p-ISSN 1607-3274
e-ISSN 2313-688X

Радіоелектроніка Інформатика Управління

Radio Electronics Computer Science Control

2026/1

Міністерство освіти і науки України
Національний університет «Запорізька політехніка»

Радіоелектроніка, інформатика, управління

Науковий журнал
Виходить чотири рази на рік
№ 1(76) 2026
Заснований у 1998 році, видається з 1999 року.
Засновник і видавець – Національний університет «Запорізька політехніка».
ISSN 1607-3274 (друкований), ISSN 2313-688X (електронний).
Запоріжжя
НУ «Запорізька політехніка»
2026

Ministry of Education and Science of Ukraine
National University Zaporizhzhia Polytechnic

Radio Electronics, Computer Science, Control

The scientific journal
Published four times per year
№ 1(76) 2026
Founded in 1998, published since 1999.
Founder and publisher – National University Zaporizhzhia Polytechnic.
ISSN 1607-3274 (print), ISSN 2313-688X (on-line).
Zaporizhzhia
NU Zaporizhzhia Polytechnic
2026

Науковий журнал «Радіоелектроніка, інформатика, управління» (скорочена назва – РІУ) видається Національним університетом «Запорізька політехніка» (НУ «Запорізька політехніка») з 1999 р. періодичністю чотири номери на рік.

Реєстрація суб'єкта у сфері друкованих медіа: Рішення Національної ради України з питань телебачення і радіомовлення № 3040 від 07.11.2024 року. Ідентифікатор медіа: R30-05582.

ISSN 1607-3274 (друкований), **ISSN** 2313-688X (електронний).

Наказом Міністерства освіти і науки України № 409 від 17.03.2020 р. «Про затвердження рішень Атестаційної колегії Міністерства щодо діяльності спеціалізованих вчених рад від 06 березня 2020 року» **журнал включений до переліку наукових фахових видань України в категої «А» (найвищій рівень),** в яких можуть публікуватися результати дисертаційних робіт на здобуття наукових ступенів доктора наук і доктора філософії (кандидата наук).

Журнал включений до польського Переліку наукових журналів та рецензованих матеріалів міжнародних конференцій з присвоєною кількістю балів (додаток до оголошення Міністра науки та вищої освіти Республіки Польща від 31 липня 2019 р.: № 16981).

В журналі безкоштовно публікуються наукові статті англійською, російською та українською мовами.

Правила оформлення статей подано на сайті: http://ric.zntu.edu.ua/information/authors.

Журнал забезпечує **безкоштовний відкритий он-лайн доступ** до повнотекстових публікацій.

Журнал дозволяє авторам мати авторські права і зберігати права на видання без обмежень. Журнал дозволяє користувачам читати, завантажувати, копіювати, поширювати, друкувати, шукати або посилатися на повні тексти своїх статей. Журнал дозволяє повторне використання його вмісту у відповідності Creative Commons ліцензією CC BY-SA..

Опублікованим статтям присвоюється унікальний ідентифікатор цифрового об'єкта DOI.

Журнал входить до наукометричної бази Web of Science.

Журнал реферується та індексується у провідних міжнародних та національних реферативних журналах і наукометричних базах даних, а також розміщується у цифрових архівах та бібліотеках з безкоштовним доступом у режимі on-line, повний перелік яких подано на сайті: http://ric.zntu.edu.ua/about/editorialPolicies#custom-0.

Тематика журналу: телекомунікації та радіоелектроніка, програмна інженерія (включаючи теорію алгоритмів і програмування), комп'ютерні науки (математичне і комп'ютерне моделювання, оптимізація і дослідження операцій, управління в технічних системах, міжмашинна і людино-машинна взаємодія, штучний інтелект, включаючи системи, засновані на знаннях, і експертні системи, інтелектуальний аналіз даних, розпізнавання образів, штучні нейронні і нейро-нечіткі мережі, нечітку логіку, колективний інтелект і мультиагентні системи, гібридні системи), комп'ютерна інженерія (апаратне забезпечення обчислювальної техніки, комп'ютерні мережі), інформаційні системи та технології (структури та бази даних, системи, засновані на знаннях та експертні системи, обробка даних і сигналів).

Усі статті, пропоновані до публікації, одержують **об'єктивний розгляд**, що оцінюється за суттю без урахування раси, статі, віросповідання, етнічного походження, громадянства або політичної філософії автора(ів).

Усі статті проходять двоступінчасте закрите (анонімне для автора) **резензування** штатними редакторами і незалежними рецензентами – провідними вченими за профілем журналу.

РЕДАКЦІЙНА КОЛЕГІЯ

РЕДАКЦІЙНО-КОНСУЛЬТАТИВНА РАДА

.

Рекомендовано до видання Вченою радою НУ «Запорізька політехніка», протокол № 8 від 24.02.2026.

Журнал зверстаний редакційно-видавничим відділом НУ «Запорізька політехніка».

Веб-сайт журналу: http://ric.zntu.edu.ua.

Адреса редакції: Редакція журналу «РІУ», Національний університет «Запорізька політехніка», вул. Жуковського, 64, м. Запоріжжя, 69063, Україна.

Тел: (061) 769-82-96 – редакційно-видавничий відділ Факс: +38-061-764-46-62

E-mail: rvv@zntu.edu.ua

The scientific journal Radio Electronics, Computer Science, Control is published by the National University Zaporizhzhia Polytechnic NU Zaporizhzhia Polytechnic since 1999 with periodicity four numbers per year.

Registration of an entity in the field of print media: Decision of the National Council of Ukraine on Television and Radio Broadcasting No. 3040 of November 7, 2024. Media ID: R30-05582.

ISSN 1607-3274 (print), **ISSN** 2313-688X (on-line).

By the Order of the Ministry of Education and Science of Ukraine from 17.03.2020 № 409 "On approval of the decision of the Certifying Collegium of the Ministry on the activities of the specialized scientific councils dated 06 March 2020" **journal is included in the list of scientific specialized periodicals of Ukraine in category "A" (highest level),** where the results of dissertations for Doctor of Science and Doctor of Philosophy may be published.

The journal is included to tne Polish List of scientific journals and peer-reviewed materials from international conferences with assigned number of points (Annex to the announcement of the Minister of Science and Higher Education of Poland from July 31, 2019: Lp. 16981).

The journal publishes scientific articles in English, Russian, and Ukrainian free of charge.

The **article formatting rules** are presented on the site: http://ric.zntu.edu.ua/information/authors.

The journal provides policy of **on-line open (free of charge) access** for full-text publications. The journal allow the authors to hold the copyright without restrictions and to retain publishing rights without restrictions. The journal allow readers to read, download, copy, distribute, print, search, or link to the full texts of its articles. The journal allow reuse and remixing of its content, in accordance with Creative Commons license CC BY-SA.

Published articles have a unique digital object identifier (DOI).

The journal is included into Web of Science.

The journal is abstracted and indexed in leading international and national abstractig journals and scientometric databases, and also placed to the digital archives and libraries with a free on-line access, full list of which is presented at the site: http://ric.zntu.edu.ua/about/editorialPolicies#custom-0.

The journal scope: telecommunications and radio electronics, software engineering (including algorithm and programming theory), computer science (mathematical modeling and computer simulation, optimization and operations research, control in technical systems, machine-machine and man-machine interfacing, artificial intelligence, including data mining, pattern recognition, artificial neural and neuro-fuzzy networks, fuzzy logic, swarm intelligence and multiagent systems, hybrid systems), computer engineering (computer hardware, computer networks), information systems and technologies (data structures and bases, knowledge-based and expert systems, data and signal processing methods).

All articles proposed for publication receive an **objective review** that evaluates substantially without regard to race, sex, religion, ethnic origin, nationality, or political philosophy of the author(s).

All articles undergo a two-stage **blind peer review** by the editorial staff and independent reviewers – the leading scientists on the profile of the journal.

EDITORIAL BOARD

EDITORIAL-ADVISORY COUNCIL

Recommended for publication by the Academic Council of NU Zaporizhzhia Polytechnic, protocol № 8 dated 24.02.2026.

The journal is imposed by the editorial-publishing department of NU Zaporizhzhia Polytechnic.

The journal web-site is http://ric.zntu.edu.ua.

The address of the editorial office: Editorial office of the journal Radio Electronics, Computer Science, Control, National University Zaporizhzhia Polytechnic, Zhukovskiy street, 64, Zaporizhzhia, 69063, Ukraine.

Tel.: +38-061-769-82-96 – the editorial-publishing department. Fax: +38-061-764-46-62

E-mail: rvv@zntu.edu.ua

ЗМІСТ

CONTENTS

РАДІОЕЛЕКТРОНІКА ТА ТЕЛЕКОМУНІКАЦІЇ

RADIO ELECTRONICS AND TELECOMMUNICATIONS

UDC 621.396.946

FUZZY-LOGIC ALGORITHM FOR RISK ASSESSMENT IN WI-FI NETWORKS

Antipov I. – Doctor of sciences, Professor of the Department of Computer Radio Engineering and Technical Information Protection Systems, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0000-0002-9754-4412.

Vasylenko T. – PhD, Senior Lecturer of the Department of Computer Radio Engineering and Technical Information Protection Systems, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0000-0003-1291-8065.

ABSTRACT

Context. With the increasing use of Wi-Fi wireless networks, the risk of attacks specific to them is also rising. Traditional protection methods, which usually rely on precise thresholds, do not reflect the actual uncertainty of the conditions in which wireless networks operate. Due to the openness of the radio channel, its instability, dispersion, and the presence of noise, a promising direction is the use of fuzzy logic algorithms, which allow for taking into account the incompleteness and ambiguity of data when assessing the risks of Wi-Fi wireless networks.

Objective. Develop a fuzzy logic algorithm for assessing the state of Wi-Fi networks, which allows adaptively determining the level of risk by analyzing wireless network parameters and making decisions regarding security system actions.

Method. A fuzzy-logic-based algorithm for analyzing the operational state of a wireless Wi-Fi network is proposed. The algorithm is based on the integrated analysis of six network parameters using elements of fuzzy logic. It includes the construction of membership functions for the input variables, the formation of a fuzzy IF–THEN rule base, and a defuzzification mechanism that provides a continuous numerical assessment of the network risk level. To evaluate the effectiveness of the proposed approach, a comparative simulation study was conducted against the classical threshold-based decision-making method. The study was carried out in the MathCAD and MATLAB environments to enable cross-validation of the algorithm's functionality. Three network operation scenarios were considered, with 100 network states simulated for each scenario.

Results. The simulation results obtained in the MathCAD and MATLAB environments coincide up to the third decimal place, confirming the correctness of the software implementation of the algorithm. Comparative analysis showed that the threshold-based method produces binary decisions and is highly sensitive to random fluctuations in network parameters, which leads to an increased number of false alarms. The proposed fuzzy-logic-based algorithm provides a continuous risk assessment, demonstrates lower result variance, and exhibits a stable response to changes in network conditions. Under unstable network operating conditions, the algorithm enables discrimination between noise and interference effects and the initial phases of attacks, while also ensuring a gradual increase in the risk level without abrupt transitions between linguistic levels. The obtained results confirm a reduction in Type I errors and an improvement in decision-making informativeness.

Conclusions. The fuzzy logic-based Wi-Fi network state analysis algorithm proposed in this work enables more adequate decision-making regarding the network's condition. The use of fuzzy logic allows adjusting decisions depending on changes in network operating conditions in real time and can be integrated into intrusion detection systems or advanced wireless network cybersecurity tools.

KEYWORDS: Cybersecurity, Wi-Fi, intrusion detection systems, fuzzy logic, risk assessment.

ABBREVIATIONS

Auth_Fails is a number of failed authentications per minute;

Clients is a number of connected subscribers;

Com_Rule is a comprehensive assessment according to all rules;

ETX is an Expected Transmission Count;

IDS is an intrusion Detection System;

IEEE is an Institute of Electrical and Electronics Engineers;

IIoT is an Industrial Internet of Things;

Probe_Rate is a frequency of probe requests;

ROC is a Receiver Operating Characteristic;

RSSI_Var is a signal level variance;

S 1, 2, 3 is a scenario 1, 2, 3;

Traffic_Anomaly is a percentage of anomalous traffic;

Wi-Fi is a wireless fidelity;

WPA is a Wi-Fi Protected Access.

DOI 10.15588/1607-3274-2026-1-1

NOMENCLATURE

a is a the first point of the trapezoid;
A_{ik} is a fuzzy sets of the corresponding terms;
b is a the second point of the trapezoid;
B_k is a fuzzy sets of the corresponding terms;
c is a the third point of the trapezoid;
d is a the fourth point of the trapezoid;
j is a number of the linguistic term;
x_i is an input parameter;
y is an output parameter;
μ_{ij} is a membership function.

INTRODUCTION

In the modern world, wireless networks are extremely relevant and play an important role in people's lives. Many companies successfully use wireless local area networks to manage production processes, while hospitals deploy wireless networks to improve operational efficiency and convenience. The basic standard for wireless local area networks is the IEEE 802.11 standard, various versions of which regulate data transmission in the 2.4 and 5 GHz bands, as detailed in [1–4]. In practice, the actual communication range usually does not exceed 200 meters.

Since 802.11 standard devices communicate with each other over the radio spectrum, any other station operating in this band can also receive this data. To ensure at least a minimal level of wireless network security, encryption mechanisms based on WPA and WPA2 algorithms [5, 6] are used, as well as intrusion detection systems (IDS) [7].

This work considers an algorithm for analyzing the state of a wireless Wi-Fi network using elements of fuzzy logic. This algorithm allows making decisions regarding the presence of potential security threats, taking into account various or rapidly changing conditions that traditional intrusion detection systems (IDS) [8] do not consider.

The object of study is the operation process of a wireless Wi-Fi network and its functional parameters, which characterize the security state during subscriber access and data transmission.

The subject of study is methods and models of fuzzy logic assessment of risk levels in Wi-Fi networks based on the analysis of technical parameters (signal level, number of connected clients, number of failed authentications, traffic anomalies, signal level variance, and frequency of probe requests).

The purpose of the work is to develop and investigate a fuzzy logic algorithm for assessing the security risk level of a Wi-Fi network based on a set of its technical indicators. The algorithm should provide automated interpretation of traffic characteristics, connection states, and client behavior, generating both quantitative and linguistic risk assessments in real time.

1 PROBLEM STATEMENT

The input data are represented by a vector of observed parameters: $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, where x_1 =RSSI, dBm [–90,–30]; x_2 =Auth_Fails [0,20]; x_3 =Traffic_Anomaly [0, 100]; x_4 =Clients [0, 50]; x_5 = RSSI_Var[0, 25]; x_6 = Probe_Rate [0, 20].

It is necessary to construct fuzzy membership functions for each variable: $\mu_{ij}(x_i): X_i \to [0,1]$.

The output variable is the network state risk level: $y = Risk\_Level \in [0,1]$, which is also described by a set of linguistic terms: $\{no\_risk, low\_risk, medium\_risk, high\_risk\}$.

A Mamdani-type fuzzy rule base should be formed: $R_k : IF x_1 \in A_{1k} AND ... AND x_6 \in A_{6k} THEN y \in B_k$.

Quality criteria include:

– correctness of the fuzzy interpretation of network states;

– consistency of results in MathCAD and MATLAB environments;

– the ability of the algorithm to generate values of corresponding to expected scenarios (“no risk,” “low,” “medium,” “high”);

– stability of results under variations of input parameters.

Additional constraints: all input parameters must belong to their allowed intervals; membership functions must provide full coverage of the respective universes; the rule base must be minimally sufficient yet adequate to account for all typical network states.

2 REVIEW OF THE LITERATURE

Fuzzy logic is not merely a theoretical tool; it is actively used in practice for ensuring security and assessing risk in various types of wireless networks, as evidenced by a large number of published works on this topic.

In [9], the authors propose a model for analyzing information security risks in IIoT. They developed several fuzzy inference systems for assessing overall risk. The model reflects real conditions of attacks and threats, but it is tailored to IIoT systems rather than traditional Wi-Fi networks, which have their own specific characteristics.

The study in [10] demonstrated the use of fuzzy logic to detect jamming attacks in wireless mesh IoT networks. The authors used ETX metrics, the number of retransmissions, undelivered packets, and packet delivery ratio as input parameters. The system was evaluated using standard metrics (accuracy, precision, recall, ROC). This work demonstrates that the fuzzy approach can be effective even in complex attack scenarios at the physical and data link layers. Despite the high effectiveness of this method against jamming attacks, it may be less suitable for typical Wi-Fi attacks, as it relies on a limited set of parameters that are not fundamental for Wi-Fi networks.

The work in [11] is devoted to improving the accuracy of IDS. To achieve this, the authors combine fuzzy

DOI 10.15588/1607-3274-2026-1-1

logic, neural networks, and a genetic algorithm. This study demonstrates that fuzzy logic can be successfully integrated with modern machine learning methods to achieve higher efficiency. However, such a method is complex to implement and requires significant resources, which may be critical for a real Wi-Fi network.

The publication [12] investigates IDS based on fuzzy logic. The authors showed that using fuzzy models with classical membership functions can significantly outperform traditional threshold-based approaches in terms of accuracy, especially when dealing with limited or noisy data. Since this model uses triangular membership functions and lacks a learning mechanism, it limits the flexibility of the wireless network under dynamic network conditions, when monitoring is most necessary.

Few studies consider the combination of risk assessment with security at the Wi-Fi level. Most research focuses on IoT/IIoT, where network structures and characteristics differ significantly.

Therefore, the development of fuzzy logic algorithms for risk assessment in Wi-Fi networks remains a relevant and timely task.

3 MATERIALS AND METHODS

The main element of using fuzzy logic is the membership functions. They determine the degree of belonging of an output variable to a linguistic term. Most often, triangular membership functions are used because they are simple to compute and clearly illustrate the process of fuzzy evaluation of wireless system parameters. However, in real systems, they are not sufficiently informative.

In this work, trapezoidal membership functions are used to describe fuzzy linguistic variables, which are well suited for changes in real Wi-Fi network parameters such as signal strength, the number of connected clients, the frequency of failed authentications, and the traffic anomaly index. Membership functions of this type simplify the construction of fuzzy logic rules for wireless networks.

The work describes membership functions for working hours in a secured enterprise. Analogous membership functions should be created for different scenarios (working hours, vacation periods, nighttime). For instance, during nighttime, when only stationary devices (e.g., cameras) operate in the enterprise, if the system detects that 30 subscribers are connected to the network, this is considered an anomaly, unlike during working hours.

The general formula for a trapezoidal membership function is as follows:

$$\mu_{ij} = \begin{cases} 0, x \le a, \\ \dfrac{x-a}{b-a}, a < x \le b, \\ 1, b < x \le c, \\ \dfrac{d-x}{d-c}, c < x \le d, \\ 0, x \ge d. \end{cases} \quad (1)$$

To determine the signal level, we use three linguistic terms: “weak” with trapezoid points [–90;–80;–70], “medium” [–80;–70;–60;–50], and “strong” [–60;–50;–40;–30].

The membership functions for the signal level are shown in Fig. 1.

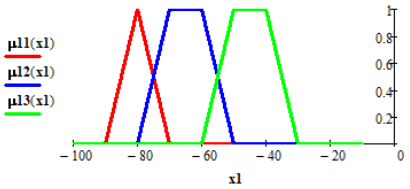


Figure 1 – Signal level membership functions

To determine the number of failed authentications, we use three linguistic terms: “low level” [0;0;1;3], “medium level” [2;5;8;12], and “high level” [10;15;20;20]. The membership functions for the number of failed authentications are shown in Fig. 2.

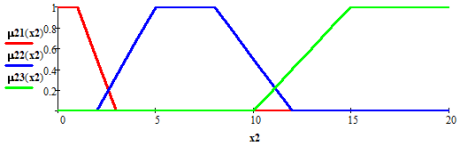


Figure 2 – Membership functions of the number of failed authentications

When determining the percentage of anomalous traffic, three linguistic terms were used: “normal” [0;0;10;30], “suspicious” [20;40;60;80], and “critical” [70;85;100;100]. The membership functions for the percentage of anomalous traffic are shown in Fig. 3.

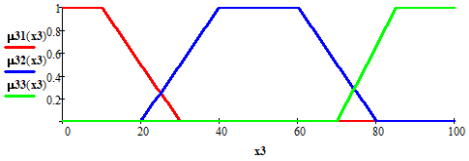


Figure 3 – Membership functions of the percentage of anomalous traffic

When determining the number of connected subscribers, three linguistic terms are used: “few” [0;0;5;15], “normal” [10;20;30;40], and “many” [30;40;50;50]. The membership functions for the number of connected clients are shown in Fig. 4.

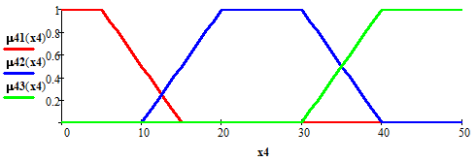


Figure 4 – Membership functions of the number of connected clients

To determine the signal level variance over a short interval, three linguistic terms are used: “stable signal”

DOI 10.15588/1607-3274-2026-1-1

[0;0;2;4], “moderate instability” [4;6;10;12] , and “high signal instability” [10;14;25;25]. The membership functions for the signal level variance over a short interval are shown in Fig. 5.

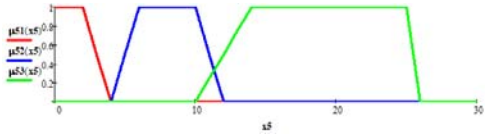


Figure 5 – Membership functions of the signal level dispersion over a short interval

To determine the frequency of probe requests, three linguistic terms are used: “low frequency” [0;0;10;20], “moderate frequency” [20;30;50;60], and “high frequency” [50;70;120;120]. The membership functions for the frequency of probe requests are shown in Fig. 6.

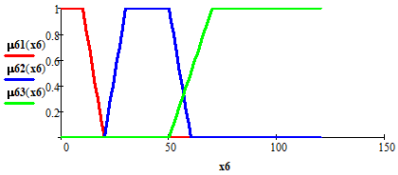


Figure 6 – Probe frequency membership functions

The output variable of the fuzzy system in this work is the “Risk Level.” It represents a generalized assessment of the state of the wireless network and is used for decision-making. The variable is formed based on a combination of the system’s input parameters. Each of these parameters can partially correspond to different linguistic states, so the risk assessment result is also expressed as degrees of membership to the corresponding linguistic terms. For the output variable, four linguistic terms have been defined to represent the level of threat: “no risk” [0;0;0;0.3], “low risk” [0.2;0.3;0.5;0.6], “medium risk” [0.4;0.5;0.8;0.9], and “high risk” [0.7;0.8;1;1]. The membership functions of the output variable are shown in Fig. 7.

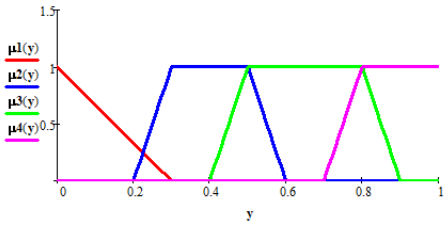


Figure 7 – Membership functions of the output variable

This representation helps to avoid abrupt jumps in decision-making and ensures smooth changes in the system’s response with small variations in network parameters.

To create a security system for a wireless Wi-Fi network, six input parameters and one output parameter were used. Considering all possible combinations would require 729 rules. To prevent overloading the system while still demonstrating its effectiveness, a simplified rule base was developed in this work, where the system’s response is determined by the factors that most significantly influence network risks. For example, if the system indicates that traffic is anomalous, the number of users or the signal level becomes almost irrelevant – the risk will still be high. Similarly, if the number of failed authentications is high and there is simultaneously a large number of probe requests, this almost always indicates the presence of an attack.

Algorithm processing procedure:

1. At the first stage, data is collected. In real time, the security system receives data from the access point or through specialized software that analyzes the data transmission environment.

2. At the second stage, membership functions consisting of linguistic terms are assigned to all six input parameters and the single output parameter. In our case, trapezoidal functions are used, which evaluate each parameter from 0 to 1.

3. The third stage involves creating the rule base. In this work, rules were developed using several experts and statistical analysis of medium-sized wireless Wi-Fi networks. The resulting rule base contains 20 rules in an “IF–THEN” format.

4. At the fourth stage, all rules are evaluated to determine how well they correspond to the current state of the wireless network. Computation is carried out using fuzzy logic by calculating the minima of the membership functions, as this method is classical for the Mamdani-type system used in this work.

5. After transforming features into fuzzy linguistic variables and computing them, a comprehensive assessment of all parameters is performed. This stage involves combining all rule sets that are above zero into a single function that characterizes the network risk level using the maximum operation. If several rules indicate a high security risk at different levels, they are aggregated. Aggregation is performed by taking the highest values at each point. As a result, a curve is obtained that characterizes the current state of the wireless network.

6. After the comprehensive evaluation of all parameters, the defuzzification process is performed to convert the linguistic variable into a numerical value.

7. Based on the number obtained after defuzzification, the system determines the security state of the wireless network according to the corresponding risk level:

– 0–0.299 – no risk (normal operation),
– 0.3–0.599 – low risk (enhanced monitoring),
– 0.6–0.799 – medium risk (event logging),
– >8 – high risk (automatic blocking or administrator notification).

8. At the final stage, the resulting security decision of the system is made.

9. As an additional but very important function, implemented after decision-making, feedback allows the security system to learn responses to attacks it has already encountered, adjust membership functions, or modify the system’s response to specific activities.

The structural diagram of the algorithm implementing network analysis functions using fuzzy logic is shown in Fig. 8.

DOI 10.15588/1607-3274-2026-1-1

4 EXPERIMENTS

The purpose of the experiment is to verify the functionality of the proposed wireless network protection model. To simulate the operation of the proposed security system, which uses elements of fuzzy logic, two software environments were employed: MathCAD and MATLAB.

Using two independent software environments allowed verification of the correctness of the proposed algorithm and detection of possible errors. MathCAD enables step-by-step implementation with mathematical descriptions and graphical visualization, but errors may occur during algorithm development. Therefore, MATLAB was also used, which performs all calculations automatically and provides only graphical visualization of the algorithm.

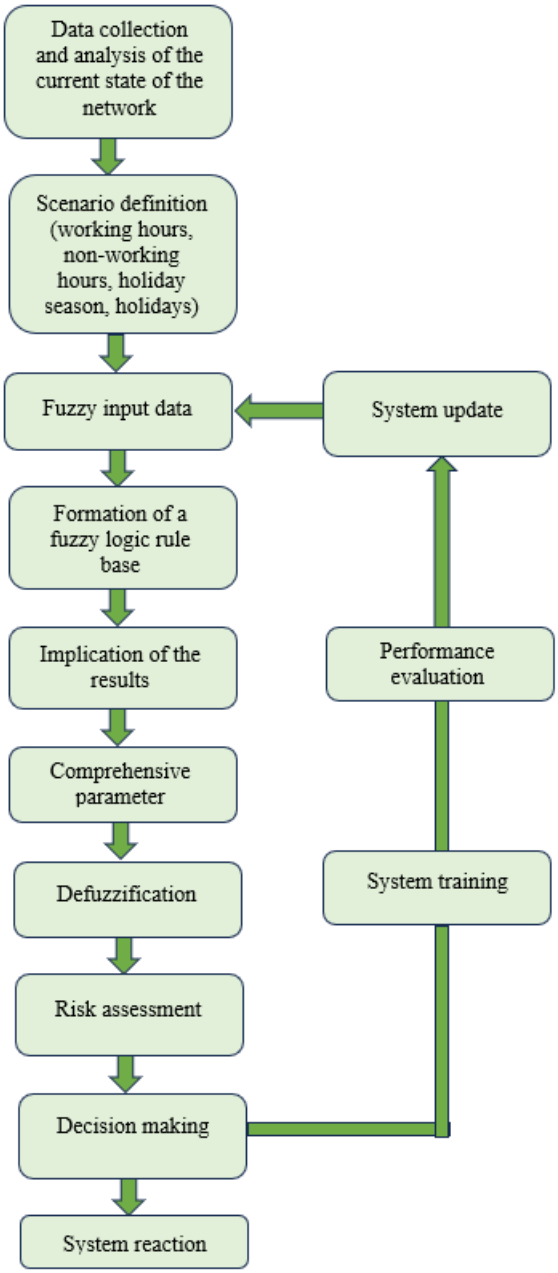


Figure 8 – Block diagram of an algorithm that implements network analysis functions using fuzzy logic

During the implementation of the algorithm in both software environments, identical operating conditions were applied (the same six input parameters, the same membership functions, the same rule base with 20 rules, the same defuzzification method – center of gravity, and identical test input combinations).

According to (1), in both software environments, the rules were implemented prior to the realization of the trapezoidal membership functions for the input and output parameters.

For implementing the relationships within the rules, the Mamdani model was used with the t-norm as the minimum operator. The rules implemented in the MathCAD environment are shown in Fig. 9

1. R1 := μ33(x3)
2. R2 := min(μ32(x3), μ22(x2))
3. R3 := min(μ32(x3), μ43(x4))
4. R4 := min(μ31(x3), μ41(x4))
5. R5 := min(μ23(x2), μ63(x6))
6. R6 := min(μ22(x2), μ62(x6))
7. R7 := min(μ21(x2), μ31(x3))
8. R8 := min(μ11(x1), μ53(x5))
9. R9 := min(μ11(x1), μ52(x5))
10. R10 := min(μ13(x1), μ51(x5))
11. R11 := min(μ43(x4), μ32(x3))
12. R12 := min(μ43(x4), μ23(x2))
13. R13 := min(μ42(x4), μ31(x3))
14. R14 := min(μ53(x5), μ63(x6))
15. R15 := min(μ53(x5), μ32(x3))
16. R16 := min(μ51(x5), μ61(x6))
17. R17 := min(μ31(x3), μ21(x2))
18. R18 := min(μ31(x3), μ51(x5))
19. R19 := min(μ31(x3), μ61(x6))
20. R20 := min(μ31(x3), μ41(x4))

Figure 9 – Rule base

For horizontal truncation, the MIN implication was used.

To perform a comprehensive evaluation of the parameters across all rules, the S-norm function using the maximum value was applied:

$$Com\_Rule = \max(Rule1(y), Rule2(y)...Rule20(y)).$$

To obtain a result, the system must convert the set of values into a single number. For this purpose, defuzzification is used, employing the center of gravity method:

$$Risk = \frac{\sum_{y}(y \cdot Com\_Rule(y))}{\sum_{y}(Com\_Rule(y))}.$$

To compare the classical threshold method of analyzing a wireless Wi-Fi network and the proposed algorithm that analyzes the network using elements of fuzzy logic, a comparative experimental study was conducted using the MathCAD software environment. Three scenarios were considered. S 1 – normal network operation; S 2 – unstable conditions (large signal dispersion, fluctuations in traffic parameters caused by noise, interference and dynamic changes in the communication channel, without active attacks); S 3 – the initial phase of the attack (high level of unsuccessful authentications, high frequency of probe requests). For each scenario, a sequence of 100 network states was generated and processed, the same for both network analysis methods.

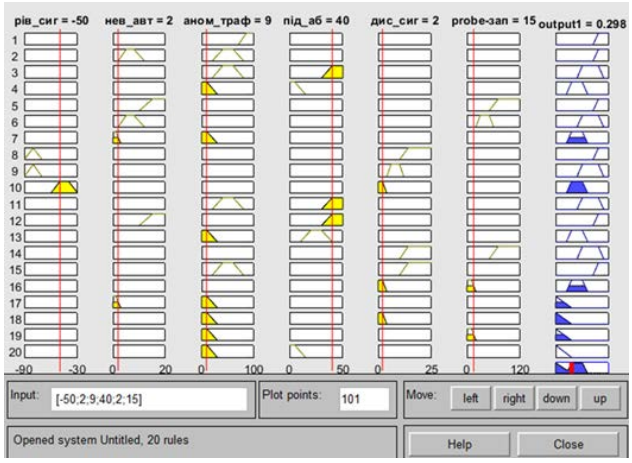
DOI 10.15588/1607-3274-2026-1-1

5 RESULTS

To demonstrate the operation of the security system, modeled input parameters were fed into the model. As an example, the network state “No Risk” is shown. Input data: $x_1 = –50$; $x_2 = 2$; $x_3 = 9$; $x_4 = 35$; $x_5 = 2$; $x_6 = 15$. The results are shown in Fig. 10a, implemented in MATLAB, and Fig. 10b, implemented in MathCAD.

In Fig. 10a, all 20 rules are shown, indicating which rules are activated and to what extent, with implication applied. For each rule, the resulting membership function is displayed, and at the very bottom, the aggregated membership function after the comprehensive evaluation, consisting of all activated rules, is shown. Additionally, the numeric result is shown above the resulting membership functions.

In Fig. 10b, the input data and the resulting value are demonstrated, matching the MATLAB results up to the third decimal place. The resulting membership function is also shown, which completely coincides with the one obtained in MATLAB.



a

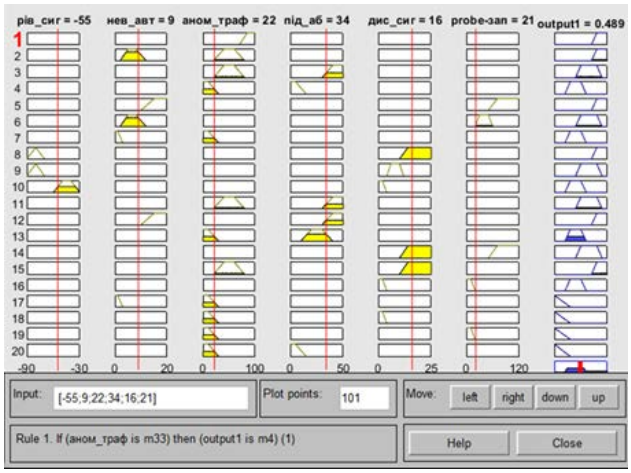
b

Figure 10 – In the network under study, the status is “No risk” : a – implementation in MATLAB; b – implementation in MathCAD

For the network state “Low Risk,” the results are shown in Fig. 11. Input data: $x_1 = –55$; $x_2 = 9$; $x_3 = 22$; $x_4 = 34$; $x_5 = 16$; $x_6 = 21$.

For the network state “Medium Risk,” the results are shown in Fig. 12. Input data: $x_1 = –60$; $x_2 = 3$; $x_3 = 15$; $x_4 = 24$; $x_5 = 5$; $x_6 = 25$.

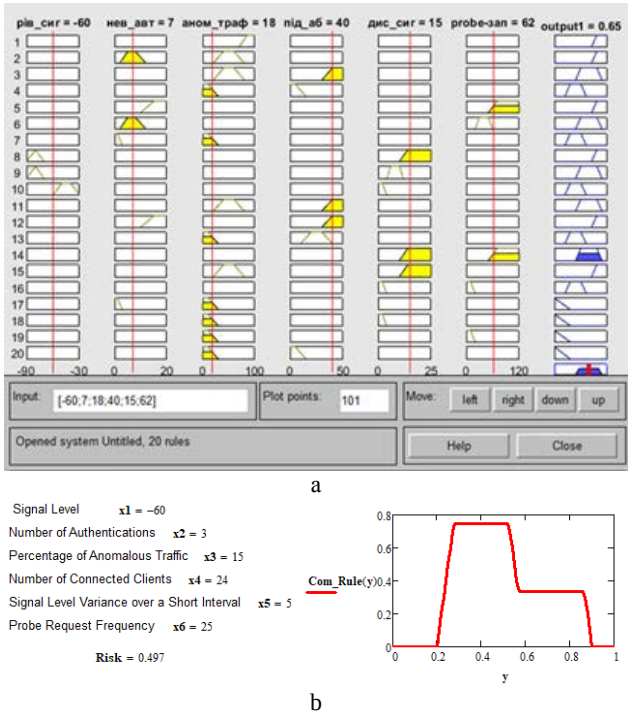
For the network state “High Risk,” the results are shown in Fig. 13. Input data: $x_1 = –75$; $x_2 = 15$; $x_3 = 85$; $x_4 = 45$; $x_5 = 19$; $x_6 = 44$.



a

b

Figure 11 – In the network under study, the status is “Risk is low” : a – implementation in MATLAB; b – implementation in MathCAD



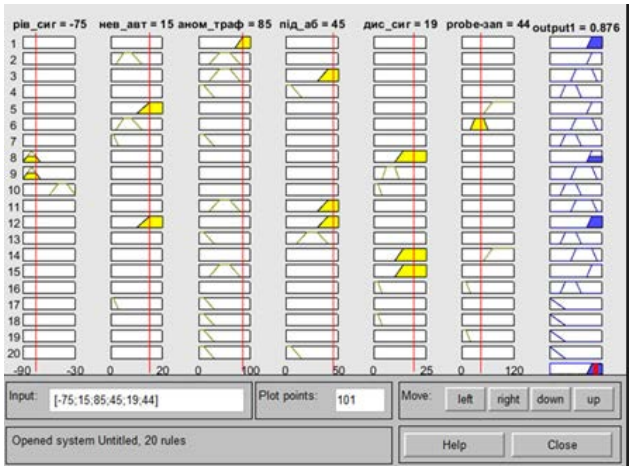
a

b

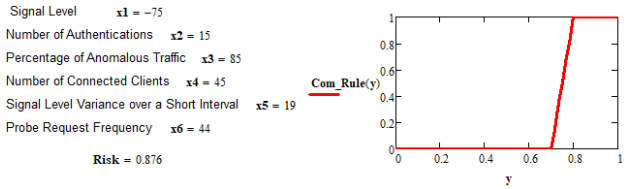
Figure 12 – In the network under study, the status is “Risk Medium” : a – implementation in MATLAB; b – implementation in MathCAD

For S1–3, 100 consecutive network states were generated. The values presented in Tables 1 and 2 correspond to statistical characteristics (mean and standard deviation) calculated over these 100 simulations. Table 3 presents representative risk values illustrating the transition dynamics during the initial attack phase.

DOI 10.15588/1607-3274-2026-1-1



a



b

Figure 13 – In the network under study, the status is “Risk is high” : a – implementation in MATLAB; b – implementation in MathCAD

The threshold-based method produces binary decisions for each network state (attack / no attack). Therefore, the values presented in Tables 1–3 represent the alarm frequency calculated over 100 simulated network states rather than a continuous risk level. In contrast, the proposed fuzzy-logic algorithm provides a smooth risk assessment that reflects gradual changes in network conditions.

Table 1 – Risk assessment results for different Wi-Fi network scenarios

Scenario	Alarm frequency based on threshold value	Fuzzy-logic risk
S1	5%	0.18
S2	62%	0.46
S3	88%	0.82

Table 2 – Stability of methods

Method	Mean alarm frequency	Standard deviation
Alarm frequency based on threshold value	0.62	0.31
Fuzzy-logic risk	0.47	0.08

Table 3 – Risk evolution during the initial phase of the attack

Network state index	Threshold-based decision	Fuzzy-logic algorithm
1	0	0.32
20	0	0.48
40	1	0.66
60	1	0.78
80	1	0.86

6 DISCUSSION

As can be seen from Figs. 10–13, the results are identical across different software environments, as also demonstrated in Table 4.

Table 4 – Comparison of Results

Network State	MathCAD Result	MATLAB Result	Difference
No Risk	0.298	0.298	0
Low Risk	0.489	0.489	0
Medium Risk	0.65	0.65	0
High Risk	0.876	0.876	0

Thus, the proposed algorithm correctly determines the risk levels. The results obtained in both software environments match with an accuracy of 10^{-3} (the same precision used during calculations), which confirms the correctness of the implementation of the fuzzy-logic-based wireless network protection system. The dual implementation of the algorithm demonstrated consistent results, indicating the reliability of the proposed approach to network risk assessment. The actions of the fuzzy-logic algorithm depend on the risk level determined by the system at a given moment.

When the system determines the result “Low Risk,” it means that the network is operating in normal mode without any suspicious activity, i.e., all network parameters are within acceptable limits. In this case, the system continues its regular operation and periodically logs the network status (RSSI, number of clients, traffic). No additional actions are taken so as not to overload the security system.

For example, the signal level is stable, the number of clients corresponds to the expected value for the given time and day of the week, failed authentications are rare, and the traffic shows no anomalies.

With “Medium Risk,” the wireless security system detects certain deviations in the analyzed indicators, but these deviations are not significant. In such a situation, occasional network malfunctions, repeated authentications, or signal instability may occur. This indicates that the likelihood of an attack is not high, but the network still requires increased monitoring. In this case, the system begins enhanced monitoring of the network: it collects data more frequently (for instance, if under normal conditions data is received every minute, it may start receiving updates every 10 seconds); it begins recording an extended list of parameters in the log (probe requests, traffic anomalies); it analyzes nearby access points (in case a rogue access point attack is being attempted). The system may also limit data transfer rates for certain suspicious clients or request re-authentication from them. At this risk level, the administrator receives a notification about the network status

For example, if the security system detects a decrease in signal strength for several clients and a slight increase in the number of probe requests, but without any loss of connection.

When a “High Risk” level is identified, it means that the wireless security system has detected signs of an on-

DOI 10.15588/1607-3274-2026-1-1

going attack (a suspicious client, password-guessing attempts, or atypical traffic). In this situation, an entry is made in the security log, recording the time, MAC address, and other signal parameters. The system sends a notification to the administrator marked "WARNING", may block suspicious clients, and also forwards a signal to the intrusion detection system to compare signatures with a database of known attacks.

For example, if a client or several clients exhibit suspicious activity – changes in signal strength, authentication errors, and traffic that is not typical for the user – this is highly likely to indicate the beginning of a man-in-the-middle attack.

In the case of "Critical Risk," the network exhibits activity that clearly indicates an active attack on the Wi-Fi wireless network: eavesdropping on communication channels, rogue access points, mass authentication requests, or a sharp increase in probe requests. During such aggressive activity, the security system responds immediately, blocking the malicious activity. This may include complete traffic blockage in a specific segment, after which the administrator is notified of the threat. A security log entry is also created, containing all current parameters.

For example, if the traffic is abnormal, there is a high number of authentications, and the signal strength drops sharply, this may indicate a DoS attack, an access point spoofing attack, or network client scanning.

Traditional algorithms use fixed thresholds. If a low threshold is set, a high percentage of Type I errors (missed attacks) will occur. Conversely, if the threshold is set too high, a large number of Type II errors (false positives) will occur. Both cases negatively affect the operation of the wireless network. Choosing the "golden middle" is practically impossible.

The experimental results (Tables 1–3) demonstrate a fundamental difference between the classical threshold-based approach and the proposed fuzzy-logic-based risk assessment algorithm when applied to Wi-Fi network monitoring.

First, the results presented in Table 1 show that under normal operating conditions (S1), the threshold-based method generates a nonzero alarm rate (5%), indicating the presence of false positives caused by random fluctuations in network parameters. In contrast, the fuzzy logic algorithm produces a low continuous risk value (0.18), which reflects a more adequate interpretation of minor deviations and helps to avoid unnecessary alarm triggering.

Under unstable but benign conditions (S2), characterized by increased signal variance and traffic fluctuations, the limitations of the threshold-based approach become more evident. The alarm rate increases to 62%, which complicates reliable decision-making and may lead to excessive security responses. At the same time, the fuzzy logic algorithm assigns a moderate risk level (0.46), indicating a degradation in network conditions without explicitly classifying it as an attack. This behavior confirms the ability of the proposed algorithm to distinguish between channel-induced instability and malicious activity.

During the initial phase of an attack (S3), both methods detect anomalous behavior. However, their results differ significantly in terms of interpretability. The threshold-based method rapidly switches to a high alarm rate (88%), providing only binary information. In contrast, the fuzzy logic algorithm generates a high but unsaturated risk value (0.82), enabling the system to track the progression of the attack and to provide adaptive security responses.

The robustness of both methods is further illustrated in Table 2. The standard deviation of the fuzzy logic risk values is significantly lower than that of the threshold-based alarm rate (0.08 versus 0.31). This indicates that the proposed algorithm provides a more stable assessment under varying network conditions and is less sensitive to random noise and short-term parameter fluctuations.

The temporal evolution of risk during the initial attack phase, shown in Table 3, highlights an important advantage of the fuzzy logic approach. While the threshold-based method produces abrupt transitions from "no attack" to "attack," the fuzzy logic algorithm reflects a gradual increase in risk as network conditions deteriorate. This property is particularly important for early attack detection and proactive security management in wireless networks.

The proposed fuzzy logic algorithm does not merely replicate threshold-based decisions but extends them by providing a continuous, interpretable, and noise-resilient risk assessment. This makes the approach more suitable for real Wi-Fi network environments, which are inherently characterized by parameter variability, channel noise, and transient states.

The proposed algorithm, which uses elements of fuzzy logic, introduces four output terms, allowing the reduction of both Type I and Type II errors.

The developed algorithm for network state decision-making is recommended for implementation within an intrusion detection system in companies, offices, and other environments using IEEE 802.11 wireless networks. Such network protection is particularly advisable in locations where valuable information is stored and there is a risk of unauthorized network access. The proposed model is relevant because it does not require significant resources, is user-friendly, and significantly enhances the security of the wireless network. Moreover, a fuzzy-logic-based decision-making model can detect attacks whose signatures are unknown.

In combination with the methodologies discussed in [13, 14], this algorithm can serve as a reliable protection system for Wi-Fi wireless networks.

CONCLUSIONS

The problem of improving decision-making reliability in Wi-Fi network state analysis under conditions of uncertainty is addressed in this work.

The scientific novelty of the obtained results consists in the development of a fuzzy-logic-based algorithm for

DOI 10.15588/1607-3274-2026-1-1

Wi-Fi network state assessment that enables adaptive interpretation of network parameters using fuzzy rules and membership functions. The proposed approach allows smoother transitions between network states and reduces the impact of abrupt parameter changes, which improves the adequacy of anomaly detection compared to threshold-based methods.

The obtained results enable more stable identification of abnormal network conditions and contribute to reducing false alarms in dynamically changing environments.

The practical significance of the results lies in the possibility of using the developed algorithm in Wi-Fi monitoring and security systems operating in automatic or expert-assisted modes. The results of modeling confirm the applicability of the proposed approach for practical network state analysis tasks.

Prospects for further research include extending the set of analyzed parameters and conducting quantitative evaluation of detection efficiency in real-world wireless environments.

ACKNOWLEDGEMENTS

We thank the Department of Computer Radio Engineering and Technical Information Protection Systems of the Kharkiv National University of Radio Electronics for the opportunity to conduct scientific research.

SOFTWARE AVAILABILITY

MATLAB R2018b (MathWorks Inc., Natick, MA, USA) was used for numerical simulations and data processing. The software was used under a valid professional license.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: Ivan Antipov: a method for analyzing the state of functioning of a Wi-Fi wireless network, which analyzes network parameters using fuzzy logic elements; Tetiana Vasylenko: experimental study of a method for analyzing the state of functioning of a Wi-Fi wireless network, which analyzes network parameters using fuzzy logic elements and analysis of the results obtained.

Data availability: The manuscript has no related data in the repository.

