial Intelligence, Machine Learning, and Deep Learning in Smart Logistics", *Sustainability*, Vol. 12, Art. 3760. DOI: <https://doi.org/10.3390/su12093760>
4. Skitsko, V. I. (2023), "Logistics 5.0: synergy of artificial intelligence and humans in the context of sustainable development", *Business Inform*, No. 11, pp. 174–179. DOI: <https://doi.org/10.32983/2222-4459-2023-11-174-179>
5. Younis, H., Sundarakani, B., Alsharairi, M. (2021), "Applications of Artificial Intelligence and Machine Learning within Supply Chains: Systematic review and future research directions", *Journal of Modelling in Management*. DOI: <https://doi.org/10.1108/JM2-12-2020-0322>
6. Beskorovainyi, V., Draz, O. (2021), "Mathematical Models of Decision Support in the Problems of Logistics Networks Optimization", *Innovative Technologies and Scientific Solutions for Industries*, No. 4 (18), pp. 5–14. DOI: <https://doi.org/10.30837/ITSSI.2021.18.005>
7. Beskorovainyi, V., Kuropatenko, O., Gobov, D. (2019), "Optimization of transportation routes in the system of closed logistics", *Innovative Technologies and Scientific Solutions for Industries*, No. 4 (10), pp. 24–32. DOI: <https://doi.org/10.30837/2522-9818.2019.10.024>
8. Wijaya, D. R., Athallah, A., Noor'afina, T. N., Telsoni, P. A., Budiwati, S. D. (2023), "Cargo route optimization using shortest path algorithms: Runtime and validity comparison", *Journal of Computer Science*, Vol. 19, No. 11, pp. 1369–1379. DOI: <https://doi.org/10.3844/jcsp.2023.1369.1379>
9. Sun, H. (2023), "Design logistics management system based on computer algorithm model", *Advances in Artificial Intelligence, Big Data and Algorithms*, IOS Press. DOI: <https://doi.org/10.3233/FAIA230856>
10. Bruzzone, A., Orsoni, A., Mosca, R., Revetria, R. (2002), "AI-based optimization for fleet management in maritime logistics", *Proceedings of the Winter Simulation Conference*, Vol. 2, pp. 1174–1182. DOI: <https://doi.org/10.1109/WSC.2002.1166375>
11. Chen, W., Men, Y., Fuster, N., Osorio, C., Juan, A. A. (2024), "Artificial Intelligence in Logistics Optimization with Sustainable Criteria: A Review", *Sustainability*, Vol. 16, Art. 9145. DOI: <https://doi.org/10.3390/su16219145>
12. Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., Fischl, M. (2021), "Artificial intelligence in supply chain management: A systematic literature review", *Journal of Business Research*, Vol. 122, pp. 502–517. DOI: <https://doi.org/10.1016/j.jbusres.2020.09.009>
13. Sentürk, C., Lindner, H.-G. (2019), "Fair Tour Planning: Managing Parcel Logistics by Neural Networks", University of Würzburg. DOI: <https://doi.org/10.20378/irb-58609>
14. Xie, T., Yao, X. (2023), "Smart Logistics Warehouse Moving-Object Tracking Based on YOLOv5 and DeepSORT", *Applied Sciences*, Vol. 13, No. 17, Art. 9895. DOI: <https://doi.org/10.3390/app13179895>



15. Khlie, K., Benmamoun, Z., Jebbor, I., Serrou, D. (2024), "Generative AI for enhanced operations and supply chain management", *Journal of Infrastructure, Policy and Development*, Vol. 8, No. 10, Art. 6637. DOI: <https://doi.org/10.24294/jipd.v8i10.6637>
16. Canhoto, A. I., Clear, F. (2020), "Artificial intelligence and machine learning as business tools: A framework for diagnosing value destruction potential", *Business Horizons*, Vol. 63, No. 2, pp. 183–193. DOI: <https://doi.org/10.1016/j.bushor.2019.11.003>
17. Baryannis, G., Dani, S., Antoniou, G. (2019), "Predicting supply chain risks using machine learning: The trade-off between performance and interpretability", *Future Generation Computer Systems*, Vol. 101, pp. 993–1004. DOI: <https://doi.org/10.1016/j.future.2019.07.059>
18. Gonçalves, R., Domingues, L. (2025), "Artificial Intelligence Driving Intelligent Logistics: Benefits, Challenges, and Drawbacks", *Procedia Computer Science*, Vol. 256, pp. 665–672. DOI: <https://doi.org/10.1016/j.procs.2025.02.165>
19. Pasupuleti, V., Thuraka, B., Kodete, C. S., Malisetty, S. (2024), "Enhancing Supply Chain Agility and Sustainability through Machine Learning: Optimization Techniques for Logistics and Inventory Management", *Logistics*, Vol. 8, No. 3, Art. 73. DOI: <https://doi.org/10.3390/logistics8030073>
20. Holovina, O. (2023), "Modern technologies in transport logistics management", *International Science Journal of Management, Economics & Finance*, Vol. 2, No. 3, pp. 35–42. DOI: <https://doi.org/10.46299/j.isjmef.20230203.04>
21. Du Plessis, M., Gerber, R., Goedhals-Gerber, L., Eeden, J. (2025), "Shaping the Future of Freight Logistics: Use Cases of Artificial Intelligence", *Sustainability*, Vol. 17, pp. 1–18. DOI: <https://doi.org/10.3390/su17041355>
22. Maniatis, P. (2025), "The Role of Artificial Intelligence in Supply Chain Management: A Quantitative Exploration of its Impact on Efficiency and Performance", *International Journal of Clinical Case Reports and Reviews*, Vol. 22, No. 4. DOI: <https://doi.org/10.31579/2690-4861/671>
23. Mozumder, M. A. I., Sumon, R. I., Khan, Z., Uddin, S. M. I., Khan, M. O., Kim, H.-C. (2024), "AI-Based Logistics System Overview and a Workflow for Digital Freight Forwarding in Logistics", *2024 International Conference on Advanced Communications Technology (ICACT)*, pp. 295–299. DOI: <https://doi.org/10.23919/ICACT60172.2024.10471983>
24. Adeoye, Y., Onotole, E. F., Ogunyankinnu, T., Aipoh, G., Osunkanmibi, A. A., Egbemhenghe, J. (2025), "Artificial Intelligence in Logistics and Distribution", *World Journal of Advanced Research and Reviews*, Vol. 25, No. 02, pp. 155–167. DOI: <https://doi.org/10.30574/wjarr.2025.25.2.0214>
25. Gidiagba, J. O., Tartibu, L. K., Okwu, M. O. (2025), "A systematic review of machine learning applications in sustainable supplier selection", *Decision Analytics Journal*, Art. 100547. DOI: <https://doi.org/10.1016/j.dajour.2025.100547>
26. Balan, G. S., Kumar, V. S., Raj, S. A. (2025), "Machine Learning and Artificial Intelligence Methods and Applications for Post-Crisis Supply Chain Resiliency and Recovery", *Supply Chain Analytics*. DOI: <https://doi.org/10.1016/j.sca.2025.100121>

Received (Надійшла) 13.03.2026

Accepted for publication (Прийнята до друку) 16.04.2026

Publication date (Дата публікації) 29.05.2026

#### Відомості про авторів / About the Authors

Островецький Сергій Володимирович – Запорізький національний університет, аспірант, Запоріжжя, Україна;  
 Serhii Ostrovetskyi – Zaporizhzhia National University, Postgraduate Student, Zaporizhzhia, Ukraine;  
 e-mail: [sergeu.ostrovetskiu.2001@gmail.com](mailto:sergeu.ostrovetskiu.2001@gmail.com)  
 ORCID ID: <https://orcid.org/0009-0005-4429-6071>

### APPLICATION OF ARTIFICIAL INTELLIGENCE IN TRANSPORT LOGISTICS: A COMPREHENSIVE REVIEW OF MODELS, METHODS AND PRACTICAL SOLUTIONS

**Subject matter** covers theoretical and applied aspects of implementing AI technologies in transport logistics, including mathematical modeling, algorithmic optimization, and decision automation in the context of the transition to Industry 5.0. **The Goal** is a comprehensive analysis of modern models, algorithms, and AI-based solutions to improve the efficiency of logistics networks. Particular emphasis is placed on the synergy of deterministic methods, machine learning, and generative models. **The Tasks:** classifying optimization approaches; comparing the efficiency of statistical methods versus ML; analyzing the impact of heuristic algorithms on routing; developing a structural-logical scheme of digital twin interaction; evaluating the potential of GenAI. **Methods:** system analysis, comparative analysis (ARIMA, ETS, LSTM), methods

of conceptual and mathematical modeling (to formalize the architecture of the logistics ecosystem). **Results.** Performance indicators are systematized: it is proved that LSTM networks reduce demand forecasting error (MAPE) to 2.3% compared to 3.7–4.1% for traditional methods. It is determined that applying CNN for transport planning increases delivery timeliness to 98%, ensuring cost savings of 6%. The feasibility of bionic methods is substantiated: ant colony algorithms reduce fuel costs by 35.25% and delivery time by 30.12%, while genetic algorithms reduce operating costs by 4.5%. A structural-logical scheme of the "Logistics 5.0" ecosystem is developed, describing the end-to-end data flow from IoT sensors through the ML core to the generative interface. A mathematical formalization of this ecosystem is proposed as an ordered tuple of sets and functional mappings, allowing to delineate the responsibilities of forecasting, optimization, and decision generation algorithms. **Conclusions.** The transition to adaptive "smart" networks is substantiated. It is confirmed that integrating generative AI contributes to the humanization of technologies by simplifying "human-machine" interaction. The economic feasibility of implementing complex AI systems and the prospects of using digital twins for risk minimization are proven.

**Keywords:** logistics; artificial intelligence; optimization; neural networks; forecasting; routing; algorithms; automation.

#### *Бібліографічні описи / Bibliographic descriptions*

Островецький С. В. Застосування штучного інтелекту в транспортній логістиці: комплексний огляд моделей, методів і практичних рішень. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 297–308. DOI: <https://doi.org/10.30837/0135-1710.2026.189.297>

Ostrovetskyi, S. (2026), "Application of artificial intelligence in transport logistics: a comprehensive review of models, methods and practical solutions", *Management Information System and Devices*, No. 2 (189), P. 297–308. DOI: <https://doi.org/10.30837/0135-1710.2026.189.297>

---

Stanislav Pohorielov, Yaroslav Balaba, Maksym Tatariiaants, Yevhen Berkunskyi, Alona Pavlenko

## A SYSTEMATIC SENSITIVITY STUDY OF PHYSICS-INFORMED NEURAL NETWORKS FOR EPIDEMIC TIME SERIES

This paper presents a systematic sensitivity analysis of Physics-Informed Neural Networks (PINN) applied to epidemic time-series modeling. **The objective** of the study is carrying out of a systematic analysis of the sensitivity of PINN models on real epidemic time series and identification of the key factors governing the quality of epidemic-wave dynamics reconstruction. **Tasks:** assessment of the impact of sliding-window length, the set of covariates, regularization parameters, and random initialization on PINN quality; benchmarking PINN against a classical baseline model without physical constraints. **Methods:** an extensive series of experiments across three countries with varying sliding-window length, window stride, fractional derivative order  $\alpha$ , stochastic stability, and sets of covariates; statistical analysis using the Wilcoxon test, OLS with HC3 correction, PCA, and Lasso regularization. **Results:** PINN almost consistently outperforms the baseline in reproducing epidemic time series, yielding higher  $R^2$  values and lower RMSE/MAE across all countries. A 14-day sliding window provides the optimal balance between model adaptability and stability, whereas extending the window length to 21 days leads to over-smoothing of the dynamics and loss of local sensitivity. The experimental results indicate that the effect of covariates does not exhibit a universally positive impact. In certain configurations and countries, additional predictors improve model quality, while in others they introduce noise owing to multicollinearity and poor data quality. The stability analysis confirms that PINN is largely insensitive to random initialization and regularization hyperparameters. The findings provide a clear characterization of PINN behavior in epidemiological modeling tasks and yield practical recommendations for parameter selection that ensure model reliability and reproducibility. **Conclusions:** PINN is a more effective alternative to classical models for epidemic modeling; a 14-day sliding window is optimal; the effect of covariates is unstable and country-dependent.

**Key words:** physics-informed neural networks (PINN); epidemic time series, sensitivity analysis; sliding window; covariates; model stability; machine learning; epidemiological modeling.

### 1. Introduction

Mathematical modeling of epidemic processes remains one of the key tools for analyzing the dynamics of infectious diseases and supporting decision-making in the field of public health [1]. During the COVID-19 pandemic, SIR and SEIR models were actively used to analyze the impact of anti-epidemic measures and to explore possible scenarios of tightening or relaxing control measures. In particular, it was shown that the use of Caputo fractional derivatives allows for the natural incorporation of memory and inertia effects [1, 2], while the application of neural networks enables the adaptive reproduction of the temporal variability of infection and recovery parameters [3–5]. These results formed the basis for the further development of models capable of combining the interpretability of analytical equations with the flexibility of data-driven methods [6–8].

### Literature Review

The classical SIR models proposed in [1] became the foundation of mathematical modeling of epidemics. The authors of [2] revisited classical SIR modeling in the

context of the COVID-19 pandemic. In [3], the SIR model is applied to forecast the spread of COVID-19 in Ukraine.

A separate line of research is related to the application of fractional derivatives in epidemic models. The authors of [9] proposed a fractional SIR model for COVID-19. In [10], the problem of optimal control based on a fractional SIR model with a Caputo derivative was investigated, and in [11], the order of the fractional derivative was optimized to improve modeling accuracy.

Physics-informed neural networks (PINNs) as an approach to solving differential equations using neural networks were systematized in [6, 7]. This approach combines the flexibility of machine learning with physical constraints, ensuring the interpretability and accuracy of models.

A comprehensive review of the current state of PINN methodology, including architectural solutions, optimization methods, and open problems, is presented in [12]. Wang et al. [13] investigated the causes of PINN training instability from the perspective of neural tangent kernels and proposed adaptive weights to balance the components of the loss function. Kharazmi et al. [14] applied PINN to the parameter identification of fractional epidemic models, demonstrating the advantages of

a physics-informed approach for forecasting the dynamics of COVID-19. The issue of self-adaptive PINN training is discussed in [15], and a review of PINN applications in continuum mechanics is provided in [16].

An evaluation of ensemble forecasts of COVID-19 mortality, highlighting the importance of validating models on real data, is presented in [17]. Theoretical guarantees of PINN generalizability for inverse problems are investigated in [18].

However, most previous studies have focused either on synthetic data or on reconstructing hidden trajectories within the framework of fractional models  $S(t)$ ,  $I(t)$ ,  $R(t)$  within the framework of fractional models. Much less attention has been paid to the practical aspects of applying Physics-Informed Neural Networks (PINNs) to real-world epidemic time series, particularly issues of stability, parameter sensitivity, and the influence of input data structure. In real-world conditions, models face a number of challenges: data noise, uneven sampling, changes in behavioral and social factors, multicollinearity of additional variables, as well as the dependence of results on the choice of the interval length over which the model is trained and the method of its progression along the time series. Analysis of these factors complicates the practical application of PINN in epidemiological modeling tasks [6–8]. The absence of such systematic checks significantly complicates the practical application of PINN.

The aim of this work is to conduct a systematic sensitivity analysis of PINN models on real epidemic time series and to identify the key factors influencing the quality of the reproduction of epidemic wave dynamics.

The article investigates the influence of three key groups of factors:

- 1) structural modeling parameters, specifically the length of the data interval used for local model training and the frequency of its shifting;
- 2) information content, i.e., the set of additional variables (covariates) that may influence the dynamics;
- 3) stochastic and regularization aspects that determine the reproducibility of results and the model's robustness to variations in initial conditions.

A large-scale series of experiments was conducted for three countries, which allowed us to assess the stability of the PINN, compare it with a classical model without physical constraints, and form a generalized picture of the model's behavior under various conditions.

Particular attention was paid to how changes in the length of the training interval, the frequency of its shift, and the set of covariates affect the accuracy of reproducing epidemic dynamics, as well as to the extent to which the results remain stable when stochastic factors change.

Thus, this work continues the logical line of the authors' previous research, moving from the theoretical analysis of fractional models and hybrid approaches to the practical evaluation of their effectiveness on real data. The results obtained allow us to formulate recommendations regarding the selection of PINN model parameters that ensure their reliability, robustness, and interpretability in epidemiological modeling tasks.

## 2. Methodology

This section describes the data structure, the preprocessing procedure, the principles for forming training samples, the model structure, optimization parameters, the approach to sensitivity analysis, and the statistical processing of results. Particular attention is paid to ensuring the reproducibility of experiments and the validity of comparisons between the classical model and physically informed neural networks (PINN).

### 2.1. Data and Rationale for Country Selection

For our study, we used open daily incidence data published by the **Our World in Data** project (University of Oxford) [19]. From this dataset, we selected three countries: Germany, Italy, and Ukraine. This selection is formally justified based on three factors.

First, these countries have different population sizes, which allows us to assess the stability of the models under various demographic conditions. Germany represents a large European population, Italy a medium-sized one, and Ukraine a medium-sized one with a different age and mobility structure.

Second, the countries demonstrate varying quality and completeness of epidemic data. Germany's data are among the most stable and systematic in Europe; Italy is characterized by high variability at the onset of waves; Ukraine exhibits significant noise and uneven testing. This allows us to assess how robust the PINN models are to varying data quality.

Third, these countries have different covariate profiles, particularly regarding vaccination, testing, and the strictness index of restrictions. This allows us to

investigate whether the effect of covariates is universal or depends on the country.

For each country, we selected one epidemic wave covering the period from the autumn surge to the spring decline. The main observed variable is the daily incidence rate per 100,000 population:

$$Y_{\text{data}}(t) = \frac{\text{new\_case\_smoothed}(t)}{N} \cdot 10000 \quad (2.1)$$

where  $\text{new\_case\_smoothed}(t)$  is the 7-day smoothed number of new cases, and  $N$  is the country's population.

The data is sorted by date and divided into training (60%) and test (40%) sets.

## 2.2. Formal Formulation of the Problem

All models operate within the framework of a discrete SIR structure [1]. Let:

$S(t)$  – the number of susceptible individuals at time  $t$ ;

$I(t)$  – the number of infected individuals;

$R(t)$  – the number of removed (recovered or deceased);

$\tilde{N} = 100000$  – effective population (scaling for stability).

Then the following holds:

$$S(t) + I(t) + R(t) = \tilde{N}. \quad (2.2)$$

In the classical SIR model, the dynamics are given by:

$$\begin{aligned} S(t+1) &= S(t) - \beta(t)S(t)I(t), \\ I(t+1) &= I(t) + \beta(t)S(t)I(t) - \gamma(t)I(t), \\ R(t+1) &= R(t) + \gamma(t)I(t) \end{aligned} \quad (2.3)$$

where  $\beta(t)$  is the transmission coefficient, and  $\gamma(t)$  is the withdrawal coefficient.

In this study,  $\beta(t)$  and  $\gamma(t)$  are assumed to be constant within each sliding window but may vary between windows.

The model incidence is defined as:

$$Y_{\text{model}}(t) = \beta(t)S(t)I(t). \quad (2.4)$$

The study employs the FSIR model, which uses the Caputo fractional derivative of order  $\alpha \in (0, 1]$ :

$${}^c D_t^\alpha f(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-\tau)^{-\alpha} f'(\tau) d\tau, \quad (2.5)$$

where  $\Gamma(\cdot)$  is the gamma function, which extends the concept of the factorial to real arguments, and  $f'(\tau)$  is the ordinary derivative of the function  $f$  with respect to time [20, 21].

This allows us to naturally account for the memory and inertia effects of the epidemic process. The fractional derivative enables smooth interpolation between classical

Markov dynamics ( $\alpha = 1$ ) and processes with long memory ( $\alpha < 1$ ) [9–11, 22]. A discrete  $L1$  approximation was used to approximate the fractional derivative. Let  $x_n = x(t_n)$  be the values of the scalar sequence at the nodes  $t_n = n\Delta t$ . According to the  $L1$  scheme, the Caputo fractional derivative of order  $\alpha$  is approximated as:

$${}^c D_t^\alpha x(t_n) \approx \sum_{k=0}^{n-1} \alpha_{n-k} (x(t_{k+1}) - x(t_k)), \quad (2.6)$$

where the weights  $\alpha_m$  are specified analytically and take the form:

$$\alpha_m = (m+1)^{1-\alpha} - m^{1-\alpha}, \quad m \geq 0. \quad (2.7)$$

For practical implementation with a fixed window of length  $L$ , we compute the weights  $\alpha^0, \dots, \alpha_{L-1}$  and for each step  $n = 1, \dots, L-1$  we use

$$D_n = \sum_{m=0}^{n-1} \alpha_m \Delta x_{n-1-m}, \quad \Delta x = x_{j+1} - x_j. \quad (2.8)$$

A naive implementation of computing all  $D_n$  has asymptotic complexity of  $O(L^2)$  for a single window. Optimized tensor approaches reduce the constants, but the order remains quadratic. For  $n = 0$ , we assume  $D^0 = 0$ .