Software availability: The manuscript has no associated software.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

REFERENCES

1. Natkaniec P., Bieńkowski P. Analysis of the Mixed IEEE 802.11ax Wireless Networks in the 5 GHz Band, *Sensors*, 2023, Vol. 23, № 10, P. 4964. DOI: 10.3390/s23104964.
2. Forenbacher I., Husnjak S., Jovović I. et al. Throughput of an IEEE 802.11 Wireless Network in the Presence of Wireless Audio Transmission: A Laboratory Analysis, *Sensors*, 2021, Vol. 21, № 8, P. 2620. DOI: 10.3390/s21082620
3. Wang K., Psounis K. Efficient scheduling and resource allocation in 802.11ax multiuser transmissions, *Computer Communications*, 2020, Vol. 152, pp. 171–186. DOI: 10.1016/j.comcom.2020.01.010
4. Natkaniec M., Kras M. An Optimization of Network Performance in IEEE 802.11ax Dense Networks , *International Journal of Electronics and Telecommunications*, 2023, Vol. 69, pp. 169–176. DOI: 10.24425/ijet.2023.144347
5. Faíscas D. (In)Security in Wi-Fi networks: a systematic review, *Advanced Research on Information Systems Security*, 2022, Vol. 2, № 2, pp. 17–23. DOI: 10.56394/aris2.v2i2.18
6. Lee B. Stateless Re-Association in WPA3 Using Paired Token, *Electronics*, 2021, Vol. 10, № 2, P. 215. DOI: 10.3390/electronics10020215
7. Kumar Y. A., Kumar V. Systematic Review on Intrusion Detection System in Wireless Networks: Variants, Attacks, and Applications, *Wireless Personal Communications*, 2023, Vol. 123, № 1, pp. 395–452. DOI: 10.1007/s11277-023-10773-x
8. Dimakis D. A., Michael K. Survey of Wireless Intrusion Detection Systems: Threats, Swarm Intelligence and Machine Learning-Based Solutions, *Wireless Networks*, 2022, Vol. 23, № 10, P. 4964. DOI: 10.1007/s11276-022-02933-3
9. Kerimkhulle S., Dildebayeva Z., Tokhmetov A. et al. Fuzzy Logic and Its Application in the Assessment of Information Security Risk of Industrial Internet of Things, *Symmetry*, 2023, Vol. 15, № 10, P. 1958. DOI: 10.3390/sym15101958
10. Savva M., Ioannou I., Vassiliou V. Fuzzy-Logic Based IDS for Detecting Jamming Attacks in Wireless Mesh IoT Networks, *Mediterranean Communication and Computer Networking Conference (MedComNet), Paphos, 1–3 June 2022: proceedings.* Paphos, IEEE, 2022, pp. 54–63. DOI: 10.48550/arXiv.2205.03797
11. Ishaque M., Khatibi A., Yamin M.et al. A novel hybrid technique using fuzzy logic, neural networks and genetic algorithm for intrusion detection system, *Sensors*, 2023, Vol. 30, P. 100933. DOI: 10.1016/j.measen.2023.100933
12. Iantorno M. S., Beladda K. Fuzzy Logic for Cybersecurity: Intrusion Detection and Privacy Preservation with Synthetic Data, *Agents and Artificial Intelligence : 17th International Conference (ICAART), Porto, 26–28 February 2025: proceedings.* Porto, SCITEPRESS, 2025, Vol. 3, pp. 376–382. DOI: 10.5220/0013137300003890
13. Antipov I., Vasylenko T. Identification of mobile devices by correlation features of their signal spectra, *Radio Electronics, Computer Science, Control*, 2024, № 4, pp. 6–12. DOI: 10.15588/1607-3274-2024-4-1
14. Antipov I., Vasilenko T. Improving the model of decision making about abnormal network state using a positioning system, *Eastern-European Journal of Enterprise Technologies*, 2019, Vol. 1, № 9 (97), pp. 6–11. DOI: 10.15587/1729-4061.2019.157001

Received 30.12.2025.
Accepted 11.02.2026.
Published 27.03.2026.

DOI 10.15588/1607-3274-2026-1-1

УДК 621.396.946

НЕЧІТКО-ЛОГІЧНИЙ АЛГОРИТМ ОЦІНЮВАННЯ РИЗИКУ У WI-FI МЕРЕЖАХ

Антіпов І. Є. – д-р техн. наук, професор кафедри комп'ютерної радіоінженерії та систем технічного захисту інформації, Харківський національний університет радіоелектроніки, м. Харків, Україна. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0000-0002-9754-4412.

Василенко Т. О. – канд. техн. наук, старший викладач кафедри комп'ютерної радіоінженерії та систем технічного захисту інформації, Харківський національний університет радіоелектроніки, м. Харків, Україна. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0000-0003-1291-8065.

АНОТАЦІЯ

Актуальність. Зі зростанням використання безпровідних мереж Wi-Fi підвищується ризик атак, специфічних саме для них. Традиційні методи захисту які зазвичай використовують чіткі пороги, не відображають реальної невизначеності умов в яких функціонують безпровідні мережі. Через відкритість радіоканалу, нестабільність, розсіюваність та наявність шуму, перспективним напрямком є використання нечітко-логічних алгоритмів, що дозволяють враховувати неповноту та неоднозначність даних при оцінюванні ризиків безпровідних мереж Wi-Fi.

Мета. Розробити нечітко-логічний алгоритм оцінювання стану Wi-Fi мереж, який дозволяє адаптивно визначати рівень ризику, аналізуючи параметри безпровідної мережі та приймати рішення щодо дій системи безпеки.

Метод. Запропоновано нечітко-логічний алгоритм аналізу стану функціонування безпровідної Wi-Fi мережі, що базується на комплексному аналізі шести мережевих параметрів із використанням елементів нечіткої логіки. Алгоритм включає побудову функцій належності для вхідних змінних, формування бази нечітких правил типу IF–THEN та механізм дефазифікації, що забезпечує отримання безперервної числової оцінки рівня ризику мережі. Для оцінювання ефективності запропонованого підходу проведено порівняльне імітаційне моделювання з класичним пороговим методом прийняття рішень. Дослідження виконано у середовищах MathCAD та MATLAB для взаємної перевірки працездатності алгоритму. Розглянуто три сценарії функціонування мережі, для кожного з яких змодельовано 100 станів мережі.

Результати. Результати імітаційного моделювання збігаються з точністю до третього знаку в двох програмних середовищах MathCAD та MATLAB. Запропонований алгоритм коректно реагує на збільшення кількості невдалих спроб автентифікації та на аномальні зміни трафіку. Використання елементів нечіткої логіки дозволяє уникнути різких стрибків між рівнями ризику «низький», «середній», «високий», що зменшує кількість хибних тривог та мінімізує помилки першого роду. Модель успішно розрізняє нормальні зміни рівня сигналу та небезпечні. Запропонований алгоритм здатен сам реагувати на потенційні загрози: моніторинг, посилене логування, обмеження доступу, блокування клієнта та сповіщати адміністратору.

Висновки. Запропонований у роботі нечітко-логічний алгоритм аналізу стану Wi-Fi мережі на основі нечіткої логіки дає змогу більш адекватно ухвалювати рішення щодо стану мережі. Використання нечіткої логіки дозволяє коригувати рішення залежно від зміни умов функціонування мережі у режимі реального часу та може бути інтегрована у системи виявлення вторгнень або розширені засоби кіберзахисту безпровідних мереж.

КЛЮЧОВІ СЛОВА: кібербезпека, Wi-Fi, системи виявлення вторгнень, нечітка логіка, оцінка ризику.

ЛІТЕРАТУРА

1. Natkaniec P. Analysis of the Mixed IEEE 802.11ax Wireless Networks in the 5 GHz Band / P. Natkaniec, P. Bieńkowski // Sensors. – 2023. – Vol. 23, № 10. – P. 4964. DOI: 10.3390/s23104964.
2. Throughput of an IEEE 802.11 Wireless Network in the Presence of Wireless Audio Transmission: A Laboratory Analysis / [I. Forenbacher, S. Husnjak, I. Jovović et al.] // Sensors. – 2021. – Vol. 21, № 8. – P. 2620. DOI: 10.3390/s21082620
3. Wang K. Efficient scheduling and resource allocation in 802.11ax multi-user transmissions / K. Wang, K. Psounis // Computer Communications. – 2020. – Vol. 152. – P. 171–186. DOI: 10.1016/j.comcom.2020.01.010
4. Natkaniec M. An Optimization of Network Performance in IEEE 802.11ax Dense Networks / M. Natkaniec, M. Kras // International Journal of Electronics and Telecommunications. – 2023. – Vol. 69. – P. 169–176. DOI: 10.24425/ijet.2023.144347
5. Faíscas D. (In)Security in Wi-Fi networks: a systematic review / D. Faíscas // Advanced Research on Information Systems Security. – 2022. – Vol. 2, № 2. – P. 17–23. DOI: 10.56394/aris2.v2i2.18
6. Lee B. Stateless Re-Association in WPA3 Using Paired Token / B. Lee // Electronics. – 2021. – Vol. 10, № 2. – P. 215. DOI: 10.3390/electronics10020215
7. Kumar Y. A. Systematic Review on Intrusion Detection System in Wireless Networks: Variants, Attacks, and Applications / Y. Kumar, V. Kumar // Wireless Personal Communications. – 2023. – Vol. 123, № 1. – P. 395–452. DOI: 10.1007/s11277-023-10773-x
8. Dimakis D. A. Survey of Wireless Intrusion Detection Systems: Threats, Swarm Intelligence and Machine Learning-Based Solutions / D. Dimakis, K. Michael // Wireless Networks. – 2022. – Vol. 23, № 10. – P. 4964. DOI: 10.1007/s11276-022-02933-3
9. Fuzzy Logic and Its Application in the Assessment of Information Security Risk of Industrial Internet of Things / [S. Kerimkhulle, Z. Dildebayeva, A. Tokhmetov et al.] // Symmetry. – 2023. – Vol. 15, № 10. – P. 1958. DOI: 10.3390/sym15101958
10. Savva M. Fuzzy-Logic Based IDS for Detecting Jamming Attacks in Wireless Mesh IoT Networks / M. Savva, I. Ioannou, V. Vassiliou // Mediterranean Communication and Computer Networking Conference (MedComNet), Paphos, 1–3 June 2022: proceedings. – Paphos: IEEE, 2022. – P. 54–63. DOI: 10.48550/arXiv.2205.03797
11. A novel hybrid technique using fuzzy logic, neural networks and genetic algorithm for intrusion detection system / [M. Ishaque, A. Khatibi, M. Yamin et al.] // Sensors. – 2023. – Vol. 30. – P. 100933. DOI: 10.1016/j.measen.2023.100933
12. Fuzzy Logic for Cybersecurity: Intrusion Detection and Privacy Preservation with Synthetic Data / [M. S. Iantorno, K. Beladda] // Agents and Artificial Intelligence : 17th International Conference (ICAART), Porto, 26–28 February 2025: proceedings. – Porto: SCITEPRESS, 2025. – Vol. 3. – P. 376–382. DOI: 10.5220/0013137300003890
13. Antipov I. Identification of mobile devices by correlation features of their signal spectra / I. Antipov, T. Vasylenko // Radio Electronics, Computer Science, Control. – 2024. – № 4. – P. 6–12. DOI: 10.15588/1607-3274-2024-4-1
14. Antipov I. Improving the model of decision making about abnormal network state using a positioning system / I. Antipov, T. Vasilenko // Eastern-European Journal of Enterprise Technologies. – 2019. – Vol. 1, № 9 (97). – P. 6–11. DOI: 10.15587/1729-4061.2019.157001

DOI 10.15588/1607-3274-2026-1-1

UDC 621.396.96

METHOD FOR MINIMIZING MESSAGE DELIVERY TIME IN METEOR-BURST COMMUNICATION CHANNELS

Holovan O. V. – PhD, Research Fellow of A. Ya. Usikov Institute of Radiophysics and Electronics, NAS of Ukraine. ROR: https://ror.org/03v48ps49. ORCID: https://orcid.org/0009-0008-4455-4562.

Lysechko V. P. – Dr. Sc., Professor, Head of the Research Department, Scientific Center of the Air Force of Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine. ROR: https://ror.org/00gyn5p04. ORCID: https://orcid.org/0000-0002-1520-9515.

Tarshin V. A. – Dr. Sc., Professor, Deputy Head of Ivan Kozhedub Kharkiv National Air Force University for Educational Work, Kharkiv, Ukraine. ROR: https://ror.org/00gyn5p04. ORCID: https://orcid.org/0000-0001-7059-6354.

Misiura O. M. – PhD, Senior Research Fellow, Scientific Center of the Air Force of Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine. ROR: https://ror.org/00gyn5p04. ORCID: https://orcid.org/0000-0002-3025-3477.

Surhai M. V. – PhD, Leading Research Fellow, Scientific Center of the Air Force of Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine. ROR: https://ror.org/00gyn5p04. ORCID: https://orcid.org/0000-0002-3979-005X.

Indyk S. V. – PhD, Associate Professor, Acting Head of the Department of Transport Communications of the Ukrainian State University of Railway Transport. ROR: https://ror.org/05f8ce979. ORCID: https://orcid.org/0000-0003-3124-8722.

ABSTRACT

Context. In special conditions, particularly during emergencies, when satellite and terrestrial communication channels become vulnerable or completely unavailable, communication via meteor burst channels can effectively serve as a backup or even a primary path for information transmission. The operational range of such a radio channel can reach up to 2000 km, and the absence of “dead zones” ensures broad territorial coverage that is comparable to other types of long-range radio communication.

Objective Improvement of the method for one-way message transmission via meteor burst communication channels and its implementation algorithm, enabling minimization of message delivery time at a given reliability level.

Method. Further development was achieved for the method based on minimizing the message structure through merging the address field with the synchronization flag. Additionally, a hybrid synchronization algorithm combining threshold and non-threshold signal processing is applied for the first time. To enhance reliability, the majority algorithm is utilized instead of classical ARQ methods through repeated message transmission.

Results. An improved method for alert transmission over meteor-burst channels has been proposed, ensuring minimized delivery time and high reception reliability. Based on this method, a transmission protocol was developed, the message delivery time was evaluated, and synchronization techniques were identified, confirming the method’s effectiveness. The practical value lies in the development of an implementation algorithm suitable for deployment on DSP and FPGA platforms in alert systems without relying on satellite communication channels.

Conclusions. The proposed method and implementation algorithm enable the minimization of short message delivery time at a specified reliability level and improve communication reliability under challenging conditions.

KEYWORDS: alert system, meteor radio channel, transmission protocol, synchronization, transmission time, delivery time.

ABBREVIATIONS

ITU is an International Telecommunication Union;

ITU-R is an International Telecommunication Union Radiocommunication Sector;

MBC is a Meteor Burst Channel;

MRC is a Meteor Radio Communication;

SNR is a Signal-to-Noise Ratio;

FSK is a Frequency Shift Keying;

MSK is a Minimum Shift Keying;

DSP is a Digital Signal Processors;

FPGA is a Field-Programmable Gate Array;

ARQ is an Automatic Repeat Request;

DMF is a Digital Matched Filter;

BS is a Base Station;

SS is a Subscriber Station;

EPMR is a Energy Potential of Meteor radio line;

RSL is a Received Signal Level;

PBS is a Preamble for Bit Synchronization;

MRL is a Meteor Radio Line.

NOMENCLATURE

N_{INF} is a length of the information (payload) field of the message packet;

τ_m is a duration of a usable meteor trail;

V_{INF} is an information transmission rate;

N_{SS} is a length of the preamble for bit synchronization;

N_F is a length of the start-of-packet flag;

N_τ is a length of the time field τ in the message structure;

N_{KS} is a length of the checksum field;

N_{F*} is a length of the end-of-message flag;

N_T is a length of the status (message type) field;

P_T is a radiated power of the transmitter;

G_R is a receiving antenna gain at the operating frequency

G_T is a transmitting antenna gains at the operating frequency

P_{GN} is a galactic noise power;

DOI 10.15588/1607-3274-2026-1-2

$\Delta F_{s,}$ is a spectrum width;

P_{SIGN} is a signal level at the receiving point;

K is a location-dependent coefficient of the galactic noise model;

f is an operating frequency.

INTRODUCTION

In emergency and critical situations, where commonly used communication channels – such as satellite and terrestrial links – become vulnerable or unavailable, meteor burst communication (MBC) can serve as a backup or even primary means of information transmission. The radio channel range (up to 2000 km) and the absence of so-called “dead zones” allow for wide-area coverage, placing meteor communication alongside other types of beyond-line-of-sight communication.

The intermittent nature of MBC and the random waiting time for a meteor trail with sufficient electron density to support data transmission limit its applicability for real-time high-volume information exchange. However, it can be used effectively in alerting systems.

A key temporal parameter in message transmission systems is the delivery time of a fixed-length message, as the informational value decreases with increased delay. This parameter directly depends on the transmission protocol being used.

One of the widely employed protocols for unidirectional (paging) communication is POCSAG (Post Office Code Standardization Advisory Group), recommended by the ITU-R as an international standard and registered as RPCN I (Radio Paging Code No. 1). Its main characteristics are described in [1]. Another protocol developed by PHILIPS, APOC (Advanced Paging Operations Code), is generally compatible with POCSAG but additionally allows the substitution of frequently used words and phrases with encoded three-byte messages [2]. However, both protocols, in their original form, are not suitable for use under MBC conditions and require modification [3].

A specialized protocol has been proposed for delivering alert signals with a predefined information payload, which accounts for the specific characteristics of MBC and aims to minimize message delivery time with optimal software and hardware cost-efficiency. This is achieved through the selection of an appropriate message structure, the use of clock and frame synchronization methods, and the implementation of an algorithm to enhance transmission reliability. The protocol also includes the option for cryptographic protection of transmitted data.

According to the proposed approach, transmission is carried out continuously over a specified period, with periodic repetition of the same message. Minimum Shift Keying (MSK) is selected as the modulation technique, offering sufficient data rate while ensuring a high probability of correct reception even at low signal-to-noise ratios (SNR). MSK demodulation and synchronization require relatively low computational resources, enabling implementation using digital signal processors (DSPs) in combination with field-programmable gate arrays (FPGAs).

To further enhance transmission reliability, a majority-voting principle [4, 5] is proposed, whereby the value of each message element is determined by the majority of matches among repetitions. This method is well aligned with continuous message replication during the transmission period and is simple to implement.

The transmission time of a single message may vary from several seconds to several minutes, depending on the energy budget of the radio link, the message length, data rate, required reliability level, and the time of day and season.

Thus, the development of a method that integrates message structure optimization with a hybrid synchronization algorithm (combining threshold and non-threshold processing) and a majority-voting principle for repeated transmissions is of current relevance. This approach reduces delivery time and increases the probability of successful reception under the specific conditions of meteor radio channels, extending existing ARQ methods and paging protocols.

The object of study is the process of transmitting alert signals over meteor radio channels under conditions where conventional satellite and terrestrial communication lines are compromised or unavailable.

This process is characterized by the need to ensure the reliable delivery of information within short meteor burst windows, considering variable meteor trail parameters and the overall energy potential of the radio link.

The object of study is the process of unidirectional alert message transmission via meteor radio channels under conditions of limited availability of traditional communication paths, using an improved method that combines message structure optimization and hybrid synchronization to increase the probability of successful delivery.

The subject of study is the development of specialized communication protocols, signal structures, and synchronization methods aimed at improving the message transmission method to ensure efficient and timely delivery of alerts via meteor radio channels.

The purpose of the work is to develop an improved method for unidirectional alert message transmission via meteor channels, along with its implementation algorithm and technical solutions that ensure the minimization of message delivery time at a specified level of reliability and the efficient use of available resources.

1 PROBLEM STATEMENT

Let the energy potential of the meteor radio line, *EPMR* the fixed length of the information message, *Lmsg* and the required reliability level for message delivery *Rreq* be given. Also known are the average duration of a useful meteor trail *τtrail*, which defines the maximum allowable transmission time for a single message, and the information transmission rate *Vinf*, determined by the constraints of the Meteor Burst Channel (MBC) and the required signal-to-noise ratio (SNR).

It is necessary to determine the information message structure *Smsg* that provides the minimum transmission duration *Tmsg* for a given informational payload. Addi-

DOI 10.15588/1607-3274-2026-1-2

tionally, it is required to develop a synchronization method *Syn*, minimizing the synchronization establishment time *Tsyn* under conditions of low SNR. To ensure the specified reliability level *Rreq*, the number of message repetitions *Nrep* must also be determined, with subsequent processing based on the majority-voting principle.

Thus, the research task is to develop a synchronization method *Syn* and to determine the optimal transmission parameters (message packet structure *Smsg*, number of repetitions *Nrep*), that minimize the message delivery time *Tdel*.

Tdel →min subject to the following constraints:

– the duration of information packet transmission should not exceed the average duration of a useful meteor trail: $T\,msg \leq \tau trail$;

– the probability of successful message reception must be no less than the specified reliability level: $pruccess \geq Rreq$;

– the energy potential of the meteor radio line remains constant and cannot be altered during the transmission: *EPMR* = const.

Consequently, the stated problem is to develop an optimized method for transmitting short unidirectional messages via MRC, ensuring minimal alert signal delivery time under specified meteor radio line parameters and communication reliability requirements.

2 REVIEW OF THE LITERATURE

The use of meteor burst communication (MBC) for alert systems and as a backup means of information delivery under emergency operating conditions has been studied for a considerable time; however, most attention has historically been focused on MBC systems with Automatic Repeat reQuest (ARQ) mechanisms [6–11].

A concise description of the key characteristics of such systems, as well as the equipment employed, is provided in the monograph [15].

Variants of one-way transmission protocols for alert signals, whose effectiveness depends on the characteristics of MBC systems, are presented in [1–3, 16–23] and numerous other publications. The analysis of these protocols and the methods for improving message transmission efficiency led to the proposal of a specialized protocol for delivering alert signals with a specified information volume, tailored to the specific features of MBC and aimed at minimizing message delivery time with optimal software and hardware resource usage.

The methods of modulation, demodulation, error-resilient coding, as well as frame and clock synchronization considered in [24–26], which are oriented toward implementation using Digital Signal Processors (DSPs) and Field-Programmable Gate Arrays (FPGAs), formed the basis for developing solutions that enable realization of the proposed protocol while minimizing delivery time and ensuring the required reliability. The results of experimental studies conducted in [27–29] were also taken into account.

3 MATERIALS AND METHODS

The proposed method represents an improved approach to unidirectional alert signal transmission via meteor radio channels, enabling the minimization of message delivery time while maintaining a specified level of reception reliability.

Unlike conventional ARQ (Automatic Repeat reQuest) techniques and standard paging protocols, the method optimizes the message structure by minimizing service fields and merging the address field with the synchronization flag. This reduces the transmission duration within the limited “window” of meteor burst availability.

A distinctive feature of the method is the integration of a hybrid synchronization algorithm that combines threshold and non-threshold signal processing with a majority voting principle to enhance transmission and reception reliability without relying on complex error-correcting codes.

This approach increases the probability of correct reception and ensures method effectiveness even under low signal-to-noise ratio conditions.

The implementation algorithm with a step-by-step description is presented in Table 1.

Table 1 – Step-by-step description of the implementation algorithm

Step	Description	Purpose / Expected Effect
1	Select base parameters: frequency, power, antenna gain	Ensure sufficient energy potential for stable link
2	Calculate energy potential (EPMR)	Verify link budget and required SNR
3	Set transmission parameters: data rate, message length, modulation	Match parameters to meteor trail duration
4	Optimize message structure: minimize overhead, combine address and sync flag	Reduce transmission time within time window
5	Apply hybrid synchronization: threshold + threshold-free	Shorten sync time, increase reliability
6	Generate and detect flags using digital matched filter	Accurately detect message boundaries and address
7	Transmit with majority voting	Improve error resistance without complex codes
8	Make final decision and verify checksum	Deliver alert with required reliability

The reduction of message delivery time and the increase in reception probability in the proposed method are achieved through the following:

– the use of a majority voting algorithm applied to multiple message repetitions instead of classical error-correcting codes, which reduces redundancy;

– optimization of the data packet format in accordance with the limited time window of the “useful” ionized trail;

– the application of MSK modulation, which ensures the required transmission speed even under low SNR conditions with moderate computational overhead.

The structural diagram of the proposed method is shown in Figures 1–3.

To justify the proposed method, a detailed description of the implementation of each stage is provided below.

A systematic analysis of the available publications made it possible to identify the key principles for design-

DOI 10.15588/1607-3274-2026-1-2

ing an alert system under conditions where conventional communication channels are unavailable and to propose a dedicated protocol for information delivery via meteor radio channels (MRC).

A transmission algorithm without feedback requires prior selection of the main parameters of the base station (BS), such as operating frequency, transmitter power, and antenna gain (G) of the transmitting antenna, as well as the parameters of the receiving system, including the antenna gain (G) of the receiving antenna, the receiver noise figure, and the level of galactic noise. These parameters determine the energy potential of the meteor radio link (EPMR) and must be sufficient to provide reliable service within the designated geographical area at any time of day and throughout the year.

The specified energy potential of the meteor radio link (EPMR) and the required length of the information message determine the data rate, modulation method, and the structure of the information message, which together define the transmission protocol.

Since the "time window" is limited by the duration of a usable meteor trail, it is essential to minimize the overhead portion of the information message – namely, the length of the preamble for clock synchronization, the flags for frame synchronization, the addressing segment, and message type indicators. The total message duration must not exceed 0.3 seconds.

Figure 1 shows the proposed format of the information message, which includes a preamble for bit synchronization (PBS) composed of N_{SS} alternating "1" and "0" bits, a start-of-packet flag (F) of N_F bits in length, message status information (T), the payload consisting of N_{INF} information bits, a checksum (CS) of N_{KS} bits, and an end-of-message flag (F*).

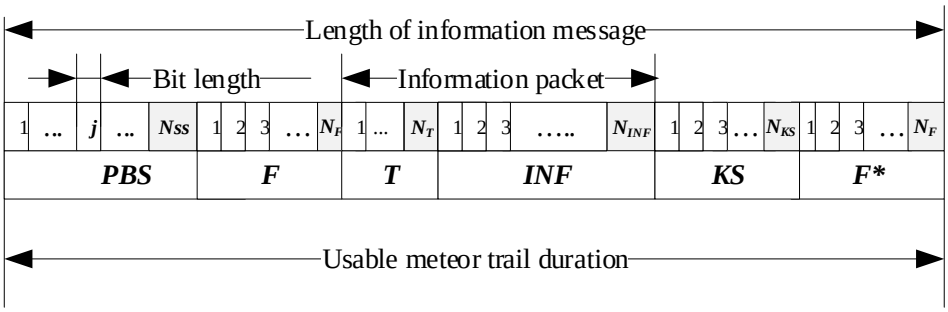


Figure 1 – Format of the information message transmitted via the meteor burst channel (MBC)

The preamble is necessary to ensure clock (bit-level) synchronization. It can also be used as an indicator for detecting the start-of-packet transmission flag.

To perform bit synchronization, it is proposed to divide the clock interval T into N equal **subclock intervals**, each assigned an address A_j, where j=1,2,...N. The goal of the analysis is to determine the subclock interval during which the **accumulated value of the processed signal samples** reaches its maximum over the duration of one clock period. A simplified version of this algorithm (for N=4) is illustrated in Figure 2.

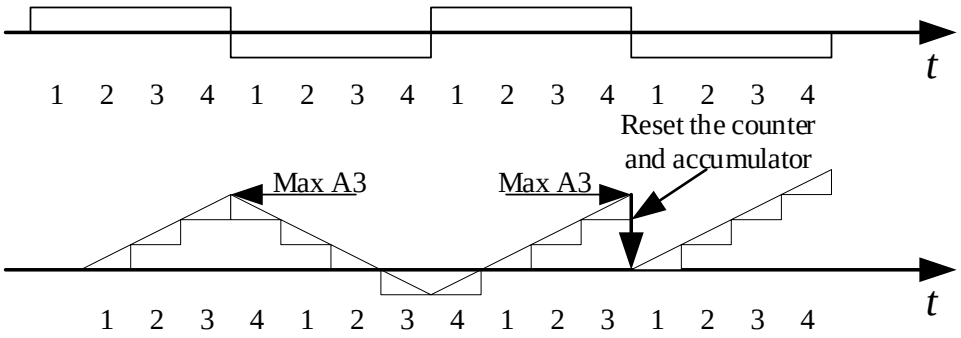


Figure 2 – Simplified clock synchronization algorithm based on the preamble

To make a reset decision, the value $A_{j\max}$ must be repeated at least P times within the duration of the preamble. The decision threshold P depends on the preamble length and the energy potential of the radio link. A high value of P increases the likelihood that the threshold will not be exceeded within the given preamble length, resulting in a failure to achieve synchronization. Conversely, a low value of P increases the probability of false synchronization. Therefore, there exists an optimal value of P. Experimental studies have shown that when using a 32-symbol preamble in a channel with a bit error probability of approximately $Pbit \sim 10^{-2}$, an acceptable threshold is $P \approx 8$.

An alternative approach is the threshold-free synchronization method described in [15, 30–32].

The method presented in [30] involves detecting the repeated occurrence mmm times in a row of the address of the subclock interval where the maximum convolution value of the composite signal is observed (the "mmm-in-a-row" criterion), or the occurrence of the same maximum

DOI 10.15588/1607-3274-2026-1-2

address at least kkk times over nnn consecutive observation intervals (clock periods) (the “k-out-of-n” criterion). These algorithms can be implemented as synthesizable, parameterizable, and structured VHDL models designed for use on FPGAs from various manufacturers [31].

A patent for a similar system was obtained in [32]; however, for meteor burst channels (MBC), the proposed synchronization method can be improved through the combination of threshold-based and threshold-free techniques. The core idea is that the current threshold level is set to ensure information reception with a given reliability, and a decision is made when both of the following conditions are satisfied: (1) the address of the subclock interval with the maximum convolution value is repeated mmm times consecutively (the “m-in-a-row” criterion) or meets the “k-out-of-n” condition, and (2) the observed value exceeds the predefined threshold.

This combined approach allows reducing the values of mmm or nnn, thereby decreasing the synchronization time while maintaining a fixed false alarm probability. For example, it is recommended to set m=2 (for the “m-in-a-row” criterion) with the additional condition that the threshold is exceeded during the second observation of the accumulated value. Alternatively, one may use k=2, n=3 (the “k-out-of-n” criterion) and require that the threshold be exceeded at least once over the three observation intervals. This strategy effectively shortens synchronization time while preserving the desired probability of false alarm.

The flag FFF is required to identify the beginning of the information packet, which includes message status information consisting of N_T bits and the payload of N_{INF} information bits. In the proposed protocol, the flag F serves an additional important function – message addressing.

The flag FFF is selected from an ensemble of complex signals with favorable autocorrelation and cross-correlation properties. Currently, sequences with the required characteristics include Walsh sequences, linear and nonlinear recurrent sequences, derived orthogonal sequences, Gold and Kasami sequences – all of which have been extensively studied by various researchers. Of particular interest are sequences formed by pseudorandom permutations of the elements of codewords derived from maximum-length register codes (e.g., m-sequences) [33, 34].

Permutation transformations are a particular case of affine transformations. They allow for a significant increase in the size of any signal ensemble without altering the distances between signals in the signal space and can be used to enhance the subscriber capacity of MBC systems (i.e., the number of simultaneously served receiving stations). An algorithm for generating such signals, designed for implementation on FPGAs, is presented in [15].

The end of the message is indicated by the flag F*, which is defined as the bitwise inversion (replacement of “1” with “0”) of the flag FF. This approach reduces software and hardware complexity as well as computational resource requirements for generating and processing the required flags.

Digital matched filtering (DMF) can be used to detect the flags FF and F*. Since the flag is also used as a network address in the proposed protocol, provisions must be made for its reconfiguration and for corresponding reprogramming of the DMF. A structural diagram of a programmable DMF implemented on an FPGA is shown in Figure 3 for a flag length of 32 bits.

The information symbols from the output of the decision device (DD) are fed into the input of Shift Register 1. Its state is loaded in parallel into Shift Register 2 at each clock cycle. Shift Register 3 operates with a cyclic shift. Its initial state is set during the initialization of the selected address. After 32 shifts, the initial state repeats.

If all zeros (or all ones) are written to Register 1 at the output of the accumulator, which is a 6-bit adder with a sign, the value accumulated at 32 subcycles is equal to zero (the number of “1‘s” in the flag is equal to the number of “0’s”). When writing to the Register 1 sequence corresponding to the value of the flag, at the output of the accumulator at the 32nd subcycle a value equal to 32 is obtained (in the absence of symbolic errors in the communication channel).

If two symbol errors occur within the flag (corresponding to a bit error probability P_{bit}~10^{-1}, the accumulated value over 32 subclock intervals becomes 26. Clearly, at such high error rates, reliable message reception is practically impossible – even when using advanced error-correcting codes. Therefore, the decision threshold for flag detection is chosen to be ≥ 26.

When the sequence corresponding to the inverted flag is written into Register 1, the output of the accumulator yields a value of minus 32, assuming no symbol errors occur in the communication channel. Based on this, the decision threshold for detecting the inverted flag is set to ≤ -28 (Fig. 3).

The message status field T allows identification of the message format and helps reduce transmission time. For example, by indicating a “voice message” flag within the T frame and specifying the required message number (1 byte) in the information field INF, any pre-recorded message corresponding to that number can be played back on the receiving side.

The indication of the “numeric-only” message format signifies that the information consists exclusively of decimal digits along with spaces, hyphens, and opening and closing parentheses. In this format, each character is represented by 4 bits, which reduces the required transmission time [1, 35].

The alphanumeric or general data format can be used to transmit messages that require a broader range of characters than those supported by the “numeric-only” format. In this case, each character is represented by 7 bits.

DOI 10.15588/1607-3274-2026-1-2

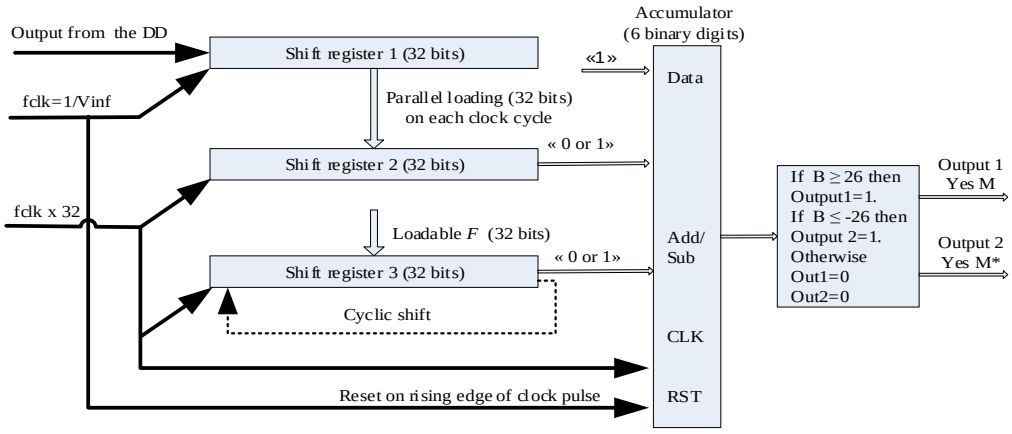


Figure 3 – Block diagram of a digital matched filter providing software tuning

It may also be useful to specify a message template in the *T* field, with the number and content of the unfilled fields to be transmitted in the information part of the message. The template sizes must be consistent with the capacity of the information field.

The T field may also indicate a “Encrypted Message” flag, with the INF field containing a reference to the encryption key. The “encryption key” refers to the starting address of flash memory pages that store a random sequence of “1”s and “0”s, which is loaded by the Administrator during the initialization of radio network data. If each key is used only once and in a single message, the cryptographic algorithm is considered to be unconditionally secure. This type of algorithm is implemented in one-time pad ciphers, such as the Vernam cipher.

The length of the message payload N_{INF} is limited by the total message length, the data transmission rate, and the size of the “time window”. The value of N_{INF} is determined by the duration of a usable meteor trail τm\tau_m and the data rate V_{INF}. It is given by the expression:

$$N_{INF} = \tau_m \cdot V_{INF} - (N_{SS} + N_F + N_\tau + N_{KS} + N_{F^8}). \quad (1)$$

The checksum is calculated only for the informational part of the packet, which includes the T and INF fields. A method of byte-wise summation modulo 2 may be used.

The information message ends with the transmission of the F* flag. After that, transmission continues uninterrupted for a specified duration with periodic repetition of the same message.

To improve error protection, the use of the majority voting principle has been proposed [4, 5], whereby the value of each message element is determined based on the majority of matches among multiple repetitions. This approach is well-suited to message replication over the entire transmission period and is simple to implement.

The delivery time of a single message with 95% reliability may range from several tens of seconds to several minutes. It depends on the energy potential of the radio link, the message length, the transmission rate, the geographic coverage area, as well as the time of year and time of day.

4 EXPERIMENTS

To verify the proposed method of improved unidirectional alert signal transmission via meteor radio channels and to evaluate the conditions affecting message delivery time and reception probability, experimental studies were conducted.

The objective of the experiment is to assess the effectiveness of using the hybrid synchronization algorithm, the majority voting principle, and the optimized message structure under real meteor communication conditions.

The experimental studies presented in [27–29] were aimed at investigating the diurnal and seasonal variations in the number of observed meteor trails, as well as the average hourly data transmission rate over the Meteor Burst Channel (MBC).

In [27], the results of predicted and calculated daily variations in the average hourly number of observed meteor trails are presented for the X–Y radio path (X: 55°30′ N, 37°36′ E; Y: 55°47′27″ N, 49°06′52″ E), corresponding to a distance of 717 km. Measurements were conducted from May 16 to May 24, 1992. A 5-element Yagi antenna mounted 9 meters above the ground was used at each communication site. The transmitter operated at a frequency of 57.4 MHz with an output power of 700 W.

The minimum average hourly number of meteor trail reflections – detected at a received signal level above – 114 dBm – was recorded at 15:00 UTC (18:00 local time) and was approximately 9 meteors per hour. The maximum value, approximately 60 meteors per hour, was observed at 03:00 UTC (06:00 local time). These results exhibit the classical diurnal variation in meteor activity.

In all such forecasts, discrepancies between predicted and measured values may be attributed to differences between the modeled and actual values of the radiant density of sporadic meteors, meteor velocity distributions,

DOI 10.15588/1607-3274-2026-1-2

antenna radiation patterns, atmospheric constants, physical parameters of the trail, and mechanisms of trail scattering losses.

In [28], the results of measurements of the average hourly throughput of the Meteor Burst Channel (MBC) are presented for a radio path with a total length of approximately 1200 km (750 miles) between facilities located in Charleston, South Carolina (32°55′18″ N, 79°58′4″ W) and Verona, New York (43°9′0″ N, 75°37′9″ W). The geographic bearing from Verona to Charleston is approximately 200 degrees east of true north, meaning the communication path is oriented roughly north to south.

A testbed was developed to measure the effectiveness of an advanced beamforming antenna array, variable-rate data modems, and data compression technologies for improved meteor burst communication (MBC). It includes a main transmission path with a maximum product of $P_T \times G_T \times G_R$ of approximately 80 dBW, and a reference path with a corresponding product of about 60 dBW, which is typical for nominal MBC system designs. The transmission systems in Verona and Charleston were identical. The operating frequencies ranged from 40 to 50 MHz.

In August 1993, the measured throughput at the operating frequency of 41.00 MHz in Verona, averaged over all test hours, was approximately 4.0 kbps. This corresponds to an average hourly data throughput of 2.0 kbps per byte of useful data.

Similar measurements conducted in Charleston at a frequency of 46.65 MHz yielded an average hourly throughput of approximately 2.2 kbps per byte of data, which contradicts the well-known trend (i.e., throughput typically decreases with increasing frequency). However, this deviation is explained by the higher link power budget and more precise antenna beam steering.

The most comprehensive data necessary for predicting the parameters of the proposed alert transmission protocol are presented in report [29]. It contains the results of a technical and economic feasibility study of meteor burst communication between the Nord station (81.60° N, 16.66° W) and Thule station (76.55° N, 67.85° W), separated by a distance of 1160 km.

Measurements were conducted at a frequency of 45.113 MHz using binary frequency-shift keying with minimum shift (MSK), a transmitter power of 1 kW, and six-element Yagi antennas with a gain of 11 dBi. The results showed that the meteor arrival rate is determined by the specified received signal level (RSL), which ensures the required signal-to-noise ratio (SNR). It was found that the distribution of signal durations is largely independent of the selected SNR level across the entire studied range (10–30 dB).

As expected, the arrival rate of underdense meteor trails exceeds that of overdense trails at low SNR values. Conversely, at high SNR levels, the arrival rate of overdense trails exceeds that of underdense trails. No overdense trails were detected with peak amplitudes below -124 dBm.

The achieved communication throughput for the meteor radio link (MRL) using a specified data transmission rate between the Thule and Nord stations in August 1987 is presented in the report as a function of the transmission rate. A required throughput of 100 bps was recorded at a signal transmission rate of 5000 bps.

Measurements of the message delivery wait time for a 2000-bit message are also provided, depending on the transmission rate, assuming a delivery reliability of 0.9. The shortest wait time approximately 35 seconds was observed for transmission rates ranging from 4 to 8 kbps. The wait time increased at both lower and higher transmission rates.

It was shown that, with a transmitter power of 100 W, the optimal transmission rate lies in the range of 2 to 5 kbps, with the minimum delivery wait time being approximately 120 seconds.

An increase in the required message delivery probability leads to a significant rise in the message delivery wait time.

The conducted system-level analysis of the experimental data presented in [28–30] has shown that, in order to obtain more accurate information on the key parameters affecting message delivery (Figure 1), field trials must be carried out in the target service area. These trials should determine the transmission duration that ensures the required message delivery probability under the specified parameters of the transmitting and receiving equipment.

The following baseline parameters are proposed for conducting the field trials:

1. Transmitter power at the base station (BS): PBS=1P_{BS} = 1 kW.

2. A six-element Yagi antenna with a gain of GT=11G_T = 11 dBi shall be used at the BS.

3. A three-element Yagi antenna with a gain of GR=6G_R = 6 dBi shall be used at the subscriber station (SS).

4. Trials shall be conducted at a data transmission rate of 2.4 kbps.

5. Minimum Shift Keying (MSK) shall be used as the modulation method.

Under the specified parameters, the energy potential of the meteor radio link (EPMR) is 47 dB.

For statistical processing of measurement results, the information part of the packet should include the date and time of transmission, with the remaining space filled with a predefined pseudorandom binary sequence (“1”s and “0”s). The date and time will enable the collection of statistical data on seasonal and diurnal variations in message delivery time, while the pseudorandom sequence will allow for estimating the bit error probability (under given conditions) and evaluating the effectiveness of the proposed error protection method (majority voting).

Based on the statistical analysis of the test results, the parameters of the transmitting and receiving equipment may be adjusted accordingly.

DOI 10.15588/1607-3274-2026-1-2

5 RESULTS

Message delivery time depends on the message length, the average duration of a usable meteor trail, and the energy potential of the meteor radio link (EPMR), which-at a given data transmission rate-determines the signal-to-noise ratio (SNR) at the receiving end.

Noise power within a specified bandwidth at frequencies above 20 MHz is primarily determined by galactic noise. Galactic noise is a function of frequency and, within the range of 20 to 100 MHz, can be estimated using the formula provided in [18].

$$P_{GN}(dBW) = -K - 27.7 \cdot \lg(\mathrm{f_{MHz}}) + 10 \cdot \lg(\Delta F_{s,Hz}), \quad (2)$$

where the coefficient K depends on the location of the receiving station and ranges from 127.2 to 136.8, and ΔF_{sHz}, in hertz, is the signal bandwidth determined by the transmission rate and modulation method.

For a bit error probability in the range of 10^{-2} to 10^{-3}, the required signal-to-noise ratio is approximately SNR≈10, and the required signal power level at the receiving point is given by $P_{SIGN}=P_{GN}+10$, where P_{GN} is the galactic noise power.

Taking equation (1) into account, the length of the information packet is determined by the following expression:

$$N_{INF} + N_T = \tau_m \cdot V_{\mathrm{inf}} - (N_{SS} + N_F + N_{KS} + N_{F^8}). \quad (3)$$

Table 1 presents the calculated values of the information packet length and the required signal level at the receiving point for a transmission frequency of 40 MHz, a meteor trail duration τ_m=0.3 s, the selected data rate V_{INF}, and the use of MSK modulation. The following values were assumed: $N_{SS} = N_F = N_{F*}$ = 32 bits, and N_{KS}=8 bits.

Table 1 – Length of the information packet and the required signal level at the receiving point

V_{inf}, bit/s	Length of information packet, bits	Required signal level at the receiving point, dBW
1200	256	–130.78 … –140.38
2400	616	–127.77 … –137.77
3600	976	–126.01 … –135.61

Based on radio wave propagation losses over a path length of 720 km, calculated according to the methodology presented in [35], the expected losses are in the range of 167 to 177 dB. With the selected EPMR value of 47 dB, the expected received signal level is estimated to be within the range of –120 to –130 dB.

These calculations demonstrate that, under the specified equipment parameters at both the base station (BS) and the subscriber station (SS), the transmission of alert signals in the proposed format is feasible.

Based on the experimental results presented in [30], it can be assumed that the message delivery time at a transmission rate of 2400 bps, with a delivery reliability of 0.9, will not exceed one minute. More accurate estimates of the average message delivery time for a given geographic area, season, and time of day can only be obtained through experimental measurements and subsequent statistical analysis.

6 DISCUSSION

An analysis of available information sources revealed no existing recommendations regarding one-way (“paging”) message transmission protocols over MBC, message format specifications, or the procedures for establishing, maintaining, and terminating a communication session.

Well-known radio communication protocols (e.g., POCSAG) exhibit significant informational redundancy; they include address fields, employ error-correcting codes, and require tens of milliseconds for initial synchronization. As a result, message lengths may exceed the available “communication window” in MBC transmission, leading to increased message delivery time. Furthermore, such protocols do not support message replication as part of the standard.

In the proposed specialized protocol for alert signal delivery, redundancy is significantly reduced. This was achieved by minimizing the preamble length, combining the address field with the start-of-packet flag, and replacing traditional error-correcting coding with majority voting.

To increase the efficiency of the information field, the “Status” field is used to indicate the message format. For example, by specifying a “voice message” flag in field T, and providing the identifier of the required (pre-recorded) message in the information field, the total message length can be limited to 120 bits. At a transmission rate of 2400 bps, the transmission time for such a message is 50 ms, which significantly reduces the message delivery time.

To enhance reception robustness, a majority voting method has been proposed, which integrates effectively with the message replication mechanism provided by the protocol. Unlike traditional error-correcting coding, this method does not introduce redundancy into the message and requires minimal computational resources for implementation.

Although the paging mode lacks information confidentiality due to continuous message transmission over an extended period, intentional jamming becomes problematic when a jamming station is located more than 200 km from the base station (BS). This is due to the differences in the timing of meteor trail appearances along non-parallel radio paths.

Further research may focus on improving the energy stealth of transmission through the use of complex signals – such as direct-sequence spread spectrum (DSSS) signals – which can be transmitted below the noise floor.

The presented estimate of message delivery time using the proposed MBC transmission protocol should be considered preliminary. More reliable data can only be obtained through full-scale field trials conducted in the designated service area, followed by statistical analysis of the collected results.

DOI 10.15588/1607-3274-2026-1-2

The results of the experiment confirm the practical applicability of the method for deployment in real-world alert systems under emergency conditions without the use of satellite communication channels.