## 2.3. Hyperparameters and Covariates

For local learning, a sliding window of length  $L$  is used. Let  $t_k$  be the start of the  $k$ -th window. Then the window covers  $t \in \{t_k, t_k + 1, t_k + L - 1\}$ . In this work, we considered two window lengths:  $L \in \{14, 21\}$ . We denote the shift step between windows by  $s$ . The default value is  $s = 7$ , and for sensitivity analysis, we additionally considered  $s = 3$  and  $s = 10$ .

All model parameters  $\beta(t)$  and  $\gamma(t)$  are considered constant within the window  $[t_k, t_k + L - 1]$ .

Let  $X(t) \in R^m$  be the vector of covariates at time  $t$ . The following variables (covariates) were used in the experiments: vaccination and its lags, the restriction stringency index, testing rates, the proportion of positive tests, and the number of new tests.

## 2.4. Models

Baseline is a classic machine learning model that approximates the relationship:

$$Y_{\text{model}}(t) = f_\theta(X(t)), \quad (2.9)$$

where  $f_\theta$  is a multilayer perceptron.



The model imposes no physical constraints and serves as a baseline for evaluating the advantages of PINN. The baseline model is implemented as a lightweight MLP head (BetaGammaHead class) with three hidden layers of 64 neurons each and tangent-hyperbolic activations; the output consists of two linear components, which are converted via softplus into non-negative sequences:  $\beta(t)$  and  $\gamma(t)$ . The head's input size corresponds to the number of predictors in the window (normalized time and covariates). For each sliding window, the head is initialized with a deterministic random seed (the seed depends on the window number), after which it is optimized separately for the baseline version, minimizing only the MSE between the forecast  $Y_{\text{hat}}$  (computed via FSIR-forward using the obtained  $\beta(t)$ ,  $\gamma(t)$ ) and the observations. Training is performed using the Adam optimizer ( $lr \approx 2 \cdot 10^{-3}$ ) for 400 epochs, with the best snapshot selected based on the minimum loss within the window. This approach ensures that the baseline and PINN start with the same initial weights for a fair comparison.

The physics-informed neural network combines trajectory approximation  $S$ ,  $I$ ,  $R$  with the execution of discretized FSIR equations [6].

Let  $\tau \in [0,1]$  be the normalized time within the window. PINN approximates:

$$(S(\tau), I(\tau), R(\tau)) = f_{\theta}(\tau), \quad (2.10)$$

where  $f_{\theta}$  is an MLP with three hidden layers.

In physically informed neural networks that approximate the trajectories  $S(t)$ ,  $I(t)$ ,  $R(t)$ , it is important to ensure that the basic structural properties of the SIR model are satisfied: non-negativity of components and approximate constancy of the total population. Standard neural networks do not guarantee these constraints, as their outputs can take on arbitrary real values. The use of hard activations (e.g., ReLU) leads to gradient issues and optimization instability, which is critical in PINNs, where physical equations are directly incorporated into the loss function [6, 7].

For this reason, the model employs softplus normalization, which ensures a smooth and guaranteed non-negative transformation of the network outputs and approximate constancy of the  $(S + I + R)$ :

$$\text{softplus}(x) = \ln(1 + e^x) > 0. \quad (2.11)$$

Based on these three values  $u$ ,  $v$ ,  $w$ , a normalized distribution is formed:

$$\begin{aligned} S &= \tilde{N} \frac{u}{u + v + w}, \\ I &= \tilde{N} \frac{v}{u + v + w}, \\ R &= \tilde{N} \frac{w}{u + v + w}, \end{aligned} \quad (2.12)$$

which guarantees  $S, I, R \geq 0$  and  $(S + I + R \approx \tilde{N})$ .

Softplus is infinitely differentiable, has no "dead" zones, and ensures stable gradients. PINN minimizes the deviation between the network output and the right-hand sides of the SIR equations:

$$\delta^2 = \left| (S, I, R) - (S^{\text{FSIR}}, I^{\text{FSIR}}, R^{\text{FSIR}}) \right|^2. \quad (2.13)$$

Loss function:

$$L = L_{\text{data}} + \lambda_{\text{phys}} L_{\text{phys}} + \lambda_{\text{reg}} L_{\text{reg}}, \quad (2.14)$$

where the first term accounts for data fit, the second for satisfying the physical equations, and the third for regularization.

## 2.5. Training, optimization, metrics, and sensitivity analysis

Training is performed separately on each window. The parameters of the next window are initialized with the values from the previous one (warm start), which ensures smooth trajectories.

The Adam optimizer was chosen, which has already become the standard in all PINN studies. Adam automatically scales gradients for each parameter, which is critically important when the loss function contains both a physical term and a data term, as well as when the neural network weights, the parameters  $\beta(t)$  and  $\gamma(t)$ , and the physical and data losses differ in magnitude by several orders of magnitude [23]. Adam's adaptability ensures robustness against noisy and poorly scaled gradients, which are characteristic of models with physical constraints. Furthermore, Adam performs well in "warm start" mode, where the parameters of each subsequent epoch are initialized with the values from the previous one, which promotes smooth trajectories and accelerates convergence.

Data for each country are first truncated to the selected wave and sorted by date; the primary unit of analysis is daily incidence per 100,000 population. For local model training, a sliding window of length  $L$  ( $L \in \{14, 21\}$  in experiments) is used with a stride step

(base case stride = 7, additional ablations 3 and 10). For each window, the baseline and PINN are trained separately with the same initialization (a deterministic seed dependent on the window number), after which forecasts from all windows covering a given time point are aggregated by averaging (taking into account the number of overlaps). After aggregation across the entire time series, a global split into training and test sets is formed: the first 60% of observations are used to evaluate training metrics, the remaining 40% for the final evaluation (train/test = 0.6/0.4). For each window, the best model snapshot is saved based on the minimum MSE loss in that window. It is these best local forecasts that are aggregated to construct the final  $Y_{\text{model}}$  and subsequently calculate  $RMSE / MAE / R^2$ .

The hyperparameters were determined by combining preliminary tests and targeted ablation studies (sensitivity analysis) rather than a single large-scale automated search. The following values were selected as working baseline values: learning rate for both models  $lr \approx 2 \cdot 10^{-3}$ , number of epochs per window for the baseline – 400, for PINN – 500; regularization term weights  $\lambda_{\text{phys}} = 0.1$ ,  $\lambda_{\text{L\_reg}} = 0.01$ ,  $\lambda_{\text{smooth\_b}} = 10^{-3}$ ; order of the fractional derivative  $\alpha \in \{0.7, 0.9, 1.0\}$ . A series of controlled experiments was conducted in which key parameters were varied sequentially (window length, stride,  $\alpha$ , covariate sets,  $\lambda$  values, and seed), and performance was evaluated using aggregated test metrics (specifically, the average RMSE across all windows and seeds). For PINN, a "warmup" is applied – the physical term  $\lambda_{\text{phys}} L_{\text{phys}}$  begins to take full effect after 40% of the epochs, which reduces the risk of the physical penalty dominating the data fitting too early. The criterion for selecting the configuration is the minimization of the test  $RMSE$  (supplemented, if necessary, by a check of result stability across different seeds and countries). The final parameter values were chosen as a compromise between forecast quality and stability/reproducibility across all three countries.

For each model, the following are calculated:  $RMSE$ ,  $MAE$ , and  $R^2$  on the training and test sets. Stability with respect to changes in random initialization and regularization parameters is evaluated separately.

The sensitivity analysis covers three groups of factors:

- 1) structural parameters (window length, shift frequency),
- 2) informational parameters (set of covariates),

- 3) stochastic parameters (initialization, regularization).

## 2.6. Statistical Analysis

Statistical hypothesis testing and statistical analysis of the obtained results include data preparation, test selection, correction for multiple comparisons, estimation of effect size, construction of confidence intervals, and additional models for generalizing conclusions.

For each group of differences (RMSE PINN minus Baseline), we test the normality of the distribution of differences using the Shapiro–Wilk test. If the Shapiro–Wilk test does not reject normality, we apply the paired  $t$ -test. If normality is rejected or there is suspicion of outliers, we apply the nonparametric Wilcoxon signed-rank test. To compare the effect of the sliding window length ( $L14$ ) versus ( $L21$ ), we use paired tests on corresponding pairs when such pairs are available; if the pairs are aggregated, we apply a *one-sample* test for differences or a nonparametric equivalent.

Next, a correction for multiple comparisons is applied [24]. To control the proportion of false-positive results, the Benjamini-Hochberg FDR is applied. To assess the effect size in paired comparisons, the paired Cohen's  $d$  is provided along with its interpretation. Statistical significance is accompanied by an effect size estimate and confidence intervals.

Due to possible heteroscedasticity of residuals in regression models, standard errors may be underestimated. To mitigate this effect, the HC3 method is used, which provides a more conservative correction, especially with small samples or in the presence of outliers [25, 26]. In all regression estimates of coefficients and when constructing mixed models, standard errors and confidence intervals are reported with HC3 correction. The tables indicate that standard errors are HC3-corrected.

For a generalized conclusion, this study constructed a mixed linear model with fixed effects for model type and window length and a random effect for country [27]. Such a model allows for accounting for the correlation of observations within a country and for estimating the overall impact of the transition from Baseline to PINN. The paper presents estimates of fixed effects, standard errors of HC3, and an interpretation of the coefficients in terms of changes in RMSE.

To check for multicollinearity among predictors, we calculated the variance inflation factor (VIF) using

the standard formula [28–30]. All values were  $VIF < 3.6$ , indicating the absence of serious multicollinearity.

All calculations were performed using Python version 3.13.5 (packages: pandas, scipy, numpy, statsmodels, seaborn, matplotlib, sklearn) [31, 32].

The main limitations of the statistical analysis are small sample sizes (interpretation should be cautious in groups with small  $n$ ); the absence of complete pairs by seed (some configurations do not have complete pairs by seed, so some comparisons were performed in an aggregated manner); sensitivity to test selection (some results depend on the choice of a paired or nonparametric test, and this is accounted for in the report and in the correction for multiple comparisons).

### 3. Results

This section presents the results of a systematic comparison of the physically-informed neural network (PINN) model with a baseline machine learning model, as well as an analysis of the PINN's sensitivity to key structural, informational, and stochastic factors. The focus is on a practical assessment of the model's stability, generalizability, and factor-dependent quality on real-world epidemic data.

The analysis covers 118 experimental configurations that differ in sliding window length, set of covariates, regularization parameters, order of the fractional derivative ( $\alpha$ ), window shift step, and random initialization.

The section is structured to sequentially answer three key questions:

**To what extent does the PINN model consistently outperform the baseline model?** This question is addressed in subsection 3.1, which presents aggregated results for all countries and configurations.

**Which structural parameters (window length, stride, fractional order) have the greatest impact on model quality?** This is the subject of Section 3.2, which analyzes the role of the sliding window and hyperparameters.