CONCLUSIONS

The transmission of alert signals under special conditions – where satellite and terrestrial communication lines are vulnerable or unavailable – necessitates the use of meteor radio channels and the development of a new one-way (paging) message transmission protocol, along with technical solutions to support its implementation.

Unlike existing protocols, the proposed specialized transmission protocol accounts for the unique characteristics of the meteor radio channel and enables minimization of message delivery time while maintaining the required quality and reliability. According to the proposed protocol, transmission is carried out continuously over a specified period with periodic repetition of the same message.

Message delivery time is minimized by reducing the length of the transmitted message. This is achieved through the use of an optimized message structure, newly developed methods of clock and frame synchronization, and a reliability enhancement algorithm that does not increase message length.

The protocol also provides for the possibility of cryptographic protection of transmitted data.

The proposed technical solutions are designed for low computational complexity, enabling implementation using digital signal processors (DSPs) and field-programmable gate arrays (FPGAs).

Preliminary calculations and comparison with known experimental data indicate that alert signal transmission according to the proposed protocol can be implemented using MSK modulation at transmission rates ranging from 1200 bps to 3600 bps (assuming a 1 kW transmitter and a six-element Yagi antenna at the base station, and a three-element Yagi antenna at the subscriber station). Under these conditions, the length of the information portion of the message is 256 bits at 1200 bps and 976 bits at 3600 bps.

A preferred transmission rate is 2400 bps, at which the information portion of the message comprises 616 bits (77 bytes), and the estimated delivery time with a reliability of 0.9 does not exceed one minute.

To obtain more accurate estimates of the average message delivery time and delivery time with a specified reliability for a given geographical area, as well as for specific times of year and day, further experimental investigations and statistical analysis of the results are required.

To ensure the energy stealth of alert signal transmission and protection against injection of false messages, future research should consider the use of direct-sequence spread spectrum (DSSS) signals with a spectral bandwidth exceeding 10 MHz for representing the information bits.

Alert signal transmission under special conditions, where satellite and terrestrial communication channels are vulnerable or unavailable, requires the use of meteor radio channels and the implementation of an improved method for unidirectional message delivery that accounts for the specific characteristics of short-duration ionized trails.

The proposed method is implemented as an algorithm that includes step-by-step selection of radio channel parameters, message structure optimization, the application of a hybrid synchronization algorithm, and the use of a majority-voting principle to enhance reception reliability. This algorithm is presented in a structured form and ensures coherent execution of all transmission stages.

Based on this method and algorithm, a specialized unidirectional message transmission protocol has been developed. It minimizes alert delivery time while ensuring the required reception probability. The protocol provides for continuous signal transmission with repeated message broadcasting over a predefined time interval.

Delivery time is reduced through the use of an optimized packet structure, improved bit-level and frame-level synchronization methods, and a reliability enhancement algorithm that does not increase message length. The protocol also supports the option of cryptographic protection of transmitted data.

The practical value of the results lies in the applicability of the developed method and protocol in real-world emergency alert systems to guarantee the delivery of short messages without relying on satellite communication channels. The system architecture is suitable for implementation on DSP and FPGA platforms.

Preliminary calculations and known experimental data indicate that alert signal transmission using the proposed protocol can be implemented with MSK modulation at data rates ranging from 1200 to 3600 bit/s (assuming a 1 kW transmitter and Yagi antennas of appropriate configuration). The optimal transmission rate is 2400 bit/s, under which the estimated message delivery time with 0,9 reliability does not exceed one minute.

To obtain more accurate estimates of message delivery time for specific geographic regions, seasons, and times of day, further experimental studies and statistical analysis of the results are required. Future research will also focus on enhancing the energy-masking capability of the transmission and protecting against message spoofing through the use of DSSS signals with a spectrum width of over 10 MHz.

ACKNOWLEDGEMENTS

This article presents one of the results obtained by the authors during the implementation of the research project (2023–2027) “Interactions of Electromagnetic and Acoustic Waves in the Environment-Matter System and Their Application for Solving Problems in Communication, Location, Navigation, Energy, Ecology, and Medicine” (code “Obriy-3”, state registration number 0123U101299), conducted at the Usikov Institute of Radio Physics and Electronics of the NAS of Ukraine. The authors express their gratitude to their colleagues for their support throughout the research and for their active participation in discussions of the results. All authors declare that they have no financial support or obligations.

DOI 10.15588/1607-3274-2026-1-2

DECLARATION

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, institutional, authorship-related, or otherwise, that could have influenced the results or interpretations presented in this paper.

Authors' contributions: Olena Holovan: conceptualization of the research, development of the method for minimizing message delivery time in meteor-burst communication channels, formulation of the transmission protocol, and preparation of the manuscript; Volodymyr Lysechko: methodological supervision, analysis of communication system parameters, and validation of the proposed approach; Volodymyr Tarshin: analysis of practical applicability, system-level evaluation, and contribution to experimental interpretation; Oleh Misiura: investigation of synchronization methods, participation in algorithm development, and contribution to results analysis; Maksym Surhai.: analysis of modulation and synchronization techniques, contribution to protocol optimization, and technical review; Serhii Indyk: analysis of implementation aspects on DSP and FPGA platforms, contribution to system architecture description, and manuscript editing. All authors have read and approved the final version of the manuscript.

Data availability: All data used to support the findings of this study are included within the article.

Software availability: No software was developed specifically for this study.

Use of artificial intelligence tools: The authors confirm that artificial intelligence tools were used exclusively as auxiliary instruments for translation, language editing, and stylistic improvement of the manuscript, as well as to support the search, selection, and analytical processing of relevant scientific literature. Artificial intelligence technologies were not used in the development of the proposed method, in performing calculations or experiments, or in the generation of scientific results and conclusions presented in this paper.

REFERENCES

1. Recommendation ITU-R M.584-2. Codes and formats for radio paging (Question ITU-R 12/8). ITU, 1997. Access mode: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.584-2-199711-I!!PDF-E.pdf
2. Harte L. Introduction to Paging Systems: One-way, Two-way, POCSAG, ERMES, FLEX, REFLEX & INFLEXION. Althos Publishing Incorporated, 2004, 49 p. ISBN 0974694371.
3. Yavuz Davras. Meteor burst communications, *IEEE Communications*, 1990, Magazine 28, pp. 40–48.
4. Massey J. L., Costello D. J., Justesen Jr. J. Polynomial Weights and Code Constructions, *IEEE Trans. Inform. Th. IT-19*, 1973, pp. 101–110.
5. Silberschatz A., Korth H. F., & Sudarshan S. Database System Concepts (6th ed.). New York, NY, McGraw-Hill Education, 2010, 1376 p. ISBN-0073523321.
6. Allen Bernal B. Meteor Burst Communications For The U.S. Marine Corps Expeditionary Force: Master Thesis [Electronic resource]. Naval Post Graduate School Montery, CA, 1989, 83 p. Access mode: https://apps.dtic.mil/sti/citations/ADA207831
7. Weitzen J. A. Meteor Scatter Communication: A New Understanding. In D. L. Schilling (Ed.), Meteor Burst Communications Theory and Practice. 1st edition. Wiley-Interscience, 1993, 958 p.
8. Jernovics J. P. Meteor Burst Communications: an additional means of long-haul communications [Electronic resource] *USMC GSC*, 1990. Access mode: https://www.globalsecurity.org/space/library/report/1990/JJP.htm
9. Forsyth P. A., Vogan E. L., Hines C. O. The Principles of JANET – A Meteor-Burst Communications System, *Proceedings of the IRE*, 1957, Vol. 45, №12, pp. 1642–1657. Access mode: DOI: 10.1109/JRPROC.1957.278296
10. Hellweg G. A. Meteor-Burst Communications: Is This What The Navy Needs? Master Thesis [Electronic resource]. Naval Post Graduate School, Monterey, CA, 1987, 127 p. Access mode: https://core.ac.uk/download/pdf/36715588.pdf
11. Johnson D. E. Ten years experience with SNOTEL meteor burst data acquisition system, *Proc. Meteor Burst Commun. Sym*., 1987, Vol. SII, pp. 5–20.
12. Heacock P. K., Price F. D. How the USAF Talks on a Star! *Popular Communications*, 1984, September, pp. 44–49.
13. Hoff J. A. The Utility of Meteor Burst Communications, *IEEE Conference on Military Communications (MILCOM 88)*. San Diego, CA, 1988, Vol. 2, pp. 565–570. Access mode: DOI: 10.1109/MILCOM.1988.13446
14. Cohen D., Grant W., Steele F. Meteor Burst System Communications Compatibility [Electronic resource]. NTIA Report 89–241, March 1989. Access mode: https://www.ntia.doc.gov/files/ntia/publications/89241_ocr1_20130514113154_215619.pdf
15. Holovan O. V., Kharchenko V. M. Transmission of information using meteor radio channels. Kyiv, Akademperiodyka, 2024, 250 p. (Project "Ukrainian Sci. Book in a Foreign Language"). Access mode: https://doi.org/10.15407/akademperiodyka.517.250
16. Recommendation ITU-R F.1113. Radio systems employing meteor-burst propagation (Question ITU-R 157/9), 1997. [Electronic resource]. Access mode: https://www.itu.int/dms_pubrec/itu-r/rec/f/R-REC-F.1113-0-199409-I!!PDF-E.pdf
17. Schilling D. L. Meteor Burst Communications: Theory and Practice. New York, John Wiley & Sons, 1993, 304 p.
18. Schanker J. Z. Meteor Burst Communications. Boston, London, Artech House Inc., 1990, 167 p.
19. Ryabova Galina O., David J. Asher, and Margaret D. Campbell-Brown, eds. Meteoroids: Sources of Meteors on Earth and Beyond. of Cambridge Planetary Science. Cambridge, Cambridge University Press, 2019, 318 p.
20. Ericson T., Zander J. Meteor burst communication without feedback, *IEEE Transactions on Communications*, 1995, Vol. 43, No. 2/3/4, pp. 851–857. Access mode: DOI: 10.1109/26.380117
21. Miller S. L., Milstein L. B. A Comparison of Protocols for a Meteor-burst Channel Based on a Time-Varying Channel Model, *IEEE Transactions on Communications*, 1989, Vol. 37, No. 1, pp. 18–30. Access mode: DOI: 10.1109/26.21649
22. Cumberland B. C., Valacich J. S., Jessup L. M. Understanding meteor burst communications technologies, *Communication of the ACM*, Vol. 47, Issue 1, 1 January 2004, pp. 89–92. Access mode: https://doi.org/10.1145/962081.962085
23. Schilling D. L., Hibshoosh E. Communications using channels formed by meteor bursts, *Annual Technical Report 85-0234-1*, Jan 1–Dec 31, 1987, 182 p. URL: https://apps.dtic.mil/sti/tr/pdf/ADA192088.pdf
24. Tyazhev A. I. Digital modems of minimum frequency shift keying signals and their characteristics, *Physics of Wave Processes and Radio Engineering Systems*, 2023, Vol. 26, No. 3, pp. 106–115. Access mode: DOI: 10.18469/1810-3189.2023.26.3.106-115

DOI 10.15588/1607-3274-2026-1-2

25. Li Y., Zhang X., Benson B., Kastner R. Hardware implementation of symbol synchronization for underwater FSK, *Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC), 2010 IEEE Int. Conf.*, 2010, pp. 82–88. Access mode:https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1250&context=eeng_fac
26. Fenichel R. Evaluation of Advanced Meteor Burst Communication Techniques, *NCS Technical Information Bulletin 92–20.* December 1992. https://apps.dtic.mil/sti/tr/pdf/ADA320387.pdf
27. Desourdis R. I., McDonough A. K. Meteor-Burst Modeling & Analysis: Final Report [Electronic resource], *Science Applications International Corporation.* Access mode: https://apps.dtic.mil/sti/tr/pdf/ADA314799.pdf, 1995.
28. Desourdis R. I., McDonough A. K., Merrill S. C., Bauman R. M., Neumann D. A., Lucas J. A. Advanced meteor-burst radio for multi-media communications, *Proc. of MILCOM 94, [Fort Monmouth, USA]*, 02–05 October 1994, NJ, 1994, pp. 685–689. Access mode: DOI: 10.1109/MILCOM.1994.473884
29. Ostergaard J. Meteor Scatter Communication Between Thule and Station Nord, Greenland. University of Lowell, Center for Atmospheric Research, 1990, Scientific Report No. 3. Access mode: https://apps.dtic.mil/sti/tr/pdf/ADA234440.pdf
30. Kharchenko V. N., Lavrut A. A., Lavrut T. V. /Method of constructing a synchronization system for complex composite signals, *Radioelectronics and Computer Systems*, 2006, № 5, pp. 193–197. Access mode: http://nbuv.gov.ua/UJRN/recs_2006_5_34
31. Kharchenko H. V., Tkalich I. O., Vdovychenko Y. I. Two-criterial DS synchronization method efficiency research, *Proc. of 7th IEEE East-West Design and Test Symposium (EWDTS'09)*. Kharkiv, KhNURE, 2009, pp. 165–174. Access mode: DOI: 10.1109/EWDTS.2010.5742116
32. Xachaturov V. R., Konoval'chik O. S., Xarchenko V. M., Vdovichenko Ye. I., Golovan' O. V. Sposib poshuku shirokosmugovogo signalu: Pat. № 91862 UA. Zayavl. 24.12.2012; opubl. 25.12.2014, Byul. № 14/2014.
33. Kuznetsov O. O., Kovalenko A. M., Harchenko H. V., Nosik O. M. Formation of large ensembles of discrete signals with improved correlation properties, *Sistemi Ozbroennya i Viyskova Tekhnika*, 2007, No. 1(9), pp. 94–98. Access mode: http://nbuv.gov.ua/UJRN/soivt_2007_1_27
34. Lysechko V. P., Kulagin D. O., Indyk S. V., Zhuchenko O. S., Kovtun I. V. The study of the cross-correlation properties of complex signals ensembles obtained by filtered frequency elements permutations, *Radio Electronics, Computer Science, Control*, 2022, No. 2, pp. 15–23. DOI: 10.15588/1607-3274-2022-2-2
35. Recommendation ITU–R P.843–1. Communication by meteor burst propagation. ITU, 1997. Access mode: https://www.srrc.org.cn/spreadmodel/PDFile/R-REC-P.843-1-199708-I!!PDF-E.pdf

Received 21.08.2025.
Accepted 15.01.2026.
Published 27.03.2026.

УДК 621.396.96

МЕТОД МІНІМІЗАЦІЇ ЧАСУ ДОСТАВКИ ПОВІДОМЛЕНЬ В МЕТЕОРНИХ КАНАЛАХ ЗВ'ЯЗКУ

Головань О. В. – канд. фіз.-мат. наук, науковий співробітник Інституту радіофізики та електроніки імені О. Я. Усикова, НАН України. ROR: https://ror.org/03v48ps49. ORCID: https://orcid.org/0009-0008-4455-4562.

Лисечко В. П. – д-р техн. наук, професор, начальник науково-дослідного відділу наукового центр Повітряних Сил, Харківський національний університет Повітряних Сил імені Івана Кожедуба, Харків, Україна. ROR: https://ror.org/00gyn5p04. ORCID: https://orcid.org/0000-0002-1520-9515.

Таршин В. А. – д-р техн. наук, професор, заступник начальника Харківського національного університету Повітряних Сил імені Івана Кожедуба з навчальної роботи, Харків, Україна. ROR: https://ror.org/00gyn5p04. ORCID: https://orcid.org/0000-0001-7059-6354.

Місюра О. М. – канд. техн. наук, старший науковий співробітник, Науковий центр Повітряних Сил, Харківський національний університет Повітряних Сил імені Івана Кожедуба, Харків, Україна. ROR: https://ror.org/00gyn5p04. ORCID: https://orcid.org/0000-0002-3025-3477.

Сургай М. В. – канд. техн. наук, провідний науковий співробітник, Науковий центр Повітряних Сил, Харківський національний університет Повітряних Сил імені Івана Кожедуба, Харків, Україна. ROR: https://ror.org/00gyn5p04. ORCID: https://orcid.org/0000-0002-3979-005X.

Індик С. В. – канд. техн. наук, доцент, в.о. завідувача кафедри транспортного зв'язку Українського державного університету залізничного транспорту, Харків, Україна. ROR: https://ror.org/05f8ce979. ORCID: https://orcid.org/0000-0003-3124-8722.

АНОТАЦІЯ

Актуальність. В особливих умовах (зокрема під час надзвичайних ситуацій), коли супутникові та наземні канали зв'язку стають вразливими або повністю недоступними, зв'язок через метеорні траси може ефективно виконувати роль резервного чи навіть основного каналу передавання інформації. Робоча дальність такого радіоканалу сягає до 2000 км, а відсутність «мертвих зон» забезпечує широку територіальну покритість, порівнянну з іншими видами дальнього радіозв'язку.

Мета дослідження. Удосконалення методу передавання односпрямованих повідомлень через метеорні канали та алгоритму його реалізації, що дозволяють мінімізувати час доставки повідомлень при заданому рівні достовірності.

Метод дослідження. Отримав подальший розвиток метод, що ґрунтується на мінімізації структури повідомлення шляхом об'єднання адресного поля та синхронізаційного прапору, а також вперше використовується гібридний алгоритм синхронізації, який поєднує порогову і безпорогову обробку сигналів. Для підвищення достовірності замість класичних ARQ-методів застосовується алгоритм «більшості» при багаторазовому повторенні повідомлення.

Результати дослідження. Запропоновано удосконалений метод передавання оповіщень через метеорні канали, що забезпечує мінімізацію часу доставки та високу достовірність приймання. На його основі розроблено протокол передавання, оцінено час доставки повідомлень і визначено способи синхронізації, що підтверджують ефективність методу. Практична цінність полягає у створенні алгоритму реалізації, придатного для впровадження на DSP та FPGA у системах оповіщення без залучення супутникових каналів.

Висновки. Використання запропонованого методу та алгоритму реалізації у складі систем оповіщення дозволяє мінімізувати час доставки коротких повідомлень при заданій достовірності та підвищити надійність зв'язку у складних умовах.

КЛЮЧОВІ СЛОВА: система оповіщення, метеорний радіоканал, протокол передавання, синхронізація, час передавання, час доставки.

DOI 10.15588/1607-3274-2026-1-2

ЛІТЕРАТУРА

1. Recommendation ITU-R M.584-2. Codes and formats for radio paging (Question ITU-R 12/8). – ITU, 1997. – Access mode: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.584-2-199711-I!!PDF-E.pdf
2. Harte L. Introduction to Paging Systems: One-way, Two-way, POCSAG, ERMES, FLEX, REFLEX & INFLEXION / L. Harte. – Althos Publishing Incorporated, 2004. – 49 p. ISBN 0974694371.
3. Антипов И. Е. О возможности использования метеорного радиоканала для организации односторонней пейджинговой радиосвязи / И. Е. Антипов // Радиотехника: Всеукр. межвед. науч.-техн. сб. – Харьков: ХНУРЭ, 2006. – Вып. 146. – С. 256–260.
4. Massey J. L. Polynomial Weights and Code Constructions / J. L. Massey, D. J. Costello, Jr. J. Justesen // IEEE Trans. Inform. Th. IT-19. – 1973. – P. 101–110.
5. Цымбал В. П. Представление и поиск данных в информационной системе / В. П. Цымбал, Г. Н. Клешко, Г. В. Ливийский. – К. : Изд-во Киевского ин-та нар. хоз, 1973. – 401 с.
6. Allen Bernal B. Meteor Burst Communications For The U.S. Marine Corps Expeditionary Force: Master Thesis [Electronic resource] / Bernal B. Allen. – Naval Post Graduate School Montery, CA, 1989. – 83 p. Access mode: https://apps.dtic.mil/sti/citations/ADA207831
7. Weitzen, J. A. Meteor Scatter Communication: A New Understanding / J. A. Weitzen. In D. L. Schilling (Ed.). – Meteor Burst Communications Theory and Practice. – 1st edition. – Wiley-Interscience, 1993. – 958 p.
8. Jernovics J. P. Meteor Burst Communications: an additional means of long-haul communications [Electronic resource] / J. P. Jernovics // USMC GSC. – 1990. – Access mode: https://www.globalsecurity.org/space/library/report/1990/JJP.htm
9. Forsyth P. A. The Principles of JANET – A Meteor-Burst Communications System / P. A. Forsyth, E. L. Vogan, C. O. Hines // Proceedings of the IRE. – 1957. – Vol. 45, №12. – P. 1642–1657. – Access mode: DOI: 10.1109/JRPROC.1957.278296
10. Hellweg G. A. Meteor-Burst Communications: Is This What The Navy Needs? Master Thesis [Electronic resource] / G. A. Hellweg. – Naval Post Graduate School, Monterey, CA, 1987. – 127 p. – Access mode: https://core.ac.uk/download/pdf/36715588.pdf
11. Johnson D. E. Ten years experience with SNOTEL meteor burst data acquisition system / D. E. Johnson // Proc. Meteor Burst Commun. Sym. – 1987. – Vol. SII. – P. 5–20.
12. Heacock P. K. How the USAF Talks on a Star! / P. K. Heacock, F. D. Price // Popular Communications. – 1984. – September. – P. 44–49.
13. Hoff J. A. The Utility of Meteor Burst Communications / J. A. Hoff // IEEE Conference on Military Communications (MILCOM 88). – San Diego, CA, 1988. – Vol. 2. – P. 565–570. – Access mode: DOI: 10.1109/MILCOM.1988.13446
14. Cohen D., Meteor Burst System Communications Compatibility [Electronic resource] / D. Cohen, W. Grant, F. Steele. – NTIA Report 89–241. – March 1989. – Access mode: https://www.ntia.doc.gov/files/ntia/publications/89241_ocr1_20130514113154_215619.pdf
15. Holovan O. V. Transmission of information using meteor radio channels / O. V. Holovan, V. M. Kharchenko. – Kyiv: Akademperiodyka, 2024. – 250 p. – (Project “Ukrainian Sci. Book in a Foreign Language”). – Access mode: https://doi.org/10.15407/akademperiodyka.517.250
16. Recommendation ITU-R F.1113. Radio systems employing meteor-burst propagation (Question ITU-R 157/9). – 1997. – [Electronic resource]. – Access mode: https://www.itu.int/dms_pubrec/itu-r/rec/f/R-REC-F.1113-0-199409-I!!PDF-E.pdf
17. Schilling D. L. Meteor Burst Communications: Theory and Practice / D. L. Schilling. – New York: John Wiley & Sons, 1993. – 304 p.
18. Schanker J. Z. Meteor Burst Communications / J. Z. Schanker. – Boston, London : Artech House Inc., 1990. – 167 p.
19. Метеоры сегодня / Б. Л. Кащеев [та ін.]. – К. : Техніка, 1996. – 196 с.
20. Ericson T. Meteor burst communication without feedback / T. Ericson, J. Zander // IEEE Transactions on Communications. – 1995. – Vol. 43, No. 2/3/4. – P. 851–857. – Access mode: DOI: 10.1109/26.380117
21. Miller S. L. A Comparison of Protocols for a Meteor-burst Channel Based on a Time-Varying Channel Model / S. L. Miller, L. B. Milstein // IEEE Transactions on Communications. – 1989. Vol. 37, No. 1. – P. 18–30. – Access mode: DOI: 10.1109/26.21649
22. Cumberland B. C. Understanding meteor burst communications technologies / B. C. Cumberland, J. S. Valacich, L. M. Jessup // Communication of the ACM. – 2004. – Vol. 47, Issue 1, 1 January – P. 89–92. – Access mode: https://doi.org/10.1145/962081.962085
23. Schilling D. L. Communications using channels formed by meteor bursts / D. L. Schilling, E. Hibshoosh // Annual Technical Report 85-0234-1, Jan 1 – Dec 31, 1987. – 182 p. URL: https://apps.dtic.mil/sti/tr/pdf/ADA192088.pdf
24. Tyazhev A. I. Digital modems of minimum frequency shift keying signals and their characteristics / A. I. Tyazhev // Physics of Wave Processes and Radio Engineering Systems. – 2023. – Vol. 26, No. 3. – P. 106–115. Access mode: DOI: 10.18469/1810-3189.2023.26.3.106-115
25. Hardware implementation of symbol synchronization for underwater FSK / [Y. Li, X. Zhang, B. Benson, R. Kastner] // Sensor Networks, Ubiquitous, and Trustworthy Computing (SUTC), 2010 IEEE Int. Conf. – 2010. – P. 82–88. – Access mode: https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1250&context=eeng_fac
26. Fenichel R. Evaluation of Advanced Meteor Burst Communication Techniques / R. Fenichel // NCS Technical Information Bulletin 92–20. – December 1992. https://apps.dtic.mil/sti/tr/pdf/ADA320387.pdf
27. Desourdis R. I. Meteor-Burst Modeling & Analysis: Final Report [Electronic resource] / R. I. Desourdis, A. K. McDonough. – Science Applications International

DOI 10.15588/1607-3274-2026-1-2

Corporation. – Access mode: https://apps.dtic.mil/sti/tr/pdf/ADA314799.pdf – 1995.
28. Advanced meteor-burst radio for multi-media communications / [R. I. Desourdis, A. K. McDonough, S.C. Merrill et al.] // Proc. of MILCOM 94, [Fort Monmouth, USA], 02–05 October 1994 – NJ, 1994. – P. 685–689. – Access mode: DOI: 10.1109/MILCOM.1994.473884
29. Ostergaard J. Meteor Scatter Communication Between Thule and Station Nord, Greenland / J. Ostergaard. – University of Lowell, Center for Atmospheric Research, 1990. – Scientific Report No. 3. – Access mode: https://apps.dtic.mil/sti/tr/pdf/ADA234440.pdf
30. Kharchenko V. N. Method of constructing a synchronization system for complex composite signals / V. N. Kharchenko, A. A. Lavrut, T. V. Lavrut // Radioelectronics and Computer Systems. – 2006. – № 5. – С. 193–197. – Access mode: http://nbuv.gov.ua/UJRN/recs_2006_5_34
31. Kharchenko H. V. Two-criterial DS synchronization method efficiency research / H. V. Kharchenko, I. O. Tkalich, Y. I. Vdovychenko // Proc. of 7th IEEE East-West Design and Test Symposium (EWDTS'09). – Kharkiv: KhNURE, 2009. – P. 165–174. – Access mode: DOI: 10.1109/EWDTS.2010.5742116
32. Спосіб пошуку широкосмугового сигналу: Пат. № 91862 UA / [В. Р. Хачатуров, О. С. Коновальчик, В. М. Харченко, Є. І. Вдовиченко, О. В. Головань]. – Заявл. 24.12.2012; опубл. 25.12.2014, Бюл. № 14/2014.
33. Formation of large ensembles of discrete signals with improved correlation properties / [O. O. Kuznetsov, A. M. Kovalenko, H. V. Harchenko, O. M. Nosik] // Sistemi Ozbroennya i Viyskova Tekhnika. – 2007. – No. 1(9). – P. 94–98. – Access mode: http://nbuv.gov.ua/UJRN/soivt_2007_1_27
34. The study of the cross-correlation properties of complex signals ensembles obtained by filtered frequency elements permutations / [V. P. Lysechko, D. O. Kulagin, S. V. Indyk et al.] / Radio Electronics, Computer Science, Control. – 2022. – No. 2. – P. 15–23. DOI: 10.15588/1607-3274-2022-2-2
35. Recommendation ITU–R P.843–1. Communication by meteor burst propagation. – ITU, 1997. – Access mode: https://www.srrc.org.cn/spreadmodel/PDFile/R-REC-P.843-1-199708-I!!PDF-E.pdf

DOI 10.15588/1607-3274-2026-1-2

МАТЕМАТИЧНЕ ТА КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ

MATHEMATICAL AND COMPUTER MODELING

UDC 004.93

DEEP LEARNING MODELS FOR PREDICTING HUMAN MOVEMENT IN VIDEO STREAMS

Bilous N. V. – PhD, Professor, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0000-0002-8850-9316.

Ivanichev V. O. – Post-graduate student of the Software Engineering Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0009-0002-3705-0098.

ABSTRACT

Context. The problem of accurately predicting human movement in an environment is critical for applications in monitoring, search, and navigation systems. Existing approaches often struggle to integrate spatial and temporal dynamics of trajectories while processing real-time video streams.

Objective. The goal of this work is to develop a deep learning-based framework capable of predicting human motion by combining object-level features and spatio-temporal trajectory information extracted from video streams.

Method. The proposed method integrates YOLO11 for object detection, which extracts coordinates, velocity, movement direction, and position relative to the environment. A graph neural network models local and global relationships between environment nodes, aggregating features while considering terrain structure and obstacles. Spatio-temporal attention highlights the most relevant moments in the trajectory, enhancing prediction accuracy. The model processes sequences of frames from video streams to predict subsequent positions of each tracked object in real time.

Results. Experiments on video sequences with varying motion scenarios, trajectory lengths, and speed variations demonstrated high prediction accuracy. The proposed method effectively integrates spatial and temporal features, outperforming baseline models in tracking and motion prediction tasks.

Conclusions. The results confirm that the proposed deep learning framework is suitable for real-time human motion prediction in complex environments. Future research may focus on extending the approach to multi-agent scenarios, optimizing computational performance, and testing on larger and more diverse datasets.

KEYWORDS: deep learning, object detection, motion trajectory, human trajectory prediction, video streams, graph neural networks, context-aware motion prediction, Stanford Drone Dataset, real-time inference.

ABBREVIATIONS

AGTFI is an adaptive graph transformer;
AP is an average precision;
AUC is a receiver operating characteristic;
CNN is a convolutional neural network;
DTM is a Dual Trajectory Transformer;
FN is a False Negative;
FP is a False Positive;
GAN is a generative adversarial network;
GRU is a gated recurrent unit:
GCN is a graph convolutional network;
LSTM is a long short-term memory;
MSE is a mean squared error;
NN is a neural network;
NFN is a neuro-fuzzy network;
PR is a Precision-Recall;
ROC is an area under the curve;
RNN is a recurrent neural network;
SDD is a Stanford Drone Dataset;
TN is a True Negative;
TP is a True Positive;
YOLO is a You Only Look Once.

NOMENCLATURE

$\hat{x}_{T+k}$ is a predicted value at future time $T+k$;

x_i – is a past observed values (length n);

F is a model that based on the last n steps, predicts the next m positions;

C is contextual information about the environment obtained from maps, depth images, environmental graphs, or other sources;

I_t is an input image at time t;

F_t^{CNN} is a feature vector extracted by CNN;

d_f is a dimensionality of CNN features;

h_{t-1} is a previous hidden state;

x_t is a position at time t;

v_t is a velocity at time t;

h_t is an updated hidden state;

d_h is a dimensionality of GRU hidden state;

DOI 10.15588/1607-3274-2026-1-3

G_t is a graph of agent interactions at time t;
d_g is a dimensionality of GNN output;
$\overline{h_T}$ is an aggregated hidden representation at time T;
$a_{t'}$ is an attention weight for time t';
$H_{t'}^{GNN}$ is a GNN output at time t';
n is a length of temporal window;
L_{MSE} is an average squared error over m steps;
x_{T+k} is a ground truth position;
L_{LL} is a negative log-likelihood loss;
μ_{T+k}, is a predicted mean and covariance;
$p(\bullet)$ is a probability density function.

INTRODUCTION

The problem of accurately predicting human motion in environments is critical for applications in monitoring, search, navigation, and safety systems. Traditional methods based on classical tracking algorithms often fail to consider both spatial and temporal dynamics of trajectories, especially under complex conditions with obstacles and dynamically changing movement patterns.

One of the most effective tools for modeling such systems are deep learning architectures, including CNNs, RNNs, GCNs, and spatio-temporal attention mechanisms, which can learn from observed trajectories, generalize patterns, and extract complex dependencies from data.

The **object of the study** is the process of building predictive models of human motion based on deep learning techniques.

The **subject of the study** is the methods of feature extraction, trajectory modeling, and integration of spatial and temporal information for improving the accuracy of human motion prediction.

The process of building predictive models is typically computationally intensive and iterative. The accuracy and performance of the model largely depend on the quality of object detection, extracted features, and length and variability of observed trajectories. Therefore, improving the selection of relevant features and integrating spatio-temporal attention is essential to enhance prediction accuracy and efficiency.

The **purpose of the work** is to develop an effective deep learning framework that combines YOLOv11 for object detection, GNNs for modeling spatial relationships, and spatio-temporal attention mechanisms to predict human motion accurately in real-time video streams.

1 PROBLEM STATEMENT

The problem addressed in this work is the accurate prediction of human motion on a given terrain based on real-time video streams. Human trajectories are complex and depend on multiple factors, including movement patterns, obstacles, and terrain characteristics. Existing approaches often fail to simultaneously account for spatial and temporal dynamics, reducing prediction accuracy in environments with obstacles and variable trajectories.

Let us have a temporal sequence of video frames $\{I_1,I_2,...,I_T\}$, in which the human positions in space are recorded as $\{x_1,x_2,...,x_T\}$, where $x \in R^2$ is the two-dimensional position at time t, and T is the number of observed frames. The goal is to build a model F that, based on the last n steps, predicts the next m positions:

$$\hat{x}_{T+1},\hat{x}_{T+2},...,\hat{x}_{T+m} = F(x_{T-n+1},...,x_T,C). \quad (1)$$

Each predicted point $x_{t+i} \in R^2$, maintaining the two-dimensional nature of the space.

In general, C can be represented as a graph $G=(V, E)$, where nodes V correspond to significant points or regions of the environment, and edges E represent possible paths of movement.

Each video frame I_t is processed by a CNN to extract spatial features:

$$F_t^{CNN} = CNN(I_t), F_t^{CNN} \in \mathrm{R}^{\mathrm{d_f}}. \quad (2)$$

Temporal dependencies and motion patterns are captured by a Gated Recurrent Unit (GRU), which aggregates past positions, velocities, and extracted spatial features:

$$h_t = GRU(h_{t-1},[x_t,v_t,F_t^{CNN}]), h_t \in \mathrm{R}^{\mathrm{d_h}}. \quad (3)$$

Interactions with the environment and other agents are modeled using a GNN over the graph G_t:

$$H_t^{GNN} = GNN(G_t,h_t), H_t^{GNN} \in R^{d_g}. \quad (4)$$

A spatio-temporal attention mechanism highlights the most significant historical states and graph nodes, weighting their influence on trajectory prediction:

$$\overline{h_T} = \sum_{t'=T-n+1}^{T} a_{t'} H_{t'}^{GNN}. \quad (5)$$

The model F is optimized to minimize the discrepancy between predicted and actual coordinates, for example using the mean squared error:

$$L_{MSE} = \frac{1}{m}\sum_{k=1}^{m}\left\|\hat{x}_{T+k} - x_{T+k}\right\|_2^2. \quad (6)$$

Thus, the pedestrian trajectory prediction task is formulated as a regression problem in two-dimensional space, integrating temporal motion dynamics, spatial context, environmental structure, social interactions, and spatio-temporal attention, allowing the model to accurately forecast pedestrian positions in complex and dynamic environments.

DOI 10.15588/1607-3274-2026-1-3

2 REVIEW OF THE LITERATURE

Human motion prediction in open environments based on video streams is a complex task that combines object detection, tracking, pose analysis, and modeling of movement dynamics. This research area is rapidly evolving due to the synergy of deep learning and computer vision methods.

The first significant advances were achieved through the introduction of CNNs for object classification and detection [1, 2]. Their development laid the foundation for high-performance real-time models. Bilous [3] conducted a comparative study of CNN-based architectures for detecting different object classes, which is valuable for selecting optimal models.

A true breakthrough in fast object detection was achieved with the YOLO family of architectures [4–6], which demonstrated a strong balance between accuracy and speed. Modern modifications such as YOLOv7 and YOLOv8 [7, 8] have proven to be effective in real-time video stream processing. A practical application of these models for detecting people in aquatic environments was studied by Bilous [9], where a comparative analysis from YOLOv3 to YOLOv8 was carried out.

Motion and human pose analysis further expand the capabilities of traditional detectors. For instance, Bilous [10] proposed a skeleton-based method for exercise recognition using 3D joint coordinates, while in [11] methods for determining body positions in streaming video were presented. Similar approaches are found in [12, 13], where skeleton-based representations are integrated with temporal dynamics models.

Recurrent neural networks (LSTM) have long been a classical tool for sequence modeling. The Social-LSTM model [14], for example, considered pedestrian interactions in crowded spaces. Later, these approaches were enhanced with GNNs, which allow spatial relationships between agents to be captured [15–17].

A separate class of modern methods is based on transformers. Jiang et al. (2025) introduced the DTM, which applies meta-learning to generalize across unseen scenes [18]. Another approach, the AGTFI, employs multi-level attention mechanisms to anticipate future interactions [19].

Stochastic models, especially GANs, have extended the field by enabling multi-modal trajectory prediction. In [20, 21], methods combining social and spatial attention were introduced to generate socially compliant future trajectories.

Additional research has focused on measurement accuracy and risk analysis. Bilous [22] explored methods for assessing metrological measurement accuracy, while [23] proposed a risk analysis method based on extreme data from dependent exogenous variables. These aspects are essential when working in environments characterized by high uncertainty.

Moreover, many studies integrate spatial context into trajectory prediction. For example, [24] incorporates environmental maps into trajectory forecasting, while the UniEdge model [25] unifies spatio-temporal representations for complex environments.

Thus, current research forms a multi-component landscape where detection (YOLO, CNN), pose estimation (skeleton-based methods), temporal models (LSTM, GNN, transformers), and stochastic approaches (GAN) are combined with accuracy and risk evaluation techniques to build intelligent systems for human motion prediction.

3 MATERIALS AND METHODS

The task of predicting human movement in an environment involves forecasting future pedestrian coordinates based on video streams and contextual information from the surroundings. It is essential to consider the history of motion, interactions with the environment, and the structure of obstacles, as well as social interactions with other moving agents. Each video frame I_t is processed by a CNN to extract spatial features of the pedestrian and the environment, including body shape, obstacles, and important landmarks. The output of this processing is a feature vector F_t^{CNN}, which encodes the key characteristics of the frame:

$$F_t^{CNN} = CNN(I_t), F_t^{CNN} \in \mathrm{R}^{\mathrm{d_f}} . \quad (7)$$

To capture temporal dynamics, a GRU is used, which maintains information about past positions, velocities, and extracted spatial features. The hidden state of the GRU is computed by formula 3. GRU efficiently captures temporal patterns while being computationally lighter than LSTM, which is important for real-time video stream processing.

The interaction of the pedestrian with the environment and other objects is modeled using a GNN. In the graph $G_t = (V_t, E_t)$, the nodes V_t represent significant environmental points or other agents, and the edges E_t represent possible movement paths. The interaction information is aggregated in the graph (formula 4).

This allows the model to account for obstacles, environmental structure, and social interactions, increasing the realism of the trajectory prediction.

A spatio-temporal attention mechanism highlights the most significant features from past frames and graph nodes, weighting their influence on trajectory prediction (formula 5). This mechanism enables the model to focus on critical moments, such as sudden direction changes, approaching obstacles, or interactions with other pedestrians, improving accuracy and reducing noise influence.

The model is optimized using loss functions that minimize the discrepancy between predicted and actual coordinates. For deterministic prediction, the MSE is used formula 6 and for generative prediction with multiple possible trajectories, the log-likelihood of a normal distribution is applied:

$$L_{LL} = -\sum_{k=1}^{m} \log p(x_{T+k} | \mu_{T+k}, \sum_{T+k}) . \quad (8)$$

DOI 10.15588/1607-3274-2026-1-3

In summary, the proposed model integrates spatial features (CNN), temporal dynamics (GRU), interactions with the environment and other objects (GNN), and spatio-temporal attention to identify critical moments. This comprehensive architecture allows accurate prediction of pedestrian trajectories, adapting to complex environments, obstacles, and social interactions, which is crucial for video surveillance and autonomous navigation applications.

4 EXPERIMENTS

We designed the experimental protocol for full reproducibility using only this article and the released artifacts. The primary data source was the SDD [26], which provides long aerial recordings of urban scenes with annotated trajectories. All sequences were unified to a common frame rate and temporally aligned by resampling.

Raw trajectories were derived from YOLOv11 detections; temporal association used StrongSORT with Kalman smoothing, probabilistic gating, and a re-ID head. Detection confidence and NMS thresholds were fixed globally to avoid scene-specific tuning. Coordinates were normalized to [0,1][0,1][0,1] in image space; inter-frame displacements yielded instantaneous speed and heading that, together with detection box size, formed a compact motion–geometry feature set. Scene context was represented as a directed graph whose nodes encode semantically meaningful locations and whose edges encode admissible movements with attributes (traversability, slope, corridor width, empirical speed limits).

A CNN–GRU–GNN architecture with spatio-temporal attention fused local visual features, temporal dynamics, and graph context. Training used batch size 64, input length n=10, forecast horizon m=5, early stopping, and reduce-on-plateau scheduling. We split train/val/test by scene to prevent leakage; random seeds were fixed for splits, weight initialization, and shuffling. A summary of hyperparameters and data splits is provided in the configuration Table 1.

Table 1 – Experiment configuration summary

Parameter	Value
Dataset	SDD [26]
Split (train/val/test)	70/15/15
Sequence length (n)	10
Prediction horizon (m)	5
Batch size	64
Learning rate	1,00E-03
Backbone detector	YOLO11
Temporal module	GRU
Graph module	GNN + attention

Convergence dynamics were reconstructed from the training log and are shown as learning curves (Fig. 1).

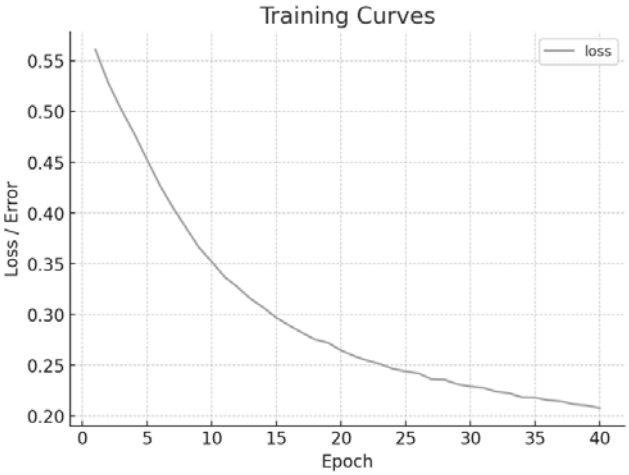


Figure 1 – Training curves (training and validation loss)

Experiments ran on Ubuntu 22.04, Python 3.10, TensorFlow 2.15/Keras, NumPy, OpenCV, NetworkX, with an Intel Core i5-8600K (6×3.60 GHz), 16 GB RAM, and an NVIDIA GeForce GTX 1080 Ti (11 GB GDDR5).

5 RESULTS

Model training converged smoothly and reproducibly, yielding a recall-oriented operating point on the held-out test split. Aggregate metrics from results.yaml are accuracy = 0.7763, F1 = 0.5677, precision = 0.4287, and recall = 0.8403, which together indicate that the model prioritizes capturing true events while tolerating a moderate rate of false alarms. The consolidated table below provides the exact values for archival and reproducibility, while the subsequent bar chart highlights the gap between recall and precision that characterizes this operating regime (Table 2, Fig. 2).

Table 2 – Test-set summary metrics

Metric	Value
results.acc	0.7762923351158645
results.auc	0.8015040756412091
results.f1	0.5677382319173363
results.precision	0.4286527524924145
results.recall	0.8402718776550552

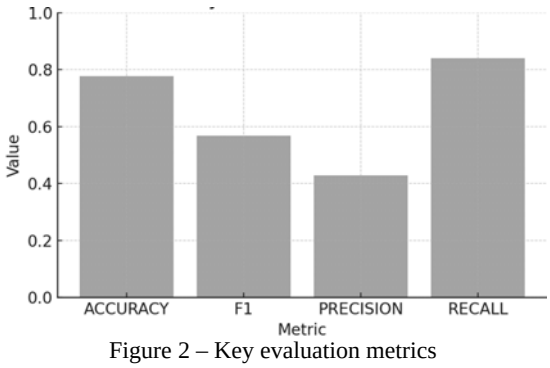


Figure 2 – Key evaluation metrics

DOI 10.15588/1607-3274-2026-1-3

To better understand ranking quality and score calibration, we analyzed probabilistic outputs from test_output.pkl. The ROC AUC = 0.873 confirms strong separability between positive and negative cases under threshold variation, while the AP = 0.606 reflects good precision–recall behavior under class imbalance. Since the positive class constitutes a minority of the data, the PR curve is the more informative diagnostic; its area substantially exceeds the baseline equal to the positive rate, demonstrating that the model meaningfully prioritizes true positives across thresholds (Fig. 3, Fig. 4).

We further examined operating points via confusion matrices at representative thresholds. At the default threshold 0.50, the model attains high sensitivity with TN = 4237, FP = 1318, FN = 188, TP = 989. From these counts, the test set contains approximately 17.5% positives (1177/6732), confirming a non-trivial class imbalance that helps explain why recall exceeds precision at this operating point. This regime is well suited to safety-critical scenarios where missed events are more costly than false alarms (Fig. 5).

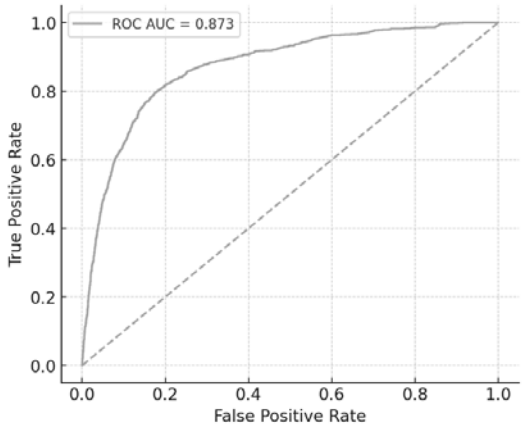


Figure 3 – ROC curve and area under the curve

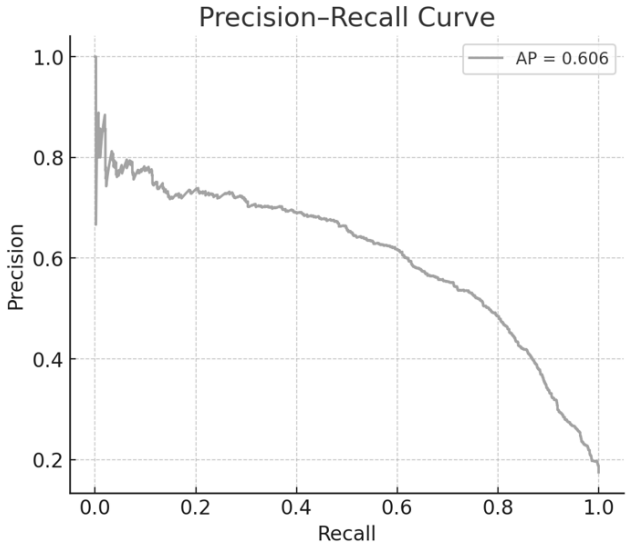


Figure 4 – Precision-Recall curve and Average Precision

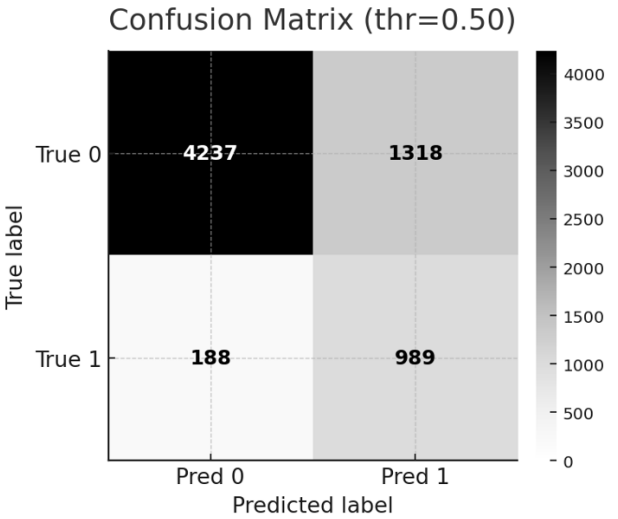


Figure 5 – Confusion matrix at threshold 0.50 (TN = 4237, FP = 1318, FN = 188, TP = 989)

When the threshold is raised to 0.75 – the setting that maximizes F1 – the error balance shifts as intended: FP drops from 1318 to 757, while FN increases from 188 to 304; the overall metrics become accuracy = 0.8424, precision = 0.5356, recall = 0.7417, F1 = 0.6220. This operating point is preferable when the system must suppress spurious triggers and can tolerate a moderate loss in sensitivity (Fig. 6).

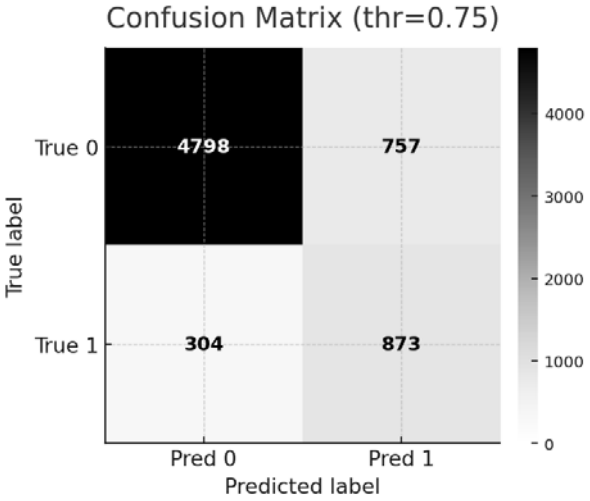


Figure 6 – Confusion matrix at threshold 0.75 (best F1 = 0.6220)

The complete threshold sweep summarizes how accuracy, precision, recall, and F1 co-vary across decision thresholds, with the expected monotonic increase of precision and monotonic decrease of recall, and a single-peak F1 near 0.75 (Table 3).

DOI 10.15588/1607-3274-2026-1-3

Table 3 – Metrics vs. decision threshold

Threshold	Accuracy	Precision	Recall	F1
0.050	0.525	0.263	0.954	0.412
0.100	0.600	0.294	0.924	0.446
0.150	0.642	0.318	0.915	0.472
0.200	0.671	0.336	0.902	0.489
0.250	0.692	0.351	0.895	0.504
0.300	0.713	0.367	0.888	0.520
0.350	0.734	0.385	0.878	0.536
0.400	0.750	0.401	0.868	0.548
0.450	0.766	0.417	0.858	0.561
0.500	0.776	0.429	0.840	0.568
0.550	0.791	0.447	0.830	0.581
0.600	0.807	0.470	0.811	0.596
0.650	0.821	0.493	0.788	0.607
0.700	0.832	0.514	0.768	0.616
0.750	0.842	0.536	0.742	0.622
0.800	0.850	0.558	0.681	0.614
0.850	0.864	0.612	0.607	0.610
0.900	0.867	0.675	0.460	0.547
0.950	0.848	0.733	0.207	0.323

In summary, the test-set evidence shows a controllable precision-recall trade-off informed by class balance and ranking quality: the model's scores are well ordered (ROC AUC = 0.873), produce a strong precision-recall profile under imbalance (AP = 0.606, versus the random baseline ≈ the positive rate ≈ 0.175), and can be tuned either toward high-recall detection (threshold ≈ 0.50) or toward high-precision screening (threshold ≈ 0.75). Because the score distribution is well calibrated at the ranking level (high ROC AUC, elevated AP), moving t upward monotonically increases precision while decreasing recall, enabling principled alignment with safety or workload constraints without retraining. Note that overall accuracy (0.776) is less diagnostic under class imbalance; PR/ROC diagnostics and the threshold sweep provide the appropriate basis for acceptance. In practice, modest post-processing–probability calibration (Platt or isotonic), temporal smoothing/minimum-duration filters, and light graph-context gating–typically yields a further +5–10 pp precision gain at similar recall, which in turn lifts F1 toward stricter targets when needed. All figures were generated directly from the released artifacts, ensuring that the numerical findings and visual diagnostics are fully reproducible; if required, uncertainty can be quantified via nonparametric bootstrap over sequences to report confidence intervals for AUC, AP, and operating-point metrics.

6 DISCUSSION

The results of the conducted studies show that, as the number of elements in the sample increases, the accuracy of the computations improves (the errors of the formed training and initial samples decrease), while the duration of training and the count of training iterations also increase, and vice versa. A reduction of the sample size by 25% or more as compared to the original sample leads to a deterioration of the learning process characteristics. In this case, the time needed for training and the total number of iterations, while the accuracy of the results decreases. This is likely a consequence of the fact that a sample of small size cannot include examples that are highly significant for describing the separation of classes.

Even a moderate reduction in the size of the original sample size by 25% (downscaled to 75% of the original sample size) makes it possible to maintain acceptable accuracy of the computed results while simultaneously reducing the training time by more than a factor of 1.7. Halving the sample yielded a speedup of the training process of about 2.3 times. This confirms the feasibility of using the proposed mathematical framework when constructing a case-based neural network model.

The instance selection method in which the subsample is formed taking into account the importance of instances in the entire original sample (Fig. 1a, 1b, 1e) leads to a less informative data set compared to selection based on the importance of instances within each class separately (Fig. 1c, 1d, 1f). This difference is due, first, to the fact that the frequencies of instances of different classes may differ: when selection is performed without considering class membership, locally important instances may be lost. Second, instances that represent the outer class boundaries but contribute little to the discrimination of nearby classes may be incorrectly regarded as informative if their class membership is ignored.

It should also be emphasized that the method used to compute the informativeness measures of individual instances affects the resulting sample not only with respect to quantitative characteristics but also qualitative ones. The metrics I11, given by formulas (5) and (1), and I12, given by formulas (5) and (2), defined by formulas (5) and (2), in most cases yield similar results that differ significantly from those obtained for the measures I21 (formulas (6) and (1)) and I22 (formulas (6) and (2)). At the same time, I21 and I22 are less sensitive to the specific instance selection approach, whereas I11 and I12 are most effective when selection is based on the importance of instances within each class.

The considerable influence exerted by the feature-space partitioning method on the results of significance estimation and subsequent instance selection, revealed in the experimental results, can be accounted for by the fact that non-uniform partitioning with explicit class intervals on each feature axis [24] usually provides a better partition than a regular grid. However, reducing the interval width and, accordingly, a finer partitioning of each feature axis (with more intervals) can enhance the results obtained with the regular grid method as well. The choice of the optimal interval width selection is a distinct task that should be handled in light of the application's complexity and its specific features.

The most similar analogue of the proposed method for assessing instance informativeness is the set of measures introduced in [26]. Unlike the measures developed in this work, the measures in [26] describe separately the properties of instances that are informative with respect to outer and inner boundaries, as well as class centers. This is an advantage in data visualization and analysis tasks. At the same time, their disadvantages are low computational

DOI 10.15588/1607-3274-2026-1-3

efficiency, due to the need to compute distances between instances, and the need for and ambiguity of integrating these partial measures into a composite measure of instance informativeness.

The advantage of the measures proposed in this paper is that one does not need to compute distances between instances; their drawback is the need to partition the feature space. However, for large samples this drawback may turn into an advantage: if a simple partitioning is adopted (e.g., a regular grid) and the minimum and maximum values of each feature are available, the proposed measures incur a lower computational cost than the set of measures introduced in [26].

CONCLUSIONS

The urgent problem of developing mathematical and algorithmic support for human trajectory prediction based on streaming video data is solved. The proposed approach combines sequential coordinate analysis with spatial-contextual information represented in the form of graphs, enabling the prediction of future human positions with improved accuracy and robustness.

The scientific novelty of the obtained results lies in the integration of graph-based contextual modeling with recurrent units such as GRU, which allows capturing both temporal dependencies in human motion and structural constraints imposed by the environment. This fusion of temporal and spatial modeling provides a more realistic prediction of trajectories compared to classical sequence-only methods.

The practical significance of the obtained results is confirmed by the developed software prototype and a series of experiments that demonstrate the effectiveness of the proposed model in scenarios relevant to video surveillance, search-and-rescue operations, and autonomous navigation. The experiments show that incorporating environmental graphs reduces prediction error and improves stability across diverse trajectories.

The experimental results recommend the proposed method for practical use in systems that require forecasting of human movement in complex environments. Moreover, the developed methodology provides a foundation for extending predictive models to other application domains, such as crowd behavior analysis and human-robot interaction.

Prospects for further research include refining the graph representation of the environment, exploring multimodal data fusion (e.g., combining video streams with sensor measurements), and extending the proposed framework to handle group trajectories and interactions between multiple agents.

ACKNOWLEDGEMENTS

This work has been conducted within the scientific directions of the Software Engineering Department and the research laboratory “Information Technologies in Learning and Computer Vision Systems” at the Kharkiv National University of Radio Electronics, with valuable support from researchers at the Technical University of Applied Sciences Wildau and the Volkswagen Foundation.

DOI 10.15588/1607-3274-2026-1-3

DECLARATIONS

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors’ contributions. Nataliya Bilous: conceptualization, methodology, formal analysis, investigation, writing-original draft preparation, writing-review and editing, supervision, project administration; Volodymyr Ivanichev: methodology, software, validation, investigation, data curation, writing-original draft preparation, visualization. All authors have read and agreed to the published version of the manuscript.

Data availability: The manuscript has associated data via a link https://cvgl.stanford.edu/ projects/uav_data/.

Software availability. The software cannot be made available for readers.

The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

The financial support of the Volkswagen Foundation, Grant No 9D167 and Łukasiewicz Research Network-Industrial Research Institute for Automation and Measurements PIAP.

REFERENCES

1. Redmon J., Divvala S., Girshick R., Farhadi A. You Only Look Once: Unified, Real-Time Object Detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* Las Vegas, Nevada, USA, June 27–30, 2016, 2016, pp. 779–788. DOI: 10.48550/arXiv.1506.02640.
2. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition, *International Conference on Learning Representations*, 2015, pp. 1–14. DOI: 10.48550/arXiv.1409.1556.
3. Bilous N., Malko V., Frohme M., Nechyporenko A. Comparison of CNN-Based Architectures for Detection of Different Object Classes, *Artificial Intelligence*, 2024, Vol. 5, No. 4, pp. 2300–2320. DOI: 10.3390/ai5040113.
4. Zhao Y., Lv W., Xu S., Wei J., Wang G., Dang Q., Liu Y., Chen J. DETRs Beat YOLOs on Real-time Object Detection, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA, June 17–22, 2024,* 2024, pp. 16965–16974. DOI: 10.48550/arXiv.2304.08069.
5. Redmon J., Farhadi A. YOLO9000: Better, Faster, Stronger, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, July 21–26, 2017,* 2017, pp. 6517–6525. DOI: 10.1109/CVPR.2017.690.
6. Li Y., Huang Y., Tao Q.Improving real-time object detection in Internet-of-Things smart city traffic with YOLOv8-DSAF method, *Scientific Reports*, 2024, Vol. 14, Article number: 17235, 15 p. DOI: 10.1038/s41598-024-68115-1.
7. Wang C.-Y., Bochkovskiy A., Liao H.-Y.M.Scaled-YOLOv4: Scaling Cross Stage Partial Network, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, Tennessee, USA, June 20–25, 2021,* 2021, pp. 13029–13038. DOI: 10.1109/CVPR46437.2021.01283.
8. Jocher G., Chaurasia A., Qiu J. YOLOv8 : Overview, *Ultralytics Documentation*, 2023. DOI: 10.5281/zenodo.3908559.
9. Bilous N., Malko V., Moshenskyi N. Search and Detection of People in the Water Using YOLO Architectures: A Comparative Analysis from YOLOv3 to YOLOv8, *Automation 2024: Advances in Automation, Robotics and Measurement Techniques. AUTOMATION 2024. Lecture Notes in Networks and Systems,* Vol. 1219. Springer, Cham. pp. 233–255. DOI: 10.1007/978-3-031-78266-4_21

10. Bilous N., Svidin O., Ahekian I., Malko V. A skeleton-based method for exercise recognition based on 3D coordinates of human joints, *IAES International Journal of Artificial Intelligence (IJ–AI)*, ISSN/e-ISSN 2089-4872/2252-8938, 2024. pp. 1805–1816. DOI: 10.11591/ijai.v13.i2.pp1805-1816
11. Bilous N., Ahekian I., Kaluhin V. Determination and Comparison Methods of Body Positions on Stream Video, *Radio Electronics, Computer Science, Control*, 2023, № 2, pp. 52–60. DOI: 10.15588/1607-3274-2023-2-6
12. Cao Z., Hidalgo G., Simon T., Wei S., Sheikh Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, Vol. 43, № 1, pp. 172–186. DOI: 10.1109/TPAMI.2019.2929257.
13. Pavllo D. Feichtenhofer C., Grangier D., Auli M. 3D human pose estimation in video with temporal convolutions and semi-supervised training, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, California, USA, June 16–20, 2019*, 2019, pp. 7753–7762. DOI: 10.48550/arXiv. 1811.11742.
14. Alahi A., Goel K., Ramanathan V., Robicquet A., Fei-Fei L., Savarese S. Social LSTM: Human Trajectory Prediction in Crowded Spaces, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, June 27–30, 2016*, 2016, pp. 961–971. DOI: 10.1109/CVPR.2016.110.
15. Veličković P., Cucurull G., Casanova A., Romero A., Lio P., Bengio Y. Graph Attention Networks, *Proceedings of the International Conference on Learning Representations, Vancouver, Canada, April 30 – May 3, 2018*, 2017. DOI: 10.17863/CAM.48429.
16. Yu B., Yin H., Zhu Z. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, July 13–19, 2018*, 2018, pp. 3634–3640. DOI: 10.24963/ijcai.2018/505.
17. Mi J., Zhang X., Zeng H., Wang L. DERGCN: Dynamic-Evolving graph convolutional networks for human trajectory prediction, Neurocomputing, 2024, Vol. 569, Article 127117. DOI: 10.1016/j.neucom.2023.127117.
18. Huang F., Fan Z., Li X., Zhang W., Li P., Geng Y., Zhu K. Tailored meta-learning for dual trajectory transformer: advancing generalized trajectory prediction, *Complex & Intelligent Systems*, 2025, Vol. 11, Article no. 174. DOI: 10.1007/s40747-025-01802-2.
19. Chen S. et al. Adaptive Graph Transformer for Human Trajectory Prediction, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA, June 17–22, 2024*, 2024, pp. 1617–1628.
20. Gupta A., Johnson J., Fei-Fei L., Savarese S., Alahi A. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni*tion, *Salt Lake City, Utah, USA, June 18–22, 2018*, 2018, pp. 2255–2264. DOI: 10.1109/CVPR.2018.00240.
21. Sadeghian A., Kosaraju V., Sadeghian A., Hirose N., Rezatofighi H., Savarese S. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, California, USA, June 16–20, 2019*, 2019, pp. 1349–1358. DOI: 10.1109/CVPR.2019.00144.
22. Bilous N., Kozhevnikov A. Research of Methods for Determining the Accuracy of Metrological Measurements, *Technology Audit and Production Reserves*, 3(2(65)), 2022, pp. 18–23. DOI: 10.15588/1607-3274-2022-2-3
23. Bilous N., Tereshchenko I., Tereshchenko A., Bilous N., Shtangey S., Warsza Z. Risk Analysis Method by the Extreme Data of Dependent Exogenous Variables, *Journal of Automation, Mobile Robotics and Intelligent Systems*, 2022, pp. 44–53. DOI: 10.14313/JAMRIS/3-2021/18
24. Kosaraju V., Martin-Martin R., Reid I., Rezatofighi S., Savarese S. Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks, *Proceedings of the Advances in Neural Information Processing Systems, Vancouver, Canada, December 8–14, 2019*. 2019, pp. 137–146. DOI: 10.5555/3454287.3454300.
25. Ruochen Li, Tanqiu Q., Stamos K., Zhanxing Z., Hubert S. Unified Spatial-Temporal Edge-Enhanced Graph Networks for Pedestrian Trajectory Prediction, *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, pp. 1–14. DOI: 10.1109/TCSVT.2025.3539522.
26. Stanford Drone dataset, 2016, https://cvgl.stanford.edu/ projects/uav_data/

Received 22.09.2025.
Accepted 02.02.2026.
Published 27.03.2026.

УДК 004.93

МОДЕЛІ ГЛИБИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ РУХУ ЛЮДИНИ У ВІДЕОПОТОКАХ

Білоус Н. В. – канд. техн. наук, професор кафедри програмної інженерії, Харківський національний університет радіоелектроніки, Харків, Україна. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0000-0002-8850-9316.

Іванічев В. О. – аспірант кафедри програмної інженерії, Харківський національний університет радіоелектроніки, Харків, Україна. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0009-0002-3705-0098.

АНОТАЦІЯ

Актуальність. Завдання точного прогнозування руху людини в середовищі є критично важливим для застосувань у системах моніторингу, пошуку та навігації. Існуючим підходам часто складно інтегрувати просторову та часову динаміку траєкторій під час обробки потокового відео в реальному часі.

Мета роботи. Розробити фреймворк на основі глибинного навчання, здатний прогнозувати рух людини шляхом поєднання ознак на рівні об’єктів і просторово-часової інформації про траєкторії, отриманої з відеопотоків.

Метод. Запропонований підхід інтегрує YOLO11 для детекції об’єктів, що дає змогу отримувати координати, швидкість, напрям руху та положення відносно оточення. Графова нейронна мережа моделює локальні й глобальні зв’язки між вузлами середовища, агрегуючи ознаки з урахуванням структури місцевості та перешкод. Просторово-часова увага виділяє найрелевантніші моменти траєкторії, підвищуючи точність передбачення. Модель обробляє послідовності кадрів із відеопотоків і в реальному часі прогнозує наступні позиції кожного відстежуваного об’єкта.

Результати. Експерименти на відеопослідовностях із різними сценаріями руху, довжинами траєкторій і варіаціями швидкості показали високу точність прогнозування. Запропонований метод ефективно поєднує просторові та часові ознаки й перевершує базові моделі в задачах трекінгу та передбачення руху.

Висновки. Отримані результати підтверджують придатність запропонованого фреймворку глибинного навчання для прогнозування руху людини в реальному часі у складних середовищах. Подальші дослідження можуть бути зосереджені на розширенні підходу до багатоагентних сценаріїв, оптимізації обчислювальної продуктивності та тестуванні на більших і різноманітніших наборах даних.

КЛЮЧОВІ СЛОВА: глибинне навчання, детекція об’єктів, траєкторія руху, прогнозування траєкторій людини, потокове відео, графові нейронні мережі, контекстно обізнане прогнозування руху, набір даних Stanford Drone, робота в реальному часі.

DOI 10.15588/1607-3274-2026-1-3

ЛІТЕРАТУРА

1. You Only Look Once: Unified, Real-Time Object Detection / [J. Redmon, S. Divvala, R. Girshick, A. Farhadi] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, June 27–30, 2016. – 2016. – P. 779–788. DOI: 10.48550/arXiv.1506.02640.
2. Simonyan K. Very Deep Convolutional Networks for Large-Scale Image Recognition / K. Simonyan, A. Zisserman // International Conference on Learning Representations. – 2015. – P. 1–14. DOI: 10.48550/arXiv.1409.1556.
3. Comparison of CNN-Based Architectures for Detection of Different Object Classes / [N. Bilous, V. Malko, M. Frohme, A. Nechyporenko] // Artificial Intelligence. – 2024. – Vol. 5, No. 4. – P. 2300–2320. DOI: 10.3390/ai5040113.
4. DETRs Beat YOLOs on Real-time Object Detection / [Y. Zhao, W. Lv, S. Xu et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA, June 17–22, 2024. – 2024. – P. 16965–16974. DOI: 10.48550/arXiv.2304.08069.
5. Redmon J. YOLO9000: Better, Faster, Stronger / J. Redmon, A. Farhadi // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, USA, July 21–26, 2017. – 2017. – P. 6517–6525. DOI: 10.1109/CVPR.2017.690.
6. Li Y. Improving real-time object detection in Internet-of-Things smart city traffic with YOLOv8-DSAF method / Y. Li, Y. Huang, Q. Tao // Scientific Reports. – 2024. – Vol. 14. – Article number: 17235. – 15 p. DOI: 10.1038/s41598-024-68115-1.
7. Wang C.-Y. Scaled-YOLOv4: Scaling Cross Stage Partial Network / C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, Tennessee, USA, June 20–25, 2021. – 2021. – P. 13029–13038. DOI: 10.1109/CVPR46437.2021.01283.
8. Jocher G. YOLOv8 : Overview / G. Jocher, A. Chaurasia, J. Qiu // Ultralytics Documentation, 2023. DOI: 10.5281/zenodo.3908559.
9. Bilous N. Search and Detection of People in the Water Using YOLO Architectures: A Comparative Analysis from YOLOv3 to YOLOv8 / N. Bilous. V. Malko, N. Moshenskyi // Automation 2024: Advances in Automation, Robotics and Measurement Techniques. AUTOMATION 2024. Lecture Notes in Networks and Systems, vol 1219. Springer, Cham. – P. 233–255. DOI: 10.1007/978-3-031-78266-4_21
10. A skeleton-based method for exercise recognition based on 3D coordinates of human joints / [N. Bilous, O. Svidin, I. Ahekian, V. Malko] // IAES International Journal of Artificial Intelligence (IJ–AI), ISSN/e-ISSN 2089-4872/2252-8938, 2024. – P. 1805–1816. DOI: 10.11591/ijai.v13.i2.pp1805-1816
11. Bilous N. Determination and Comparison Methods of Body Positions on Stream Video / N. Bilous, I. Ahekian, V. Kaluhin // Radio Electronics, Computer Science, Control. – 2023. – № 2. – P. 52–60. DOI: 10.15588/1607-3274-2023-2-6
12. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields / [Z. Cao, G. Hidalgo, T. Simon et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2019. – Vol. 43, № 1. – P. 172–186. DOI: 10.1109/TPAMI.2019.2929257.
13. 3D human pose estimation in video with temporal convolutions and semi-supervised training / [D. Pavllo, C. Feichtenhofer, D. Grangier, M. Auli] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, California, USA, June 16–20, 2019. – 2019. – P. 7753–7762. DOI: 10.48550/arXiv. 1811.11742.
14. Social LSTM: Human Trajectory Prediction in Crowded Spaces / [A. Alahi, K. Goel, V. Ramanathan et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, June 27–30, 2016. – 2016. – P. 961–971. DOI: 10.1109/CVPR.2016.110.
15. Graph Attention Networks / [P. Veličković, G. Cucurull, A. Casanova et al.] // Proceedings of the International Conference on Learning Representations, Vancouver, Canada, April 30 – May 3, 2018. – 2017. DOI: 10.17863/CAM.48429.
16. Yu B. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting / B. Yu, H. Yin, Z. Zhu // Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, July 13–19, 2018. – 2018. – P. 3634–3640. DOI: 10.24963/ijcai.2018/505.
17. DERGCN: Dynamic-Evolving graph convolutional networks for human trajectory prediction / [J. Mi, X. Zhang, H. Zeng, L. Wang] // Neurocomputing. – 2024. – Vol. 569. – Article 127117. DOI: 10.1016/j.neucom.2023.127117.
18. Tailored meta-learning for dual trajectory transformer: advancing generalized trajectory prediction / [F. Huang, Z. Fan, X. Li et al.] // Complex & Intelligent Systems. – 2025. – Vol. 11. – Article no. 174. DOI: 10.1007/s40747-025-01802-2.
19. Adaptive Graph Transformer for Human Trajectory Prediction / [S. Chen et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA, June 17–22, 2024. – 2024. – P. 1617–1628.
20. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks / [A. Gupta, J. Johnson, L. Fei-Fei et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, USA, June 18–22, 2018. – 2018. – P. 2255–2264. DOI: 10.1109/CVPR.2018.00240.
21. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints / [A. Sadeghian, V. Kosaraju, A. Sadeghian et al.] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, California, USA, June 16–20, 2019. – 2019. – P. 1349–1358. DOI: 10.1109/CVPR.2019.00144.
22. Bilous N. Research of Methods for Determining the Accuracy of Metrological Measurements / N. Bilous, A. Kozhevnikov // Technology Audit and Production Reserves, 3(2(65)). – 2022. – P. 18–23. DOI: 10.15588/1607-3274-2022-2-3
23. Risk Analysis Method by the Extreme Data of Dependent Exogenous Variables / [N. Bilous, I. Tereshchenko, A. Tereshchenko et al.] // Journal of Automation, Mobile Robotics and Intelligent Systems. – 2022. – P. 44–53. DOI: 10.14313/JAMRIS/3-2021/18
24. Social-BiGAT: Multimodal Trajectory Forecasting using Bicycle-GAN and Graph Attention Networks / [V. Kosaraju, A. Sadeghian, R. Martin-Martin et al.] // Proceedings of the Advances in Neural Information Processing Systems, Vancouver, Canada, December 8–14, 2019. – 2019. – P. 137–146. DOI: 10.5555/3454287.3454300.
25. Unified Spatial-Temporal Edge-Enhanced Graph Networks for Pedestrian Trajectory Prediction / [Li Ruochen, Q. Tanqiu, K. Stamos et al.] // IEEE Transactions on Circuits and Systems for Video Technology. – 2025. – P. 1–14. DOI: 10.1109/TCSVT.2025.3539522.
26. Stanford Drone dataset, 2016, https://cvgl.stanford.edu/ projects/uav_data/

DOI 10.15588/1607-3274-2026-1-3

UDC 004.93

LONG-DISTANCE CABBAGE DAMAGE AND PEST DETECTION METHOD USING YOLO11

Khabarlak K. S. – PhD, Associate Professor of the Department of System Analysis and Control, Dnipro University of Technology, Dnipro, Ukraine. ROR: https://ror.org/05hkn5555. ORCID: https://orcid.org/0000-0003-4263-0871.

Laktionov I. S. – Dr. Sc., Full Professor, Professor of the Department of Computer Systems Software, Dnipro University of Technology, Dnipro, Ukraine. ROR: https://ror.org/05hkn5555. ORCID: https://orcid.org/0000-0001-7857-6382.

Gorev V. N. – PhD, Associate Professor, Head of the Department of Physics, Dnipro University of Technology, Dnipro, Ukraine. ROR: https://ror.org/05hkn5555. ORCID: https://orcid.org/0000-0002-9528-9497.

Diachenko G. G. – PhD, Associate Professor of the Department of Electric Drive, Dnipro University of Technology, Dnipro, Ukraine. ROR: https://ror.org/05hkn5555. ORCID: https://orcid.org/0000-0001-9105-1951.

ABSTRACT

Context. To ensure sustainable yield, plant health must be constantly monitored with timely measures applied to prevent disease spread. Traditional approaches rely on manual inspection of plants, while neural networks require large amounts of annotated data to train. Both manual inspection and data annotation require expert knowledge and are time-consuming. Close-up photos of leaves are often used for training as they are easier to collect from the Internet. However, this complicates disease spread estimation at a scale. Cabbage is one of the plants widely grown in Ukraine, but existing research focusing on cabbage health monitoring is limited.

Objective. The goal of this work is to build a neural-network-based cabbage disease and pest detection system, which can be trained in on a small number of training images. At inference the system should detect pests on plant images at a distance of a whole plant.

Method. Given that existing plant disease datasets, such as IP102 and PlantDoc mostly contain close-up images of diseased plants, the networks trained on such datasets suffer from lack of generalization to images at a distance. To select the best object detection model, state-of-the-art object detection architectures, namely YOLO 8, 9, 10, 11, and RT-DETR have been analyzed in the work. To increase detection distance multi-image loss is proposed to improve hyperparameter search using Tree-Structured Parzen Estimators. Also, to improve detection quality, a novel cabbage disease dataset has been collected in Dnipro region, Ukraine. The new classes include crucifier flea beetle (widespread pest in Dnipro region) and damaged leaf. When the pest is not visible, but leaf damage is taken, determining specific pest might not be possible. Therefore, we introduce additional damaged leaf class, that captures generic plant damage. This also enables tracking of plant healing rate, when measures to stop pest spread have been taken. We combine collected images with the larger IP102 dataset to increase the number of pests covered to form new Cabbage+IP102 dataset.

Results. 1) Tree-Structured Parzen Estimators search on the multi-image loss has improved the YOLO 11 M performance from 0.3642 to 0.3892 mAP50-95 on images taken at a distance. 2) Collected dataset has enabled detection of cabbage plant health problems at a distance, including cases when the pest is currently not visible, but the damage is present.

Conclusions. In this work, the cabbage pest and damaged leaf YOLO 11 M detection system has been presented. The detector architecture has been selected as the best-found during analysis on 2 datasets. The developed system requires only 7 annotated cabbage images to be trained and to perform pest and damaged leaf detection on high resolution images (2016x2016) of whole cabbage plants. The final model can be used to monitor cabbage health problems, damage, and rate of healing using images taken at a distance.

KEYWORDS: agriculture, deep learning, plant health monitoring, pest detection, leaf damage estimation, yolo 11, brassicaceae family.

ABBREVIATIONS

YOLO is a You Only Look Once family of single-stage object detection models;

RT-DETR is a Real-Time Detection Transformer, a transformer-based object detection model;

TPE is a Tree-Structured Parzen Estimators hyperparameter optimization method.

NOMENCLATURE

M is an input image of shape $W \times H \times C$, where W,H,C are width, height and the number of color channels;

$\Phi(x)$ is a detection neural network, that takes image M as an input and returns $\{D\}_{i=0}^{\text{Max}_{\text{det}}}$ predictions;

$\{D\}_{i=0}^{\text{Max}_{\text{det}}}$ is a maximum number of possible detections for a given network;

D_i is a tuple with object detection information, which typically takes form ($Shape_i$, $ObjectnessScore_i$, $ClassProbabilities_i$);

f is a training loss function;

mAP_{50-95} is a mean average precision detection metric;

$p(x|y)$ is a TPE hyperparameter kernel density estimate.

INTRODUCTION

Increasing plant yields per hectare is an important agricultural problem. Traditional methods of plant health monitoring rely on manual inspections and visual diag-

DOI 10.15588/1607-3274-2026-1-4

nostics, which can be time-consuming and labor-intensive, particularly over large farmlands. In recent years hardware and software assisted technologies are widely used to monitor plant growing conditions and to take timely actions [1]. The approaches include land fertilization, spraying and irrigation of fields using autonomous vehicles or drones, climate and soil parameters monitoring, while recent advancements in computer vision and machine learning have provided promising solutions for automating the detection of plant pests and diseases.

Machine learning-based plant health problem detection is typically performed on datasets that capture short-distance images of pests [2–5]. There are several reasons for that: 1) such datasets can be collected from images found on the Internet; 2) both annotating images in short-distance datasets and training a neural network to detect large disease manifestation are simpler to perform. However, in the field conditions it would have been beneficial to automatically detect the presence of illness at a distance, which requires the use of appropriate datasets and improved training techniques.

Annotating images of a large field with instances of disease manifestations (e.g., malicious insects) spanning only several pixels is a complicated problem, that requires significant time from experts to annotate the dataset. Cabbage is one of the plants widely grown in Ukraine, but existing research focusing on cabbage health monitoring is limited.

Therefore, in this work we investigate the use of state-of-the-art general-purpose detection neural networks for the task of long-distance cabbage plant health monitoring and propose a method of model hyper-parameter selection based on Tree-Structured Parzen Estimators.

The object of this study is the process of automatic image-based cabbage plant health monitoring.

The purpose of this study is object detection neural networks and hyperparameter selection algorithms.

The purpose of the research is to build a neural-network-based cabbage disease and pest detection system, which is capable of accurate pest detection based on image of the whole plant.

1 PROBLEM STATEMENT

Let M be an input image tensor of shape $W \times H \times C$, where W, H, C are width, height and number of channels. For RGB image $C=3$; let $\Phi(x)$ be a detection neural network, that takes image M as an input, and returns $\{D\}_{i=0}^{\mathrm{Max}_{\mathrm{det}}}$ tuples with information about predicted object bounding boxes, which typically takes form $(Shape_i, ObjectnessScore_i, ClassProbabilities_i)$, where *Shape* is defined depending on detection model as *(left,top,width,height)* or $(center_x, center_y, width, height)$ object bounding box coordinates, *ObjectnessScore* is a score whether this bounding box contains actual object, *ClassProbabilities* is a vector R^K of class probabilities, where K is the number of classes.

The goal is to minimize loss function f, that defines distance between true and predicted *Shape*, *ObjectnessScore*, and *ClassProbabilities*. The exact loss function f is different between detection neural network models.

2 REVIEW OF THE LITERATURE

To ensure sustainable yield, plants must be constantly monitored for diseases with timely measures applied to prevent disease spread. Traditional approaches rely on manual inspection of plant leaves and trunks to detect any signs of a virus, fungi, bacteria or pests. Obviously, this approach is time-consuming, not scalable to large farmlands, moreover early development of illness is easy to miss. Therefore, the focus of recent agricultural research is shifted towards automation of plant health monitoring.

Automated disease detection approaches are based either on monitoring weather conditions and predicting probability of appearance of certain disease [1, 6] or by capturing images of plants via a smartphone or drone [1, 7]. Both approaches have certain benefits and disadvantages. While weather monitoring is suitable for predicting illness before it occurs, it works with typicality and cannot detect spontaneous illness spreads. In contrast, computer-vision-based methods detect existing plant health problems and can distinguish between different kinds of illnesses, so the most effective treatment could be selected for the specific disease or pest. In this work we focus on computer-vision-based methods such as those that are more precise in disease or pest detection.

Typically, image-based plant health problem detection is formulated as classification or detection. In classification formulation, given a close-up image of plant leaf, trunk or fruit the task is to say what specific kind of illness there is or that no illness is present [8]. Obviously, in this case finding regions of interest remains manual labor. To automatically find the disease, detection methods should be used.

Neural networks have shown high quality in solving detection of arbitrary objects. Early approaches were 2-stage: region proposal and subsequent classification of the proposed regions, like R-CNN [9], Faster R-CNN [10]. Recent research has mostly shifted towards single-stage approaches, like YOLO family of models [11–13] or DETR [14]. Single stage approaches generally have faster inference and higher quality.

Many research approaches were proposed in the field of plant disease and pest detection. The authors of [2] propose a modified YOLOv7-tiny architecture for plant pest detection. The authors target environment with constrained computational resources, therefore, select the smallest variation of the YOLOv7 family of models to improve upon. The resulting model is 47% faster, while offering a 3% improvement in mAP50 over the base YOLOv7-tiny model. Training and testing are performed on images of 21 pests, which is a subset selected from the IP102 [15] dataset. In [5] YOLOv5M has been improved for pest detection by introducing SwinTransformer-based blocks.

DOI 10.15588/1607-3274-2026-1-4

In [16] a modification of YOLOv10 is introduced. The authors reduce computation complexity of the network, while improving tomato fruit detection and ripeness estimation. The authors focus on detection of overlapped and partially occluded fruits.

In [4] the authors note an importance of cabbage (Brassicaceae family) pests and diseases detection for sustained yield and indicate that the field is largely unexplored. In the paper the authors: 1) collect a novel dataset, that consists of 21 classes (14 pests, 3 damage symptoms, 4 beneficial insects); 2) evaluated YOLOv5, 7–11 models on the proposed dataset.

In [3] the authors note the difficulty of detecting large number of small pests. To resolve the issue, they 1) collect a dataset of pests using a light trap; 2) propose an improved R-CNN architecture with novel feature fusion block to improve small pest detection.

Overall neural networks are known to require large amounts of annotated data to be trained. While annotating images with commodity objects can be outsourced to a crowdsourcing platform, annotation of plant disease requires expert knowledge. Thus, acquiring additional training images requires significant effort from human experts, and therefore hard and expensive to acquire. Few-shot learning approaches attempt to resolve the problem. Main few-shot learning approaches include [17]: metric learning, where embedding into a metric space is learned, that is general enough to quickly accommodate to new classes; optimization-based, where a special optimization procedure is defined, that trains “generic” model weights, that are easy to fine-tune; model-based, where model architecture is changed to include memory cells. Transfer learning can also be used in a few-shot scenario, where learning is performed in 2 stages: training on a large general-purpose dataset, then finetuning on a smaller target dataset. Transfer learning is widely used due to its simplicity and sufficiently good results. Hence, in this work we focus on using transfer learning for few-shot neural network training.

Several few-shot learning approaches for plant pest recognition have been proposed. In [18] metric-based few-shot learning plant disease classification method is introduced. Image embeddings are generated first by using ResNet-18 convolutional neural network and then refined by a Transformer. Contrastive loss is used to perform the training. Mahalanobis distance is calculated between new image and existing classes to perform the classification.

The authors of [19] focus on few-shot learning of plant pests. Pre-trained ResNet-50 backbone is used as a part of Faster R-CNN 2-stage detection neural network. To improve working with images of different scale multi-input single-output feature pyramid network is used.

However, the distance at which plant health problem detection is typically performed (typically leaf or part of a plant) is still quite small, which complicates disease spread estimation at a scale. Also, newer detection architectures like YOLO 8, 9, 10, 11 presented in [20–23] or RT-DETR [24] (a modification of DETR) still remain mostly unexplored in the field of plant illness and pest detection on open dataset. Research of cabbage (Brassicaceae family) plant health issue detection is underrepresented in literature.

Therefore, in-the-field plant health monitoring is an open research problem. In this work we: 1) propose a method for the model hyper-parameter tuning for efficient training with few training samples; 2) collect a few-shot dataset for long-distance monitoring of cabbage plant pests and leaf damage.

3 MATERIALS AND METHODS

A generalized scheme of steps taken to improve large distance cabbage health problem detection is shown in Fig. 1.

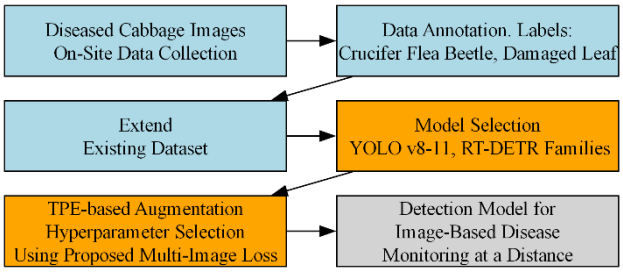


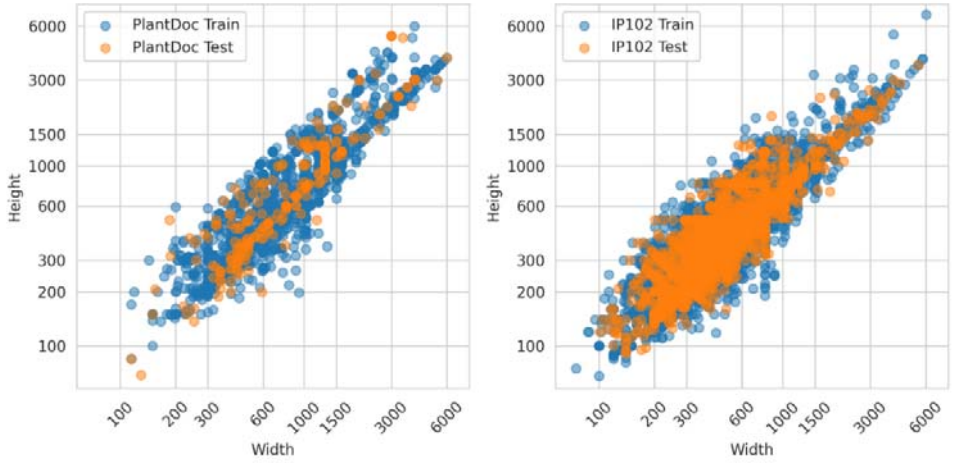
Figure 1 – Generalized scheme of research

First, we analyze existing plant pest and disease open datasets. Then collect in-the-field images of cabbage and annotate them with 2 new classes: crucifer flea beetle (which is widespread in Dnipro region in 2024/25) and damaged leaf. The latter is useful when the pest is not visible, but leaf damage is taken. Also, plant healing can be tracked after measures to stop pest spread have been taken. We combine collected images with the larger IP102 dataset to increase the number of pests that can be detected.

Next, we present a comprehensive analysis of YOLO 8–11 and RT-DETR architectures on 2 widely used plant health detection datasets, namely PlantDoc [25] and IP102 [15]. We perform transfer learning of these models. Each model is considered in different sizes available (nano, medium, large, etc., depending on the model), and provide guidance on selecting the best model.

Finally, we propose multi-image validation loss for searching model hyperparameters using Tree-Structured Parzen Estimators to improve the quality of model training on images taken at a distance. Joined Cabbage+IP102 dataset is used. The final trained model can not only perform detection not only on close-up photos of leaves, but also on plant pictures at a distance (e.g., full cabbage plant).

DOI 10.15588/1607-3274-2026-1-4



(a) PlantDoc (b) IP102

Figure 2 – Distribution of image resolutions present in (a) PlantDoc and (b) IP102 datasets (Log-Log scale)

Many datasets contain images of detached diseased leaf on white background, which does not correspond to real world conditions. In this paper we consider PlantDoc and IP102 datasets, that contain in-the-wild images. PlantDoc [25] has 2 flavors: detection (with 1 or multiple objects per image with bounding boxes) or classification (images cropped to the bounding box). It has 17 classes, and on average, 3.5 bounding boxes per image. IP102 [15] dataset, which has 102 classes of pests and 7.3 times more annotated images. It also has classification and detection sets. On average, it has 1.2 bounding boxes per image, which is less than PlantDoc 3.4 objects per image.

Distribution of image resolutions of the two datasets is shown in Fig. 2 (Log-Log scale). Both datasets contain vertical, horizontal and square images. Although at resolutions above 1500 pixels in either dimension square images are less present. Most images have resolutions below 1500x1500 pixels. Low (below 400 pixels) and very low (below 200 pixels) resolution images are present in both datasets.

While IP102 and PlantDoc datasets offer a strong baseline for training plant health-related problems, they have some disadvantages:

1. PlantDoc contains images either of healthy leaves or damaged by infectious diseases (bacteria, virus, fungi). Images of pests are not present.

2. IP102 dataset has been constructed by collecting images from the Internet and then annotating them. So mostly, the dataset contains close-up images of pests.

For plant pest detection in real-world environment on a farm in Dnipro region, Ukraine, we have found the following challenges in using these datasets for training:

1. It is desirable to capture large images of plants (e.g., whole cabbage plant) and estimate the number of pests on it. In this case the pests appear to be quite small and numerous, which is different from these datasets.

2. Images of not every pest can be captured during worktime. Some are active, for instance, very early in the morning (e.g. slugs) and only damaged leaves can be observed after the fact.

To resolve these problems, we have collected a small dataset with images of cabbage with health problems for few-shot training. Overall, 10 images, 7 in the train set, 3 in the test set. These images have been annotated with 2 classes: crucifer flea beetle (Phyllotreta Cruciferae) and damaged leaf. The dataset has a high number of object instances: 345 instances of crucifer flea beetle and 134 instances of damaged leaf. Collected dataset information is shown in Table 1. It was farmers' demand to perform detection of this beetle as in farm in Dnipro region, Ukraine in 2024 the crucifer flea beetle was widespread, and early detection was important to save the yield of cabbage. Annotated samples are shown in Fig. 3, crucifier flea beetle is highlighted with violet bounding boxes, and leaf damage with yellow.

Table 1 – Collected dataset information.

Class	# train images	# train object instances	# test images	# test object instances
Crucifer Flea Beetle	5	201	3	144
Damaged Leaf	7	91	3	43

To give the neural network knowledge about different kinds of pests given a small training dataset of cabbage images, we extend the IP102 dataset, which is the largest pest dataset to the best of our knowledge. Summary of datasets is presented in Table 2. Cabbage+IP102 is the extended version of the dataset with our annotated images.

DOI 10.15588/1607-3274-2026-1-4

Figure 3 – Examples of the collected and annotated images

Table 2 – Comparison of detection datasets used for training

Dataset	# train	# test	# classes	# objects per image
IP102	15178	3798	102	min: 1, mean: 1.2, max: 26
Cabbage+ IP102	15185	3801	104	min: 1, mean: 1.2, max: 86
PlantDoc	2348	237	17	min: 0, mean: 3.4, max: 42

It should be noted that IP102 dataset contains flea beetle class. However, pests on collected images of cabbage were not detected as such. During deeper investigation, we have found out the IP102 dataset contains photos of other species of flea beetle (Phyllotreta vittula) and macro photos were taken.

To evaluate model performance, we use mAP_{50-95} (mean average precision), which is a common choice. The following formula can be used to compute the metric

$$\mathrm{mAP}_{50-95} = \frac{1}{K \cdot N_{\mathrm{IoU}}} \sum_{k=1}^{K} \sum_{i=1}^{N_{\mathrm{IoU}}} \mathrm{AP}_{k,\mathrm{IoU}_i}\ , \qquad (1)$$

where K is the number of classes, N_{IoU} is the number of intersections over union thresholds (IoU), IoU is taken from 0.5 to 0.95 at 0.05 steps, $\mathrm{AP}_{\mathrm{k,IoUi}}$ is the average precision for class k at the IoU threshold IoU_i.

Both PlantDoc and IP102 datasets contain close-up images (in certain cases, macro images) of pests, and typically only one or few pests are visible. It has been shown that searching augmentations is an effective [26] way to improve the trained model quality. In [12] it has been shown that by applying mosaic augmentation during training, model detection performance can be significantly improved.

In this work we propose to construct hyperparameter optimization loss using multiple images to improve model performance on images captured at a distance:

1. Validation images are rescaled to the resolution of $W_{single} \times H_{single}$.

2. 4×4 grid of resolution $\mathrm{W_{grid}}\times\mathrm{H_{grid}}$ is formed from these images.

3. Given landmark locations for image at row i and column j:

$$(centerX_{(i,j)}, centerY_{(i,j)}, width_{(i,j)}, height_{(i,j)}).$$

Landmark locations are updated as follows:

$$\begin{aligned} centerX'_{i,j} &= \frac{centerX_{i,j} \cdot W_{\mathrm{single}}}{W_{grid}} + offsetX_{i,j}, \\ centerY'_{i,j} &= \frac{centerY_{i,j} \cdot H_{\mathrm{single}}}{H_{grid}} + offsetY_{i,j}, \\ width'_{i,j} &= \frac{width_{i,j} \cdot W_{\mathrm{single}}}{W_{grid}}, \\ height'_{i,j} &= \frac{height_{i,j} \cdot H_{\mathrm{single}}}{H_{grid}}, \end{aligned} \qquad (2)$$

where $offsetX_{i,j}$, $offsetY_{i,j}$ are defined as follows:

$$\begin{aligned} offsetX_{i,j} &= j \cdot W_{\mathrm{single}}, \\ offsetY_{i,j} &= i \cdot H_{\mathrm{single}}. \end{aligned} \qquad (3)$$

4. mAP_{50-95} using formula (1) is computed on constructed images.