**Do the covariates (vaccination, testing, severity index, etc.) have a statistically significant additional effect?** This question is addressed in Section 3.3, which presents the results of nonparametric tests, robust regression, decomposition ( $R^2$ ), PCA, and the regularized Lasso model.

Particular attention is paid to the fact that, despite theoretical expectations, none of the covariates demonstrated a consistent statistically significant effect on PINN returns, and the effect of covariates varies by country. In some configurations, additional variables ( ) yield a slight improvement, while in others they worsen performance due to noise or multicollinearity. This result warrants separate discussion, as it contradicts the intuitive assumption that adding information on vaccination, testing, or the strictness index should improve model quality.

#### 3.1. General Comparison of PINN and the Baseline Model

The first stage of the analysis is a comparison of the physically informed neural network (PINN) with the baseline machine learning model (Baseline), which contains no physical constraints and serves as a reference point for evaluating the effectiveness of PINN. For each of the 118 experimental configurations, the root mean square error (*RMSE*), *mean absolute error* (*MAE*), and  $R^2$  were calculated on the test dataset, along with the performance gain of PINN over the baseline model. The difference between the models' errors was used as the comparison metric:

$$\Delta = RMSE_{\text{baseline}} - RMSE_{\text{PINN}}. \quad (3.1)$$

where the  $\Delta > 0$  indicates the advantage of PINN. This metric allows us to quantitatively assess the gain provided by the model's physical structure, i.e., to what extent the physical constraints and structural properties of PINN improve the model's ability to reproduce epidemic dynamics.

In all configurations, PINN outperforms the standard baseline model and does so consistently across all three countries. The average gain is ( $\bar{\Delta} = 1.3301$ ), with a 95% confidence interval of [1.1920; 1.4682]. The Wilcoxon nonparametric test for paired samples yields  $p \ll 0.001$ , which corresponds to a very large effect on the Cohen scale ( $d = 2.12$ ). This indicates a systematic advantage of PINN that is not an artifact of specific configurations or random initialization.

Figure 3.1 shows the distributions of *the* test *RMSE* for Baseline and PINN separately by country. Violin plots have been constructed for each country, showing the shape of the distribution, median values, and spread. It is evident that for all countries, PINN demonstrates lower median *RMSE* values

and a smaller spread, indicating higher accuracy and model stability. The greatest improvement is observed in Italy, where the epidemic wave has a complex shape with pronounced asymmetry.

In Ukraine, the improvement is moderate but stable, while in Germany it is the smallest, although PINN still outperforms Baseline in every configuration.

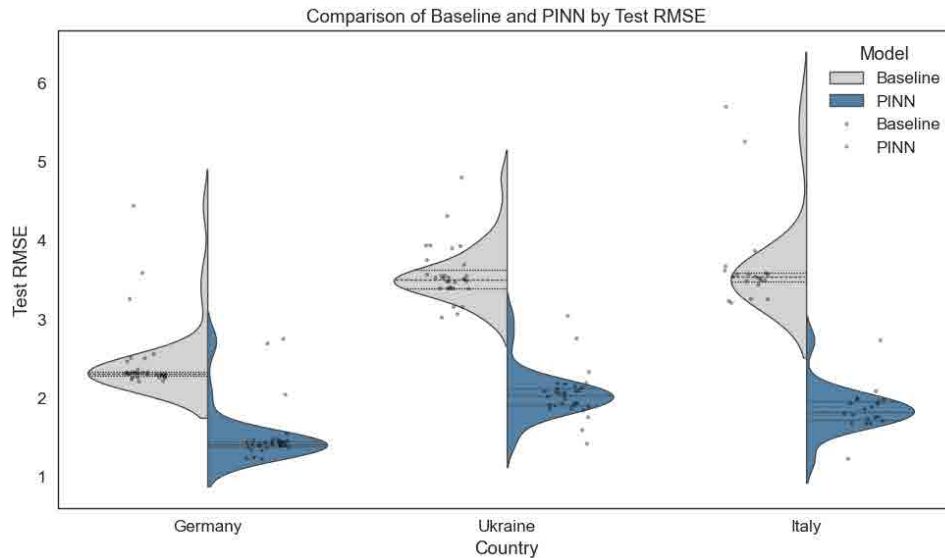


**Fig. 3.1.** Distribution of test RMSE for Baseline and PINN across three countries

To confirm these findings, Figure 3.2 presents a pairwise comparison of the test RMSE for PINN and Baseline in each configuration. Each point corresponds to a distinct combination of model parameters, and the diagonal line ( $y = x$ ) indicates

equality of errors. All points lie below this line, confirming PINN's consistent advantage over Baseline. The color coding shows that the largest gain is observed in Italy, and the smallest in Germany, which is consistent with the results in Figure 3.1.

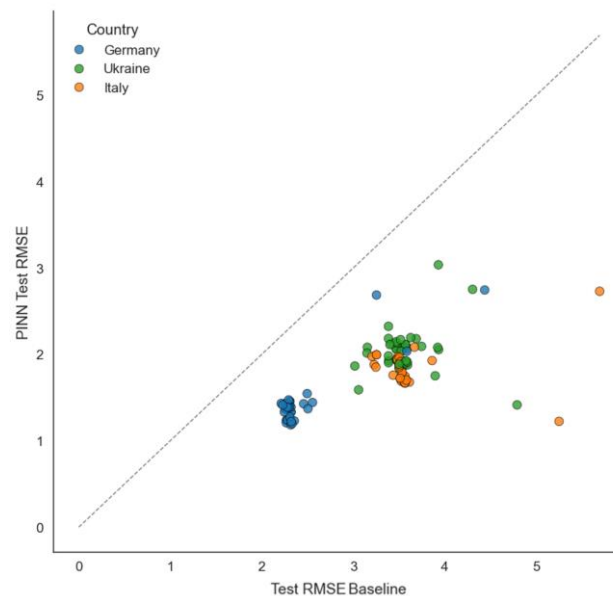


**Fig. 3.2.** Paired comparison of test RMSE for PINN and Baseline

A summary of the results is presented in Table 3.1, where for each country and value of the parameter  $k$

(number of covariates), the mean values of RMSE and the coefficient of determination  $R^2$  are given

for both models. It can be seen that regardless of the number of covariates, PINN demonstrates lower *RMSE* and higher  $R^2$ , confirming its ability to better align with the data.

Statistical analysis confirmed that PINN models yield significantly lower test *RMSE* values compared to the baseline (paired tests with FDR correction; all key comparisons have adjusted  $p < 0.01$ ). Experiments with models having one and eight covariates confirm that switching to PINN is associated with a reduction in *RMSE* of approximately 1.42 units from the baseline ( $p < 0.001$ ; Table 3.2).

**Table 3.1.** Comparison of the quality of the Baseline and PINN models based on test *RMSE* and  $R^2$  for different countries and numbers of covariates  $k$

| Country | k | Baseline RMSE | PINN RMSE | Baseline $R^2$ | PINN $R^2$ |
|---------|---|---------------|-----------|----------------|------------|
| Germany | 0 | 2.35          | 1.42      | 0.87           | 0.95       |
| Germany | 1 | 2.60          | 1.67      | 0.84           | 0.93       |
| Germany | 8 | 2.61          | 1.59      | 0.85           | 0.94       |
| Italy   | 0 | 3.32          | 1.89      | 0.91           | 0.97       |
| Italy   | 1 | 4.24          | 1.92      | 0.85           | 0.97       |
| Italy   | 8 | 3.66          | 1.90      | 0.89           | 0.97       |
| Ukraine | 0 | 3.42          | 2.02      | 0.91           | 0.97       |
| Ukraine | 1 | 3.96          | 2.05      | 0.87           | 0.96       |
| Ukraine | 8 | 3.63          | 2.21      | 0.89           | 0.96       |

**Table 3.2.** Statistical analysis of *RMSE* comparisons in PINN and baseline models

| country         | Germany  |          | Italy    |          | Ukraine  |          |          |
|-----------------|----------|----------|----------|----------|----------|----------|----------|
| k               | 1        | 8        | 1        | 1        | 1        | 1        | 8        |
| window_len      | 21       | 21       | 14       | 21       | 14       | 21       | 21       |
| n               | 16       | 9        | 9        | 10       | 9        | 10       | 5        |
| mean_pinn       | 1.472    | 1.628    | 1.652    | 2.011    | 1.852    | 2.167    | 2.287    |
| mean_base       | 2.371    | 2.684    | 3.732    | 3.671    | 3.680    | 3.624    | 3.659    |
| mean_diff       | -0.899   | -1.056   | -2.080   | -1.660   | -1.828   | -1.457   | -1.372   |
| std_diff        | 0.112    | 0.336    | 0.730    | 0.480    | 0.583    | 0.100    | 0.343    |
| test            | wilcoxon | wilcoxon | wilcoxon | wilcoxon | wilcoxon | paired_t | paired_t |
| stat            | 0        | 0        | 0        | 0        | 0        | -46.187  | -8.946   |
| pval            | 0.00043  | 0.003906 | 0.00391  | 0.001953 | 0.0039   | 5.23E-12 | 0.00086  |
| cohens_d_paired | -8.02671 | -3.145   | -2.848   | -3.460   | -3.133   | -14.606  | -4.001   |
| ci_diff_lo      | -0.94861 | -1.300   | -2.572   | -1.973   | -2.228   | -1.520   | -1.632   |
| ci_diff_hi      | -0.84362 | -0.869   | -1.821   | -1.451   | -1.605   | -1.402   | -1.091   |
| shapiro_p       | 0.00143  | 0.009    | 1.03E-06 | 0.00012  | 4.21E-06 | 0.802    | 0.922    |
| pval_adj        | 0.00115  | 0.0039   | 0.00391  | 0.00313  | 0.00391  | 4.19E-11 | 0.0017   |
| reject_fdr      | True     | True     | True     | True     | True     | True     | True     |

The results demonstrate that PINN not only reduces the test error but does so consistently across all countries and all model configurations. This provides a basis for further analysis of which specific factors – window length, hyperparameters, covariates, etc. – influence the magnitude of the gain, and whether the model can be further improved through the use of information variables.

### 3.2. The Effect of Sliding Window Length and the Stride Parameter

One of the key structural decisions in constructing the PINN model is the choice of the sliding window length ( $L$ ), on which local model training is performed. The window defines the time interval during which the model parameters (specifically  $\beta(t), \gamma(t)$ ) are considered constant or slowly changing. On the one hand,

a short window allows the model to quickly adapt to local changes in epidemic dynamics; on the other hand, it can lead to instability in estimates due to an insufficient number of data points. A long window, conversely, provides smoothing but may lose sensitivity to short-term changes.

This study examines two window lengths:  $L = 14$  and  $L = 21$  days. This choice represents a compromise between practical constraints and epidemiological soundness. The 14-day window corresponds to the typical incubation period for COVID-19 and is widely used in statistical reports and quarantine policy. The 21-day value allows us to assess whether a longer window provides additional stability or smoothing. Other values were not systematically considered, as increasing  $L$  beyond 21 days leads to a loss of local sensitivity, while decreasing it below 14 days leads to excessive data



fragmentation and optimization instability. Furthermore, the limit on the number of windows in each wave does not allow for a full-scale test of shorter intervals without losing statistical power. We investigated how the choice of window length  $L=14$  or  $L=21$  days affects model quality, as well as how the result changes when varying the window stride parameter  $s$  (stride), which determines the frequency of window updates along the time series.

For  $L=14$ , the average test  $RMSE$  drops to 1.72. And for  $L=21$ , it rises to 1.94, and the model no longer captures rapid wave changes as well. The corresponding

$R$ -squared values are  $R^2 = 0.965$  and  $R^2 = 0.952$ . The difference is statistically significant (Wilcoxon test,  $p \ll 0.01$ ) and practically important, especially for waves with rapid phase transitions. This is consistent with the hypothesis that a shorter window better captures the local features of the curve, specifically the peak asymmetry, the rate of rise, and the decay.

Fig. 3.3 shows a comparison of the  $RMSE$  distributions for the PINN model using  $L=14$  and  $L=21$  days. It can be seen that the short window not only yields lower median values but also reduces the spread, indicating higher model stability.

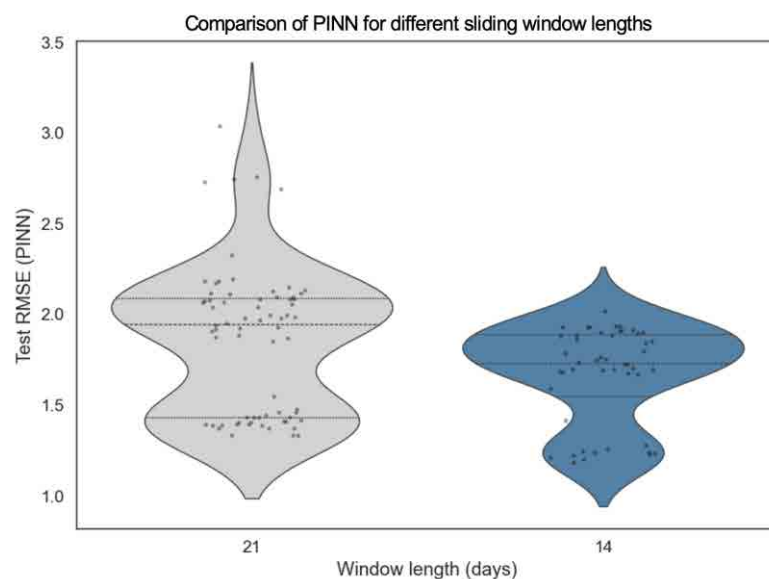


Fig. 3.3. Distribution of test  $RMSE$  for PINN with different window lengths

The effect of the stride parameter ( $s$ ), which determines how often the window moves along the time series, was investigated separately. The default value of  $s=7$  corresponds to a weekly update, and  $s=3$  and  $s=10$  were additionally considered. Changing from 7 to 3 or 10 alters the  $RMSE$  by only  $\pm 0.05$ , and  $R^2$  remains stable. This indicates that the model is robust to the choice of window update frequency, and the parameter  $s$  can be adapted to computational constraints without losing accuracy.

In summary, it can be stated that the choice of sliding window length is a critical factor for model quality. A short window  $L=14$  ensures better adaptation to local dynamics, reduces overfitting, and improves the model's alignment with the data. The *stride* parameter has a secondary effect and can vary depending on computational resources. These results confirm that the structural organization of training

intervals is a key element in constructing effective PINN models for epidemic time series.

### 3.3. The Impact of Covariates on Model Quality

The objective of this section is to quantitatively assess the role of informational covariates in explaining the difference in model quality, defined as the  $\Delta$  (3.1). Positive values of  $\Delta$  indicate an advantage for PINN. For a systematic analysis, we applied the following sequence of methods:

- 1) a basic linear model with control variables;
- 2) an extended model with indicators of covariate presence and an assessment of their combined contribution via partial- $R^2$ ;
- 3) dimensionality reduction of the covariate block using principal component analysis (PCA);
- 4) a regularized Lasso model as a stability check.

This structure allows to separate the influence of structural parameters and cross-country differences from the contribution of informational covariates.

The main analysis was performed for three values of  $k$ : a model without covariates ( $k = 0$ ), a model with one covariate ( $k = 1$ ), and a model with the full set of eight variables ( $k = 8$ ). It is important to emphasize that the case ( $k = 1$ ) is not a single model: eight separate models were constructed, each containing only one of the eight available covariates. This allows us to assess the individual contribution of each variable without the influence of multicollinearity and to identify the potentially most informative one.

In addition, for each country, an additional series of "cumulative" models was built, in which covariates were added gradually: first two, then three, and so on up to eight. These models used a standard set of other parameters (window length, stride,  $\alpha$ ), allowing us to assess whether model quality improves as the information space increases, or whether adding new variables leads to overfitting and a decline in fit.

Fig. 3.4 shows the distribution of test  $RMSE$  values for PINN models with different numbers of covariates  $k$ . Fig. 3.5 shows the distribution of test  $RMSE$  values for PINN models for each of the eight covariates individually.

Changes in the median and density of the distributions help to visually assess the impact of additional covariates on forecast quality.

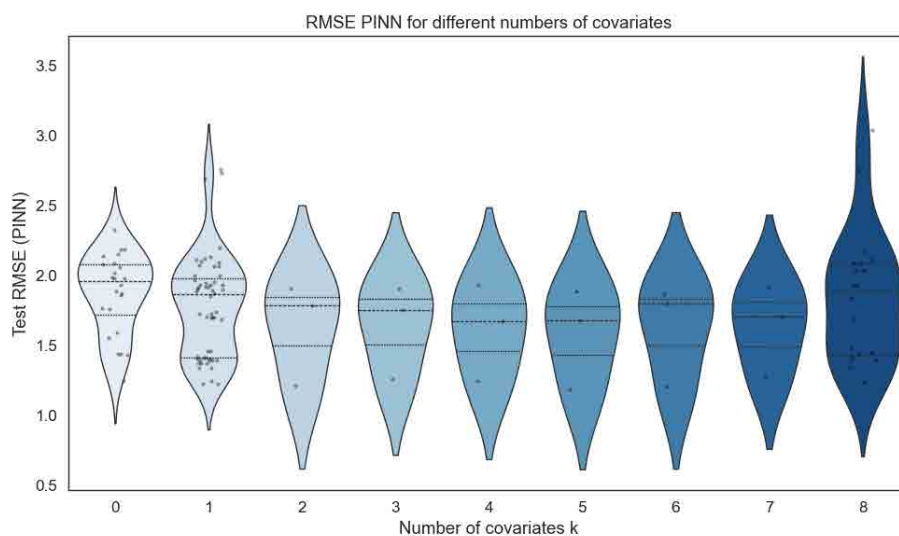


Fig. 3.4. Distribution of PINN test  $RMSE$  by number of covariates

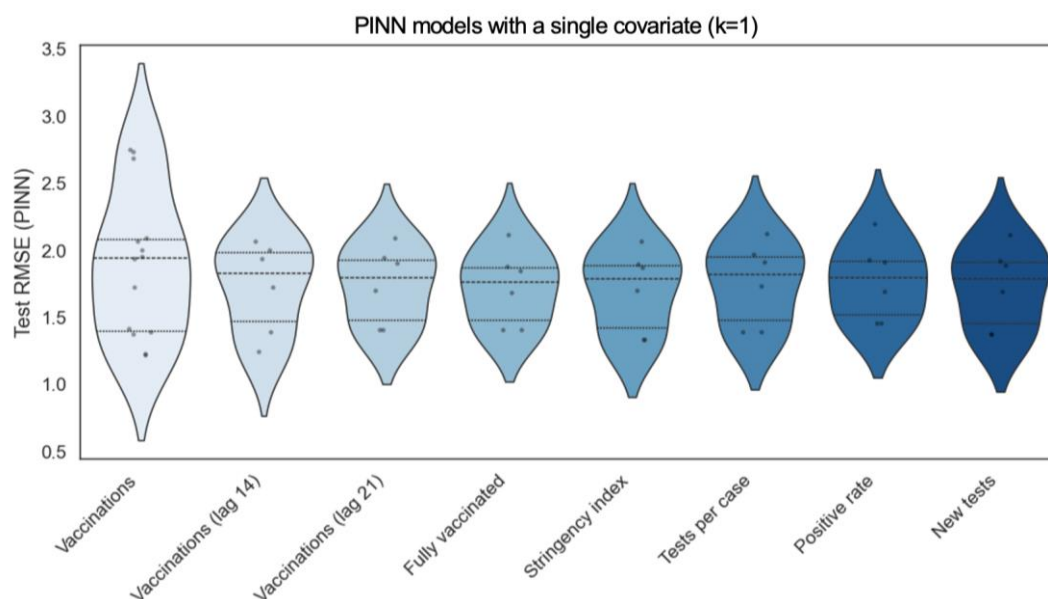


Fig. 3.5. Test  $RMSE$  of PINN for models with a single covariate

Covariates have a weak and often varying effect in each country. In Germany, adding variables often leads to a slight deterioration in *RMSE*, which may be due to the high quality of the baseline data and the relatively smooth waveform, which PINN reproduces well even without additional factors. In Italy and Ukraine, the picture is mixed: some individual covariates

(primarily vaccination indicators) sometimes yield a slight improvement, but in other configurations the effect disappears or changes sign.

### Baseline linear model

The initial baseline estimation was performed using a linear model:

$$\Delta_i = \beta_0 + \gamma_{\text{Italy}} \cdot 1\{\text{Italy}_i\} + \gamma_{\text{Ukraine}} \cdot 1\{\text{Ukraine}_i\} + \beta_L \cdot \text{window\_len}_i + \beta_\alpha \cdot \alpha_i + \beta_s \cdot \text{stride}_i + \varepsilon_i \quad (3.2)$$

where  $\varepsilon_i$  is the random error. Parameter estimation was performed using the least squares method:  $\hat{\beta} = (X^T X)^{-1} X^T y$ .

Due to possible heteroscedasticity of the residuals (Table 3.3), we used robust standard errors of the HC3 type [25, 33]. All standard errors and confidence

intervals presented in the tables are adjusted specifically for HC3.

The results of the baseline model are presented in Table 3.4. Country effects and the *window\_len* parameter remain significant, with the latter having a consistently negative effect on the  $\Delta$ , i.e., longer windows reduce the PINN gain.