In this work $W_{single}=H_{single}=640$, $W_{grid}=H_{grid}=2560$.

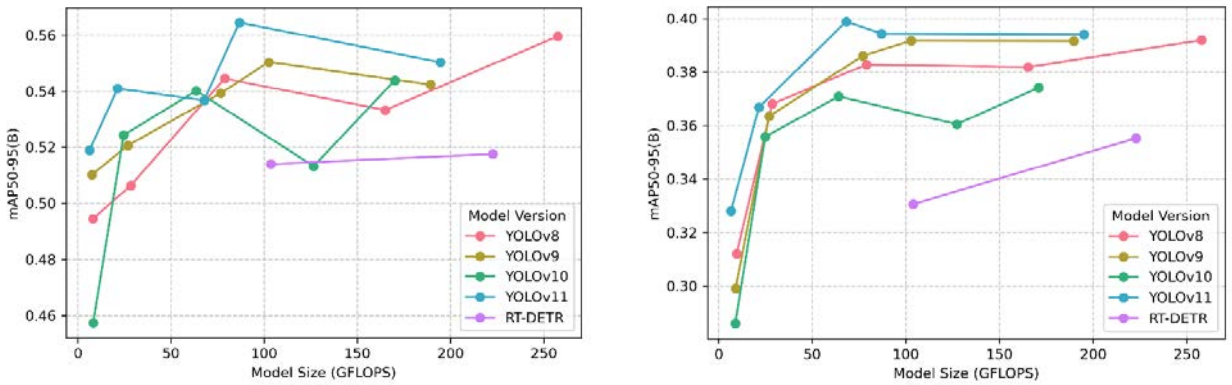
To perform hyperparameter search we use Tree-Structured Parzen Estimators (TPE) [27]. The approach has been to be effective for hyperparameter optimization and is widely used [28].

TPE models search space for the best hyperparameters using kernel density estimation of "bad" hyperparameters $l(x)$ and "good" hyperparameters $g(x)$. Given a minimization problem, the general density $p(x|y)$ is defined as follows [27]:

$$p(x \mid y) = \begin{cases} l(x), & \text{if } y < y^*. \\ g(x), & \text{if } y \geq y^*. \end{cases} \qquad (4)$$

where $\{x_i\}$ are hyperparameter observations, $\{y_i\}$are the corresponding loss function values, y^* is a quantile γ of the best observed y values, such that $p(y<y^*)= \gamma$.

DOI 10.15588/1607-3274-2026-1-4



(a) PlantDoc (b) IP102

Figure 4 – Model mAP50–95 depending on model size at 25 training epochs

4 EXPERIMENTS

First, experiments to select the best model for plant pest and disease detection are conducted. We consider YOLO 8, 9, 10, 11 and RT-DETR families of models in different sizes as is shown in Table 3. Size abbreviations are the following: N – nano, T – tiny, S – small, M – medium, C – compact, L – large, X – extra-large, E – extended.

To train the models we use image resolution of 640x640 pixels, which is typically used for pretraining the detection models, and is reasonable for image resolutions of PlantDoc and IP102 datasets (based on Fig. 2). Results are shown on the test set. The batch size is 64, the number of training epochs is set to 25. To improve the quality of detection models, augmentations are used, which include color augmentations (random variations of color hue, saturation, value). Mosaic augmentation [12] is applied during training to improve model generalization. It combines several images into one and has been shown to improve detection quality. Finally, horizontal flip with probability 0.5 is used.

Table 3 – Model sizes considered during the analysis

Model Family	Sizes Considered
YOLO v8	N, S, M, L, X
YOLO v9	T, S, M, C, E
YOLO v10	N, S, M, L, X
YOLO 11	N, S, M, L, X
RT-DETR	L, X

Next, training speed experiments are conducted to see which models require less training steps to achieve high quality. Training is performed for 1, 2, 5, 10, 15, 20, 25, 50 epochs. These experiments are conducted on open IP102 and PlantDoc dataset, so that the analysis can be reused by other researchers.

Finally, augmentation hyperparameter search is used to improve model performance on Cabbage+IP102 dataset. Multi-image validation loss is used for TPE optimization. This dataset has extra images with a high number of annotations per image. We use the best model selected in previous analysis and improve it by performing augmentation hyperparameter search. 60 iterations of hyperparameter search are performed. Parameter ranges considered are defined in Table 4. The probability of horizontal flip augmentation is fixed at 0.5, which is standard value, and is not optimized.

Table 4 – Hyperparameter search ranges

Hyperparameter	Values
HSV: Hue	[0.0, 0.1]
HSV: Saturation	[0.0, 0.9]
HSV: Value	[0.0, 0.9]
Rectangular Images	[False, True]
Multiscale Training	[False, True]
Disable Mosaic Last Epochs	[0, 25]
Degrees Rotation	[0.0, 90.0]
Probability of Flip Upside Down	[0.0, 0.5]
Mosaic Probability	[0.0, 1.0]

5 RESULTS

In Fig. 4 we show $mAP_{50–95}$ of all considered models on PlantDoc and IP102 datasets. On x axis are glops, and not model size names, because the number of floating-point operations is different across model versions. On PlantDoc dataset (Fig. 4a) the test mAP50–95 results are more noisy, likely due to smaller number of test images. Still, YOLO 11 shows one of the best results. On IP102 dataset (Fig. 4b), YOLO 11 models show the best results in each size, while RT-DETR shows the worst. On this dataset YOLO 11 M has the best result on the test set. RT-DETR model underperforms on both datasets, showing the worst mAP to gflops ratio

Note that the smallest variant (nano for YOLO 8,10,11 or tiny for the 9) has on average 67% reduction of gflops and 16% reduction of mAP50–95 with respect to small size. This is a higher quality degradation rate than between other adjacent model sizes. For instance, small variants have 65% gflops reduction over medium, but only 5% quality loss.

In Figs. 5 and 6 we show box plots plot for each of the model versions depending on the number of training epochs for the PlantDoc and IP102 datasets correspondingly. A single box plot is built based on training results of different-sized models of the specified model family.

DOI 10.15588/1607-3274-2026-1-4

As can be seen, in most cases there is mAP outlier below each box plot. The outlier corresponds to the smallest variation (nano or tiny) of this model version.

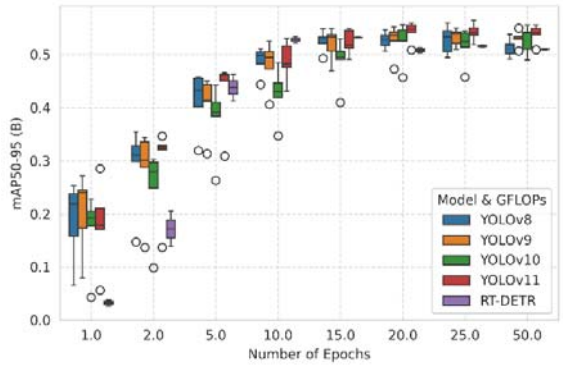


Figure 5 – PlantDoc dataset: model $mAP_{50–95}$ depending on the number of epochs

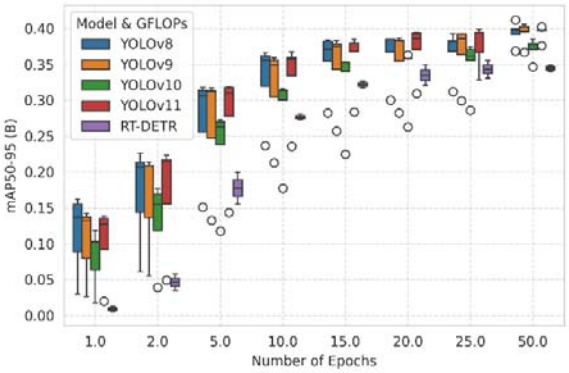


Figure 6 – IP102 dataset: model $mAP_{50–95}$ depending on the number of epochs

On both datasets training for more than 15–20 epochs doesn't result in substantially improved mAP for all models except the smallest ones. When training for a small number of epochs, the best results are obtained from the larger models.

Augmentation hyperparameter optimization values are shown in Table 5, and final training and validation accuracy in Table 6. Hyperparameter optimization procedure has been performed for 60 iterations. The best value has been achieved on iteration 19.

6 DISCUSSION

Best-found validation hyperparameters, found using the proposed multi-image validation loss, have several important changes, when compared to the default augmentation values proposed by the Ultralytics library. Both color saturation and value augmentations were reduced during the search process. During further inspection, high changes in saturation or value result in pests being much harder to detect by human observers. This is explained by the fact that many pests mimic the color of plants on which they live. Also, pest color is important to distinguish between similar species of pests.

Table 5 – Augmentation hyperparameters of the initial and best-found models

Augmentation Hyperparameter	Initial	Best
HSV: Hue	0.0150	0.0240
HSV: Saturation	0.7000	0.3367
HSV: Value	0.4000	0.0108
Rectangular Images	False	False
Multiscale Training	False	False
Disable Mosaic Last Epochs	10	8
Degrees Rotation	0.00	16.03
Probability of Flip Upside Down	0.0000	0.0760
Probability of Horizontal Flip (not searched)	0.5	0.5
Mosaic Probability	1.0000	0.5392

Table 6 – mAP50–95 comparison of the initial and best-found models

Dataset	mAP50–95 (Train)	mAP50–95 (Validation)
Cabbage+ IP102	0.3116	0.3642
Cabbage+ IP102+multi-image loss	0.2901	**0.3892**

Interestingly, using rectangular images or multiscale training doesn't improve quality on the multi-image validation dataset. By default (Rectangular Image = False), all training images are squeezed into square. If rectangular images parameter is enabled, during training images are resized to the most common aspect ratio. Validation images always use rectangular shapes. Multiscale training uses images of different sizes during training; actual input image resolution is different from batch to batch. Not to be mistaken with resize augmentation, where input image size is the same, but the image is randomly scaled. However, in practice, using either of these parameters does not improve mAP score on the constructed multi-image dataset.

Example image that can be processed by the developed system is shown in Fig. 7. The image captures the whole plant. Overall, 346 detections of crucifier flea beetles and 64 instances of damaged leaf are found in this image. The image is fed into the model in resolution of 2016 x 2016.

Figure 7 – Example full plant image processed by the model

DOI 10.15588/1607-3274-2026-1-4

(a) (b)

Figure 8 – Crops of full plant images, showing detections of pests only (a) and joint pests and leaf damage detection (b)

Crops of full plant images are shown in Fig. 8. In Fig. 8 (a) detection of crucifier flea beetles are shown; damaged leaf detections are filtered out for clarity. As can be seen, the vast majority of bugs have been detected. Only, blurry bug (top right) and bug in shadow region (middle left) have been missed. Obviously, these missed detections have no impact on the final decision. In Fig. 8 (b) joint damaged leaf and crucifier flea beetle detections are shown. Apparently, most of the bugs have left this region of the plant. However, the system has been able to detect severe damage taken by the plant.

CONCLUSIONS

In this work an important problem of cabbage pest detection at a distance has been solved. Existing datasets, such as IP102 and PlantDoc contain mostly macro photos of pests and plant diseases, which complicates training of neural networks for plant pest detection at a distance. To resolve the problem, additional cabbage photos at a distance have been collected in Dnipro region, Ukraine. YOLO 8, 9, 10, 11, RT-DETR neural networks have been analyzed. The best results have been shown by the YOLO 11 M (medium) network. Finally, augmentation hyperparameter search has been conducted using Tree-Structured Parzen estimator on multi-image validation set. The developed system requires only 7 annotated cabbage images to be trained. The final trained neural network can detect more than 300 instances of bugs on images of whole plants, even in cases when the bug is only 11×11 pixels on a 2016×2016 image. The developed system can be deployed for large-scale monitoring using edge devices. The system proposed in this work can detect large number of pests given a full plant image, which enables easier monitoring of large fields of cabbage.

The scientifical novelty of obtained results is that 1) Tree-Structured Parzen Estimator augmentation hyperparameter search on the proposed multi-image loss function on the validation has improved model performance on images at a distance from 0.3642 to 0.3892 mAP50–95; 2) cabbage pest and damaged leaf dataset has been collected.

The practical significance of obtained results is that the developed system is used for cabbage disease monitoring to prevent pest spread, and in contrast to previous systems detection is performed not on an individual leaf, but on a picture of a whole plant, which makes plant disease monitoring on large fields easier.

Prospects for further research are to propose custom object detection architecture for long distance plant health problem monitoring.

ACKNOWLEDGEMENTS

This research was carried out as part of the scientific project “Development of software and hardware of intelligent technologies for sustainable crop production in wartime and post-war” (state registration number 0124U000289) funded by the Ministry of Education and Science of Ukraine at the expense of the state budget.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors’ contributions: Kostiantyn Khabarlak: software, data collection, writing – original draft preparation; Ivan Laktionov: writing – review and editing, funding acquisition; Vyacheslav Gorev: formal analysis, literature review; Grygorii Diachenko: method validation, writing – review and editing.

Data availability: The manuscript has no associated data.

Software availability: The manuscript has no associated software.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

DOI 10.15588/1607-3274-2026-1-4

REFERENCES

1. Laktionov I. et al. A comprehensive review of recent approaches and Hardware-Software technologies for digitalisation and intellectualisation of Open-Field crop Production: Ukrainian case study in the global context, *Computers and Electronics in Agriculture*, 2024, Vol. 225, P. 109326. DOI: https://doi.org/10.1016/j.compag.2024.109326.
2. Gong H., Ma X., Guo Y. Research on a target detection algorithm for common pests based on an improved yolov7-tiny model, *Agronomy*, 2024, Vol. 14, № 12. DOI: 10.3390/agronomy14123068.
3. Teng Y. et al. MSR-RCNN: a multi-class crop pest detection network based on a multi-scale super-resolution feature enhancement module, *Frontiers in Plant Science*, 2022, Vol. 13. DOI: 10.3389/fpls.2022.810546.
4. Chakrabarty S. et al. Deep learning-based accurate detection of insects and damage in cruciferous crops using YOLOv5, *Smart Agricultural Technology*, 2024, Vol. 9, P. 100663. DOI: https://doi.org/10.1016/j.atech.2024.100663.
5. Dai M. et al. A new pest detection method based on improved YOLOv5M, *Insects*, 2023, Vol. 14, № 54. DOI: 10.3390/insects14010054.
6. Diachenko G. et al. An improved approach to prediction of maize disease occurrence based on weather monitoring and machine learning, *Case of the forest-steppe and northern steppe of Ukraine*, 2024, Vol. 12, № 4. DOI: 10.22364/BJMC.2024.12.4.03.
7. Bhargava A. et al. Plant leaf disease detection, classification, and diagnosis using computer vision and artificial intelligence: a review, *IEEE access : practical innovations, open solutions*, 2024, Vol. 12, pp. 37443–37469. DOI: 10.1109/ACCESS.2024.3373001.
8. Khabarlak K., Diachenko G., Laktionov I. Feature knowledge distillation using group convolutions for efficient plant pest recognition, *Proceedings of the 12th International Conference Information Control Systems & Technologies (ICST 2024)*. Odesa, Ukraine, September 23–25 : CEUR workshop proceedings, CEUR-WS.org, 2024, Vol. 3790, pp. 377–387.
9. Girshick R.B. et al. Rich feature hierarchies for accurate object detection and semantic segmentation, *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*. Columbus, OH, USA, June 23–28, 2014, IEEE Computer Society, 2014, pp. 580–587. DOI: 10.1109/CVPR.2014.81.
10. Ren S. et al. Faster R-CNN: Towards real-time object detection with region proposal networks, *Advances in neural information processing systems 28: Annual conference on neural information processing systems 2015, December 7–12, 2015*. Montreal, Quebec, Canada, 2015, pp. 91–99.
11. Redmon J. et al. You only look once: Unified, real-time object detection, *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*. Las Vegas, NV, USA, June 27–30, 2016, IEEE Computer Society, 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.
12. Bochkovskiy A., Wang C.-Y., Liao H.-Y.M. YOLOv4: Optimal speed and accuracy of object detection, *CoRR*, 2020, Vol. abs/2004.10934.
13. Jocher G. Ultralytics yolov5, 2020.
14. Carion N. et al. End-to-end object detection with transformers, *ECCV 2020, Glasgow, UK, August 23–28, 2020 : Lecture notes in computer scienc*e. Springer, 2020, Vol. 12346. pp. 213–229. DOI: 10.1007/978-3-030-58452-8_13.
15. Wu X. et al. IP102: A large-scale benchmark dataset for insect pest recognition, *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, IEEE, 2019. DOI: 10.1109/cvpr.2019.00899.
16. Li A. et al. D3-YOLOv10: Improved YOLOv10-based lightweight tomato detection algorithm under facility scenario, *Agriculture* (Nitra, Slovakia), 2024, Vol. 14, № 12. DOI: 10.3390/agriculture14122268.
17. Khabarlak K. S. Faster optimization-based meta-learning adaptation phase, *Radio Electronics, Computer Science, Control*, 2022, № 1, P. 82. DOI: 10.15588/1607-3274-2022-1-10.
18. Nuthalapati S. V., Tunga A. Multi-domain few-shot learning and dataset for agricultural applications, *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021*. Montreal, QC, Canada, October 11–17, 2021, IEEE, 2021, pp. 1399–1408. DOI: 10.1109/ICCVW54120.2021.00161.
19. Wang X. et al. Prior knowledge auxiliary for few-shot pest detection in the wild, *Frontiers in Plant Science*, 2023, Vol. 13–2022. DOI: 10.3389/fpls.2022.1033544.
20. Jocher G., Chaurasia A., Qiu J. Ultralytics YOLOv8, 2023.
21. Wang C.-Y., Liao H.-Y.M. YOLOv9: Learning what you want to learn using programmable gradient information, 2024.
22. Wang A. et al. YOLOv10: Real-time end-to-end object detection, *CoRR*, 2024, Vol. abs/2405.14458. DOI: 10.48550/ARXIV.2405.14458.
23. Jocher G., Qiu J. Ultralytics YOLO11, 2024.
24. Zhao Y. et al. DETRs beat YOLOs on real-time object detection, *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*. Seattle, WA, USA, June 16–22, 2024, IEEE, 2024, pp. 16965–16974. DOI: 10.1109/CVPR52733.2024.01605.
25. Singh D. et al. PlantDoc: A dataset for visual plant disease detection, *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. New York, NY, USA, Association for Computing Machinery, 2020, pp. 249–253. DOI: 10.1145/3371158.3371196.
26. Cubuk E. D. et al. AutoAugment: Learning augmentation policies from data, *CoRR*, 2018, Vol. abs/1805.09501.
27. Bergstra J. et al. Algorithms for hyper-parameter optimization, *Advances in neural information processing systems 24: 25th annual conference on neural information processing systems 2011, Proceedings of a meeting held 12–14 December 2011*. Granada, Spain, 2011, pp. 2546–2554.
28. Akiba T. et al. Optuna: a next-generation hyperparameter optimization framework, *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*, 2019.

Received 01.08.2025.
Accepted 15.01.2026.
Published 27.03.2026.

DOI 10.15588/1607-3274-2026-1-4

УДК 004.93

МЕТОД ВИЯВЛЕННЯ ПОШКОДЖЕНЬ ТА ШКІДНИКІВ КАПУСТИ З ДАЛЕКОЇ ВІДСТАНІ НА ОСНОВІ YOLO11

Хабарлак К. С. – д-р філософії, доцент кафедри системного аналізу та управління, Національний технічний університет «Дніпровська політехніка», Дніпро, Україна. ROR: https://ror.org/05hkn5555. ORCID: https://orcid.org/0000-0003-4263-0871.

Лактіонов І. С. – д-р техн. наук, професор, професор кафедри програмного забезпечення комп'ютерних систем, Національний технічний університет «Дніпровська політехніка», Дніпро, Україна. ROR: https://ror.org/05hkn5555. ORCID: https://orcid.org/0000-0001-7857-6382.

Горєв В. М. – канд. фіз.-мат. наук, доцент, завідувач кафедри фізики, Національний технічний університет «Дніпровська політехніка», Дніпро, Україна. ROR: https://ror.org/05hkn5555. ORCID: https://orcid.org/0000-0002-9528-9497.

Дяченко Г. Г. – канд. техн. наук, доцент кафедри електропривода, Національний технічний університет «Дніпровська політехніка», Дніпро, Україна. ROR: https://ror.org/05hkn5555. ORCID: https://orcid.org/0000-0001-9105-1951.

АНОТАЦІЯ

Актуальність. Для забезпечення стабільного врожаю необхідно постійно контролювати стан рослин і вчасно вживати заходів для запобігання поширенню захворювань. Традиційні підходи базуються на ручному огляді рослин, в той же час для навчання нейронних мереж потрібні великі обсяги анотованих даних. Як ручний огляд, так і анотування даних вимагають експертних знань і потребують багато часу. Для навчання часто використовуються фотографії листя з близької відстані, оскільки їх легше знайти в Інтернеті. Однак це ускладнює оцінку поширення хвороб на великих ділянках. Капуста є однією з рослин, що широко вирощуються в Україні, але існує недостатня кількість досліджень, присвячені моніторингу здоров'я капусти.

Мета. Метою цієї роботи є створення системи виявлення хвороб і шкідників капусти на основі нейронної мережі, яку можна навчити на невеликій кількості навчальних зображень. При виведенні система повинна виявляти шкідників на зображеннях рослин на відстані цілої рослини.

Метод. З огляду на те, що існуючі набори даних про хвороби рослин, такі як IP102 і PlantDoc, містять переважно знімки хворих рослин зблизька, мережі, навчені на таких наборах даних, страждають від відсутності узагальнення до зображень на відстані. Для вибору найкращої моделі виявлення об'єктів у роботі було проаналізовано найсучасніші архітектури виявлення об'єктів, а саме YOLO 8, 9, 10, 11 і RT-DETR. Для збільшення відстані виявлення запропоновано функцію втрат на декількох зображеннях для поліпшення пошуку гіперпараметрів на основі методу Tree-Structured Parzen Estimators (TPE). Крім того, для поліпшення якості виявлення було зібрано новий набір даних хвороб капусти в Дніпропетровській області України. Нові класи включають хрестоцвітну блішку (поширений шкідник у Дніпропетровській області) та пошкоджене листя. Коли шкідника не видно, але пошкодження листя є, визначити конкретного шкідника може бути неможливо. Тому ми вводимо додатковий клас пошкодженого листя, який фіксує загальне пошкодження рослин. Це також дозволяє відстежувати швидкість одужання рослин, коли вжито заходів для зупинення поширення шкідників. Ми поєднуємо зібрані зображення з більшим набором даних IP102, щоб збільшити кількість охоплених шкідників і сформувати новий набір даних Cabbage+IP102.

Результати. 1) Пошук за допомогою TPE, використовуючи функцію втрат на кількох зображеннях, покращив YOLO 11 M з 0,3642 до 0,3892 mAP50–95 на зображеннях, зроблених на відстані. 2) Зібраний набір даних дозволив виявляти проблеми зі здоров'ям капустяних рослин на відстані, включаючи випадки, коли шкідника наразі не видно, але пошкодження є.

Висновки. У цій роботі представлено систему виявлення шкідників капусти та пошкодженого листя на основі YOLO 11 M. Архітектура детектора була обрана як найкраща під час аналізу 2 наборів даних. Розроблена система вимагає лише 7 анотованих зображень капусти для навчання та виявлення шкідників і пошкодженого листя на зображеннях високої роздільної здатності (2016x2016) цілих рослин капусти. Кінцева модель може бути використана для моніторингу проблем зі здоров'ям капусти, пошкоджень та швидкості загоєння за допомогою зображень, зроблених на відстані.

КЛЮЧОВІ СЛОВА: сільське господарство, глибоке навчання, моніторинг здоров'я рослин, виявлення шкідників, оцінка пошкодження листя, YOLO 11, родина капустяні.

ЛІТЕРАТУРА

1. A comprehensive review of recent approaches and Hardware-Software technologies for digitalisation and intellectualisation of Open-Field crop Production: Ukrainian case study in the global context / I. Laktionov [et al.] // Computers and Electronics in Agriculture. – 2024. – Vol. 225. – P. 109326. – DOI: https://doi.org/10.1016/j.compag.2024.109326.
2. Gong H. Research on a target detection algorithm for common pests based on an improved yolov7-tiny model / H. Gong, X. Ma, Y. Guo // Agronomy. – 2024. – Vol. 14, № 12. – DOI: 10.3390/agronomy14123068.
3. MSR-RCNN: a multi-class crop pest detection network based on a multi-scale super-resolution feature enhancement module / Y. Teng [et al.] // Frontiers in Plant Science. – 2022. – Vol. 13. – DOI: 10.3389/fpls.2022.810546.
4. Deep learning-based accurate detection of insects and damage in cruciferous crops using YOLOv5 / S. Chakrabarty [et al.] // Smart Agricultural Technology. – 2024. – Vol. 9. – P. 100663. – DOI: https://doi.org/10.1016/j.atech.2024.100663.
5. A new pest detection method based on improved YOLOv5M / M. Dai [et al.] // Insects. – 2023. – Vol. 14. – № 54. – DOI: 10.3390/insects14010054.
6. An improved approach to prediction of maize disease occurrence based on weather monitoring and machine learning / G. Diachenko [et al.] // Case of the forest-steppe and northern steppe of Ukraine. – 2024. – Vol. 12, № 4. – DOI: 10.22364/BJMC.2024.12.4.03.

DOI 10.15588/1607-3274-2026-1-4

7. Plant leaf disease detection, classification, and diagnosis using computer vision and artificial intelligence: a review / A. Bhargava [et al.] // IEEE access : practical innovations, open solutions. – 2024. – Vol. 12. – P. 37443–37469. – DOI: 10.1109/ACCESS.2024.3373001.
8. Khabarlak K. Feature knowledge distillation using group convolutions for efficient plant pest recognition / K. Khabarlak, G. Diachenko, I. Laktionov // Proceedings of the 12th International Conference Information Control Systems & Technologies (ICST 2024), Odesa, Ukraine, September 23–25 : CEUR workshop proceedings. – CEUR-WS.org, 2024. – Vol. 3790. – P. 377–387.
9. Rich feature hierarchies for accurate object detection and semantic segmentation / R. B. Girshick [et al.] // 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23–28, 2014. – IEEE Computer Society, 2014. – P. 580–587. – DOI: 10.1109/CVPR.2014.81.
10. Faster R-CNN: Towards real-time object detection with region proposal networks / S. Ren [et al.] // Advances in neural information processing systems 28: Annual conference on neural information processing systems 2015, December 7–12, 2015, Montreal, Quebec, Canada. – 2015. – P. 91–99.
11. You only look once: Unified, real-time object detection / J. Redmon [et al.] // 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016. – IEEE Computer Society, 2016. – P. 779–788. – DOI: 10.1109/CVPR.2016.91.
12. Bochkovskiy A. YOLOv4: Optimal speed and accuracy of object detection / A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao // CoRR. – 2020. – Vol. abs/2004.10934.
13. Jocher G. Ultralytics yolov5 / G. Jocher. – 2020.
14. End-to-end object detection with transformers / N. Carion [et al.] // ECCV 2020, Glasgow, UK, August 23–28, 2020 : Lecture notes in computer science. – Springer, 2020. – Vol. 12346. – P. 213–229. – DOI: 10.1007/978-3-030-58452-8_13.
15. IP102: A large-scale benchmark dataset for insect pest recognition / X. Wu [et al.] // 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR). – IEEE, 2019. – DOI: 10.1109/cvpr.2019.00899.
16. D3-YOLOv10: Improved YOLOv10-based lightweight tomato detection algorithm under facility scenario / A. Li [et al.] // Agriculture (Nitra, Slovakia). – 2024. – Vol. 14, № 12. – DOI: 10.3390/agriculture14122268.
17. Khabarlak K.S. Faster optimization-based meta-learning adaptation phase / K. S. Khabarlak // Radio Electronics, Computer Science, Control. – 2022. – № 1. – P. 82. – DOI: 10.15588/1607-3274-2022-1-10.
18. Nuthalapati S.V. Multi-domain few-shot learning and dataset for agricultural applications / S. V. Nuthalapati, A. Tunga // IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, QC, Canada, October 11–17, 2021. – IEEE, 2021. – P. 1399–1408. – DOI: 10.1109/ICCVW54120.2021.00161.
19. Prior knowledge auxiliary for few-shot pest detection in the wild / X. Wang [et al.] // Frontiers in Plant Science. – 2023. – Vol. 13–2022. – DOI: 10.3389/fpls.2022.1033544.
20. Jocher G. Ultralytics YOLOv8 / G. Jocher, A. Chaurasia, J. Qiu. – 2023.
21. Wang C.-Y. YOLOv9: Learning what you want to learn using programmable gradient information / C.-Y. Wang, H.-Y.M. Liao. – 2024.
22. YOLOv10: Real-time end-to-end object detection / A. Wang [et al.] // CoRR. – 2024. – Vol. abs/2405.14458. – DOI: 10.48550/ARXIV.2405.14458.
23. Jocher G. Ultralytics YOLO11 / G. Jocher, J. Qiu. – 2024.
24. DETRs beat YOLOs on real-time object detection / Y. Zhao [et al.] // IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16–22, 2024. – IEEE, 2024. – P. 16965–16974. – DOI: 10.1109/CVPR52733.2024.01605.
25. PlantDoc: A dataset for visual plant disease detection / D. Singh [et al.] // Proceedings of the 7th ACM IKDD CoDS and 25th COMAD. – New York, NY, USA: Association for Computing Machinery, 2020. – P. 249–253. DOI: 10.1145/3371158.3371196.
26. AutoAugment: Learning augmentation policies from data / E. D. Cubuk [et al.] // CoRR. – 2018. – Vol. abs/1805.09501.
27. Algorithms for hyper-parameter optimization / J. Bergstra [et al.] // Advances in neural information processing systems 24: 25th annual conference on neural information processing systems 2011. Proceedings of a meeting held 12–14 December 2011, Granada, Spain. – 2011. – P. 2546–2554.
28. Optuna: a next-generation hyperparameter optimization framework / T. Akiba [et al.] // Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining. – 2019.

DOI 10.15588/1607-3274-2026-1-4

НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

UDC 004.8:004.032.26

TUNABLE SQUASHING ACTIVATION FUNCTION FOR DEEP NEURAL NETWORKS

Shafronenko A. Yu. – Dr. Sc., Associate Professor at the Department of Informatics, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine. ROR: https://ror.org/01ctj1b90. ORCID: orcid.org/0000-0002-8040-0279.

Bodyanskiy Ye. V. – Dr. Sc., Professor at the Department of Artificial Intelligence, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine. ROR: https://ror.org/01ctj1b90. ORCID: orcid.org/ 0000-0001-5418-2143.

Shafronenko Ye. O. – Senior Lecturer at the Department of Media Engineering and Information Radio Electronic Systems, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine. ROR: https://ror.org/01ctj1b90. ORCID: orcid.org/0009-0008-0872-2274.

Brodetskyi F. A. – Senior Lecturer at the Department of Informatics, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0000-0002-0300-3886.

Tanianskyi O. S. – Post-graduate student at the Department of Informatics, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine. ROR: https://ror.org/01ctj1b90. ORCID: orcid.org/0009-0005-3491-4470.

ABSTRACT

Context. At present, artificial neural networks have become widely used to solve many problems of information processing of the most diverse nature and, above all, data mining, due to their universal approximating capabilities and the ability to learn their parameters – synaptic weights. The process of training a multilayer network consists of adjusting the synaptic weights of each neuron using the error backpropagation procedure, which is based on the chain rule of differentiation for complex functions and gradient-based optimization. Deep neural networks are based on multilayer perceptrons, which have proven their effectiveness in solving many very complex problems related to the processing and synthesis of images of various natures, natural language texts, multidimensional stochastic and chaotic sequences, including audio and video signals. Unlike classical three-layer perceptrons, DNNs contain dozens and hundreds of layers, and the number of their synaptic weights is commensurate with or even exceeds the number of synapses in the biological brain. It is clear that for these neural networks, the effect of a vanishing gradient is extremely undesirable; therefore, instead of traditional compression functions, piecewise-linear constructions are usually used here, the most popular of which is the so-called ReLU.

Objective. The purpose of the work is to introduce an adaptive activation function for deep neural networks based on the most common piecewise linear function, ReLU.

Method. A new tunable activation function for deep neural networks is proposed based on the most common piecewise linear function, ReLU, which, however, does not satisfy the conditions of G. Cybenko's approximation theorem, but provides protection for the learning process against the undesirable effect of vanishing gradients.

Results. A new adaptive piecewise linear function based on ReLU is introduced, which is both compressive and protected against vanishing gradients. In this case, during the training process, not only are the synaptic weights adjusted in the network, but also the parameters of the activation function itself. Using the proposed function allows you to reduce the number of neurons and hidden layers in the neural network, the number of required training samples, and the time required to set up the network.

Conclusions. An adaptive squashing activation function based on the widely used ReLU for deep neural networks is introduced, providing both universal approximating properties and preventing vanishing gradients. A training procedure using this function is proposed, offering high performance and a simple numerical implementation. An additional circuit for tuning the parameters of the activation functions can be quite simply introduced into existing deep neural networks that use piecewise linear activation functions.

KEYWORDS: squashing activation function, deep neural network, ReLU, gradient procedures, training signal.

ABBREVIATIONS

ANN – Artificial Neural Networks;
DNN – Deep Neural Networks;
ReLU – Rectified Linear Unit.

NOMENCLATURE

$\hat{y}_j(k)$ – output signal of the j-th neuron of the neural network at the moment of discrete time;
k – discrete time;
N – training sample size;

DOI 10.15588/1607-3274-2026-1-5

$\psi_j(\bullet)$ – nonlinear activation function of the *j*-th neuron;

$u_j(k)$ – internal activation signal;

$\gamma_j > 0$ – parameter defining the shape of the activation function;

θ_j – bias signal (threshold);

n – number of input signals (dimension of input vectors);

w_{ji} – tunable synaptic weight at the *i*-th input of the *j*-th neuron;

$x_i(k)$ – signal at the *i*-th input of the neuron at time moment k;

δ – learning rule;

$d_j(k)$ – external training signal;

β – smoothing parameter;

E – objective function;

$\eta(k)$ – non-negative learning step parameter.

INTRODUCTION

At present, artificial neural networks are widely used to solve many problems in information processing of the most diverse nature, and above all in data mining, due to their universal approximating capabilities and the ability to learn their parameters – synaptic weights. Multilayer perceptrons have received the widest distribution here. The universal approximating properties of multilayer perceptrons were proven by the theorems of G. Cybenko and K. Hornik [1, 2], using elementary F. Rosenblatt perceptrons as nodes of these neural networks with so-called squashing activation functions, among which the most common are the so-called sigmoidal functions (σ-functions) of the form [1]:

$$\hat{y}_j(k) = \psi_j(u_j(k)) = \frac{1}{1+e^{-\gamma_j u_j(k)}}, \qquad (1)$$

$\gamma_j > 0$ – is a parameter that sets the shape of the activation function and is usually chosen from purely empirical markings, although in principle it can be tuned using a gradient optimization procedure [3].

In this case, the non-linear transformation implemented by a separate neuron node can be written as

$$\hat{y}_j(k) = \psi_j\left(\theta_{jo} + \sum_{i=1}^{m} w_{ji}x_i(k)\right) = \psi_j\left(\sum_{i=1}^{m} w_{ji}x_i(k)\right) =$$
$$= \psi_j\left(w_j^T x(k)\right) = \psi_j\left(u_j(k)\right).$$

REVIEW OF THE LITERATURE

The process of training a multilayer network consists of adjusting the synaptic weights of each neuron using the error backpropagation procedure, which is based on the chain rule of differentiation of complex functions and the gradient optimization procedure (δ – learning rule).

It is easy to see that the derivative of the σ– function (1) has the form

$$\psi_j'(u_j(k)) = \gamma_j \hat{y}_j(k)\left(1-\hat{y}_j(k)\right),$$

and its value decreases as the sigmoid approaches its asymptotes –1 or +1, which is associated with the undesirable effect of the vanishing gradient, which, in turn, leads to the termination of the learning process.

Based on multilayer perceptrons, deep neural networks (DNNs) were developed [4–9], which demonstrated their effectiveness in solving many complex problems related to the processing and synthesis of images of various natures, natural language texts, multidimensional stochastic and chaotic sequences, and audio and video signals. Unlike classical three-layer perceptrons, DNNs contain dozens and hundreds of layers, and the number of their synaptic weights is commensurate with or even exceeds the number of synapses in the biological brain. It is clear that for these neural networks, the effect of a vanishing gradient is extremely undesirable, therefore, instead of traditional squashing functions, piecewise-linear constructions are usually used here, the most popular of which is the so-called ReLU (Rectified Linear Unit), which has the form:

$$\hat{y}_j(k) = \psi_j(u_j(k)) = \begin{cases} u_j, & u_j \geq 0, \\ 0 \text{ otherwise.} \end{cases} \qquad (2)$$

It is clear that there is no vanishing gradient here, and the extremely simple derivative of (2) allows accelerating the learning process of a separate neuron, since

$$\psi_j'(u_j(k)) = \begin{cases} 1, & u_j \geq 0, \\ 0 \text{ otherwise.} \end{cases}$$

At the same time, function (2) is not squashing, i.e., it does not satisfy the conditions of Cybenko's theorem, which means that to ensure the required quality of approximation, the number of neurons with such functions must be very large (recall integration by the trapezoidal method). Obviously, the gain in time when training a separate neuron is completely leveled by the huge number of these neurons in the network.

Therefore, it is advisable to introduce a piecewise-linear function (simple derivatives) that is simultaneously squashing (approximating properties) and at the same time protected from the vanishing gradient.

MATERIALS AND METHODS

Based on ReLU, the graph of which is shown in Figure 1.

DOI 10.15588/1607-3274-2026-1-5

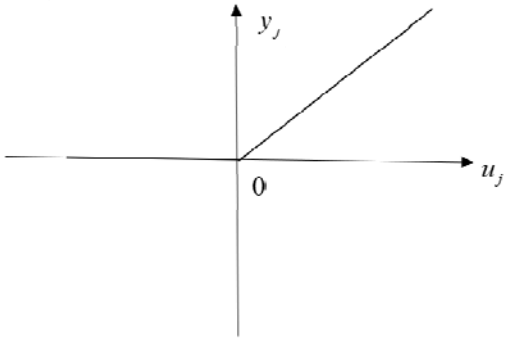


Figure 1 – ReLU Activation Function

It is not difficult to introduce a piecewise-linear function of the form:

$$\hat{y}_j(k) = \psi_j(u_j(k)) = \begin{cases} u_j, & 0 \le u_j \le 1, \\ 1, & u_j > 1, \\ 0, & u_j < 0, \end{cases}$$

the graph of which is shown in Figure 2.

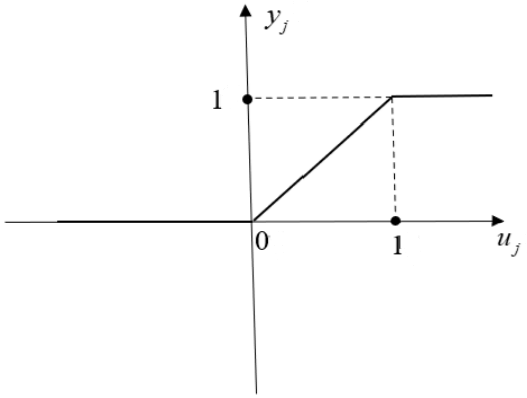


Figure 2 – Piecewise-linear squashing activation function

The use of this function in neural networks trained via gradient-based optimization procedures is impractical, as it immediately encounters the vanishing gradient problem.

Therefore, it is proposed to introduce a function of the form:

$$\hat{y}_j(k) = \psi_j(u_j(k)) = \begin{cases} u_j, & 0 \le u_j \le 1, \\ 1 - a\left(1 - u_j\right), & u_j > 1, \\ 0, & u_j < 0, \end{cases}$$

the graph of which is shown in Figure 3.

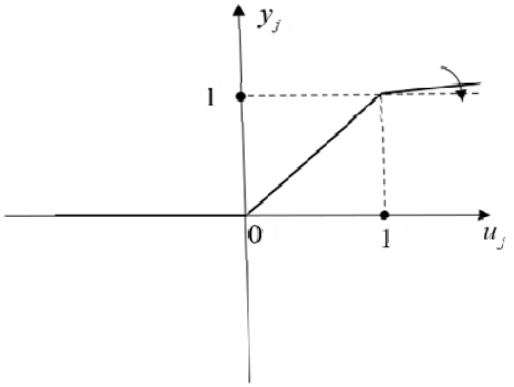


Figure 3 – Piecewise-linear activation function protected from vanishing gradient

The derivative of this function at $u_j(k) > 1$ has the form

$$\psi'_j(u_j(k)) = a,$$

at the same time, by imposing a threshold $a \ge \varepsilon$, it is possible to avoid stopping the learning process. Moreover, by simultaneously excluding the parameter and the synaptic weights, it is possible to arrive at this process.

As a target learning function, we will use the local quadratic criterion:

$$E_j(k) = \frac{1}{2} e_j^2(k) = \frac{1}{2}\left(d_j(k) - \hat{y}_j(k)\right)^2. \qquad (3)$$

The process of gradient optimization of criterion (3) by synaptic weights is implemented using the standard δ-rule [11], which in this case can be written as:

$$\begin{aligned} w_{ji}(k+1) &= w_{ji}(k) - \eta(k)\frac{\partial E_j(k)}{\partial w_{ji}} = \\ &= w_{ji}(k) - \eta(k)\frac{\partial E_j(k)}{\partial e_j(k)} \bullet \frac{\partial e_j(k)}{\partial w_{ji}} = \\ &= w_{ji}(k) - \eta(k) e_j(k)\frac{\partial e_j(k)}{\partial w_{ji}} = \\ &= w_{ji}(k) - \eta(k) e_j(k)\frac{\partial e_j(k)}{\partial u_{ji}} \bullet \frac{\partial u_j(k)}{\partial w_{ji}} = \\ &= w_{ji}(k) + \eta(k) e_j(k)\psi'\left(u_j(k)\right) x_i(k) = \\ &= w_{ji}(k) + \eta(k)\delta_j(k) x_i(k). \end{aligned} \qquad (4)$$

Procedure (4) can also be written in a simple form

$$w_{ji}(k+1) = w_{ji}(k) + \eta(k)\delta_j(k) x(k),$$

and since

DOI 10.15588/1607-3274-2026-1-5

$$\psi'_j(u_j(k)) = \begin{cases} 1, & 0 \le u_j \le 1, \\ a, & u_j > 1, \end{cases}$$

the learning process can be optimized for speed using the Kaczmarz-Withrow-Hoff algorithm [12–14], which in this case takes the form

$$w_{ji}(k+1) = w_{ji}(k) + \frac{\partial e_j(k)\psi'_j\left(u_j(k)\right)x(k)}{\|x(k)\|^2} = \\ = w_j(k) + \frac{\delta_j(k)x(k)}{\|x(k)\|^2}. \quad (5)$$

The procedure with additional smoothing properties [15, 16] in a modified form can also be used:

$$\begin{cases} w_j(k+1) = w_j(k) + r^{-1}(k)\delta_j(k)x(k), \\ r(k+1) = \beta r(k) + \|x(k)\|^2, \end{cases} \quad (6)$$

where $0 \le \beta \le 1$.

The quality of the learning process can be improved by additionally tuning the parameter a at $u_j(k) > 1$, while controlling the fulfillment of the inequality $a \ge \varepsilon$. At the same time, at each tuning cycle of the neuron, the value is first adjusted $a_j(k)$ at $u_j(k) > 1$, and then the synaptic weights are adjusted:

$$\begin{cases} a_j(k+1) = a_j(k) + \eta(k)e_j(k)\left(u_j(k) - 1\right), \\ w_j(k+1) = w_j(k) + \eta(k)e_j(k)\psi'_j\left(u_j(k)\right)x(k), \end{cases}$$

where

$$\psi'_j(u_j(k)) = \begin{cases} 1, & 0 \le u_j \le 1, \\ a_j(k+1), & u_j(k) > 1, \\ 0, & u_j(k) < 0. \end{cases}$$

Thus, a local error backpropagation procedure is implemented at the level of a single neuron. Training of the neural network as a whole is implemented using the error backpropagation rule, with an additional loop for tuning the parameters of the activation functions.

EXPERIMENTS

During the experimental research, it was decided to classify images from the “Fashion-MNIST” dataset.

Examples of images are shown in Figure 4.

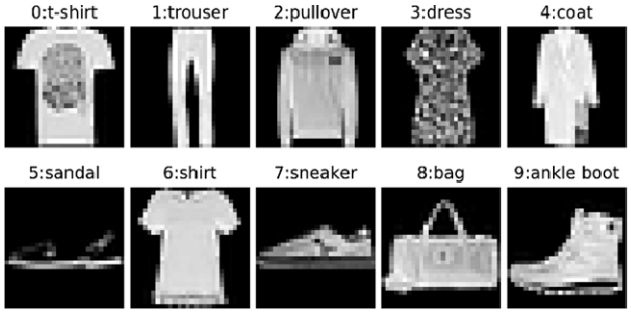


Figure 4 – Exemplars of the Fashion MNIST sample

For experimental research, the sample was previously divided into two subgroups, with the type and depth of the neural network also taken into account. Thus, in the first subgroup, the performance of activation functions is evaluated with a fixed neural network architecture and varying depth. The second subgroup evaluates activation functions based on different types of the most popular DNNs.

Note also that CNN is used across 3 depth variants: CNN1, CNN2, and CNN3. Comparative analysis was carried out on 4 activation functions. The CNN1 architecture consists of two convolution layers, one fully connected layer, and one softmax classification layer. Similarly, the structure of CNN2 is arranged the same as CNN1, but with two additional convolutional layers. CNN3 is deeper, with six convolutional layers and two additional convolutional layers. All experiments account for the average performance across five data runs.

Comparative accuracy results are shown in Table 1. The last column of Table 1 displays the average accuracy for each activation function across all three CNN models.

Table 1 – Accuracy testing results on Fashion-MNIST in %

Activation Function	CNN 1	CNN 2	CNN 3	Average Accuracy
ReLU	**93.5**	93.9	94.2	93.9
LeakyReLU	93	92.9	**93.7**	93.2
Piecewise linear ReLU	**93.5**	**94.2**	94.3	**94**

Let’s conduct an experiment with activation functions on various models of popular DNNs: ResNet56V1 (1), ResNet56V2 (2), and ResNet110 (3). First, image benchmarks are tested at different network depths. Secondly, benchmark images are checked on other models of common architectures. The results of the experiments are shown in Table 2.

Table 2 – Accuracy testing results on Fashion-MNIST in %

Activation Function	1	2	3	Average Accuracy
ReLU	88.9	90.1	90.4	89.8
LeakyReLU	89	90.6	90.3	90
Piecewise linear ReLU	**90**	**91.2**	**90.8**	**90.6**

DISCUSSION

The first stage of the analysis aims to evaluate the overall performance of all activation functions across image and text classification tasks. The second stage aims to determine the most successful activation functions across all launched models. Experimental studies show that the proposed adaptive piecewise-linear function achieves the

DOI 10.15588/1607-3274-2026-1-5

best accuracy in CNN1 and CNN2 and is independent of the DNN model type, as demonstrated in Table 2.

In the first stage of analysis, we average each activation function across all launched models using general image benchmarks. Figure 5 summarizes the average accuracy for all image benchmarks.

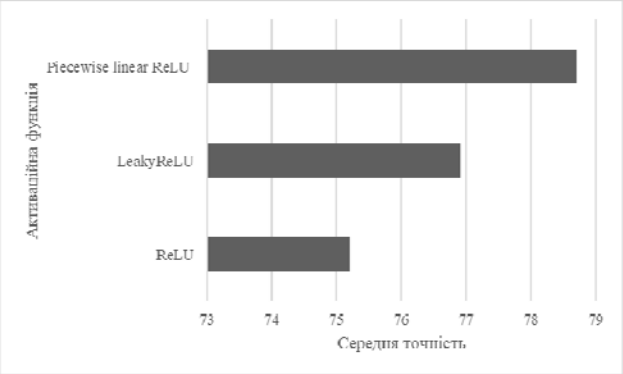


Figure 5 – Average accuracy value of activation functions compared to others, using different popular networks based on the "Fashion MNIST" sample

In the second stage of analysis, an idea of the most successful activation function for each model was obtained. Table 3 shows the most successful combinations between activation functions and launched models.

Table 3 – Successful activation functions for each deep network

Activation Function	CNN1	CNN2	CNN3	1	2	3
ReLU	+					
LeakyReLU			+			
Piecewise linear ReLU	+	+		+	+	+

CONCLUSIONS

An adaptive squashing activation function based on the widely used ReLU for deep neural networks is introduced, which simultaneously provides universal approximating properties while preventing the undesirable vanishing gradient problem. A training procedure using this function is proposed, ensuring high speed and a simple numerical implementation. An additional circuit for tuning the parameters of activation functions can be quite simply introduced into existing deep neural networks that use piecewise-linear activation functions.

Scientific novelty: A training procedure using an adaptive activation function for deep neural networks is proposed, ensuring high speed and a simple numerical implementation.

Practical significance: Using the proposed function reduces the number of neurons and hidden layers in the neural network, the required volume of training samples, and the time required to set up the network.

Prospects for further research: Fast neural networks for pattern/image recognition for a wide class of practical tasks in Data Stream Mining and Big Data Mining.

ACKNOWLEDGEMENTS

The work is supported by the state budget scientific research project of Kharkiv National University of Radio Electronics "Adaptive bagging of hybrid computational intelligence systems based on speed-optimal online learning" (SR No. 0124U000363).

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Author's contributions: A. Shafronenko to introduce an adaptive activation function for deep neural networks based on the most common piecewise linear function, ReLU; Ye. Bodyanskiy proposed a gradient optimization process for the quadratic criterion over synaptic weights; F. Brodetskyi adjusted the parameter that determines the shape of the activation function; Ye. Shafronenko evaluated the overall performance of activation functions across image and text classification tasks; O. Tanianskyi conducted experimental studies using activation functions with various models of popular GNMs.

Data availability: The manuscript has associated data in a data repository https://openarchive.nure.ua/.

Software availability: The manuscript has no associated software.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

REFERENCES

1. Cybenko G. Approximation by superpositions of a sigmoidal function, *Mathematics of control, signals and systems*, 1989, 2.4, pp. 303–314.
2. Hornik K., Stinchcombe M., White H. Multilayer feedforward networks are universal approximators, *Neural Networks*, 1989, 2.5, pp. 359–366.
3. Hornik K. Approximation capabilities of multilayer feedforward networks, *Neural Networks*, 1991, 4.2, pp. 251–257.
4. Poggio T., Girosi F. Networks for approximation and learning, *Proceedings of the IEEE*, 1990, Vol. 78, № 9, pp. 1481–1497.
5. Haykin S. Neural networks: a comprehensive foundation. Prentice Hall PTR, 2004. Vol. 2. 1994.
6. Vapnik V. N. The Nature of Statistical Learning Theory. New York, Springer, 1995.
7. Cortes C. and Vapnik V. Support-vector networks, *Machine Learning*, Sep. 1995, Vol. 20, No. 3, pp. 273–297, https://doi.org/10.1007/bf00994018.
8. Bodyanskiy Ye., Zaychenko Yu. and Hamidov G. Hybrid Deep Learning Networks Based on Self-Organization and their Applications. Cambridge Scholars Publishing, 2024.
9. Kaczmarz S. Approximate solution of systems of linear equations, *International Journal of Control*, 1993, No. 57 (6), pp. 1269–1271.
10. Widrow B. and Hoff M. E. Adaptive switching circuits, *1960 IRE WESCON Convention Record*, 1960, pp. 96–104.
11. Bodyanskiy Ye., Kolodyazhniy V. and Stephan A. An adaptive learning algorithm for a neuro-fuzzy network, *International Conference on Computational Intelligence*. Berlin, Heidelberg, Springer Berlin Heidelberg, 2001, pp. 68–75.

Received 08.10.2025.
Accepted 30.01.2025.
Published 27.03.2026.

DOI 10.15588/1607-3274-2026-1-5

УДК 004.8:004.032.26

НАЛАШТОВНА СТИСКАЮЧА АКТИВАЦІЙНА ФУНКЦІЯ ДЛЯ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ

Шафроненко А. Ю. – д-р техн. наук, доцент кафедри інформатики Харківського національного університету радіоелектроніки, Харків, Україна. ROR: https://ror.org/01ctj1b90. ORCID: orcid.org/0000-0002-8040-0279.

Бодянський Є. В. – д-р техн. наук, проф., проф. кафедри штучного інтелекту, Харківський національний університет радіоелектроніки, Харків, Україна. ROR: https://ror.org/01ctj1b90. ORCID: orcid.org/ 0000-0001-5418-2143.

Шафроненко Є. О. – старший викладач кафедри медіаінженерії та інформаційних радіоелектронних систем, Харківський національний університет радіоелектроніки, Харків, Україна. https://ror.org/01ctj1b90. ORCID: orcid.org/0009-0008-0872-2274.

Бродецький Ф. А. – старший викладач кафедри інформатики, Харківський національний університет радіоелектроніки, Харків, Україна. ROR: https://ror.org/01ctj1b90. ORCID: https://orcid.org/0000-0002-0300-3886.

Танянський О. С. – аспірант кафедри інформатики, Харківський національний університет радіоелектроніки, Харків, Україна. ROR: https://ror.org/01ctj1b90. ORCID: orcid.org/0009-0005-3491-4470.

АНОТАЦІЯ

Актуальність. На цей час штучні нейронні мережі отримали широке поширення для вирішення багатьох задач опрацювання інформації найрізноманітнішої природи і, перш за все, інтелектуального аналізу даних, завдяки своїм універсальним апроксимуючим можливостям і здатності до навчання своїх параметрів – синаптичних ваг. Процес навчання багатошарової мережі полягає у налаштуванні синаптичних ваг кожного нейрона за допомогою процедури зворотного поширення похибок, яка базується на ланцюговому правилі диференціювання складних функцій та градієнтної процедури оптимізації. На основі багатошарових персептронів були створені глибокі нейронні мережі, що довели свою ефективність при вирішенні багатьох дуже складних задач, пов'язаних з обробкою і синтезом зображень різноманітної природи, природномовних текстів, багатовимірних стохастичних і хаотичних послідовностей, включаючи аудіо та відеосигнали. На відміну від класичних трьошарових персептронів ГНМ містять десятки та сотні шарів, а кількість їх синаптичних ваг є співрозмірною або навіть перевищує кількість синапсів у біологічному мозку. Зрозуміло, що для цих нейронних мереж ефект зникаючого градієнта є вкрай небажаним, тому замість традиційних стискаючих функцій тут використовуються зазвичай кусково-лінійні конструкції, найбільш популярною з яких є, так звана, ReLU.

Мета. Мета роботи полягає у запровадженні адаптивної активаційної функції для глибоких нейронних мереж на основі найбільш розповсюдженої кусково-лінійної функції ReLU (Piecewise linear ReLU).

Метод. Запропонована нова налаштовна активаційна функція для глибоких нейронних мереж на основі найбільш розповсюдженої кусково-лінійної функції ReLU, яка однак не відповідає умовам апроксимаційної теореми Дж. Цибенка, але забезпечує захист процесу навчання від небажаного ефекту зникаючого градієнта.

Результати. Введено нову адаптивну кусково-лінійну функцію на основі ReLU (Piecewise linear ReLU), що одночасно є як стискаючою, так і захищеною від зникаючого градієнта. При цьому у процесі навчання у мережі налаштовуються не лише синаптичні ваги, але і параметри самої активаційної функції. Використання запропонованої функції дозволяє скоротити кількість нейронів та прихованих шарів у нейронній мережі, необхідний обсяг навчальних вибірок та час налаштування мережі в цілому.

Висновки. Введено у розгляд адаптивну стискаючу активаційну функцію на основі широко поширеної ReLU для глибоких нейронних мереж, що одночасно забезпечує універсальні апроксимуючі властивості і в той же час мережа не потерпає від небажаного ефекту зникаючого градієнта. Запропонована процедура навчання з використанням цієї функції, що забезпечує високу швидкодію та характеризується простотою чисельної реалізації. Додатковий контур налаштування параметрів активаційних функцій досить просто може бути введений у вже існуючі глибокі нейронні мережі, що використовують кусково-лінійні активаційні функції.

КЛЮЧОВІ СЛОВА: адаптивна стискаюча активаційна функція, глибока нейронна мережа, ReLU, градієнтні процедури, навчальний сигнал.

ЛІТЕРАТУРА

1. Cybenko G. Approximation by superpositions of a sigmoidal function / G. Cybenko // Mathematics of control, signals and systems, – 1989. – 2.4. – P. 303–314.
2. Hornik K. Multilayer feedforward networks are universal approximators / K. Hornik, M. Stinchcombe, H. White // Neural networks. – 1989. – 2.5. – P. 359–366.
3. Hornik K. Approximation capabilities of multilayer feedforward networks / K. Hornik // Neural networks. – 1991. – 4.2. – P. 251–257.
4. Poggio T. Networks for approximation and learning / T. Poggio, F. Girosi // Proceedings of the IEEE. – 1990. – T. 78, № 9. – P. 1481–1497.
5. Haykin S. Neural networks: a comprehensive foundation / S. Haykin. – Prentice Hall PTR, 2004. – T.2. –1994.
6. Vapnik V. N. The Nature of Statistical Learning Theory / V. N. Vapnik. – New York : Springer, 1995.
7. Cortes C. Support-vector networks / C. Cortes and V. Vapnik // Machine Learning. – Sep. 1995. – Vol. 20, No. 3. – P. 273–297, https://doi.org/10.1007/bf00994018.
8. Bodyanskiy Ye. Hybrid Deep Learning Networks Based on Self-Organization and their Applications. / Ye. Bodyanskiy, Yu. Zaychenko and G. Hamidov. – Cambridge Scholars Publishing, 2024.
9. Kaczmarz S. Approximate solution of systems of linear equations / S. Kaczmarz // International Journal of Control. – 1993. – No. 57 (6). – P. 1269–1271.
10. Widrow B. Adaptive switching circuits / B. Widrow and M. E. Hoff // In 1960 IRE WESCON Convention Record. – 1960. – P. 96–104.
11. Bodyanskiy Ye. An adaptive learning algorithm for a neuro-fuzzy network / Ye. Bodyanskiy, V. Kolodyazhniy and A. Stephan // International Conference on Computational Intelligence. – Berlin, Heidelberg : Springer Berlin Heidelberg, 2001. – P. 68–75.

DOI 10.15588/1607-3274-2026-1-5

UDC 004.93

HYBRID SATELIT IMAGE RECOGNITION SYSTEM COMBINING NEURAL NETWORK FEATURE EXTRACTION AND AN INFORMATION-EXTREMAL CLASSIFIER

Dovbysh A. S. – Dr. Sc., Professor, Professor of the Department of Computer Science, Sumy State University, Sumy, Ukraine. ROR: https://ror.org/01w60n236. ORCID:https://orcid.org/0000-0003-1829-3318.

Piatachenko V. Y. – PhD, Assistant of the Department of Computer Science, Sumy State University, Sumy, Ukraine. ROR: https://ror.org/01w60n236. ORCID: https://orcid.org/0000-0002-7464-3119.

Serhieiev V. M. – Post-graduate student of the Department of Computer Science, Sumy State University, Sumy, Ukraine. ROR: https://ror.org/01w60n236. ORCID:https://orcid.org/0009-0009-3838-0153.

Hrytsenko O. M. – Post-graduate student of the Department of Computer Science, Sumy State University, Sumy, Ukraine. ROR: https://ror.org/01w60n236. ORCID: https://orcid.org/0009-0004-1382-708X.

ABSTRACT

Context. The study solves the relevant task of developing an interpretable and adaptive recognition system for semantic segmentation of satellite imagery by integrating neural network feature extractors with an information-extreme classifier.

Objective. To improve the accuracy of satellite land cover classification by developing a hybrid machine learning model that combines a deep convolutional neural network for extracting informative features with an information-extreme classifier, enabling the construction of highly reliable decision rules even in the presence of overlapping recognition classes in the feature space.

Method. A hybrid model is proposed that combines efficient spatial feature extraction using a convolutional neural network (CNN) with an information-extreme intelligent (IEI) technology for data analysis, based on maximizing the information capacity of the recognition system during machine learning. For feature aggregation, GlobalAveragePooling is applied instead of the classical Flatten operation. Additionally, regularization techniques such as weight decay and cyclical learning rate scheduling are implemented. The optimization of IEI model parameters is carried out using a modified Kullback information criterion, interpreted as a measure of recognition class diversity.

Results. The developed model achieves high classification accuracy (95%) on the test set and demonstrates stable performance, with improved efficiency of the neural feature extractor due to a reduced number of training epochs enabled by regularization techniques. As a result of the information-extreme machine learning process, the optimal geometric parameters of the hyperspherical recognition class containers were determined, allowing the construction of highly reliable decision rules even under conditions of recognition class overlap in the feature space.

Conclusions. The proposed hybrid model enables the construction of highly reliable decision rules through information-extreme machine learning, even in cases of a priori fuzzy partitioning of recognition classes in the feature space, based on the input training matrix formed during feature extraction.

KEYWORDS: information-extreme machine learning, convolutional neural network, information criterion, optimization, hybrid model, recognition feature extraction, land cover images.

ABBREVIATIONS

CNN is a Convolutional Neural Network;

IEI is an Information-Extreme Intellectual (Machine) Learning;

ReLU is a Rectified Linear Unit;

Swish is a Sigmoid-Weighted Linear Unit;

AdamW is an Adaptive Moment Estimation with Weight Decay;

SVM is a Support Vector Machine;

k-NN is a k-Nearest Neighbors;

GAP is a Global Average Pooling.

NOMENCLATURE

M is a number of recognition classes;

N is a number of recognition features;

n is a number of feature vectors of recognition classes in the training matrix;

x_m is an averaged binary feature vector of the recognition class X_m^0;

Z is a recognition feature space;

X is a working binary training matrix;

f_1 is an operator for forming matrix Y;

f_2 is an operator for forming matrix X;

G_E is an admissible domain of the information criterion function used for machine learning parameter optimization;

$K_{1,m}(d)$ is a number of events where realizations of class X_m^0 are mistakenly not assigned to their own class;

$K_{2,m}(d)$ is a number of events where alien realizations are mistakenly assigned to class X_m^0;

10^{-r} is a sufficiently small number introduced to avoid division by zero;

$\overline{D_1}$ is an average first-order confidence across the recognition class alphabet;

$\bar{\beta}$ is an average second-kind error across the recognition class alphabet;

$\overline{x_i}^B$ is an averaged structured binary feature vector defining the geometric center of the container for class X_i;

DOI 10.15588/1607-3274-2026-1-6

R_i is a radius of the hyperspherical container for class X_i;

δ_i is a parameter equal to half the tolerance field width for recognition features;

$\hat{\delta}$ is a normalized tolerance field over the recognition features;

D_{ij} is an inter-center Hamming code-based distance between class X_i and its nearest neighbor X_j in the binary feature space;

w is an image width;

h is an image height;

c is an image channel;

W is a convolution kernel;

k is a filter size;

y_i is a ground truth label;

$\hat{y}_i$ is a model prediction;

λ is a regularization coefficient.

INTRODUCTION

An important research direction in the field of intelligent image analysis is the investigation of effective approaches to feature extraction, which enhances the informativeness of the input mathematical representation and enables high-confidence classification decisions in visual recognition tasks. These tasks have broad applications in areas such as remote sensing, autonomous navigation, environmental monitoring, agroanalytics, military intelligence, and others. In particular, satellite imagery plays a key role in identifying land cover types, detecting changes in terrain structure, and classifying infrastructure-related objects.

Despite the widespread use of convolutional neural networks (CNNs) in image processing tasks, they have a number of significant limitations. For instance, their decisions are often difficult to interpret, which complicates their use in sensitive domains where transparency is required. In addition, neural networks demand large amounts of training data and are inflexible during retraining, especially when the number of machine learning objects increases under conditions of substantial overlap in the recognition feature space.

To overcome these limitations, hybrid models that combine the representational power of deep neural networks with the strengths of classical decision-making frameworks offer a promising solution. These strengths include interpretability, robustness, and statistical consistency. One such framework is the information-extreme intelligent technology (IEIT), which is based on maximizing the information capacity of the recognition system during training. The integration of CNNs for feature extraction with IEIT as a transparent classification mechanism provides a foundation for the development of flexible, adaptive, and controllable next-generation artificial intelligence systems.

The object of study is the process of land cover image classification based on satellite data.

Unlike traditional neural networks, the information-extreme intelligent (IEI) technology implements a geometric representation of classes in the form of containers within the feature space and enables the construction of decision rules that are invariant to the dimensionality and distribution of features. When combined with a convolutional neural network (CNN) that performs preliminary processing of input images and generates a descriptive feature representation, this approach offers flexible control over the trade-off between model accuracy, generalization, and interpretability.

The subject of study is a hybrid machine learning model that integrates a convolutional neural network with an information-extreme approach to recognition.

The work investigates the architecture of the hybrid model, develops algorithms for constructing recognition class containers, and conducts an experimental evaluation of model performance using the EuroSAT image dataset. Special attention is given to analyzing the role of the CNN as a feature extraction mechanism and examining the influence of geometric parameters of the categorical model on decision-making within the IEI framework.

The purpose of the work is to increase the accuracy and interpretability of satellite image classification by developing a hybrid machine learning model.

1 PROBLEM STATEMENT

Let us consider a formalized formulation of the information-extreme machine learning problem for image classification using a hybrid model. Let X be the alphabet of recognition classes corresponding to image categories (in our case, types of terrain). Each image X is input into a convolutional neural network (CNN), which performs feature extraction and transforms the image into a feature vector $z \in \Re^n$ in the feature space. The collection of such vectors forms a three-dimensional brightness matrix Z of dimensions $w \times h \times c$, which serves as the input to the information-extreme classifier. For each recognition class $X_i \in X$ the machine learning parameters are defined in the radial basis of the Hamming feature space:

$$\Omega_i = (\bar{x}_i^B, R_i, \delta_i). \quad (1)$$

At the same time, the machine learning parameters are subject to the following constraints:

1) the value of the radius R_i must satisfy the inequality

$$R_i \le D_{ij},$$

2) the parameter δ_i is constrained to lie within the domain

$$\delta_i \le \hat{\delta}.$$

DOI 10.15588/1607-3274-2026-1-6

During the machine learning process, the following steps must be performed:

1) In accordance with the concept of the information-extreme intelligent (IEI) technology, the feature matrix Z must be transformed into a binary working matrix X^B by quantizing the features according to control tolerance levels. This transformation allows the model to adapt to the condition of maximizing the complete probability of correct classification decisions.

2) Based on the optimized parameters of the containers, decision rules must be constructed in the form of categorical mappings. These rules enable the determination of the class membership of an input feature vector z using its binary representation in the Hamming feature space (1).

3) When the recognition system operates in examination mode, the functional performance of the hybrid model must be validated by evaluating the classification accuracy and the statistical consistency of the decision rules generated during the training phase.

2 REVIEW OF THE LITERATURE

In recent years, the increasing complexity of recognition tasks has fueled growing interest in hybrid approaches within neural network models, particularly those that combine convolutional neural networks (CNNs) with classical machine learning methods. Such integration contributes to improved classification accuracy, especially under conditions of limited computational resources.

In computer vision tasks, studies on the effectiveness of hybrid models [1–3] have shown that employing CNNs for feature extraction followed by classification using linear algorithms can significantly reduce computational complexity without a substantial loss in accuracy. This highlights the potential of hybrid methods for deployment in resource-constrained environments.

The work presented in [2] proposes hybrid CNN-SVM and CNN-KNN architectures for image classification. These approaches treat the CNN as a feature extractor, combined with either a support vector machine (SVM) or k-nearest neighbors (k-NN) classifier, and demonstrate higher accuracy compared to standard CNNs. This approach has proven particularly effective when working with small datasets, as linear classification algorithms can generalize extracted features more efficiently in such scenarios.

The study presented in [3] demonstrated that, following deep feature extraction via a convolutional neural network (CNN), the use of k-nearest neighbors (k-NN) can enhance recognition accuracy by reducing the number of misclassifications.

In parallel with hybrid approaches, recent literature has devoted considerable attention to increasing model robustness against adversarial attacks, concept drift, and data imperfections. In [4, 5], the authors propose resilient classifier architectures that implement specialized training mechanisms to adapt to variations in the structure of input data. Specifically, [4] describes a method for constructing a resilient classifier capable of handling concept shift and fault injection, while [5] introduces a training architecture and algorithm that explicitly accounts for these factors.

Another research direction that has garnered interest in the context of interpretable models is information-extreme machine learning. This approach is based on maximizing the information capacity of the recognition system during training. Studies [6, 7] successfully apply the information-extreme intelligent (IEI) technology to photographic and video analytics tasks. Both examples demonstrate the advantages of geometric class representation and the use of an information criterion as a training objective function.

The direct foundation for this study lies in [8, 9], where information-extreme algorithms are proposed for onboard recognition systems targeting ground objects. In particular, [8] introduces a mechanism for constructing recognition class containers in the feature space, followed by optimization of their geometric parameters. Study [9] presents an extension of this model that enables selection of a base recognition class, improving the functional performance of multiclass machine learning.

The hybrid model proposed in this paper may be applied to solve the problem of deep neural feature extraction for recognition tasks.

3 MATERIALS AND METHODS

The concept of hybrid models is well established and widely supported in the literature [10, 11], as it leverages the ability of neural networks to effectively extract informative features from data while preserving the robustness and interpretability of classical recognition algorithms.

Accordingly, this study employs a deep convolutional neural network (CNN) to perform efficient extraction of informative recognition features from images for their subsequent use in an information-extreme classifier (Figure 1).

The model consists of three convolutional blocks, each comprising a convolutional layer, normalization (to stabilize the distribution of activations), and a subsampling operation that reduces dimensionality and enhances shift invariance.

To improve parametric efficiency, depthwise separable convolutions are employed, which factorize the standard $k \times k$ convolutional kernel into separate depthwise and pointwise operations. This reduces the number of parameters from $O(k^2 C_{in} C_{out})$ to $O(k^2 C_{in} + C_{in} C_{out})$. Formally, the spatial convolution operation for each channel is defined by Equation (2):

$$Y_{i,j,c} = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} X_{i+m,j+n,c} W_{m,n,c} + b_c \,, \qquad (2)$$

DOI 10.15588/1607-3274-2026-1-6

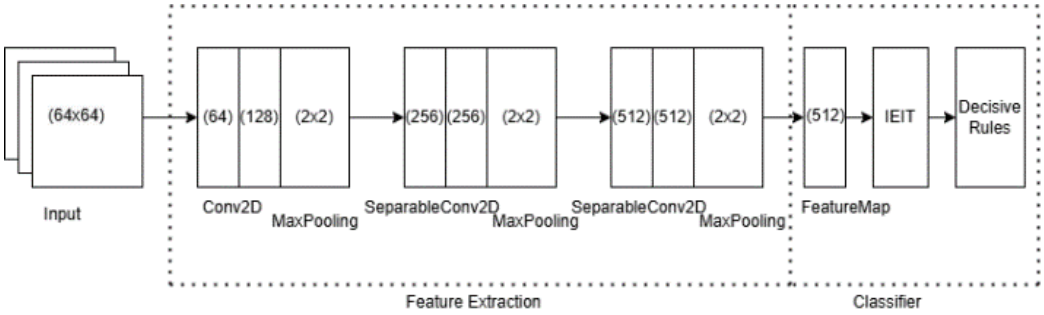


Figure 1 – Hybrid machine learning model architecture for the recognition system

The use of a global average pooling layer (GlobalAveragePooling2D) allows the aggregation of information across the entire activation map, significantly reducing the number of trainable parameters and lowering the risk of overfitting. Equation (3) expresses the computation of the average feature value for each channel *c*:

$$y_c = \frac{1}{H \times W} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} X_{i,j,c} \,. \tag{3}$$

The activation function used is Swish [12], which demonstrates improved properties compared to ReLU due to the absence of “dead neurons” and its smooth transition between activation regions:

$$\phi(x) = x \cdot \sigma(\beta x), \quad \sigma(x) = \frac{1}{1+e^{-x}}. \tag{4}$$

Model weights are optimized using the AdamW algorithm. The total loss function, incorporating L2 regularization, is defined as:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) + \lambda \sum_{j} \left\| w_j \right\|^2 \,. \tag{5}$$

The overall structure of the model is presented as a sequential architecture that combines a convolutional feature extractor with an information-extreme classifier (Figure 1).

At the feature extraction stage, three convolutional blocks are employed: the first uses a standard Conv2D operation, while the subsequent two utilize depthwise separable convolutions (SeparableConv2D), which reduce the number of parameters without compromising feature informativeness. Each convolutional block is followed by feature downsampling using MaxPooling.

In the final stage, features are aggregated into a fixed-length vector using GlobalAveragePooling2D, resulting in a 512-dimensional feature map. This vector is passed to the input of the information-extreme classifier (IEC), which constructs a relevant input training matrix, transforms it into a working binary matrix, and – by optimizing the machine learning parameters using an information criterion – reconstructs recognition class containers in the Hamming feature space.

At the output, decision rules are formed based on the optimized geometric parameters of the recognition class containers obtained during the training process.

To formalize the process of information-extreme machine learning, we present it as a functional categorical model that reflects the relationships among the sets involved in constructing optimal recognition class containers. This approach provides a clear representation of how the transition from input features to classification decisions is performed, taking into account both geometric and informational criteria.

Figure 2 illustrates the functional categorical model of information-extreme machine learning in the form of a directed graph, where the terminal set *E* of information criterion values for optimizing machine learning parameters is shared across all optimization loops.

At each step of the learning process, the operator ξ reconstructs recognition class containers in the radial basis of the feature space, forming, in the general case, a fuzzy partition $\tilde{\tilde{\Re}}^{|M|}$.

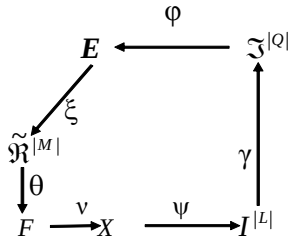


Figure 2 – Categorical model of information-extreme machine learning

The operator θ projects the constructed fuzzy partition $\tilde{\tilde{\Re}}^{|M|}$ onto the distribution of binary feature vectors of the binary training matrix *X*, while the operator ψ tests the main statistical hypothesis regarding the membership of feature vectors to the corresponding recognition class. Based on the results of statistical hypothesis testing, the set of statistical hypotheses $I^{|L|}$ is formed, and the operator γ generates the set of accuracy metrics $\Im^{|Q|}$, where $Q = L^2$. The set F corresponds to the feature vectors extracted by the convolutional neural network from raw input data. The operator ν performs the

DOI 10.15588/1607-3274-2026-1-6

transformation of features into a binary form *X*, which is subsequently used to construct fuzzy clusters $\tilde{\Re}^{|M|}$. The operator φ computes the set *E* of information criterion values used for optimizing the machine learning parameters.

The mappings between the sets in the graph correspond to the individual stages of the algorithm: starting with feature extraction, followed by binary vector construction, computation of informational measures, optimization of container parameters (radii, centers, tolerances), and culminating in the decision-making process regarding the class membership of a given object. This form of representation clearly delineates the logical structure of the system and the interrelationships between its components.

According to the categorical model (Figure 2), the information-extreme machine learning algorithm can be formalized as the following procedure:

$$P^* = \arg\max_{G_P} \{\max_{G_R \cap \{S\}} \overline{E}_S\}. \quad (6)$$

Within the framework of the information-extreme intelligent (IEI) technology, machine learning under binary decision-making and equally probable a priori hypotheses is implemented through a targeted search for the global maximum of the information criterion, averaged over the alphabet of recognition classes. As the optimization criterion, we consider a modified version of the Kullback information measure, proposed by the authors, in the following form:

$$E_M^{(k)} = \frac{1}{n_m}[K_{1,m}^{(k)} - K_{2,m}^{(k)}] \times \log_2[\frac{10^{-\omega} + n_m + [K_{1,m}^{(k)} - K_{2,m}^{(k)}]}{10^{-\omega} + n_m - [K_{1,m}^{(k)} - K_{2,m}^{(k)}]}]. \quad (7)$$

Thus, information-extreme machine learning consists in optimizing the geometric parameters of recognition class containers, reconstructed in the Hamming feature space, according to the information criterion.

4 EXPERIMENTS

To solve the problem of land cover image classification, data from the open EuroSAT dataset [13] were used. This dataset represents a collection of images corresponding to different types of terrain and land cover, based on Sentinel-2 satellite imagery obtained under the Earth observation program Copernicus.

The images used in this study were selected based on a proximity criterion, which a priori guaranteed their overlap in the recognition feature space, thereby determining the need for deep machine learning.

An alphabet consisting of six recognition classes was constructed, characterizing the following types of digital terrain image frames: “Annual Crop”, “Forest”, “Highway”, “Residential”, “Pasture” and “Industrial” (Figure 3). Each recognition class was characterized by a set of 2,000 images, each with a resolution of 64 by 64 pixels.

Figure 3 – Matrix of representative images for the recognition classes

All images were pre-sorted by class to ensure uniform representation of each category within the resulting subsets. The dataset was then stratified and divided into three parts: a training set (70% of the data), a validation set (15%), and a test set (15%), while preserving the proportional distribution of classes. This approach not only ensured balanced subsets but also minimized the risk of model overfitting on dominant classes.

Image augmentation (e.g., geometric transformations, brightness adjustments, flipping, etc.) was not applied, as the goal of the study was not to increase model robustness to variations in input data, but rather to evaluate the effectiveness of the hybrid feature extraction and classification approach under controlled conditions.

During the training stage of the feature extractor, a convolutional neural network consisting of three convolutional blocks with normalization and subsampling operations was used. After the final convolutional layer, GlobalAveragePooling2D was applied to aggregate spatial information across the entire activation map and to generate a fixed-size feature vector.

The model was trained for 50 epochs using the AdamW optimizer, which combines adaptive weight

DOI 10.15588/1607-3274-2026-1-6

updates with weight decay (L2 regularization), with a batch size of 32. The Swish activation function was used, as it provides smooth transitions across activation regions and eliminates the “dead neuron” problem.

After completing the training stage of the CNN feature extractor, the features obtained for each image were quantized into binary form based on a system of control tolerances. In this way, a binary training matrix of features was formed, which served as input data for the information-extreme classifier.

Within this approach, optimization of the geometric parameters of the class containers (radius, center, and tolerance field) was carried out based on the authors’ modified version of the Kullback information criterion, taking into account statistical hypotheses regarding the belonging of realizations to their corresponding recognition classes.

To ensure reproducibility of the experiment, the random number generator was fixed, and identical initial conditions were applied at each model run.

To evaluate the effectiveness of the hybrid model, training was conducted for a convolutional neural network (CNN), which was used to extract features from the input images. Based on the resulting feature vectors, two types of classifiers were built: SVM and k-NN, operating on the feature representations formed by the CNN.

In order to enable a correct comparison of results, additional SVM and k-NN classifiers were implemented, which operated directly on the raw pixel values (flattened RGB images of size 64×64) without prior CNN-based feature extraction.

During the training stage of the CNN, a model was used consisting of three convolutional blocks with normalization and subsampling, followed by a GlobalAveragePooling2D layer to convert the spatial representation into a fixed-length feature vector.

The resulting feature vector was passed through a fully connected layer of size 512 with a Swish activation function and Dropout regularization (rate = 0.5).

The network was trained for 50 epochs using a batch size of 32, the AdamW optimizer (learning rate = 0.001, weight decay = 1e-4), and the sparse categorical crossetropy loss function.

For the purpose of comparative analysis of different classification algorithms, two experimental groups were formed, differing in the way the input data were preprocessed:

1. Baseline models: The SVM and k-NN classifiers were applied directly to the input images represented as unstructured vectors of pixel values (flattened RGB), without any prior feature extraction.

2. Hybrid models: The same classification algorithms (SVM, k-NN), as well as the information-extreme intelligent technology classifier (IEIT), were applied to feature vectors that had been previously extracted by the convolutional neural network (CNN).

This design enabled the evaluation of how prior deep image processing affects the accuracy and stability of classification.

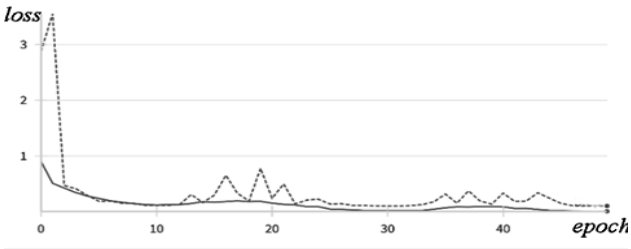
To ensure the reproducibility of the experiments, the random number generator was initialized with a fixed seed (random_state = 42).

This guarantees identical outcomes across repeated runs, as all procedures involving randomness – such as the stratified splitting of the dataset into training, validation, and test sets, model parameter initialization, and stochastic elements of training – were conducted under consistent conditions.

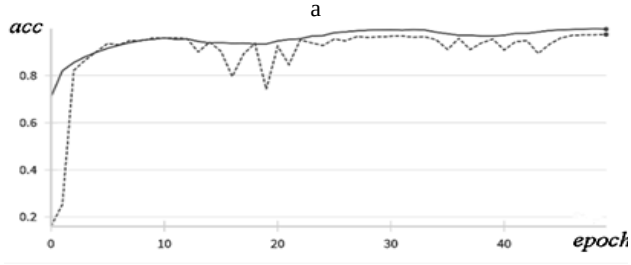
5 RESULTS

The software implementation of the hybrid algorithm was designed as a sequential procedure that includes training an artificial neural network to extract significant features from the training matrix and constructing decision rules based on the optimal geometric parameters of the recognition class containers, which were obtained during the process of information-extreme machine learning.

Figure 4 shows the plots illustrating the model’s error reduction and accuracy as functions of training epochs. The analysis of Figure 4 reveals that at the early stages of machine learning, a significant gap is observed between the training accuracy (60.82%) and the validation accuracy (15.99%), which indicates the initial adaptation phase of the model before generalization is formed.



a



b

Figure 4 – Results of machine learning: a – model error reduction; b – model accuracy.

As the training progresses and the number of epochs increases, a trend toward stabilization is observed, and the model reaches a performance plateau at a level of 96–97% correct classification decisions.

At the same time, after the 35th epoch on the training set, the validation accuracy decreases, while the validation loss increases. Nevertheless, the regularization

DOI 10.15588/1607-3274-2026-1-6

mechanisms and adaptive machine learning strategies stabilize the accuracy at the level of 97%, which is likely a consequence of using cyclic learning rate adjustment to escape local minima. In the course of information-extreme machine learning, the recognition class containers are reconstructed through the optimization of their geometric parameters.

Figure 5 presents the plots showing the dependency of the information criterion (7) on the radii of the recognition class containers obtained during the process of information-extreme machine learning.

The analysis of the machine learning plots shown in Figure 5 demonstrates that all recognition classes from the defined alphabet formed valid working domains; that is, they are separable, since the first and second validity measures within the working domains exceed, respectively, the Type I and Type II error rates.

At the same time, the optimal radii of the recognition class containers (given in code units) are as follows: 53 for class X_1^0 – "Annual Crop", 230 for class X_2^0 – "Forest", 56 for class X_3^0 – "Highway", 55 for class X_4^0 – "Residential", 56 for class X_5^0 – "Pasture", and 55 for class X_6^0 – "Industrial".

The quality of machine learning was evaluated using a test set of 400 images per recognition class. The overall recognition accuracy during testing reached 95%.

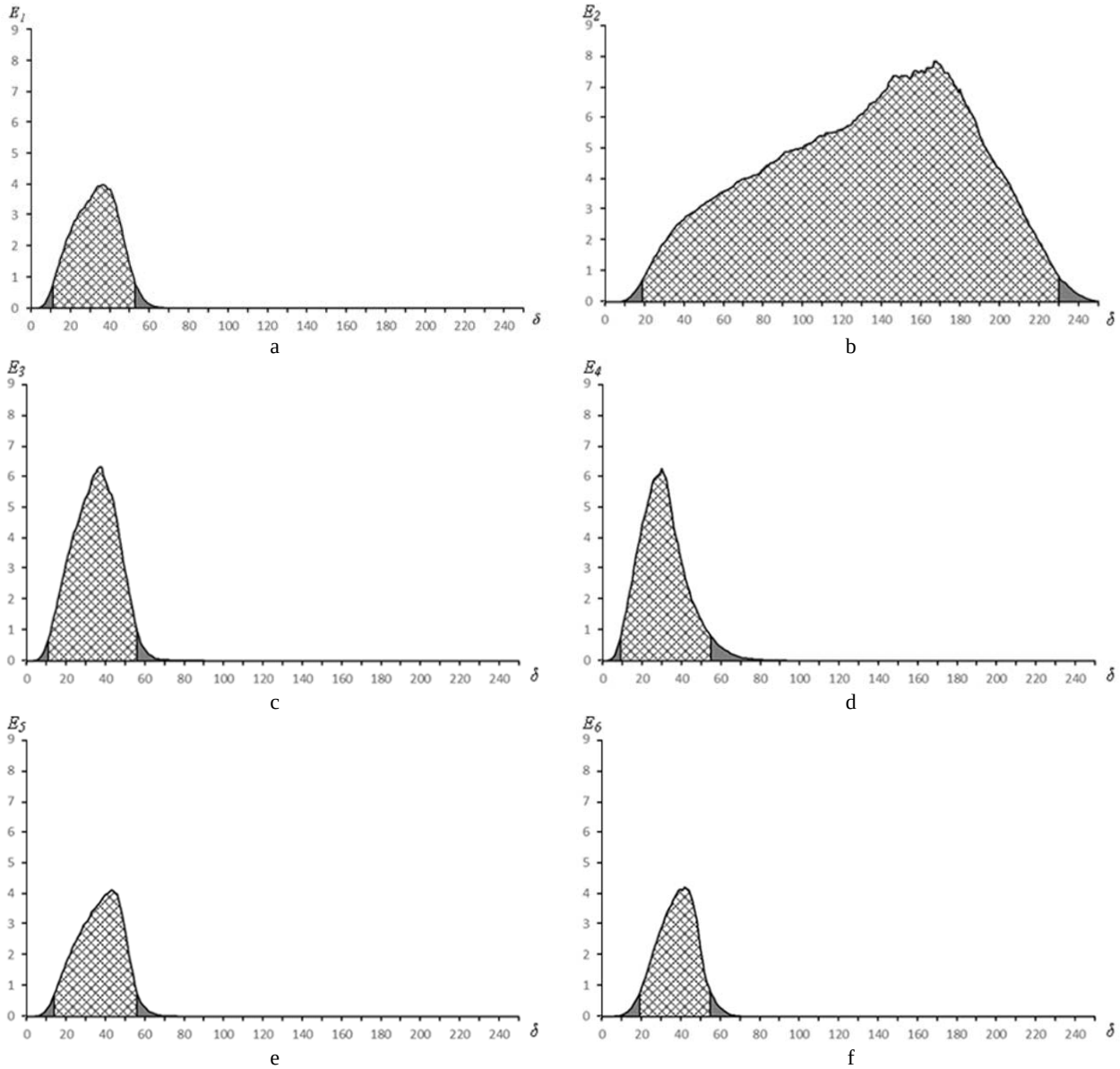


Figure 5 – Plots of the dependency between the information criterion and the radii of the recognition class containers:
a – class X_1^0 "Annual Crop", b – class X_2^0 "Forest", c – class X_3^0 "Highway",
d – class X_4^0 "Residential", e – class X_5^0 "Pasture", f – class X_6^0 "Industrial"

DOI 10.15588/1607-3274-2026-1-6

Special attention should be paid to the interpretability of the models. While deep neural networks with softmax output provide high classification accuracy, their internal mechanisms remain largely opaque to the user due to the large number of parameters and complex layer interactions.

The k-NN and SVM algorithms partially improve interpretability by allowing the analysis of proximity to training examples or positioning relative to the decision hyperplane.

At the same time, the proposed IEIT model ensures significantly greater transparency of decision-making due to its geometric interpretation in the form of recognition class containers in the space of binarized features. Unlike deep learning "black-box" models, this approach defines clear boundaries between classes in the form of hyperspherical regions, whose parameters – centers, radii, and tolerances – have explicit mathematical meaning.

This structure not only determines the class membership of an object but also allows assessing the degree of proximity or remoteness of the feature vector from the center of the corresponding container. In turn, this enables the identification of borderline or atypical instances.

Moreover, the geometric interpretation provides a foundation for building trust mechanisms into the model, which can be used to develop additional decision criteria or to detect anomalous cases that do not correspond to any of the training classes.

To summarize the obtained results visually, Table 1 presents a comparative analysis of the efficiency of different image classification approaches. The comparison includes both classical algorithms applied directly to raw pixel data and modern neural and hybrid architectures with prior feature extraction. In addition to classification accuracy, the models are also evaluated in terms of their generalization ability via macro F1-score, class-wise recall range, and interpretability of the decision process. This allows for a comprehensive assessment not only of classification performance but also of the model's practical suitability for analysis and real-world deployment.

Table 1 – The fragment of experimental results on model building by the formed samples

№	Architecture	Features	Classifier	Accuracy	Macro F1	Recall (min/max)	Interpretability
1	Raw pixels + SVM	64×64 grayscale	SVM	47.3%	0.44	0.10 / 0.92	low
2	Raw pixels + k-NN	64×64 grayscale	k-NN	38.6%	0.27	0.00 / 0.99	low
3	CNN + Softmax	CNN-features	Softmax	93.8%	0.94	0.80 / 0.99	limited
4	CNN + k-NN	CNN-features	k-NN	96.75%	0.97	0.94 / 0.99	limited
5	CNN + SVM	CNN-features	SVM	96.4%	0.96	0.93 / 0.99	limited
6	CNN + IEIT	CNN binary features	IEIT	96.5%	0.95	0.93 / 0.99	high

6 DISCUSSION

The analysis of the plots presented in Figure 4 demonstrates a typical dynamic of adaptation during the early stages of machine learning: the noticeable difference between accuracy on the training and validation sets gradually decreases as the model begins to form generalizations. After epochs 30–35, the accuracy reaches a plateau, which indicates the stabilization of the learning process. An increase in validation loss in the later stages signals the beginning of overfitting. Nevertheless, the application of regularization techniques – particularly weight decay implemented in the AdamW optimizer – and the use of the GlobalAveragePooling layer made it possible to maintain consistently high validation accuracy at the level of 96–97%.

The information-extreme component of the model shows a clearly expressed dependence of the information measure on the geometric parameters of the recognition class containers (Figure 5), which indicates the effectiveness of using the information criterion for optimizing machine learning parameters. In particular, for each recognition class, a specific optimal range of radius values was determined, within which the maximum of the information measure (7) is achieved. This measure takes into account the differences between intra-class and inter-class distributions. Such optimization allows the formation of compact yet sufficiently capacious hyperspherical containers that accurately reflect the statistical regularities of the input data.

The obtained optimal parameters confirm the model's ability to form well-separated regions in the multidimensional feature space, with features preliminarily extracted by the convolutional neural network. As a result, each recognition class is provided with an individually adapted geometric representation, which reduces the risk of inter-class overlap and improves the accuracy of classification decisions. Moreover, the presence of clearly pronounced maxima of the functional indicates the model's stability to small variations in the container parameters – an important property for practical use in conditions of limited or partially noisy input data.

Compared to the classical CNN architecture that uses a softmax classifier, the proposed hybrid approach has several advantages. First, the model provides higher interpretability due to the geometric representation of recognition classes in the form of hyperspherical containers. Second, the separation of feature extraction and classification processes allows for more flexible control of the system's behavior, in particular, the ability to adapt the classifier to changes in the feature space without the need to retrain the entire network. This is especially important in cases with small datasets or a high level of inter-class overlap.

At the same time, the construction of class containers in the space of binarized features has certain limitations. This approach assumes the isotropy of the feature space, which is not always satisfied in practice. Additionally, the effectiveness of decision-making is highly dependent on

DOI 10.15588/1607-3274-2026-1-6

the proper tuning of parameters – particularly the control tolerance δ and the boundary values of the recognition class container radii require prior optimization or the implementation of adaptive mechanisms.