**Table 3.3.** Statistics on the normality of residuals and identification of potential outliers

| Metric                                                | Value    |
|-------------------------------------------------------|----------|
| Omnibus                                               | 122.684  |
| Prob(Omnibus)                                         | 0.000    |
| Jarque-Bera                                           | 1840.597 |
| Skew                                                  | 3.641    |
| Kurtosis                                              | 20.926   |
| Max leverage                                          | 0.1517   |
| Number of observations with $ \text{std\_resid}  > 3$ | several  |

**Table 3.4.** Results of the baseline OLS model with HC3 correction

| Parameter           | Estimate | SE (HC3) | <i>t</i> | <i>p</i> -value |
|---------------------|----------|----------|----------|-----------------|
| Intercept           | 0.8820   | 0.736    | 1.198    | 0.231           |
| C(country)[Italy]   | 0.7272   | 0.088    | 8.240    | 0.000           |
| C(country)[Ukraine] | 0.5038   | 0.056    | 9.072    | 0.000           |
| <i>window_len</i>   | -0.0367  | 0.009    | -3.952   | 0.000           |
| $\alpha$            | 0.1666   | 0.793    | 0.210    | 0.834           |
| <i>stride</i>       | 0.0936   | 0.031    | 2.996    | 0.003           |

Based on our data, the covariate block yields  $R^2_{\text{partial}} \approx 0.154$ , and the 95% bootstrap confidence interval for the partial  $R^2$  is [0.054, 0.342] [34].

This means that the covariates make a moderate but not dominant contribution to explaining  $\Delta$ . The coefficients of the regression model are presented in Table 3.5.

**Table 3.5.** Regression coefficients of the full model with indicators (HC3)

| Parameter             | Estimate | SE (HC3) | <i>t</i> | <i>p</i> -value |
|-----------------------|----------|----------|----------|-----------------|
| Intercept             | 0.9518   | 0.738    | 1.290    | 0.197           |
| C(country)[Italy]     | 0.7135   | 0.080    | 8.903    | 0.000           |
| C(country)[Ukraine]   | 0.5134   | 0.059    | 8.716    | 0.000           |
| <i>window_len</i>     | -0.0452  | 0.012    | -3.712   | 0.000           |
| $\alpha$              | 0.1753   | 0.738    | 0.237    | 0.812           |
| <i>stride</i>         | 0.0902   | 0.048    | 1.891    | 0.059           |
| Vaccinations          | 0.3933   | 0.210    | 1.869    | 0.062           |
| Vaccinations (lag 14) | -0.3730  | 0.175    | -2.127   | 0.033           |
| Fully vaccinated      | 0.1209   | 0.050    | 2.395    | 0.017           |
| Stringency index      | 0.1061   | 0.039    | 2,753    | 0.006           |

### PCA of the covariate block

To reduce multicollinearity in the covariate block, we applied the principal component analysis (PCA) method and used the first two components (PC1 and PC2), which explain  $\approx 76\%$  of the variance [35]. This allowed us to stabilize the estimates without losing information.

The first component, PC1, explains  $\approx 62\%$  of the block's variance and has positive loadings (approximately 0.28–0.39) on all indicators, allowing it to be interpreted as an aggregated index of covariate presence. The second principal component, PC2 ( $\approx 14\%$ ), reflects the contrast between types of covariates (e.g., vaccination and testing indicators). Using PC1/PC2

as predictors in the regression reduces VIF and stabilizes the estimates of structural parameters.

### Lasso Regularized Model

To test the robustness of the results, we applied the **Lasso** regularized model. The objective function is written as:

$$\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \left\{ \sum_i (y_i - X_i \beta)^2 + \lambda_{\text{lasso}} |\beta|_1 \right\}. \quad (3.3)$$

Lasso selected a very small  $\lambda_{\text{lasso}}$  and set almost all indicators to zero, leaving only `window_len` as a stable predictor (Table 3.6). This confirms that structural parameters have a greater and more stable influence on  $\Delta$  than individual informational covariates.

**Table 3.6.** Results of regularized estimates: optimal parameters and stable coefficients

| Method  | Best value of the regularization hyperparameter | Key coefficients                                                                    |
|---------|-------------------------------------------------|-------------------------------------------------------------------------------------|
| LassoCV | $\lambda_{\text{lasso}} = 0.0340$               | $\text{window\_len} = -0.1608$ ;<br>$\alpha = 0.0000$ ;<br>$\text{stride} = 0.0244$ |

### Diagnosis of residuals and multicollinearity

Diagnosis of residuals and influential observations was performed using standard statistics: Omnibus, Jarque-Bera (JB), skewness and kurtosis coefficients, as well as leverage and standardized residuals (`std_resid`).

In the experiments, Omnibus and JB had very small p-values, skewness  $\approx 2.5$ – $3.7$ , and kurtosis  $\approx 15$ – $21$ , indicating strong skewness and heavy tails in the residual distribution. There are also several observations with high residual standard deviations ( ) – potential outliers with ( $|\text{std\_resid}| > 3$ ). Therefore, we applied HC3 and conducted a sensitivity analysis: we repeated the

OLS (HC3) without the observations – potential outliers with ( $|\text{std\_resid}| > 3$ ). As a result, the changes in the main coefficients and partial- $R^2$  s were insignificant; thus, the conclusions are considered robust to the influence of individual outliers.

Fig. 3.6 shows the dependence of leverage on standardized residuals for the full model with indicators (HC3 estimates) and allows for the visual identification of potentially influential observations (points with  $|\text{std\_resid}| > 3$ ) and increased leverage.

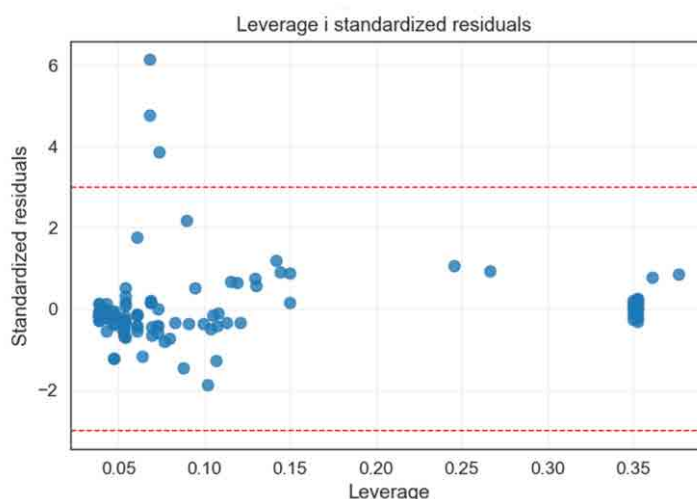


**Fig. 3.6.** Leverage and standardized residuals – diagnosis of influential observations

To diagnose multicollinearity, we used the variance inflation factor (VIF), which is defined as

$$VIF_j = \frac{1}{1 - R_j^2}, \quad (3.4)$$

where  $R_j^2$  is the coefficient of determination for the regression of  $X_j$  on other predictors.

VIF (Table 3.7) does not exceed  $\approx 3.6$ , indicating the absence of serious multicollinearity.

**Table 3.7.** VIF analysis of multicollinearity after adjustments

| Variable                  | VIF    |
|---------------------------|--------|
| Vaccinations (week 21)    | 3.5703 |
| Vaccinations (14-day lag) | 3.5638 |
| Fully vaccinated          | 2.9364 |
| Stringency index          | 2.7500 |
| Vaccinations              | 2.5938 |
| Tests per case            | 2.5814 |
| Positive rate             | 2.3759 |
| New tests                 | 2.0448 |
| window_len                | 1.2885 |
| $\alpha$                  | 1.0325 |
| stride                    | 1.0010 |

#### 4. Conclusion

Thus, statistical tests provide a consistent picture: **PINN almost always outperforms Baseline**; structural parameters, primarily sliding window length and inter-region differences, are the main factors behind the difference in model quality and PINN's advantage. The block of information covariates has a moderate partial contribution ( $R^2 \approx 0.154$ ), but individual covariates provide a weak and unstable signal. PCA and Lasso consistently confirm these conclusions.

These results are consistent with the assumption that the local form of the epidemic wave (the FSIR-PINN window structure) already contains integrated information about external influences through the shape of the epidemic wave, specifically vaccination, changes in mobility, and behavioral factors. PINN reproduces

these effects by adapting the parameters  $\beta(t)$  and  $\gamma(t)$  within each sliding window. Noise and multicollinearity negate the potential informational contribution of covariates; therefore, their explicit inclusion does not provide a significant improvement in model quality.

Further research should focus on extending the PINN sensitivity analysis to other infectious diseases and epidemic waves with different dynamics. A promising direction is the development of adaptive methods for automatically selecting the sliding window length based on local dynamics characteristics. It is also worth investigating alternative approaches to integrating covariates, particularly through attention mechanisms, which can mitigate the negative impact of multicollinearity. Furthermore, it is advisable to apply PINN in multi-population models that account for spatial interactions between regions, as well as to compare it with other classes of physically informed architectures, such as DeepONet and the Fourier Neural Operator.

#### Conflict of Interest

The authors confirm that the study was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

#### Funding

The study was conducted without financial support.

#### Data Availability

The manuscript contains data in the form of electronic supplementary material.

#### Use of Artificial Intelligence

The author confirms that no AI technologies were used in the preparation of this work.

#### References

1. Kermack, W. O., McKendrick, A. G. (1927), "A contribution to the mathematical theory of epidemics", *Proceedings of the Royal Society*, Vol. 115, pp. 700–721. DOI: <https://doi.org/10.1098/rspa.1927.0118>
2. Kalachev, L., et al. (2023), "Revisiting classical SIR modelling in light of the COVID-19 pandemic", *Infectious Disease Modelling*, Vol. 8, No. 3, pp. 72–83. DOI: <https://doi.org/10.1016/j.idm.2022.12.002>
3. Nesteruk, I. (2020), "Simulations and Predictions of COVID-19 Pandemic With the SIR Model", *Innovative Biosystems and Bioengineering*, Vol. 4(1), pp. 20–29. DOI: <https://doi.org/10.20535/ibb.2020.4.2.204274>



4. Mohajan, H. (2022), "Mathematical Analysis of SIR Model for COVID-19 Pandemic", *MPRA Paper*, No. 114390. DOI: <https://mpra.ub.uni-muenchen.de/114390/>
5. Voronin, A., Akhiezer, O., Galuza, A., Lebedeva, I., Zaitsev, Yu., Lebedev, S. (2023), "Modeling Competitive Interaction «Predator-Prey» on the Example of Two Innovative Processes", *13th International Conference on Advanced Computer Information Technologies (ACIT), Wroclaw, Poland*. pp. 131–134. DOI: <https://doi.org/10.1109/ACIT58437.2023.10275538>
6. Raissi, M., Perdikaris, P., Karniadakis, G. E. (2019), "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations", *Journal of Computational Physics*, Vol. 378, pp. 686–707. DOI: <https://doi.org/10.1016/j.jcp.2018.10.045>
7. Raissi, M., Perdikaris, P., Karniadakis, G.E. (2017), "Physics Informed Deep Learning (Part II): Data-driven Discovery of Nonlinear Partial Differential Equations", *Artificial Intelligence*, DOI: <https://doi.org/10.48550/arXiv.1711.10566>
8. Lyubchik, L., Grinberg, G., Lyubchik, M., Galuza, A., Akhiezer, O. (2020), "Interval Evaluation of Stationary State Probabilities for Markov Set-Chain Models", *10th IEEE International Conference on Advanced Computer Information Technologies (ACIT), Deggendorf, Germany*, pp. 82–85. DOI: <https://doi.org/10.1109/ACIT49673.2020.9208932>
9. Kozioł, K., Latosińska, J. N., Kozioł, M. (2020), "Fractional-Order SIR Epidemic Model for Transmission of COVID-19", *Applied Sciences*, Vol. 10(23), pp. 8316. DOI: <https://doi.org/10.3390/app10238316>
10. Banerjee, R., Bhattacharyya, R., Roy, T. K. (2022), "Fractional optimal control of compartmental SIR model for COVID-19 with Caputo's fractional derivative", *IFAC-PapersOnLine*, Vol. 55(2), pp. 349–354. DOI: <https://doi.org/10.1016/j.ifacol.2022.04.101>
11. Alshomrani, A. S., Khan, A., Ahmad, I., Baleanu, D. (2021), "Caputo SIR model for COVID-19 under optimized fractional order", *Chaos, Solitons & Fractals*. Vol. 143, pp. 110619. DOI: <https://doi.org/10.1186/s13662-021-03345-5>
12. Cuomo, S., Di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., Piccialli, F. (2022), "Scientific Machine Learning Through Physics-Informed Neural Networks: Where we are and what's next", *Journal of Scientific Computing*, Vol. 92(3), Art. 88. DOI: <https://doi.org/10.1007/s10915-022-01939-z>
13. Wang, S., Yu, X., Perdikaris, P. (2022), "When and why PINNs fail to train: A neural tangent kernel perspective", *Journal of Computational Physics*, Vol. 449, Art. 110768. DOI: <https://doi.org/10.1016/j.jcp.2021.110768>
14. Kharazmi, E., Cai, M., Zheng, X., Zhang, Z., Lin, G., Karniadakis, G. E. (2022), "Identifiability and predictability of integer- and fractional-order epidemiological models using physics-informed neural networks", *Nature Computational Science*, Vol. 2, pp. 744–753. DOI: <https://doi.org/10.1101/2021.04.05.21254919>
15. McClenny, L. D., Braga-Neto, U. M. (2023), "Self-Adaptive Physics-Informed Neural Networks", *Journal of Computational Physics*, Vol. 474, Art. 111722. DOI: <https://doi.org/10.1016/j.jcp.2022.111722>
16. Cai, S., Mao, Z., Wang, Z., Yin, M., Karniadakis, G. E. (2021), "Physics-informed neural networks (PINNs) for fluid mechanics: A review", *Acta Mechanica Sinica*, Vol. 37, pp. 1727–1738. DOI: <https://doi.org/10.1007/s10409-021-01148-1>
17. Cramer, E. Y., et al. (2022), "Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States", *Proceedings of the National Academy of Sciences*, Vol. 119(15), Art. e2113561119. DOI: <https://doi.org/10.1073/pnas.2113561119>
18. Mishra, S., Molinaro, R. (2022), "Estimates on the generalization error of physics-informed neural networks for approximating a class of inverse problems for PDEs", *IMA Journal of Numerical Analysis*, Vol. 42(2), pp. 981–1022. DOI: <https://doi.org/10.1093/imanum/drab032>
19. Our World in Data (OWID) – COVID-19 dataset (data source). IEEE DataPort entry (DOI for dataset snapshot). DOI: <https://doi.org/10.21227/2n61-4965>
20. Caputo, M. (1967), "Linear models of dissipation whose Q is almost frequency independent", *Geophysical Journal International*. DOI: <https://doi.org/10.1111/j.1365-246X.1967.tb02303.x>
21. Diethelm, K. (2010), "The Analysis of Fractional Differential Equations", *Springer*. DOI: <https://doi.org/10.1007/978-3-642-14574-2>
22. Danane, J., Hammouch, Z., Allali, K., Rashid, S., Singh, J. (2021), "A fractional-order model of coronavirus disease 2019 (COVID-19) with governmental action and individual reaction", *Math Methods Appl Sci*. DOI: <https://doi.org/10.1002/mma.7759>
23. Kingma, D. P., Ba, J. (2013), "A method for stochastic optimization", *ICLR (arXiv)*. DOI: <https://doi.org/10.48550/arXiv.1412.6980>
24. Benjamini, Y., Hochberg, Y. (1995), "Controlling the false discovery rate: A practical and powerful approach to multiple testing", *JRSS*. DOI: <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
25. MacKinnon, J. G., White, H. (1985), "Some heteroscedasticity – consistent covariance matrix estimators with improved finite sample properties", *Journal of Econometrics*. DOI: [https://doi.org/10.1016/0304-4076\(85\)90158-7](https://doi.org/10.1016/0304-4076(85)90158-7)
26. White, H. (1980), "A heteroscedasticity – consistent covariance matrix estimator and a direct test for heteroskedasticity", *Econometrica*. DOI: <https://doi.org/10.2307/1912934>
27. Bates, D., Mächler, M., Bolker, B., Walker, S. (2015), "Fitting linear mixed – effects models using lme", *Journal of Statistical Software*, Vol. 67, Issue 1. DOI: <https://doi.org/10.18637/jss.v067.i01>
28. O'Brien, R. M. (2007), "A Caution Regarding Rules of Thumb for Variance Inflation Factors", *Quality & Quantity*. DOI: <https://doi.org/10.1007/s11135-006-9018-6>
29. *Applied Regression Analysis and Generalized Linear Models* (3rd ed.). John Fox, Sage.

30. James, G. et al. (2023), "An Introduction to Statistical Learning (2nd ed.)", *Springer*.
31. Pedregosa, F. et al. (2011), "Scikit – learn: Machine learning in Python", *JMLR*. DOI: <https://doi.org/10.48550/arXiv.1201.0490>
32. Hastie, T., Tibshirani, R., Friedman, J. (2009), "The Elements of Statistical Learning", *Springer*. DOI: <https://doi.org/10.1007/978-0-387-84858-7>
33. Long, J. S., Ervin, L. H. (2000), "Using heteroscedasticity consistent standard errors in the linear regression model", *The American Statistician*, Vol. 54, No. 3. pp. 217–224.
34. Efron B. (1979), "Bootstrap methods: Another look at the jackknife", *Ann. Statist.*, Vol. 7, No. 1, pp. 1–26. DOI: <https://doi.org/10.1214/aos/1176344552>
35. Jolliffe, I. T., Cadima, J. (2016), "Principal component analysis: a review and recent developments", *Philosophical Transactions of the Royal Society*, No. 374(2065), pp. 20150202

Received (Надійшла) 27.02.2026

Accepted for publication (Прийнята до друку) 01.04.2026

Publication date (Дата публікації) 29.05.2026

#### Відомості про авторів / About the Authors

**Погорелов Станіслав Вікторович** – доктор фізико-математичних наук, професор, Національний технічний університет "Харківський політехнічний інститут", професор кафедри комп'ютерної математики і аналізу даних, Харків, Україна;  
**Stanislav Pohorielov** – Doctor of physical and mathematical sciences, Professor, National Technical University "Kharkiv polytechnic institute", Professor at the Department of Computer Mathematics and Data Analysis, Kharkiv, Ukraine;  
 e-mail: [Stanislav.Pohorielov@khpi.edu.ua](mailto:Stanislav.Pohorielov@khpi.edu.ua)  
 ORCID ID: <https://orcid.org/0000-0002-0189-8655>

**Балаба Ярослав Андрійович** – Національний технічний університет "Харківський політехнічний інститут", аспірант кафедри комп'ютерної математики і аналізу даних, Харків, Україна;  
**Yaroslav Balaba** – National Technical University "Kharkiv polytechnic institute", Postgraduate Student at the Department of Computer Mathematics and Data Analysis, Kharkiv, Ukraine;  
 e-mail: [Yaroslav.Balaba@cs.khpi.edu.ua](mailto:Yaroslav.Balaba@cs.khpi.edu.ua)  
 ORCID ID: <https://orcid.org/0009-0005-9710-2971>

**Татар'янц Максим Сергійович** – Національний технічний університет "Харківський політехнічний інститут", старший викладач кафедри комп'ютерної математики і аналізу даних, Харків, Україна;  
**Maksym Tatariants** – National Technical University "Kharkiv polytechnic institute", Senior Lecturer at the Department of Computer Mathematics and Data Analysis, Kharkiv, Ukraine;  
 e-mail: [Maksym.Tatariants@cs.khpi.edu.ua](mailto:Maksym.Tatariants@cs.khpi.edu.ua)  
 ORCID ID: <https://orcid.org/0000-0002-5952-9801>

**Беркунський Євген Юрійович** – Національний університет кораблебудування імені адмірала Макарова, старший викладач кафедри інформаційних управляючих систем та технологій, Миколаїв, Україна;  
**Yevhen Berkunskyi** – Admiral Makarov National University of Shipbuilding, Senior Lecturer at the Department of Information Control Systems and Technology, Mykolaiv, Ukraine;  
 e-mail: [yevgen.berkunskyi@nuos.edu.ua](mailto:yevgen.berkunskyi@nuos.edu.ua)  
 ORCID ID: <https://orcid.org/0000-0001-8876-1784>

**Павленко Альона Юріївна** – Національний університет кораблебудування імені адмірала Макарова, старший викладач кафедри інформаційних управляючих систем та технологій, Миколаїв, Україна;  
**Alona Pavlenko** – Admiral Makarov National University of Shipbuilding, Senior Lecturer at the Department of Information Control Systems and Technology, Mykolaiv, Ukraine;  
 e-mail: [alona.pavlenko@nuos.edu.ua](mailto:alona.pavlenko@nuos.edu.ua)  
 ORCID ID: <https://orcid.org/0000-0002-3776-8377>