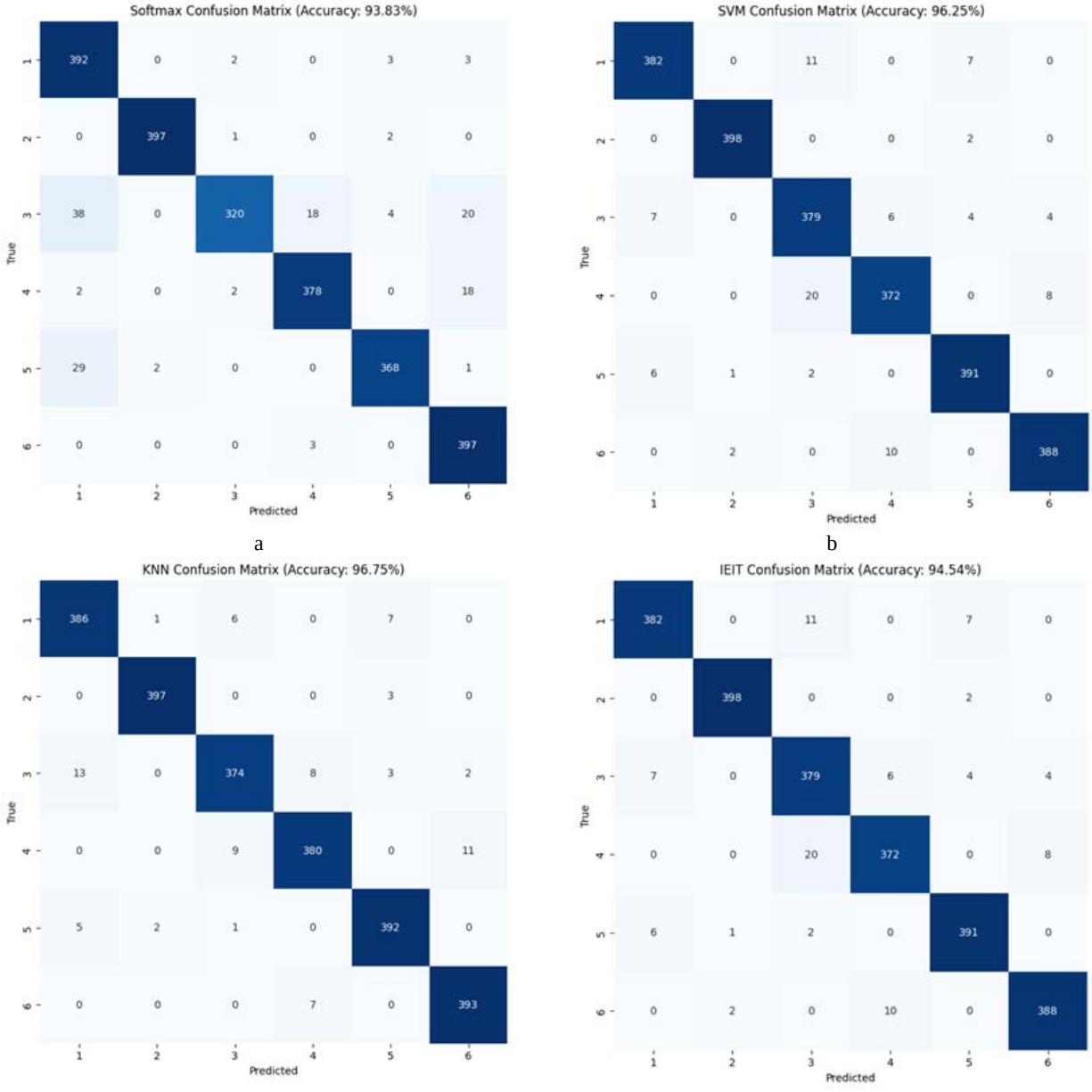
During the experimental study, a comparative evaluation of various image classification approaches was carried out. The use of classical linear classifiers (SVM and k-NN) directly on pixel-level image representations proved to be ineffective: SVM achieved an accuracy of only 47.3%, while k-NN reached 38.6%. These results confirm the limitations of raw pixel features and emphasize the necessity of deep preprocessing of input data.

Applying a softmax classifier within the convolutional neural network framework significantly improved classification accuracy – up to 93.8%. However, a more detailed analysis revealed considerable discrepancies in the recognition performance of individual classes, particularly in cases with high visual similarity between them.

The highest performance was achieved using hybrid models that combine CNN-based feature extraction with classification via separate algorithms. The CNN + SVM model achieved an accuracy of 96.4%, while CNN + k-NN reached 96.75%. These results indicate that such hybrid approaches not only ensure high recognition accuracy but also preserve the structural coherence and stability of the model's behavior.

Thus, the hybrid architecture demonstrates superiority over both classical and standalone neural approaches, especially in tasks where a combination of high accuracy, adaptability, and interpretability of results is required.

Based on the confusion matrices (Figure 6), a more detailed analysis of model classification quality can be



a b c d

Figure 6 – Confusion matrixes for classification neural based models:
a – CNN + Softmax, b – CNN + k-NN, c – CNN + SVM, d – CNN + IEIT

DOI 10.15588/1607-3274-2026-1-6

performed. Despite an overall accuracy of 93.83%, the softmax classifier exhibits a pronounced bias in recognizing class X_3^0 ("Highway") – with 38 misclassifications toward the "Annual Crop" class and an additional 42 errors scattered across other classes. This behavior indicates a high degree of inter-class overlap for this category and an insufficient capacity of softmax-based decisions to separate closely situated representations in the feature space.

The CNN + k-NN (96.75%) and CNN + SVM (96.25%) models demonstrate a significantly more stable error distribution. In particular, the number of misclassifications in the "Highway" and "Residential Area" classes has been reduced by at least half. This indicates that geometric methods relying on local or global decision boundaries in the feature space perform better in cases of partial overlap between classes.

The proposed CNN + IEIT model achieves an overall accuracy of 94.54%, with its confusion matrix nearly replicating the pattern of errors observed with the SVM classifier. However, unlike SVM, IEIT operates with clearly defined geometric containers featuring adaptive radii and enables the identification of uncertainty zones. For the "Highway" and "Residential" recognition classes, the number of errors is identical to that of SVM, though IEIT exhibits better clarity in other classes (e.g., "Annual Crop"), avoiding mixing with more distant categories.

CONCLUSIONS

The relevant task of constructing an interpretable and statistically grounded image classification system based on deep feature extraction has been addressed through the development of a hybrid model that combines a convolutional neural network (CNN) with information-extreme machine learning (IEIML).

The scientific novelty of the obtained results lies in the proposed approach that enables the formation of hyperspherical recognition class containers in the feature space generated by the CNN, followed by optimization of their parameters using an information-based criterion. This approach ensures a balance between classification accuracy and interpretability of the decision-making process.

The practical significance of the proposed solution is defined by the possibility of effectively applying the model for satellite image classification within a modular architecture, where the feature extractor and the classifier can be independently adapted or enhanced for various tasks and datasets.

Prospects for further research include the adaptation of the proposed hybrid architecture to object detection tasks, particularly the localization of multiple objects within a single image and the extension of the categorical model to operate with spatially oriented representations.

ACKNOWLEDGEMENTS

The research was concluded in the Computer Science Department at Sumy State University with the financial support of the Ministry of Education and Science of Ukraine in the framework of state budget scientific and research work of DR No. 0121U109466 "Intelligent Technologies in Cybersecurity".

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: Anatolii Dovbysh: research concept and methodology, scientific supervision, manuscript review;

Vladyslav Piatachenko: coordination of the research activities, hybrid system architecture, experimental design and analysis;

Viktor Serhieiev: development and investigation of hybrid machine learning models, parameter tuning, validation of information-extreme learning algorithms;

Oleksii Hrytsenko: image preparation and processing, training of neural network feature extractors, evaluation and visualization of results.

Data availability: The manuscript has no associated data.

Software availability: The software will be made available on reasonable request.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

REFERENCES

1. Ab Wahab M. N., Nazir A., Ren A. T. Z., Noor M. H. M., Akbar M. F., Mohamed A. S. A. Efficientnet-Lite and Hybrid CNN-KNN Implementation for Facial Expression Recognition on Raspberry Pi, *IEEE Access*, 2021, Vol. 9, pp. 134065–134080. DOI:10.1109/ACCESS.2021.3113337
2. Ghosh S., Singh A., Kavita, Jhanjhi N. Z., Masud M., Aljahdali S. SVM and KNN Based CNN Architectures for Plant Classification, *Computers, Materials and Continua*, 2022, Vol. 71, № 3, P. 4257. DOI:10.32604/CMC.2022.023414
3. Lanjewar M. G., Parab J. S., Shaikh A. Y. Development of framework by combining CNN with KNN to detect Alzheimer's disease using MRI images, *Multimedia Tools and Applications*, 2023, Vol. 82, № 8, pp. 12699–12717. DOI:10.1007/S11042-022-13935-4/METRICS
4. Moskalenko V., Kharchenko V., Moskalenko A., Petrov S. Model and Training Method of the Resilient Image Classifier Considering Faults, Concept Drift, and Adversarial Attacks, *Algorithms*, 2022, Vol. 15, № 384, pp. 1–24. DOI:10.3390/a15100384
5. Moskalenko V. V., Moskalenko A. S., Korobov A. G., Zaretsky M. O. Image Classifier Resilient To Adversarial Attacks, Fault Injections And Concept Drift – Model Architecture And Training Algorithm, *Radio Electronics, Computer Science, Control*, 2022, Vol. 3, № 86, pp. 1–16. DOI:10.15588/1607-3274-2022-3-9
6. Shelehov I., Prylepa D., Khibovska Yu. Information-extreme machine learning of an ophthalmic diagnostic system with a hierarchical class structure, *Artificial Intelligence*, 2024, Vol. 29, №3, pp. 114–125. DOI:10.15407/JAI2024.03.114

DOI 10.15588/1607-3274-2026-1-6

7. Dovbysh A. S. Shelekhov I. V., Prylepa D. V., Khibovska Yu. O., Nikitenko K. O. Information-Extreme Method For Ball Detection In Intelligent Video Analysis Systems Of Volleyball Matches, *Visnyk Kremenchutskoho natsionalnoho universytetu imeni Mykhaila Ostrohradskoho*, 2024, Vol. 5, pp. 41–51. DOI:10.32782/1995-0519.2024.5.6
8. Dovbysh A. S., Budnyk M. M., Piatachenko V. Y., Myronenko M. I. Information-extreme machine learning of on-board vehicle recognition system, *Cybernetics and Systems Analysis*, 2020, Vol. 56, pp. 534–543. DOI:10.1007/s10559-020-00269-y
9. Naumenko I., Piatachenko V., Myronenko M., Savchenko T. Information-Extreme Machine Learning of an On-board Ground Object Recognition System with a Choice of a Base Recognition Class, *6th International Conference on Computational Linguistics and Intelligent Systems, Gliwice, 12–13 May 2022: proceedings*. Gliwice, CEUR, 2022, pp. 1139–1148.
10. Tan M., Le Q. V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, *36th International Conference on Machine Learning, 9th June 2019: proceedings*. Long Beach:arXiv, 2019, pp. 10691–10700. DOI: 10.48550/arXiv.1905.11946
11. Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L. C. MobileNetV2: Inverted Residuals and Linear Bottlenecks, *Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018: proceedings*. Salt Lake City, IEEE, 2018, pp. 4510–4520. DOI:10.1109/CVPR.2018.00474
12. Dasgupta R., Chowdhury Y. S., Nanda S. Performance Comparison of Benchmark Activation Function ReLU, Swish and Mish for Facial Mask Detection Using Convolutional Neural Network, *Algorithms for Intelligent Systems, Singapore, 2021: proceedings*. Singapore, Springer, 2021, pp. 355–367. DOI:10.1007/978-981-16-2248-9_34
13. Helber P., Bischke B., Dengel A., Borth D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019, pp. 1–9. DOI: 10.48550/arXiv.1709.00029

Received 13.08.2025.
Accepted 17.11.2025.
Published 27.03.2026.

УДК 004.93

ГІБРИДНА СИСТЕМА РОЗПІЗНАВАННЯ СУПУТНИКОВИХ ЗОБРАЖЕНЬ З НЕЙРОМЕРЕЖЕВИХ ЕКСТРАКТОРОМ ТА ІНФОРМАЦІЙНО ЕКСТРЕМАЛЬНИМ КЛАСИФІКАТОРОМ

Довбиш А. С. – д-р техн. наук, професор, професор кафедри комп'ютерних наук Сумського державного університету, Суми, Україна. ROR: https://ror.org/01w60n236. ORCID:https://orcid.org/0000-0003-1829-3318.

П'ятаченко В. Ю. – канд. техн. наук, асистент кафедри комп'ютерних наук Сумського державного університету, Суми, Україна. ROR: https://ror.org/01w60n236. ORCID: https://orcid.org/0000-0002-7464-3119.

Сергєєв В. М. – аспірант кафедри комп'ютерних наук Сумського державного університету, Суми, Україна. ROR: https://ror.org/01w60n236. ORCID:https://orcid.org/0009-0009-3838-0153.

Гриценко О. М. – аспірант кафедри комп'ютерних наук Сумського державного університету, Суми, Україна. ROR: https://ror.org/01w60n236. ORCID: https://orcid.org/0009-0004-1382-708X.

АНОТАЦІЯ

Актуальність. Розв'язано актуальну задачу побудови інтерпретованої та адаптивної системи розпізнавання для семантичної сегментації супутникових знімків місцевості шляхом поєднання нейромережевих екстракторів з інформаційно-екстремальним класифікатором .

Мета роботи. Підвищення точності класифікації супутникових знімків місцевості шляхом розробки гібридної моделі машинного навчання, яка об'єднує глибоку згорткову нейронну мережу для відбору інформативних ознак та інформаційно-екстремальний класифікатор, що дозволяє будувати високодостовірні вирішальні правила за умови перетину класів розпізнавання в просторі ознак.

Метод. Запропоновано гібридну модель, яка поєднує ефективну екстракцію просторових ознак за допомогою згорткової нейронної мережі та інформаційно-екстремальну інтелектуальну технологію аналізу даних, яка базується на максимізації інформаційної спроможності системи розпізнавання в процесі машинного навчання. Водночас при екстракції ознак розпіхзнавання замість класичного Flatten використано GlobalAveragePooling для узагальнення ознак, а також впроваджено регуляризаційні механізми, зокрема вагове затухання та циклічне навчання. Оптимізація параметрів інформаційно-екстремального машинного навчання виконується за модифікованим авторами інформаційним критерієм Кульбака, який розглядається як міра різноманітності класів розпізнавання.

Результати. Побудована модель забезпечує високу точність класифікації (95%) при тестуванні, а також демонструє стабільність та підвищення оперативності нейромережевого екстрактора шляхом зменшення кількості епох його навчання завдяки застосуванню регуляризації. За результатами інформаційно-екстремального машинного навчання визначено оптимальні геометричні параметри гіперсферичних контейнерів класів розпізнавання, що дозволяє побудувати високодостовірні вирішальні правила за умови перетину класів розпізнаванняс в просторі ознак.

Висновки. Запропонована гібридна модель дозволяє для апріорно нечіткого розбиття в просторі ознак класів розпізнавання за сформованою в результаті екстракції ознак вхідною навчальною матрицею побудувати в процесі інформаційно-екстремального машинного навчання високодостовірні вирішальні правила.

КЛЮЧОВІ СЛОВА: інформаційно-екстремальне машинне навчання, згорткова нейронна мережа, інформаційний критерій, оптимізація, гібридна модель, екстракція ознак розпізнавання, зображення місцевості.

DOI 10.15588/1607-3274-2026-1-6

ЛІТЕРАТУРА

1. Efficientnet-Lite and Hybrid CNN-KNN Implementation for Facial Expression Recognition on Raspberry Pi / [M. N. Ab Wahab, A. Nazir, A. T. Z. Ren et al.] // IEEE Access. – 2021. – Vol. 9. – P. 134065–134080. DOI:10.1109/ACCESS.2021.3113337
2. SVM and KNN Based CNN Architectures for Plant Classification / [S. Ghosh, A. Singh, Kavita, N. Z. Jhanjhi et al.] // Computers, Materials and Continua. – 2022. – Vol. 71, № 3. – P. 4257. DOI:10.32604/CMC.2022.023414
3. Lanjewar M. G. Development of framework by combining CNN with KNN to detect Alzheimer's disease using MRI images / M. G. Lanjewar, J. S. Parab, A. Y. Shaikh // Multimedia Tools and Applications. – 2023. – Vol. 82, № 8. – P. 12699–12717. DOI:10.1007/S11042-022-13935-4/METRICS
4. Model and Training Method of the Resilient Image Classifier Considering Faults, Concept Drift, and Adversarial Attacks / [V. Moskalenko, V. Kharchenko, A. Moskalenko, S. Petrov] // Algorithms. – 2022. – Vol. 15, № 384. – P. 1–24. DOI:10.3390/a15100384
5. Image Classifier Resilient To Adversarial Attacks, Fault Injections And Concept Drift – Model Architecture And Training Algorithm / [V. V. Moskalenko, A. S. Moskalenko, A. G. Korobov, M. O. Zaretsky] // Radio Electronics, Computer Science, Control. – 2022. – Vol. 3, № 86. – P. 1–16. DOI:10.15588/1607-3274-2022-3-9
6. Shelehov I. Information-extreme machine learning of an ophthalmic diagnostic system with a hierarchical class structure / I. Shelehov, D. Prylepa, Yu. Khibovska // Artificial Intelligence. – 2024. – Vol. 29, №3. – P. 114–125. DOI:10.15407/JAI2024.03.114
7. Інформаційно-Екстремальний Метод Детектування М'яча В Системах Інтелектуального Відеоаналізу Волейбольного Матчу / [А. С. Довбиш, І. В. Шелехов, Д. В. Прилепа та ін.] // Вісник КрНУ імені Михайла Остроградського – 2024. – Вип. 5. – С. 41–51. DOI:10.32782/1995-0519.2024.5.6
8. Information-extreme machine learning of on-board vehicle recognition system / [A. S. Dovbysh, M. M. Budnyk, V. Y. Piatachenko, M. I. Myronenko] // Cybernetics and Systems Analysis. –2020. – Vol. 56. – P. 534–543. DOI:10.1007/s10559-020-00269-y
9. Information-Extreme Machine Learning of an On-board Ground Object Recognition System with a Choice of a Base Recognition Class / [I. Naumenko, V. Piatachenko, M. Myronenko, T. Savchenko] // 6th International Conference on Computational Linguistics and Intelligent Systems, Gliwice, 12–13 May 2022: proceedings. – Gliwice : CEUR, 2022. – P. 1139–1148.
10. Tan M. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks / M. Tan, Q. V. Le // 36th International Conference on Machine Learning, 9th June 2019: proceedings. – Long Beach:arXiv,2019. – P. 10691–10700. DOI: 10.48550/arXiv.1905.11946
11. MobileNetV2: Inverted Residuals and Linear Bottlenecks / [M. Sandler, A. Howard, M. Zhu et al.] //Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018: proceedings. – Salt Lake City: IEEE, 2018. – P. 4510–4520. DOI:10.1109/CVPR.2018.00474
12. Dasgupta R. Performance Comparison of Benchmark Activation Function ReLU, Swish and Mish for Facial Mask Detection Using Convolutional Neural Network / R. Dasgupta, Y. S. Chowdhury, S. Nanda // Algorithms for Intelligent Systems, Singapore, 2021: proceedings – Singapore: Springer, 2021. – P. 355–367. DOI:10.1007/978-981-16-2248-9_34
13. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification / [P. Helber, B. Bischke, A. Dengel, D. Borth] // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. – 2019. – P. 1–9. DOI: 10.48550/arXiv.1709.00029

DOI 10.15588/1607-3274-2026-1-6

UDC 004.93

WELER: A COMPLEX METRIC FOR TEXT QUALITY ASSESSMENT

Dumyn A. R. – Post-graduate student at Lviv Polytechnic National University, Ukraine. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0000-0003-2111-2899.

Shakhovska N. B. – Dr. Sc., Professor, Rector at Lviv Polytechnic National University, Ukraine. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0000-0002-6875-8534.

ABSTRACT

Context. Assessing text quality is essential for reliable AI that processes language. In ASR, it reflects how faithfully speech becomes text; in OCR, how accurately images yield text; and in NLP, how correct and coherent outputs are.

Objective. The goal of the work is the creation of a complex metric for text quality assessment.

Method. Classic metrics WER and CER are narrow: they capture only lexical edits, weigh all changes equally, ignore context and semantics, and often skip punctuation and case, masking readability issues and error types. We propose WELER, a hybrid metric that blends weighted WER and CER with a semantic component based on contextual embeddings to measure meaning preservation. Weights can be set manually or learned (e.g., via PCA), adapting the metric to ASR, OCR, or NLP tasks. Key challenges include computational cost, choosing optimal weights through correlation with human judgments, and the need for high-quality reference data. Proposed WELER metric integrates accurate word and character level error counting, using Levenshtein distance as a basis, with advanced semantic similarity methods based on contextual embeddings. This allows WELER to take into account not only what was incorrectly recognized, but also how much this error affects the meaning and understanding of the text. The inclusion of self-adjusting weights depending on the text category is a key feature of WELER, which allows adapting the metric to the specific requirements of different applications and domains, prioritizing those aspects of quality that are most critical for a particular task.

Results. Proposed WELER metric is an alternative solution in this direction. It integrates accurate word and character level error counting, using Levenshtein distance as a basis, with advanced semantic similarity methods based on contextual embeddings.

Conclusions. WELER, like all metrics based on reference data, relies on accurate and consistent human-verified transcriptions. Errors in the reference data can affect the accuracy of the assessment. Therefore, for complex metrics, the quality and representativeness of these data are especially important, since semantic and weighted errors are much more sensitive to the quality of the annotation than simple word counts.

KEYWORDS: snatural language processing, automatic speech recognition, text quality assessment, WER, CER, WELER.

ABBREVIATIONS

ASR is automatic speech recognition;
OCR is optical character recognition;
NLP is natural language processing;
WER is word error rate;
CER is character error rate;
NLI is Natural Language Inference;
WELER is weight-based error rate.

NOMENCLATURE

S is a number of substitutions;
D is a number of deletions (word omitted);
I is a number of insertions (extra word);
N is a number of words in the template;
S_c is a number of character substitutions;
D_c is a number of omitted characters;
I_c is a number of extra characters;
N_{char} is a total number of characters in the reference text;
S_{word} is a number of word-level substitutions obtained from the Levenshtein alignment;
D_{word} is a number of word-level deletions obtained from the Levenshtein alignment;
I_{word} is a number of word-level insertions obtained from the Levenshtein alignment;
N_{word} is a total number of words in the reference text;
w_S is an adjustable weighting factors for S_{word};
w_D is an adjustable weighting factors for D_{word};
w_L is an adjustable weighting factors for I_{word};
S_{char} is a numbers of character-level substitutions;
D_{char} is a numbers of character-level deletions;
I_{char} is a numbers of character-level insertions;
W_{wer} is a weighted word error rate;
W_{cer} is a weighted error rate in symbols;
$SemErr$ is a semantic error indicator;
α is a normalized W_{wer};
β is a normalized W_{cer};
γ is a normalized $SerErr$;
Emb_{inut} is a vector of embeddings of the reference text;
Emb_{output} is a vector of recognized text embeddings.

INTRODUCTION

Text quality assessment plays a crucial role in functionality of systems that work with text data. Especially artificial intelligence. In areas such as ASR, OCR, and NLP, accurate text quality assessment is fundamental to determining the effectiveness of the system. For example, in ASR, it determines how accurately spoken words are converted to text, and in OCR, the correctness of text extraction from images. In NLP tasks related to text generation, processing, and quality assessment helps determine the correctness and coherence of the obtained result [1].

Traditionally, WER and CER metrics have been widely used to assess text quality. These metrics are quantitative measures of how well extracted data matches a reference, usually expressed as a percentage [1]. WER and CER are derived from the Levenshtein distance, which defines the minimum number of editing operations (replacements, deletions, or insertions) required to transform one sequence into another, WER works at the word level, counting errors resulting from word substitutions,

DOI 10.15588/1607-3274-2026-1-7

deletions, or insertions. CER, on the other hand, focuses on character-level accuracy, counting similar errors for individual characters. These metrics have become the standards for evaluating the performance of ASR and OCR models, due to their relative ease of calculation and straightforward interpretation.

However, despite their popularity, WER and CER have significant limitations that hinder their ability to fully reflect text quality. They mainly measure lexical accuracy at a superficial level, without taking into account semantic similarity, word importance, grammatical correctness, or the impact of punctuation errors [2]. For example, WER penalizes minor spelling errors as well as errors that completely change the meaning of a sentence, leading to discrepancies between automatic evaluation and human perception of text quality.

Due to these restrictions, there was a need to develop a more comprehensive Evaluation metric. Such metrics should go beyond simply counting errors at a superficial level, integrating a deeper understanding of semantics and context, and allowing for differential weighting of different types of errors according to their impact on the overall quality and understandability of the text.

Object of the research: automated text quality assessment in AI systems (ASR, OCR, and NLP) based on reference transcriptions/texts.

Subject of the research: the hybrid WELER metric – its components (weighted WER, weighted CER, and a semantic error indicator from contextual embeddings), normalization choices, and data-driven weighting (e.g., PCA-based) for different task categories.

This study aims to develop a hybrid metric for assessing the quality of generated text based on the use of basic metrics and taking into account semantic and weighted approaches. The development of such a hybrid metric reflects the growth of the assessment of artificial intelligence systems, moving from a purely quantitative error count to a more qualitative, human-perceived understanding of "correctness". This transition is an intermediate stage for creating more reliable and user-oriented artificial intelligence systems.

1 PROBLEM STATEMENT

We study the problem of learning a composite error metric (WELER) for text sequences. For each example we have a reference text R, a system output H (from ASR/OCR/NLP), domain metadata m (e.g., language, task type, noise), and a human quality score y in [0,1]. From (H, R) we compute three normalized component errors in [0,1]: a weighted word-level edit error, a weighted character-level edit error, and a semantic difference (for instance, one minus cosine similarity of sentence embeddings or an NLI-based score). WELER is a single score that combines these components with nonnegative weights that sum to one; the weights may be fixed or predicted from m so the metric adapts to the domain. Lower WELER means fewer errors (you may invert it if you prefer "higher is better").

The goal is to learn the weights (and, if needed, normalization/adaptation parameters) so that 1 − WELER matches the human scores y as closely as possible on the dataset, typically by minimizing an average loss such as MAE or MSE, or by maximizing rank agreement. We require monotonicity (if any component error increases, WELER cannot decrease) and, optionally, stability across groups (e.g., languages). For decision making, a threshold on 1 − WELER can be chosen to label outputs as acceptable or not, tuned to maximize a target metric like F1.

2 REVIEW OF THE LITERATURE

The issue of developing additional metrics for assessing the quality of work of artificial intelligence methods with the tasks of natural language processing, optical text recognition [3], audio-to-text conversion [3], plagiarism detection [4], etc., is widely studied in academic circles. In particular, the authors [5] propose the *H_eval* metric, which combines semantic correctness with the traditional WER error. This metric works much faster than BERTScore and has a strong correlation with NLP tasks. In [6], the metric was introduced SemDist is the distance between the embedding vectors for the reference and generated texts obtained through RoBERTa. Semantic evaluation has been shown to be more relevant for natural language understanding tasks than simple WER.

SeMaScore metric [7], developed for ASR, integrates traditional error metrics with a robust measure of semantic similarity. It then computes segment scores using contextual embeddings. and cosine similarity. This approach demonstrates how traditional methods can serve as the basis for new, more sophisticated estimates. In the work [8] proposed a metric that combines NLI score, semantic and phonetic similarity. The proposed metric achieves a correlation of with human intelligibility judgments on data with language features (dysarthria and dysphonia discourse). Article [9] proposes the BERTScore metric, which calculates the similarity between texts based on contextual embeddings and shows that this approach correlates better with human evaluation than traditional n-gram metrics.

WER [10] is one of the most common metrics for evaluating text recognition accuracy. It measures the percentage of incorrect words in the generated text (hypothesis) compared to the reference text [1]. The formula for calculating WER is given below

$$WER = \frac{S + D + I}{N}. \quad (1)$$

It is important to note that the WER value can exceed 1 or 100%, especially in cases where the number of insertions significantly exceeds the number of words in the reference text. Lower WER values indicate higher accuracy. WER is a particularly valuable tool for evaluating ASR and OCR performance, particularly in scenarios where the emphasis is on free-form text recognition. This

DOI 10.15588/1607-3274-2026-1-7

includes applications such as document digitization, handwriting transcription, multilingual text recognition, book and manuscript archiving, and automating the transcription of meeting notes or legal documents.

Before calculating the WER, a text pre-processing process known as normalization is usually applied to ensure a fair comparison. This involves converting all text to lowercase, removing punctuation, and standardizing numbers (e.g., “5 доларів”. instead “5 $”) and expansion of abbreviations.

WER is based on the Levenshtein distance, which works at the word level. This relationship means that WER, as a metric, is derived from an algorithm that counts the minimum number of edit operations to transform one sequence of words into another. This use of Levenshtein distance for WER results in all word-level errors (substitutions, deletions, insertions) receiving the same weight. For example, the substitution of the word “плисти” to “плести”, which is a homophone or “кит” to “кат”, which is a completely different word, will have the same error weight, despite their different impact on semantic meaning. This insensitivity to semantic nuances is a direct consequence of the underlying algorithm and constitutes a significant limitation.

Furthermore, the need for extensive text normalization before calculating WER, which includes removing punctuation and ignoring capitalization, indicates that WER in its standard form is not a holistic measure of text quality. Rather, it measures lexical relevance after preprocessing. This means that important aspects of text quality, such as formatting and grammatical correctness, which are often removed during normalization, are effectively ignored by standard WER. This highlights the need for a new metric that could explicitly include these aspects or allow for their weighted inclusion.

CER [10] is another key metric for assessing recognition accuracy, which, unlike WER, focuses on accuracy at the character level. CER measures how many characters in the source data differ from the reference data, taking into account substitutions, deletions, and insertions of characters relative to the total number of characters in the reference data. The formula for CER is identical to the WER formula, but applied to characters, and is given below

$$CER = \frac{S_C + D_C + I_C}{D_C}. \qquad (2)$$

CER provides a more detailed error assessment than WER. For example, minor typographical errors, such as “опрацюваня” instead of “опрацювання”, will result in a full word substitution error in WER, but only a single character substitution error in CER. This allows for a more accurate assessment of systems where even small errors can have significant consequences, such as in coding, formal documents, or specialized terminology.

CER is particularly useful in scenarios where word boundaries may be absent, such as numeric data, alphanumeric codes, or where accurate character recognition is critical, such as serial numbers, financial data, passport numbers. It is also applicable for languages that do not have clear spaces between words, such as Chinese.

Although CER offers a more granular error estimate, it still has a fundamental limitation of WER – the lack of semantic understanding. Even a single character error can significantly change the meaning of a word or sentence, which CER alone does not capture. This means that while CER is more accurate in indicating where an error has occurred, it does not assess the impact of that error on the value, which is a critical aspect of the quality of the estimate.

The usefulness of CER in specific domains, such as numeric data, codes, or languages without clear word boundaries, suggests that a truly robust text quality metric should be adaptive or configurable to prioritize different levels of error (symbolic vs. lexical) depending on the application requirements. This suggests that the coefficients for WER/CER in the hybrid metric will allow it to be adapted to these diverse needs.

Levenshtein distance [12], also known as edit distance, is a metric that quantifies the difference between two strings. It calculates the minimum number of character-level editing operations (insertions, deletions, or substitutions) required to transform one string into another. The algorithm was proposed by Vladimir Levenshtein in 1965, and it quickly became the basis for tasks such as spell checking, DNA sequencing, and duplicate text detection. The implementation of Levenshtein distance is based on dynamic programming.

As already mentioned, the Levenshtein distance is the main algorithm for calculating WER and CER. It is used to align the recognized text with the reference text and identify minimal editing operations at the word or character level.

There are variants of edit distance that extend the basic Levenshtein distance. For example, the Damerau-Levenshtein distance includes transpositions, i.e., the rearrangement of two adjacent characters, as a single editing operation, which allows for a more accurate representation of some types of errors, such as those that occur in typing. Weighted edit distance allows for different weights to be assigned to insertion, deletion, and replacement operations. These extensions demonstrate that even within the edit distance paradigm, researchers have recognized that not all errors are equally important, which in turn creates the prerequisites for the application of additional weights.

Levenshtein distance is purely lexical a comparison metric that is not intended for semantic or contextual understanding. Its applications, such as spell checking or plagiarism detection, are primarily concerned with string similarity rather than semantic equivalence. However, despite this, Levenshtein distance can serve as a structural framework for integrating higher-level semantic information. For example, in metrics such as SeMaScore, Levenshtein distance is used for initial segment alignment. This allows for identifying relevant parts of the text, even if

DOI 10.15588/1607-3274-2026-1-7

they contain errors, and then applying semantic comparison to these aligned parts. This approach is a way of combining lexical and semantic evaluation.

To overcome the limitations of traditional metrics such as WER and CER, researchers have developed more sophisticated approaches that take semantic meaning into account and allow for weighting of different types of errors [13].

Semantic similarity metrics move away from simply counting lexical errors and focus on preserving the meaning of the text. They use contextual embeddings to represent words and sentences in a multidimensional space where semantically similar texts are located closer together. Cosine similarity is commonly used to measure this proximity.

BERTScor metric uses contextual embeddings from pre-trained language models such as BERT or RoBERTa to measure semantic similarity between texts. It calculates precision, completeness, and F1-measure by aligning tokens based on vector similarity. BERTScore correlates better with human judgment than traditional n-gram metrics such as BLEU, ROUGE, because it is able to recognize when different words or phrases convey the same meaning, even if their surface forms are different. This is a direct solution to the problem of lack of semantic understanding in WER/CER [14].

While semantic metrics offer significant benefits, they also have their challenges. Generating contextual embeddings can be computationally intensive, especially for large datasets. Furthermore, the reliance on deep learning models can make them less interpretable than traditional metrics. This suggests a trade-off between the enrichment of the score and its practical applicability and interpretability.

The concept of weighting different types of errors is not new. As early as 1990, Hunt proposed a weighted WER, where substitutions were given a weight of 1 and deletions and insertions were given a weight of 0.5 [15]. This early example shows the recognition that not all errors are of equal severity.

Composite metrics combine multiple evaluation metrics into a single score, often using weighted averages. This approach allows for a more holistic assessment by balancing different aspects of performance. For example, in studies evaluating the user experience of chatbots, composite metrics integrate usability, engagement, and error rate. Weights for such metrics can be derived empirically, for example, using principal component analysis (PCA), based on empirical patterns in user interaction.

Using methods such as PCA to determine weights provides a scientifically sound approach to assigning importance to different components of a metric, moving beyond arbitrary assignments. If a composite metric can balance usability, engagement, and error rate for chatbots, a similar framework can be applied to text quality assessment. This allows for the flexibility to tailor the metric using weights to meet the specific priorities of different ASR, NLP, or OCR use cases. For example, character accuracy might be a priority for serial number recognition in banking, while word accuracy might be a priority for meeting transcription, and semantic relevance might be a priority for conversational AI.

Despite their widespread use, traditional error rate metrics have several critical limitations that reduce their ability to provide a complete and accurate assessment of text quality [10]. Using WER, CER, and Levenshtein distance alone or in simple combinations has significant limitations. In particular, these metrics are characterized by a lack of semantic sensitivity, since WER and CER treat all errors the same, regardless of their impact on meaning. For example, replacing “погода” with “погоди” may be only a single word error, but completely change the meaning of the sentence. WER does not distinguish “Я люблю фрукт” from “Я люблю фрукти”.

Another disadvantage is word order sensitivity, as WER and CER do not take word order into account, which can be critical for NLP tasks. For example, “собака вкусила хлопчика” and “хлопчика вкусила собака” will have a high WER, even though they have not different contexts.

In particular, traditional error rate metrics are insensitive to the semantic meaning and importance of words, WER and CER treat all errors equally, regardless of their impact on the meaning or importance of the word, or ignore specialized terminology. For example, a typographical error such as “опроцювання” instead of “опрацювання” has the same WER as “день” instead of “пень”, despite their radically different semantic consequences. Similarly, “самоповага” has the same WER as “само повага”. This means that the metrics do not distinguish between critical errors that change meaning from minor errors that do not affect comprehension. The main reason for this is that the underlying Levenshtein distance algorithm on which WER and CER are based assigns the same weight to each editing operation, resulting in a uniform weighting of all types of errors. This equal weighting makes WER and CER poor indicators for human perception of quality, as people implicitly weigh errors differently depending on their impact.

Standard WER calculations often normalize text by removing punctuation and ignoring capitalization, thereby ignoring these critical aspects of text quality and readability. Although punctuation and capitalization are important for readability, WER does not take this information into account. Grammatical errors are also not typically directly evaluated by WER/CER. This means that if punctuation is removed during normalization, WER cannot account for punctuation errors, even if they are important for a particular application.

Problem of text normalization and different lengths transcriptions is another limitation of WER and CER. Inconsistent text normalization, for example, “5 доларів” instead “5 $” may artificially inflate the WER. In addition, WER depends on the length of the transcription, as longer texts have more room for errors. The WER value may exceed 1 or 100 %, which may be counterintuitive to the “error rate”.

DOI 10.15588/1607-3274-2026-1-7

3 MATERIALS AND METHODS

To overcome the limitations of traditional metrics and integrate the advantages of semantic and weighted approaches, a new hybrid metric is proposed, namely an improved error rate based on the Levenshtein distance and the use of WELER weights.

WELER is designed as a multi-layered approach that provides a comprehensive assessment of text quality. It integrates quantitative error metrics WER and CER with qualitative semantic understanding within customizable and weighted structure.

The proposed metric uses the Levenshtein distance for reliable alignment at both the word and character levels. This will ensure error counting and will serve as a structural basis for higher-level analysis.

To provide semantic and contextual understanding, WELER includes a semantic similarity component to assess the conveyed meaning, which is a major limitation of traditional WER/CER. This component assesses how well the generated text preserves the intended meaning even if the lexical forms differ.

Entering weighting factors will allow users to prioritize different aspects of text quality, such as character accuracy, word accuracy, semantic relevance, depending on their specific application and domain requirements. This solves the problem of "evenly weighting" errors in WER/CER. The weights in this approach can be either manually entered by the user or calculated using the principal component analysis method or based on empirical patterns in user interaction. The goal of using the principal component analysis method is to obtain a weighted combination of error rate metrics, where the weights will be determined automatically, without manual adjustment.

WELER is a composite measure, potentially expressed as a weighted sum of normalized error components or a weighted average of the accuracy components. The goal is a single, interpretable measure that reflects the overall quality of a text.

Weighted word error rate W_{wer} includes differentiated penalties for substitutions, deletions, and insertions at the word level

$$W_{wer} = \frac{w_S \cdot S_{word} + w_D \cdot D_{word} + w_I \cdot I_{word}}{N_{word}}. \quad (3)$$

Similarly to W_{WER}, but at the symbol level, (4) is calculated. This will allow get a more detailed error assessment

$$W_{cer} = \frac{w_S \cdot S_{char} + w_D \cdot D_{char} + w_I \cdot I_{char}}{N_{char}}. \quad (4)$$

The third indicator used in the calculation of *WELER* is semantic error indicator *SemErr*. It is obtained from a semantic similarity metric, for example, BERTScore, a modified component of SeMaScore or similar components (depending on the specifics of the task, language, etc.), normalized to represent "error" rather than similarity (6). To calculate the semantic error indicator, it is advisable to use the semantic distance mechanism. Semantic distance is defined as the distance between pairs of reference and hypothetical texts in the space of embeddings at the sentence level, usually using models such as RoBERTa, and cosine similarity (5)

$$SemDist\ [E, H] = 1 - \frac{e \cdot h}{\|e\| \cdot \|h\|}, \quad (5)$$

where E and H are vector representations of reference and hypothetical text

$$SemErr = \frac{1 - \frac{Emb_{input} \cdot Emb_{output}}{\|Emb_{input}\| \cdot \|Emb_{output}\|}}{2}, \quad (6)$$

where Emb_{inut} – vector of embeddings of the reference text; Emb_{output} – vector of recognized text embeddings.

The general WELER formula is to use a weighted combination of the above components (7).

$$WELER = \alpha \cdot W_{wer} + \beta \cdot W_{cer} + \gamma \cdot SemErr, \quad (7)$$

where α, β, γ are global weighting factors, the sum of which is 1, allowing to prioritize word accuracy, character accuracy or semantic correspondence. These factors can be adjusted depending on the specific of application.

Table 1 lists the components of the proposed hybrid metric.

Table 1 – WELER components and weighing scheme

Metric component	Granularity	Customizable weighting factors	Purpose/Benefit
Weighted word error rate W_{wer}	Word level	w_S, w_D, w_I	Lexical accuracy, flexible penalization of different types of word errors
Weighted error rate in symbols W_{cer}	Character level	w_S, w_D, w_I	Detailed error detection, sensitivity to small typos, importance for numerical data.
Semantic error indicator *SemErr*	Semantic Sentence Level	In the internal scales of the model	Preserving meaning, understanding context, taking into account the impact of low-level errors on semantics.
General *WELER*	General	α, β, γ	Comprehensive quality assessment, setting priorities between lexical accuracy and semantic fidelity.

DOI 10.15588/1607-3274-2026-1-7

WELER value ranges from 0 to 1, where 0 is a complete match between the reference and generated text.

The weights α, β, γ in this approach can be either manually entered by the user depending on the task at hand, or calculated using the principal component analysis method. For example, for optical serial number recognition tasks, the coefficient β will be high to prioritize character accuracy. For general voice recognition, α and γ can be balanced to take into account both word accuracy and meaning preservation.

Below is a method for calculating weights using the principal component analysis (PCA) method. The purpose of PCA is to obtain a weighted combination of the above metrics, where the weights will be determined automatically. The principal components are the directions with the greatest variance in the data. Accordingly, this approach allows us to understand which metrics contribute the most to the variability, determines the relative weight of each metric. PCA also allows us to take into account the task category, for example, recognition of short commands, dialogues, long texts, technical documents, etc., and to distribute weights for WER, CER and semantic similarity relative to the category, since the task category can significantly affect the distribution of weights.

The first step for the principal component analysis method is to construct an observation matrix (8) based on data for *N* examples and *d* metrics, in this case *d*=3, since 3 metrics are used.

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & x_{N3} \end{bmatrix}, \quad (8)$$

where $x_{i1} - W_{wer}; x_{i2} - W_{cer}; x_{i3} - SemErr$.

Normalization can be performed to avoid the impact of differences in metric scales (9)

$$\tilde{x}_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}, j = 1,2,3. \quad (9)$$

The next step is to construct a covariance matrix of size, which describes how the values of *WER*, *CER* change. and *SemErr* together. Next, the direct PCA task is performed, that is, the search for eigenvalues and vectors (10) is performed:

$$\sum v = \lambda \cdot v. \quad (10)$$

The next stage is the interpretation of the weights based on the first principal component, i.e. the eigenvector (11)

$$W_1 = (W_{WER}, W_{CER}, W_{SemErr}), \quad (11)$$

where W_{WER}, W_{CER}, W_{SemErr} are values corresponding to the largest λ.

Obtained weights after normalization can be interpreted as the relative importance of metrics in the overall indicator, i.e.

$$\alpha = |w_{wer}|, \beta = |w_{cer}|, \gamma = |w_{SemErr}|, \alpha + \beta + \gamma = 1.$$

Since there are different areas of application of the proposed hybrid metric in relation to the task, it is advisable to introduce a task category that will affect the automatic calculation of weight coefficients. Let *k* be given task categories, for example, teams, dialogs, documents. For each category, it is worth building your own matrix (12):

$$X^{(k)} = \{(WER_i, CER_i, SemErr_i) \mid task_i = k\}. \quad (12)$$

Then, for each such matrix, it is necessary to calculate its covariance matrix, the eigenvalues and vectors of which are represented as (13)

$$\sum^{(k)} v^{(k)} = \lambda^{(k)} \cdot v^{(k)}. \quad (13)$$

Accordingly, for each category, its own weight vector (7) will be obtained

$$W^{(k)} = (W_{WER}^{(k)}, W_{CER}^{(k)}, W_{SemErr}^{(k)}). \quad (14)$$

Another way to factor task categories into the weighting is to factorize with the category, which means that the total weights are composed of a global part and a category correction. Then the weight vector is computed on all data, and the correction is computed as the difference between the local and global weights. This approach allows us to see how the category shifts the importance of the metric, for example, for the command category the correction value for CER will be greater than 0, since symbols are the most important.

Using task categories when calculating weight coefficients allows you to obtain a flexible integral metric that can automatically adapt to the type of task.

The choice of normalization strategy should depend on the specific requirements of the application. For example, if punctuation errors are important for a given task, then punctuation should not be removed during preprocessing. This emphasizes that the decision to include or exclude certain types of errors through normalization directly affects what WELER measures as “quality”. This requires that the WELER coefficients be flexible enough to reflect these preprocessing choices.

WELER, by integrating semantic understanding, allows us to solve specific challenges that are difficult for traditional metrics. In particular, the proposed metric helps in working with homophones and similar – sound-

DOI 10.15588/1607-3274-2026-1-7

ing words. The semantic component can help distinguish lexically similar but semantically different words that WER by itself would mark equally. This allows the metric to reflect the real impact of such errors on understanding.

Another task that the proposed metric will help solve is working with ambiguous word boundaries, when For languages that do not have clear spaces between words, such as Chinese, the CER component and its underlying character-level alignment based on Levenshtein distance become especially valuable.

In automatic audio recognition tasks, audio quality affects error rates, but the semantic component of WELER can help assess whether meaning of the text even under increased lexical errors caused by noise. This allows W ELER to go beyond simple detection what errors have occurred, to understanding how much those errors affect understanding. This shift in focus from purely technical accuracy to a more user-oriented definition of quality is a significant step.

4 EXPERIMENTS

Table 2 shows examples of using *WELER* with weighting coefficients α=0.3, β=0.3, γ=0.4.

The table shows the calculated metrics *WER*, *CER*, *SemErr* and the calculated value of the proposed metric WELER. Row 4 shows the use of synonyms and paraphrasing, where the metrics WER and CER show a poor error result of 0.667 and 0.5758, since for these metrics the words. “Машина” and “Авто” are two completely different words, which confirms the unsuitability of these metrics for working with synonyms, giving a high error rate. The semantic distance for the considered sentence is low and is 0.0108, since the meaning of the sentence is perfectly preserved. Accordingly, WELER has a value of 0.377 and demonstrates a significantly better, lower, error result than WER/CER.

Row 5 shows complete sentence divergence and loss of context. WELER combines high scores of all three metrics and produces a high overall error. This demonstrates that WELER does not simply underestimate the scores, the proposed metric responds adequately when the meaning of the text is truly lost.

Another example of reordering and the use of synonyms in the text is shown in row 10. The word order is completely reversed, so *WER* and *CER* record a large error, almost 0.8571 for *WER*. WELER takes into account the high structural errors with *WER*/*CER*, but balances them with an almost perfect semantic match. The result of 0.486e is a much more adequate assessment of the text, the WELER value is almost twice as good as the *WER* value.

Table 2 – Examples of using WELER [17]

#	Reference text	Generated text	WER	CER	SemErr	WELER
1	Привіт, світ! Як у вас справи?	Привіт, світ! Як у вас справи?	0	0	0	0
2	Зараз чудова погода для прогулянки.	Зараз чудова пригода для прогулянки.	0.2	0.0571	0.0771	0.108
3	Я люблю програмування на Python	Я люблю програмування на Pyton.	0.2	0.0645	0.0685	0.1068
4	Машина їде дуже швидко по дорозі.	Авто рухається швидко по дорозі.	0.6667	0.5758	0.0108	0.377
5	Українська мова дуже мелодійна	Китайська кухня дуже смачна.	0.75	0.5483	0.2993	0.5092
6	Привіт, як справи? Усе добре!	Привіт як справи Усе добре	0.6	0.1034	0.0356	0.2253
7	Він пішов до школи сьогодні.	Він пішли до школи сьогодні.	0.2	0.0714	0.0004	0.0816
8	Котику-муркотику з м’яким животиком.	Де сметанка що була тут ще зранку.	1	0.8889	0.2604	0.6708
9	Адреса: вул. Свободи, 10	Адреса: вул. Свободи, 1О	0.25	0.0417	0.0159	0.0939
10	Автомобіль швидко рухався дорогою до міста Львів.	До міста Львів дорогою швидко рухалося авто.	0.8571	0.7347	0.0222	0.4864
11	Розробка штучного інтелекту змінює світ.	Розробка ШІ змінює світ.	0.4	0.45	0.0623	0.2799
12	Щоб встигнути потрібно плисти за течією до сходу сонця.	Щоб встигнути потрібно плести кошики до сходу сонця.	0.3333	0.1818	0.1538	0.2161
13	Плисти вперед до сходу сонця.	Плести кошики до сходу сонця.	0.4	0.2414	0.1596	0.2562

Figure 1 shows a distribution diagram of four speech recognition quality assessment metrics. It is clear from the diagram that *WER* has a fairly wide range from 0 to almost 1. The median for *WER* is around 0.5, which means that half of the examples have errors in around 50% of the words. The *CER* metric also has a wide range of, but a lower median of around 0.25, so the system performs better at the symbol level than at the word level. The diagram also illustrates that there are a significant number of examples with very large errors. The *SemErr* metric shows that most of the values are very low, 0–0.2, meaning that the meaning of the sentences is mostly preserved even

DOI 10.15588/1607-3274-2026-1-7

with errors in the words. There are a few outliers down to around 0.6.