## СИСТЕМАТИЧНИЙ АНАЛІЗ ЧУТЛИВОСТІ ФІЗИЧНО-ІНФОРМОВАНИХ НЕЙРОННИХ МЕРЕЖ У ЗАДАЧАХ МОДЕЛЮВАННЯ ЕПІДЕМІЧНИХ ЧАСОВИХ РЯДІВ

У статті подано систематичний аналіз чутливості фізично-інформованих нейронних мереж (PINN) у задачах моделювання епідемічних часових рядів. **Мета дослідження** – проведення систематичного аналізу чутливості PINN-моделей на реальних епідемічних часових рядах і визначення ключових факторів, що впливають на якість відтворення динаміки епідемічних хвиль. **Завдання:** оцінити вплив довжини ковзного вікна, набору коваріат, параметрів регуляризації та випадкової ініціалізації на якість PINN; порівняти PINN із класичною *baseline*-моделлю без фізичних обмежень. **Методи:** масштабна серія експериментів для трьох країн із варіюванням довжини ковзного вікна, кроку зсуву, порядку дробової похідної  $\alpha$ , стохастичної стабільності та набору коваріат; статистичний аналіз із застосуванням тесту Вілкоксона, OLS із HC3-корекцією, PCA та *Lasso*-регуляризації. **Результати:** PINN майже завжди дає кращі метрики, ніж *baseline*, за точністю відтворення епідемічних часових рядів і демонструє вищі значення  $R^2$  та нижчі RMSE/MAE у всіх країнах. Установлено, що 14-денне ковзне вікно забезпечує оптимальний баланс між адаптивністю та стійкістю моделі, тоді як збільшення довжини ковзного вікна до 21 дня призводить до згладжування динаміки та втрати локальної чутливості. Проведені експерименти продемонстрували, що ефект коваріат не має універсального позитивного результату. У деяких конфігураціях додаткові змінні в певних країнах покращують якість моделі, тоді як в інших додають шуму через мультиколінеарність і низьку якість даних. Дослідження стабільності підтвердило низьку чутливість PINN до випадкової ініціалізації та параметрів регуляризації. Досягнуті результати дають змогу чітко побачити поведінку PINN-моделей у задачах епідеміологічного моделювання та визначають практичні рекомендації щодо вибору параметрів, які забезпечують надійність і відтворюваність моделі. **Висновки:** PINN є більш ефективною альтернативою класичним моделям для епідемічного моделювання; 14-денне ковзне вікно є оптимальним; вплив коваріат не стабільний і залежить від країни.

**Ключові слова:** фізично-інформовані нейронні мережі (PINN); епідемічні часові ряди; аналіз чутливості; ковзне вікно; коваріати; стабільність моделі; машинне навчання; нейронні мережі; епідеміологічне моделювання.

### Бібліографічні описи / Bibliographic descriptions

Погорелов С. В., Балаба Я. А., Татар'янц М. С., Беркунський Є. Ю., Павленко А. Ю. Систематичний аналіз чутливості фізично-інформованих нейронних мереж у задачах моделювання епідемічних часових рядів. *Автоматизовані системи управління та прилади автоматики*. 2026. № 2 (189). С. 309–324. DOI: <https://doi.org/10.30837/0135-1710.2026.189.309>

Pohorielov, S., Balaba, Y., Tatariants, M., Berkunskyi, Y., Pavlenko, A. (2026), "A systematic sensitivity study of physics-informed neural networks for epidemic time series", *Management Information System and Devices*, No. 2 (189), P. 309–324. DOI: <https://doi.org/10.30837/0135-1710.2026.189.309>

## АЛФАВІТНИЙ ПОКАЖЧИК

|                    |     |
|--------------------|-----|
| Андрусевич А. О.   | 259 |
| Ахїєзер О. Б.      | 47  |
| Балаба Я. А.       | 309 |
| Беркунський Є. Ю.  | 309 |
| Білак Ю. Ю.        | 05  |
| Бобровнікова К. Ю. | 95  |
| Бомбандьоров Ю. П. | 259 |
| Віпшовський Ю. А.  | 128 |
| Вовк О. В.         | 138 |
| Глюза М. П.        | 138 |
| Григор'єв О. В.    | 138 |
| Григоренко С. М.   | 17  |
| Грицюк Ю. І.       | 128 |
| Гурман І. В.       | 95  |
| Далека В. Х.       | 259 |
| Даценко С. С.      | 153 |
| Демчук В. Г.       | 165 |
| Димерцов Д. О.     | 267 |
| Дяченко Д. О.      | 109 |
| Євтушенко О. С.    | 276 |
| Заковоротний О. Ю. | 276 |
| Каук В. І.         | 211 |
| Кириченко І. В.    | 165 |
| Кирий В. В.        | 196 |
| Кобилін І. О.      | 284 |
| Кобилін О. А.      | 284 |
| Колбасін В. О.     | 47  |
| Кучапін М. Ю.      | 182 |
| Кучук Г. А.        | 153 |
| Кушнір І. С.       | 196 |
| Луценко В. В.      | 165 |
| Лученко С. В.      | 182 |
| Маковецький С. О.  | 211 |
| Менделєва М. В.    | 138 |
| Метешкін К. О.     | 259 |
| Міхаль О. П.       | 109 |
| Нечипуренко С. І.  | 34  |
| Ніколайчук А. І.   | 284 |
| Олексюк Д. А.      | 95  |
| Островецький С. В. | 297 |
| Павленко А. Ю.     | 309 |
| Пахомов Ю. В.      | 47  |
| Погорєлов С. В.    | 309 |
| Просолов В. В.     | 224 |
| Протасова Л. І.    | 17  |
| Путятіна О. Є.     | 284 |
| Решетнікова С. М.  | 47  |

## ALPHABETICAL INDEX

|                        |     |
|------------------------|-----|
| Andrusevich Anatoly    | 259 |
| Akhiiezer Olena        | 47  |
| Balaba Yaroslav        | 309 |
| Berkunskyi Yevhen      | 309 |
| Bilak Yurii            | 05  |
| Bobrovnikova Kira      | 95  |
| Bombandyorov Yuri      | 259 |
| Vipshovsky Yuri        | 128 |
| Vovk Oleksandr         | 138 |
| Hliuza Mariia          | 138 |
| HryhoryevOleksandr     | 138 |
| Hryhorenko Svitlana    | 17  |
| Hrytsiuk Yurii         | 128 |
| Hurman Ivan            | 95  |
| Daleka Vasyl           | 259 |
| Datsenko Serhii        | 153 |
| Demchuk Vadym          | 165 |
| Dymertsov Dmytro       | 267 |
| Diachenko Dmytro       | 109 |
| Yevtushenko Oleksandr  | 276 |
| Zakovorotnyi Oleksandr | 276 |
| Kauk Viktor            | 211 |
| Kyrychenko Iryna       | 165 |
| Kyrii Valentyna        | 196 |
| Kobylin Ilya           | 284 |
| Kobylin Oleg           | 284 |
| Kolbasin Vladyslav     | 47  |
| Kuchapin Matvii        | 182 |
| Kuchuk Heorhii         | 153 |
| Kushnir Ivan           | 196 |
| Lutsenko Vitalii       | 165 |
| Luchenko Sergiy        | 182 |
| Makovetskyi Serhii     | 211 |
| Mendielieva Mariia     | 138 |
| Meteshkin Kostyantyn   | 259 |
| Mikhal Oleg            | 109 |
| Nechypurenko Stanislav | 34  |
| Nikolaichuk Anna       | 284 |
| Oleksiuk Dmytro        | 95  |
| Ostrovetskyi Serhii    | 297 |
| Pavlenko Alona         | 309 |
| Pakhomov Yuriy         | 47  |
| Pohorielov Stanislav   | 309 |
| Prosolov Vladyslav     | 224 |
| Protasova Liudmyla     | 17  |
| Putyatina Oleksandra   | 284 |
| Reshetnikova Svitlana  | 47  |

## АЛФАВІТНИЙ ПОКАЖЧИК

|                  |     |
|------------------|-----|
| Роговий А. І.    | 47  |
| Савельєва О. С.  | 17  |
| Сєверінов О. В.  | 235 |
| Сидоренко З. М.  | 235 |
| Ситнік О. О.     | 247 |
| Склярів І. І.    | 57  |
| Смеляков К. С.   | 182 |
| Сорокін А. Р.    | 34  |
| Татар'янц М. С.  | 309 |
| Ткачов В. М.     | 72  |
| Халімов Г. З.    | 224 |
| Чеботарьов Р. І. | 138 |
| Чумаченко І. В.  | 259 |
| Чуприна А. С.    | 182 |

## ALPHABETICAL INDEX

|                       |     |
|-----------------------|-----|
| Rogovyi Anton         | 47  |
| Savielieva Oksana     | 17  |
| Sievierinov Oleksandr | 235 |
| Sydorenko Zoia        | 235 |
| Sytnik Oleh           | 247 |
| Skliarov Illia        | 57  |
| Smelyakov Kyrylo      | 182 |
| Sorokin Anton         | 34  |
| Tatariaiants Maksym   | 309 |
| Tkachov Vitalii       | 72  |
| Khalimov Hennadii     | 224 |
| Chebatarov Ruslan     | 138 |
| Chumachenko Igor      | 259 |
| Chupryna Anastasiya   | 182 |

---



НАУКОВЕ ВИДАННЯ

SCIENTIFIC PUBLICATION

## Автоматизовані системи управління та прилади автоматики

## Management Information System and Devices

*Всеукраїнський міжвідомчий  
науково-технічний збірник*

*All-Ukrainian Interdepartmental  
Scientific and Technical Collection*

**Випуск 189, 2026**

**Issue 189, 2026**

Відповідальний секретар журналу

*О.М. Цимбал*

Відповідальний за ліцензування *В.В. Косенко*

Редактор *Л.В. Кузьміна*

Комп'ютерна верстка *Л.Ю. Свєтайло*

Responsible secretary of journal

*O. Tsymbal*

Responsible for licensing *V. Kosenko*

Editor *L. Kuzmina*

Computer layout *L. Svietailo*

Формат 60×84/8. Умов. друк. арк. 38,0.  
Тираж 150 прим.

Format 60×84/8. Conventional printed sheets 38,0.  
Edition of 150 copies.

Віддруковано з готових оригінал-макетів  
в типографії ФОП Андреев К.В.  
Єдиний державний реєстр юридичних осіб  
та фізичних осіб-підприємців.  
Запис №24800170000045020 від 30.05.2003.

Printed from ready-made original layouts  
in the printing house  
of Individual Entrepreneur Andreev K.V.  
Unified State Register of Legal Entities  
and Individual Entrepreneurs.  
Entry No. 24800170000045020 of 30.05.2003.

61157, Харків, вул. Акад. Богомольця, 9, кв. 50,  
тел. +38 (063) 993-62-73  
e-mail: ep.zakaz@gmail.com

fl. 50, 9, Acad. Bogomolets Str., Kharkiv, 61157,  
тел. +38 (063) 993-62-73  
e-mail: ep.zakaz@gmail.com