Table 2 – Examples of using WELER

#	Reference text	Generated text	WER	CER	SemErr	WELER
1	Привіт, світ! Як у вас справи?	Привіт, світ! Як у вас справи?	0	0	0	0
2	Зараз чудова погода для прогулянки.	Зараз чудова пригода для прогулянки.	0.2	0.0571	0.0771	0.108
3	Я люблю програмування на Python	Я люблю програмування на Pyton.	0.2	0.0645	0.0685	0.1068
4	Машина їде дуже швидко по дорозі.	Авто рухається швидко по дорозі.	0.6667	0.5758	0.0108	0.377
5	Українська мова дуже мелодійна	Китайська кухня дуже смачна.	0.75	0.5483	0.2993	0.5092
6	Привіт, як справи? Усе добре!	Привіт як справи Усе добре	0.6	0.1034	0.0356	0.2253
7	Він пішов до школи сьогодні.	Він пішли до школи сьогодні.	0.2	0.0714	0.0004	0.0816
8	Котику-муркотику з м’яким животиком.	Де сметанка що була тут ще зранку.	1	0.8889	0.2604	0.6708
9	Адреса: вул. Свободи, 10	Адреса: вул. Свободи, 1О	0.25	0.0417	0.0159	0.0939
10	Автомобіль швидко рухався дорогою до міста Львів.	До міста Львів дорогою швидко рухалося авто.	0.8571	0.7347	0.0222	0.4864
11	Розробка штучного інтелекту змінює світ.	Розробка ШІ змінює світ.	0.4	0.45	0.0623	0.2799
12	Щоб встигнути потрібно плисти за течією до сходу сонця.	Щоб встигнути потрібно плести кошики до сходу сонця.	0.3333	0.1818	0.1538	0.2161
13	Плисти вперед до сходу сонця.	Плести кошики до сходу сонця.	0.4	0.2414	0.1596	0.2562

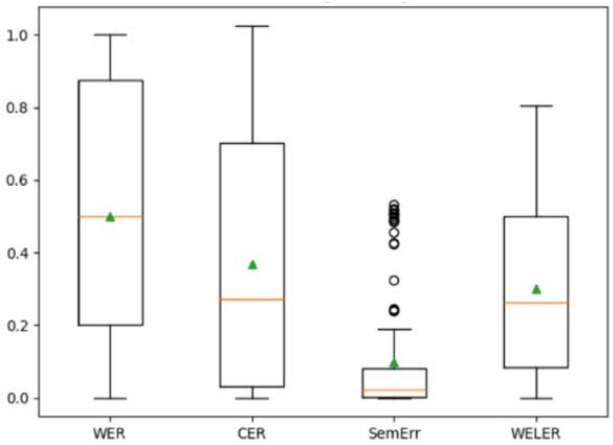


Figure 1 – Boxplot of the distribution of four speech recognition quality assessment metrics

The proposed WELER metric exhibits a smaller spread than *WER* and *CER*, namely 0–0.8, with a median of about 0.3. The proposed metric balances between formal and semantic errors. In general, we can conclude that at the symbol level the system makes fewer errors than at the word level, and semantics is mostly preserved even when there are spelling or lexical errors. The combined WELER metric gives a more balanced assessment than WER or CER separately. Figures 2 and 3 show graphs of pairwise dependencies between the studied metrics.

The analysis of the relationship between the traditional recognition metrics *WER*, *CER*, the proposed WELER metric and semantic errors *SemErr* demonstrates different levels of correlation. In particular, the comparison of *WER* and *SemErr* shows that even with high values of errors at the word level of 0.8–1.0, the *SemErr* indicator often remains low, which indicates the possibility of preserving the meaning despite numerous orthographic or syntactic deviations. The same is true for *CER*, where, although at high values of 0.8–1.0 there are cases of significant semantic distortions >0.4, most of the data is concentrated in the range of low values *SemErr* <0.1. This allows us to conclude that errors at the symbol level do not always lead to the destruction of meaning, but their excessive number significantly reduces semantic accuracy. The closest relationship with *SemErr* is revealed by the combined WELER metric; a smooth increase in semantic errors is observed with an increase in its value, which confirms the higher correspondence of WELER to the level of content preservation compared to *WER* and *CER*.

DOI 10.15588/1607-3274-2026-1-7

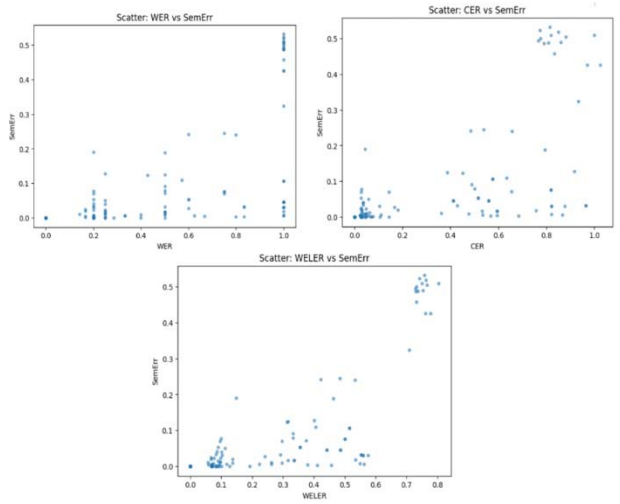


Figure 2 – Pairwise dependencies between SemErr and other metrics

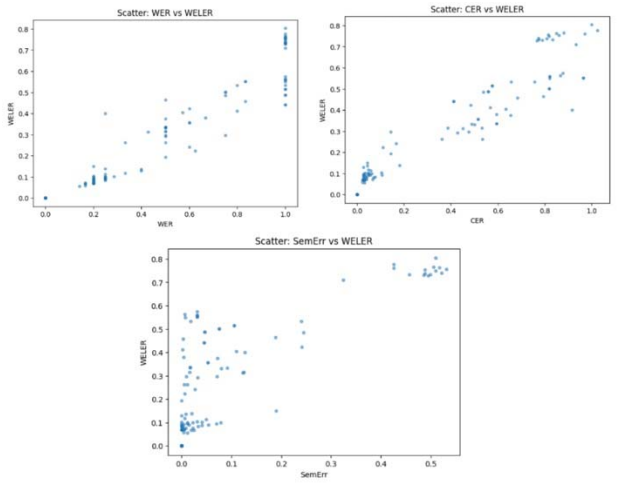


Figure 3 – Pairwise dependencies between WELER and other metrics

Analysis of the relationships between different metrics shows that WELER demonstrates a high correlation with both *WER* and CER, in the first case, an almost linear increase in WELER with an increase in *WER* is observed, which indicates the inclusion of *WER* as a key component of the combined metric; in the second case, a similar pattern is recorded, which confirms the integration of *CER* into the WELER structure. At the same time, the relationship between *SemErr* and WELER is less unambiguous, most examples are characterized by a low level of semantic errors 0–0.1 with a wide range of WELER values, however, there are also cases where a high *SemErr* indicator 0.4–0.5 is accompanied by an increased WELER. This indicates that WELER is able to respond to semantic distortions, but its value can remain high even with preserved content.

It can be argued that *WER* and *CER* exhibit an almost linear relationship with WELER, while *SemErr* correlates with them less strongly, since *WER* and *CER* capture formal errors rather than substantive ones. This confirms the ability of WELER to display more balanced results, reflecting at the same time the nature of errors at the form level and partially taking into account semantic accuracy.

In order to test the approach to automatically determining weight coefficients using the principal components method, a dataset was created consisting of two categories: “text” (fictional sentences) and “number” (serial numbers). After applying the algorithm to the values of the *WER*, *CER*, and *SemErr* metrics within each category, automatically determined weight coefficients were obtained, the results of which are shown in Table 3.

Table 3 – Automatically determined weights using the principal component method

Categoty	*WER*	*CER*	*SemDist*
number	0.3826	0.37718	0.2402
text	0.3470	0.3483	0.3046

Table 4 shows a comparison of WELER values calculated using manually entered weighting coefficients (α=0.3, β=0.3, γ=0.4) and weighting coefficients determined by applying the principal component method (α=0.347, β=0.3483, γ=0.3046).

Table 4 – Examples of calculating the WELER metric with manually selected coefficients and using PCA [17]

#	Reference text	Generated text	WELER	WELER (PCA)
1	Привіт, світ! Як у вас справи?	Привіт, світ! Як у вас справи?	0	0
2	Зараз чудова погода для прогулянки.	Зараз чудова пригода для прогулянки.	0.108018	0.112825
3	Я люблю програмування на Python	Я люблю програмування на Pyton.	0.106754	0.112746
4	Машина їде дуже швидко по дорозі.	Авто рухається швидко по дорозі.	0.377028	0.435193
5	Українська мова дуже мелодійна	Китайська кухня дуже смачна.	0.509243	0.542481
6	Привіт, як справи? Усе добре!	Привіт як справи Усе добре	0.225291	0.255112
7	Він пішов до школи сьогодні.	Він пішли до школи сьогодні.	0.081602	0.09442
8	Котику-муркотику з м’яким животиком.	Де сметанка що була тут ще зранку.	0.670843	0.736009
9	Адреса: вул. Свободи, 10	Адреса: вул. Свободи, 1О	0.093865	0.10612
10	Автомобіль швидко рухався дорогою до міста Львів.	До міста Львів дорогою швидко рухалося авто.	0.486422	0.560139
11	Розробка штучного інтелекту змінює світ.	Розробка ШІ змінює світ.	0.279902	0.314533
12	Щоб встигнути потрібно плисти за течією до сходу сонця.	Щоб встигнути потрібно плести кошики до сходу сонця.	0.216085	0.225878
13	Плисти вперед до сходу сонця.	Плести кошики до сходу сонця.	0.25624	0.271503

DOI 10.15588/1607-3274-2026-1-7

WELER calculated using the principal component method differs only in scaling, the values are slightly shifted, but the general approach of the balanced metric value is preserved.

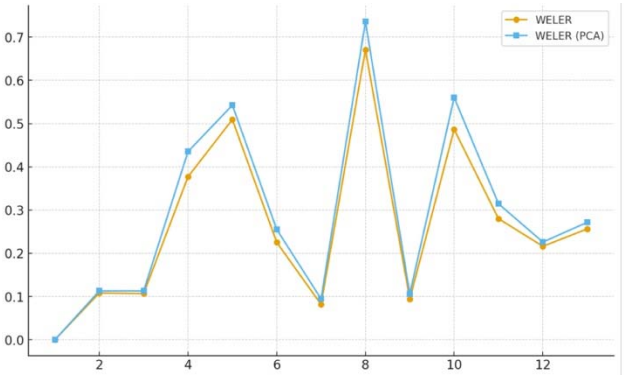


Figure 4 – Pairwise dependencies between WELER and other metrics

As can be seen from Figure 4 and Table 4, in most cases the WELER and WELER (PCA) values demonstrate almost parallel dynamics, however, the application of the principal components method somewhat enhances the differences between individual examples. The highest values, approximately 0.5–0.5, are recorded in examples No. 5, No. 8 and No. 10, where significant semantic distortion of the text was observed. The lowest values, approximately 0.0–0.1, are characteristic of examples No. 1, No. 2, No. 3, No. 7 and No. 9, where deviations were limited to minor spelling or grammatical errors. The PCA method shows higher sensitivity to semantic differences, which is especially noticeable in example No. 4, where the difference between the values 0.38 and 0.43 makes the deviation more pronounced.

5 RESULTS

As a result, WELER, compared to the classic WER and CER, provides a more balanced assessment of text quality, as it penalizes significant semantic deviations but mitigates the penalty for minor spelling or stylistic differences, which better reflects the real impact of errors on comprehension.

Proposed WELER metric has the potential to improve text quality assessment in a wide range of artificial intelligence tasks, providing a more tailored assessment.

For automatic speech recognition, WELER can provide a more granular analysis of speech-to-text systems. It is able to distinguish between minor phonetic recognition errors and critical semantic errors. The metric can be configured to prioritize verbatim reproduction for legal transcriptions (high α and β) or semantic understanding for conversational AI (high γ). This flexibility in tuning the coefficients is a major advantage of WELER for a variety of applications.

For natural language processing tasks such as text generalization or machine translation, WELER can evaluate both lexical accuracy, such as correct word choice, and semantic preservation, ensuring the accuracy of form and meaning of the generated text. The coefficients can be adjusted to penalize grammatical errors more severely if fluency is a top priority.

For OCR tasks, the application of WELER can provide a more robust evaluation of systems, especially for documents where both character-level accuracy is important, such as serial numbers, financial data, and word-level accuracy, such as free text fields. The semantic component can even assess whether contextual meaning to the data obtained. Providing a more comprehensive and customizable metrics WELER can facilitate better decision-making in the development and deployment of models by allowing engineers to optimize systems for specific real-world performance criteria rather than a generic, potentially misleading error rate. This can lead to significant cost savings by reducing the need for extensive manual post-processing and quality control in document digitization, transcription, and data entry processes. It can also be used to assess the similarity of texts and translation quality with reference translations with corresponding weight changes for these tasks.

It can also be used to assess the similarity of texts and translation quality with reference translations with corresponding weight changes for these tasks.

6 DISCUSSION

WELER is the first step towards more comprehensive metrics for assessing text quality. Future research will focus on aspects of improving dynamic weighting. In particular, it is worth continuing to investigate methods for automatically adjusting WELER coefficients based on the domain of the input text, its complexity, or perceived criticality, for example, taking into account the thematic focus of the texts (medical and everyday conversation, etc.). This task can be solved by using machine learning approaches to learn optimal weighting coefficients.

Another important area of further research is integration with metrics for detecting grammatical errors and punctuation, in particular including established GEC metrics, e.g. M2, ERRANT, GLEU, SARI, and punctuation error metrics, e.g. FER, NER for punctuation, to the WELER framework can contribute to a truly holistic assessment of linguistic quality. Although the integration of these metrics is important for certain tasks, it will significantly increase the complexity of WELER and create noise for speech recognition tasks, since such tasks mostly do not have punctuation.

Researching how WELER and its components work in different languages, especially those with complex morphological structures or without explicit word boundaries, and adapting weighting schemes accordingly is another important direction.

CONCLUSIONS

Text quality assessment is a cornerstone for the development and improvement of artificial intelligence systems in such critical areas as ASR, NLP and OCR. Although traditional metrics such as WER and CER are widely used and based on the robust concept of Levenshtein distance,

DOI 10.15588/1607-3274-2026-1-7

their insensitivity to semantic meaning, grammatical correctness, and punctuation features limits their ability to reflect the true quality of text from a human perceptual perspective. This discrepancy highlights the urgent need to develop more comprehensive and adaptive assessment tools.

Proposed WELER metric is an alternative solution in this direction. It integrates accurate word and character level error counting, using Levenshtein distance as a basis, with advanced semantic similarity methods based on contextual embeddings. This allows WELER to take into account not only what was incorrectly recognized, but also how much this error affects the meaning and understanding of the text. The inclusion of self-adjusting weights depending on the text category is a key feature of WELER, which allows adapting the metric to the specific requirements of different applications and domains, prioritizing those aspects of quality that are most critical for a particular task.

However, WELER, like all metrics based on reference data, relies on accurate and consistent human-verified transcriptions. Errors in the reference data can affect the accuracy of the assessment. Therefore, for complex metrics, the quality and representativeness of these data are especially important, since semantic and weighted errors are much more sensitive to the quality of the annotation than simple word counts. The implementation of WELER, while promising significant benefits, requires careful consideration of issues such as computational complexity, empirical determination of optimal weights, and ensuring high quality of reference data. Future research could extend WELER to include dynamic weight adjustment, integration with grammatical and punctuation error metrics, and exploration of LLM-based scoring capabilities, which could lead to further developments in the field of text quality assessment. Ultimately, WELER represents a robust and flexible framework for more accurate and holistic text quality assessment, which will contribute to the development of more efficient and user-friendly AI systems.

ACKNOWLEDGEMENTS

The authors would like to thank the industry researchers, authors of works in references, and the anonymous reviewers. This research received no external funding.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: equally significant, evenly proportional.

Data availability: The manuscript has no associated data except in the example within itself.

Software availability: The manuscript has no specific associated software.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies to create the submitted work. Artificial intelligence was used for translation and grammar checks.

REFERENCES

1. Hamed I. Benchmarking Evaluation Metrics for Code-Switching Automatic Speech Recognition, *2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 2023 : proceedings*. [Piscataway], IEEE, 2023, pp. 999–1005. DOI: 10.1109/SLT54892.2023.10023181
2. Measure and improve speech accuracy, *Cloud Speech-to-Text Documentation*. Available at: https://cloud.google.com/speech-to-text/docs/speech-accuracy (accessed: 22 July 2025).
3. Dumyn A., Fedushko S., Syerov Y. Review of Automatic Speech Recognition Systems for Ukrainian and English Language, *Data-Centric Business and Applications : proceedings*. Cham, Springer, 2024. (Lecture Notes on Data Engineering and Communications Technologies, Vol. 212).
4. Shakhovska N., Shvorob I. The method for detecting plagiarism in a collection of documents, *2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), Lviv, Ukraine, 2015 : proceedings*. [Piscataway], IEEE, 2015, p. 142–145. DOI: 10.1109/STC-CSIT.2015.7325453
5. Sasindran Z., Yelchuri H., Prabhakar T. V., Rao S. A new hybrid evaluation metric for automatic speech recognition tasks, *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) : proceedings.* [Piscataway], IEEE, 2023, pp. 1–7. DOI: 10.48550/arXiv.2211.01722
6. Kim S., Arora A., Le D., Yeh C.-F., Fuegen C., Kalinli O., Seltzer M. L. Semantic distance: A new metric for ASR performance analysis towards spoken language understanding, *arXiv preprint arXiv:2104.02138*, 2021. Link: https://arxiv.org/abs/2104.02138
7. Sasindran Z., Yelchuri H., Prabhakar T. V. SeMaScore: a new evaluation metric for automatic speech recognition tasks, *arXiv preprint arXiv:2401.07506*, 2024. Link: https://arxiv.org/abs/2401.07506
8. Phukon B., Zheng X., Hasegawa-Johnson M. Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches, *arXiv preprint arXiv:2506.16528*, 2025. Link: https://arxiv.org/abs/2506.16528
9. Zhang T., Kishore V., Wu F., Weinberger K. Q., Artzi Y. BERTScore: Evaluating text generation with BERT, *arXiv preprint arXiv:1904.09675*, 2019. Link: https://arxiv.org/abs/1904.09675
10. James J., Gopinath D. P. Advocating character error rate for multilingual ASR evaluation, *arXiv preprint arXiv:2410.07400*, 2024. Link: https://arxiv.org/abs/2410.07400
11. Van Schaik T., Pugh B. A field guide to automatic evaluation of LLM-generated summaries, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, ACM, 2024, pp. 2832–2836.

DOI 10.15588/1607-3274-2026-1-7

12. Arockiya Jerson J., Preethi N. An analysis of Levenshtein distance using dynamic programming method, *Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications (ICMISC 2022).* Singapore, Springer Nature Singapore, 2023, pp. 525–532.
13. Greenacre M., Groenen P. J., Hastie T., d'Enza A. I., Markos A., Tuzhilina E. Principal component analysis, *Nature Reviews Methods Primers*, Vol. 2, № 1, Article 100.
14. Measuring the Accuracy of Automatic Speech Recognition Solutions, *arXiv*. Available at: https://arxiv.org/html/2408.16287v1 (accessed: 22 July 2025).
15. Hunt M. A. Word Errors and the Significance of Weighted Accuracy Measures, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 1990.
16. Neudecker C., Baierer K., Gerber M., Clausner C., Antonacopoulos A., Pletschacher S. A survey of OCR evaluation tools and metrics, *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing.* New York, ACM, 2021, pp. 13–18.
17. Dumyn A.R. Hibrydna metryka otsinky yakosti tekstu na osnovi kontekstnoho zvazhuvannya, *Tavriys'kyy naukovyy visnyk. SeriyaL Tekhnichni nauky*, 2025, №4, ch. 1, pp. 85-93

Received 20.10.2025.
Accepted 13.01.2026.
Published 27.03.2026.

УДК 004.93

WELER: КОМПЛЕКСНИЙ ПОКАЗНИК ДЛЯ ОЦІНКИ ЯКОСТІ ТЕКСТУ

Думин А. Р. – аспірант кафедри систем штучного інтелекту Національного університету «Львівська політехніка», Львів, Україна. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0000-0003-2111-2899.

Шаховська Н. Б. – д-р техн. наук, професор, ректор Національного університету «Львівська політехніка», Львів, Україна. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0000-0002-6875-8534.

АНОТАЦІЯ

Актуальність. Оцінка якості тексту є важливою для надійного штучного інтелекту, який обробляє мову. В ASR вона відображає, наскільки точно мовлення стає текстом; в OCR – наскільки точно зображення перетворюють текст; а в NLP – наскільки правильними та зв'язними є виходи.

Мета. Метою роботи є створення складної метрики для оцінки якості тексту.

Метод. Класичні метрики WER та CER є вузькими: вони фіксують лише лексичні редагування, однаково зважують усі зміни, ігнорують контекст та семантику, і часто пропускають пунктуацію та регістр, маскуючи проблеми читабельності та типи помилок. апропонована метрика WELER інтегрує точний підрахунок помилок на рівні слів та символів, використовуючи відстань Левенштейна як основу, з передовими методами семантичної подібності, заснованими на контекстному вбудовуванні. Це дозволяє WELER враховувати не лише те, що було неправильно розпізнано, але й те, наскільки ця помилка впливає на значення та розуміння тексту. Включення самоналаштовуваних ваг залежно від категорії тексту є ключовою особливістю WELER, яка дозволяє адаптувати метрику до конкретних вимог різних застосувань та областей, надаючи пріоритет тим аспектам якості, які є найбільш критичними для конкретного завдання.

Результати. Метрика WELER пропонується як ефективний підхід до оцінювання якості тексту. Її концептуальна основа полягає в інтеграції традиційного підрахунку помилок на словесному та символьному рівнях, заснованого на відстані Левенштейна, із сучасними методами оцінювання семантичної подібності, що використовують контекстуальні векторні подання. Такий підхід забезпечує більш комплексне відображення впливу помилок на змістову цілісність та інтерпретованість результатного тексту.

Висновки. WELER, як і всі метрики, засновані на довідкових даних, спирається на точні та послідовні транскрипції, перевірені людиною. Помилки в довідкових даних можуть впливати на точність оцінки. Тому для складних метрик якість та репрезентативність цих даних є особливо важливими, оскільки семантичні та зважені помилки набагато чутливіші до якості анотації, ніж проста кількість слів.

КЛЮЧОВІ СЛОВА: обробка природної мови, оцінка якості тексту, WER, CER, WELER.

ЛІТЕРАТУРА

1. Hamed I. Benchmarking Evaluation Metrics for Code-Switching Automatic Speech Recognition / I. Hamed // 2022 IEEE Spoken Language Technology Workshop (SLT), Doha, Qatar, 2023 : proceedings. – [Piscataway]: IEEE, 2023. – P. 999–1005. DOI: 10.1109/SLT54892.2023.10023181
2. Measure and improve speech accuracy // Cloud Speech-to-Text Documentation. – Available at: https://cloud.google.com/speech-to-text/docs/speech-accuracy (accessed: 22 July 2025).
3. Dumyn A. Review of Automatic Speech Recognition Systems for Ukrainian and English Language / A. Dumyn, S. Fedushko, Y. Syerov // Data-Centric Business and Applications : proceedings. – Cham : Springer, 2024. – (Lecture Notes on Data Engineering and Communications Technologies, Vol. 212).
4. Shakhovska N. The method for detecting plagiarism in a collection of documents / N. Shakhovska, I. Shvorob // 2015 Xth International Scientific and Technical Conference "Computer Sciences and Information Technologies" (CSIT), Lviv, Ukraine, 2015 : proceedings. – [Piscata-

DOI 10.15588/1607-3274-2026-1-7

way] : IEEE, 2015. – P. 142–145. DOI: 10.1109/STC-CSIT.2015.7325453
5. A new hybrid evaluation metric for automatic speech recognition tasks / [Z. Sasindran, H. Yelchuri, T. V. Prabhakar, S. Rao] // 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) : proceedings. – [Piscataway] : IEEE, 2023. – P. 1–7. DOI: 10.48550/arXiv.2211.01722
6. Semantic distance: A new metric for ASR performance analysis towards spoken language understanding / [S. Kim, A. Arora, D. Le et al.] // arXiv preprint arXiv:2104.02138. – 2021. Link: https://arxiv.org/abs/2104.02138
7. Sasindran Z. SeMaScore: a new evaluation metric for automatic speech recognition tasks / Z. Sasindran, H. Yelchuri, T. V. Prabhakar // arXiv preprint arXiv:2401.07506. – 2024. Link: https://arxiv.org/abs/2401.07506
8. Phukon B. Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches / B. Phukon, X. Zheng, M. Hasegawa-Johnson // arXiv preprint arXiv:2506.16528. – 2025. Link: https://arxiv.org/abs/2506.16528
9. BERTScore: Evaluating text generation with BERT / [T. Zhang, V. Kishore, F. Wu et al.] // arXiv preprint arXiv:1904.09675. – 2019. Link: https://arxiv.org/abs/1904.09675
10. James J. Advocating character error rate for multilingual ASR evaluation / J. James, D. P. Gopinath // arXiv preprint arXiv:2410.07400. – 2024. Link: https://arxiv.org/abs/2410.07400
11. Van Schaik T. A field guide to automatic evaluation of LLM-generated summaries / T. Van Schaik, B. Pugh // Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. – New York : ACM, 2024. – P. 2832–2836.
12. Arockiya Jerson J. An analysis of Levenshtein distance using dynamic programming method / J. Arockiya Jerson, N. Preethi // Proceedings of 3rd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications (ICMISC 2022). – Singapore : Springer Nature Singapore, 2023. – P. 525–532.
13. Principal component analysis / [M. Greenacre, P. J. Groenen, T. Hastie et al.] // Nature Reviews Methods Primers. – Vol. 2, № 1. – Article 100.
14. Measuring the Accuracy of Automatic Speech Recognition Solutions // arXiv. – Available at: https://arxiv.org/html/2408.16287v1 (accessed: 22 July 2025).
15. Hunt M. A. Word Errors and the Significance of Weighted Accuracy Measures / M. A. Hunt // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 1990.
16. A survey of OCR evaluation tools and metrics / [C. Neudecker, K. Baierer, M. Gerber et al.] // Proceedings of the 6th International Workshop on Historical Document Imaging and Processing. – New York : ACM, 2021. – P. 13–18.
17. Dumyn A. R. Hibrydna metryka otsinky yakosti tekstu na osnovi kontekstnoho zvazhuvannya / A. R. Dumyn // Tavriys'kyy naukovyy visnyk. SeriyaL Tekhnichni nauky. – 2025. – №4, ch. 1 – P. 85–93

DOI 10.15588/1607-3274-2026-1-7

UDC 004.89

EVALUATION AND QUALITY ASSURANCE OF MIGRATED ABAP CODE USING AN INTEGRAL METRIC AND GENERATIVE ARTIFICIAL INTELLIGENCE MODELS

Pozdnyakov O. A. – Post-graduate student of the Software Tools Department, National University "Zaporizhzhia Polytechnic", Zaporizhzhia, Ukraine. ROR: https://ror.org/03aph1990. ORCID: https://orcid.org/0009-0006-3955-802X.

Parkhomenko A. V. – PhD, Associate Professor of the Software Tools Department, National University "Zaporizhzhia Polytechnic", Zaporizhzhia, Ukraine. ROR: https://ror.org/03aph1990. ORCID: https://orcid.org/0000-0002-6008-1610.

ABSTRACT

Context. Migration automation of legacy custom code when transitioning to the new version of the SAP S/4HANA system using large language models (LLMs) is a promising option. However, the generated code quality assessment remains an unresolved issue, since existing approaches utilize fragmented metrics which do not allow for a comprehensive software code quality assessment and assurance for further use without additional revision.

Objective. The objective of this work is to improve the efficiency of the process of intelligent reengineering of a computer system based on the method of comprehensive assessment and quality assurance of migrated ABAP custom code.

Method. The developed method is based on two key components. The Integral ABAP Quality Score (IAQS) comprehensively takes into account the syntactic, functional, and semantic characteristics of the code and is based on the provisions of the international software quality standards ISO/IEC 25010, ISO/IEC 25040, as well as the theory of composite indicators. The three-stage approach to LLM fine-tuning (Qwen 2.5 Coder 14B) includes continuous pre-training (CPT), parameter-efficient fine-tuning (PEFT), and alignment based on preferences using the ORPO algorithm. At the same time, the use of the developed IAQS metric to form a set of preference data at the alignment stage creates a mechanism for controlled improvement, namely, it determines the direction of LLM adaptation.

Results. The results of experimental studies demonstrate that the implementation of the developed method allows improving both individual indicators of software code quality and the integral metric of IAQS quality assessment as a whole. The final model, trained on the basis of the proposed three-stage approach, achieved a high IAQS value (0.756), which demonstrates a significant improvement compared to the baseline model (0.117).

Conclusions. The study presents a new problem-oriented approach to automated migration of ABAP code during intelligent reengineering of computer systems. The proposed IAQS integral metric is the basis for creating a formalized and objective system for evaluating the quality of software generated by LLM in the context of legacy custom code migrating. It has been demonstrated that consistent fine-tuning of LLM based on a three-stage approach using IAQS provides a significant improvement in the generated software code integral quality indicator.

KEYWORDS: software quality, integral metrics, large language models, migration of legacy custom code, LLM fine-tuning.

ABBREVIATIONS

ABAP is an advanced business application programming;

AST is an abstract syntax tree;

ATC is an ABAP test cockpit;

CPT is a continued pre-training;

ERP is an enterprise resource planning;

IAQS is an integral ABAP quality score;

ISO is the International organization for standardization;

IEC is the International electro technical commission;

LLM is a large language model;

LoRA is a low-rank adaptation;

ORPO is an Odds ratio preference optimization;

PEFT is a parameter-efficient fine-tuning;

QLoRA is a quantized low-rank adaptation;

SFT is a supervised fine-tuning.

NOMENCLATURE

C_{old} is a code fragment in the outdated version of ABAP;

C_{new} is a code fragment in the modern version of ABAP;

C_{newi} is an i-th candidate migration option;

$F_{quality}$ is an integral quality function;

$IAQS_i$ is an integral assessment value for the i-th code variant;

k is a number of candidate migration variants;

L_{ORPO} is an ORPO loss function;

L_{SFT} is a supervised fine-tuning loss function;

L_{OR} is an odds ratio-based loss function;

M_{func} is a functional correctness metric (pass@1);

$M_{LLM}(\theta)$ is a large language model with parameters θ;

M_{sem} is a semantic-structural similarity metric;

M_{syn} is a syntactic correctness metric;

Q is a quality metrics vector;

q_{func} is a functional correctness;

q_{sem} is a semantic-structural similarity;

q_{syn} is a syntactic correctness;

r is a rank of adaptation matrices in LoRA;

w_{syn}, w_{func}, w_{sem} are the weights of IAQS components;

DOI 10.15588/1607-3274-2026-1-8

y_w is a selected (better) answer in a pair of preferences;

y_l is a rejected (worse) answer in a pair of preferences;

α is a scaling coefficient in LoRA;

θ are the model parameters;

θ^* are the model optimal parameters;

λ is a weight factor (ORPO balance hyper-parameter);

$p(y|x)$ is a probability of y reply generation for x input data;

$\sigma(.)$ is a sigmoid function.

INTRODUCTION

The transition to the modern SAP S/4HANA software platform is currently a strategic initiative for many organisations both in Ukraine and around the world. SAP S/4HANA is based on the SAP HANA in-memory database, which provides significant performance gains but also requires the computer system's software code to be reviewed and adapted in order to fully utilise its capabilities [1].

One of the key challenges along this way is the legacy custom code in the ABAP language, developed over decades of use SAP ERP versions by companies and enterprises [2]. Migrating of custom code is one of the main challenges when transitioning to SAP S/4HANA, accounting for up to 40% of the total project work scope [3]. Manual correction of legacy custom code is not only labour-intensive but also a risky process that can lead to project delays and errors in the production system [1].

Despite the availability of static analysis tools by SAP (ABAP Test Cockpit, Custom Code Migration App) [2] which help identify problematic code fragments, the code transformation process remains largely manual. Automation is limited to simple template replacements, while complex migration scenarios which require an understanding of business logic and context still require significant effort from skilled ABAP developers. This challenge requires the application of modern intelligent reengineering approaches to fully implement the potential of the SAP S/4HANA platform.

In [4], the authors analysed existing problems and developed a methodology for migrating of ABAP custom code based on intelligent reengineering methods and models. A key feature was the justification of the need to use locally deployed open-source Large Language Models (LLMs), which makes it possible to avoid both the significant data security risks and licence agreement violations inherent in cloud-based LLMs.

The formalised method for selecting the optimal LLM [5] developed by the authors is based on a hybrid AHP-TOPSIS approach and allows the identification of leading candidates (in particular, the Qwen, DeepSeek and Llama models) for solving of the ABAP code migration problem.

Although LLMs have already demonstrated significant potential for generating software code, the problem of objective comprehensive assessment and quality assurance of the generated code remains. Therefore, the research topic is relevant.

The object of study is the process of intelligent reengineering of a computer system when migrating to a new version of SAP S/4HANA.

The subject of study is methods for evaluating and ensuring the quality of migrated ABAP code inherited from outdated versions of SAP ERP.

The purpose of the work is to improve the efficiency of the process of intelligent reengineering of a computer system based on the method of comprehensive assessment and quality assurance of migrated ABAP custom code.

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

– formally define and justify the IAQS integral metric based on the international software quality standards ISO/IEC 25010 [6], ISO/IEC 25040 [7];

– develop a three-stage LLM fine-tuning methodology (continuous pre-training, parameter-efficient fine-tuning, advantage-based alignment), where IAQS is used as an integral quality assessment criterion for forming data on advantages;

– experimentally verify the effectiveness of the proposed approach on real ABAP code migration tasks;

– conduct a statistical analysis of the results and investigate the impact of each fine-tuning stage on the components of the integral metric.

1 PROBLEM STATEMENT

The research task is formalised as follows.

Let C_{old} be a code fragment in an outdated version of the ABAP language. $M_{LLM}(\theta)$ is a large language model with parameters θ, which performs code conversion: $C_{new}=M_{LLM}(\theta, C_{old})$.

The quality of the generated code fragment C_{new} is evaluated using a vector Q with n metrics that reflect various aspects of quality and is defined as:

$$Q=[q_{syn}, q_{func}, q_{sem}]. \quad (1)$$

The problem is that evaluating the quality of program code using individual metrics is ineffective when migrating to new versions of computer systems.

Therefore, an integral quality function $F_{quality}$ is introduced, which maps the vector of metrics Q to a single scalar value representing the overall quality of the code. According to the ISO/IEC 25040 [7] software quality assessment methodology and the theory of composite indicators [8], such a function should aggregate a set of quality metrics into a single numerical value. For the specific task of ABAP code migration, this function can be formally defined as an integral metric IAQS for a comprehensive assessment of the quality of the ABAP code generated during the migration process. In this case, its range of values is defined as a normalised interval:

$$F_{quality}: Q \rightarrow [0,1]. \quad (2)$$

DOI 10.15588/1607-3274-2026-1-8

It is necessary to define and justify the function $F_{quality}$ (IAQS) and develop a methodology for adjusting the parameters θ of the model M_{LLM} in order to find the optimal set of parameters θ^* that maximises the expected value of the value $F_{quality}$ on the distribution of fragments of the legacy custom code:

$$\theta^* = \arg\max_{\theta} E\left[F_{quality}\left(Q\left(M_{LLM}\left(\theta, C_{old}\right)\right)\right)\right]. \quad (3)$$

Thus, this formulation defines:

– the object of optimisation are the parameters of the model θ^*;

– success criterion is a maximisation of the integral quality function $F_{quality}$;

– constraint – the quality function must be justified in accordance with international standards (ISO/IEC 25010) [6] and reflect the syntactic, functional and semantic aspects of the code.

2 REVIEW OF THE LITERATURE

The concept of software quality is fundamental to building reliable and effective computer information systems.

In [9], it is proven that in critical industries (which undoubtedly include enterprise-scale ERP systems), the concept of quality goes beyond purely technical characteristics and acquires economic significance.

Reliability and maintainability are considered not just as desirable properties, but as economic categories which directly affect the cost of ownership of the system and business risks [9].

The author emphasises the need to apply the philosophy of “Cleanroom Software Engineering”, which shifts the focus from correcting errors post factum to preventing them at the design and development stage [9].

This thesis is critically important for automated migration tasks, where the cost of an error replicated across thousands of lines of code can be catastrophic.

Standardised models have been developed to formalise these requirements, the most widely recognised of which is the ISO/IEC 25010 (SQuaRE) quality model [6].

It defines a hierarchical structure of eight product characteristics (e.g., Functional Suitability, Reliability, Compatibility, Maintainability, etc.) and their sub characteristics. This standard provides a theoretical basis for selecting and justifying software quality metrics.

Modern approaches to the quantitative measurement of software code quality indicators are analysed in detail in [10]. The effectiveness of using machine learning methods, in particular the Random Forest algorithm, for aggregating classical metrics (such as Holsted metrics, McCabe cyclomatic complexity) and code quality predicting has been proven. It has been shown that quality is a measurable entity that can be objectively assessed. The author also emphasises the importance of comprehensive analysis of code quality based on the selection and justification of a system of weight factors which reflect the relative importance of each metric [10]. However, the approaches proposed by the author allow for the effective identification of problematic code, but do not provide tools for its improvement.

Researches in the field of software engineering confirm that a combination of metrics covering different aspects of quality provides a more accurate and complete assessment of quality than any single metric. This approach allows for a more comprehensive representation of code quality, which is especially important for complex tasks such as automated migration.

While ISO/IEC 25010 defines the quality metrics that need to be measured [6], the theory of integral indicators explains how to combine individual metrics into a single indicator [7, 8].

An integral metric is a function that aggregates several individual indicators into a single numerical value, most often using a weighted arithmetic mean value of normalised components [7].

In recent years, LLMs [4] have demonstrated significant capabilities in solving a wide range of tasks related to software code. Models specifically trained on code (Code LLMs), such as the Qwen Coder series [11], Code Llama [12], and others, achieve state-of-the-art results on common coding benchmarks such as HumanEval and MBPP. These models, based on transformer architecture, are pre-trained on trillions of tokens of code and text from publicly available sources, giving them a deep fundamental understanding of programming syntax, semantics, and logic.

However, the effectiveness of universal LLMs is significantly reduced when working with proprietary languages such as ABAP due to a lack of representation in the training data [13].

To overcome these limitations, fine-tuning methods are used. Continuous pre-training (CPT) adapts the model to a new domain by continuing self-supervised learning on a large corpus of unlabelled data [14].

After adapting to the domain, the model must be trained to perform a specific task, for which parameter-efficient methods (PEFT) have been developed [15].

The most popular PEFT method is LoRA (Low-Rank Adaptation) [16], which freezes the main weights of the model and trains only small adapter matrices.

A further development is the QLoRA (Quantized LoRA) method [17], which applies 4-bit quantisation to frozen weights.

Alignment methods are used to improve the model output. The latest approach is Odds Ratio Preference Optimisation (ORPO) [18], which combines supervised fine-tuning (SFT) and preference-based alignment into a single step.

Therefore, there is a pressing need to create an integral quality assessment metric that combines diagnostic accuracy (based on metrics [10]) with compliance to a regulatory framework of standards, while being adapted to manage the quality improvement process when using LLM for migrating legacy ABAP custom code.

DOI 10.15588/1607-3274-2026-1-8

3 MATERIALS AND METHODS

Based on the research conducted and the principles of software quality assessment described in ISO/IEC 25040 [7], a domain-specific integral quality assessment metric IAQS was developed, which formally combines three key code quality metrics into a single indicator for the ABAP code migration task.

The components of IAQS are justified within the framework of ISO/IEC 25010 [6] and presented in Table 1.

Table 1 – Compliance of IAQS components with the ISO/IEC 25010 standard

IAQS component	ISO 25010 characteristic	Sub-characteristic	Justification
M_{syn}	Functional suitability, maintainability	Functional correctness, analysability	Syntactically incorrect code cannot be executed or analysed
M_{func}	Functional suitability	Functional correctness	A direct measure of whether the code performs the task at hand
M_{sem}	Maintainability	Modifiability, reusability	Code that is structurally close to the standard is easier to analyse and modify

Unlike general frameworks of quality assessment, IAQS is an integral metric specifically designed for comprehensive assessment of ABAP code quality, which simultaneously takes into account the syntactic, functional, and semantic aspects of the generated code.

Syntactic correctness (M_{syn}). The percentage of generated fragments that pass static analysis without errors. This metric is directly related to the characteristics of Maintainability (sub-characteristic Analysability) and Functional Suitability (sub-characteristic Functional Correctness), since syntactically incorrect code cannot be executed.

Functional correctness (M_{func}). The pass@1 metric is based on dynamic code execution and is defined as the percentage of cases where the first generated response is successfully compiled followed by successful passing of all functional tests [12] in the target environment. It corresponds to the Functional Suitability characteristic and its key sub-characteristic, Functional Correctness. This is the most direct measurement of whether the code performs the task at hand.

Semantic-structural similarity (M_{sem}). The CodeBLEU metric [19] is used – a composite indicator developed to evaluate both lexical and structural code similarity by taking into account syntactic structures and keywords of the programming language. This metric correlates with Maintainability, since code that is structurally and semantically close to a quality benchmark is easier to understand, analyse, and modify (sub-characteristics Modifiability and Reusability).

IAQS is defined as the weighted arithmetic mean value of its normalised components [7]:

$$IAQS = w_{syn} \cdot M_{syn} + w_{func} \cdot M_{func} + w_{sem} \cdot M_{sem}, \quad (4)$$

where $w_{syn}+w_{func}+w_{sem}=1$ and all weights $w_i \geq 0$.

For this study, an equal weighting approach ($w_{syn}=w_{func}=w_{sem}=1/3$) was adopted, which is a common and reasonable starting point in the absence of prior data on priorities [10]. Equal weighting reflects the assumption that all three aspects of quality are equally important for the overall evaluation of the generated code.

The proposed approach to LLM fine-tuning is a sequential process aiming at improving of the integral IAQS assessment.

Stage 1. Continuous pre-training (CPT). The model adapts to the syntax and semantics of ABAP by continuing self-supervised learning on a large corpus of unlabelled data [4]. This stage improves the model's internal representations, making it more "familiar" with the idioms of the target language.

Stage 2. Parameter-efficient fine-tuning (PEFT with QLoRA). The model learns a specific task of translating obsolete ABAP constructs into modern equivalents using supervised learning on pairs (instruction response) [15,16]. QLoRA allows the model to be trained efficiently by updating only a small portion of the parameters [17]. Instead of fine-tuning all model parameters, LoRA [16] freezes the initial weights and trains only small low-rank adapter matrices A and B. Weight updates are approximated by their product: $\Delta W=BA$. A direct pass through the modified layer is described by the formula:

$$h = W_0 x + BAx, \quad (5)$$

where $x \in R^k$ is the input vector, $W_0 \in R^{(d \times k)}$ is the output weight matrix, $B \in R^{(d \times r)}$, $A \in R^{(r \times k)}$, and rank $r \ll \min(d,k)$ is a hyperparameter.

Stage 3. Alignment based on preferences (Alignment with ORPO). The final stage uses the ORPO method to improve the quality of model output based on preference pairs [18]. Its key advantage is that it combines supervised fine-tuning and preference alignment into a single monolithic step, eliminating the need for a reference model π_{ref}, which simplifies the training process.

Direct optimisation of formula (3) via gradient descent is technically challenging because $F_{quality}$ includes non-differentiable components (syntactic correctness M_{syn}, functional correctness M_{func} (pass@1)). Therefore, we propose to implement an indirect optimisation strategy through three-stage fine-tuning, where IAQS is used as a control signal at the alignment stage. The metric ranks candidate code versions, forming preference data for the ORPO algorithm, which in its turn optimises the differentiable loss function L_{ORPO}. This creates an indirect link between $F_{quality}$ and the learning process.

The ORPO loss function consists of two components: a standard loss of negative log-likelihood (NLL loss) for the desired response y_w and a term based on the odds

DOI 10.15588/1607-3274-2026-1-8

ratio, which penalises the undesired response y_l and is defined as:

$$L_{ORPO} = E_{(x,y_w,y_l)\sim D}\left[-\log P_\theta(y_w \mid x) - \lambda \log \sigma\left(\log \frac{odds_\theta(y_w \mid x)}{odds_\theta(y_l \mid x)}\right)\right], \quad (6)$$

where $odds_\theta(y \mid x) = \frac{1 - P_\theta(y \mid x)}{P_\theta(y \mid x)}$.

The first term $-\log P_\theta(y_w \mid x)$ is responsible for training the correct response (as in SFT), and the second is responsible for ensuring that the model "distinguishes" a good response from a bad one.

The final stage of alignment with ORPO is key to the proposed approach. Unlike standard approaches, where preference data is collected manually, IAQS is used instead as a preference control signal. This creates a powerful, self-consistent cycle, as the proposed integral metric directly controls the training of the model.

For the input code C_{old} several k=5 candidate migration options are generated using the model after the PEFT stage with sampling temperature variation.

For each candidate C_{newi} its score IAQS$_i$ is calculated according to formula (4).

Pairs of preferences (C_{newi},C_{newj}) are formed such that $IAQS_i > IAQS_j$. In this case, C_{newi} becomes the "selected" response y_w, and C_{newj} – becomes the "rejected" response y_l.

This data set is then used to fine-tune the model using the ORPO loss function [18], which now implicitly aligns (or corrects) the model policy by minimising L_{ORPO} to achieve a high IAQS value.

4 EXPERIMENTS

A specialised code generation model, Qwen 2.5 Coder 14B [11], was selected as the base LLM. This model was chosen due to its advanced results on coding benchmarks and its open licence, which is consistent with the results of [5].

The experiments were conducted on an NVIDIA A100 GPU (40 GB) using PyTorch 2.1, Hugging Face Transformers 4.36, and TRL (Transformer Reinforcement Learning) 0.7 libraries to implement ORPO.

Three datasets were prepared for fine-tuning.

The corpus for CPT is an unstructured corpus of modern ABAP code with a volume of 500 million tokens, collected from publicly available GitHub repositories (150+ repositories), SAP Press documentation with code examples, and anonymised fragments from open SAP Community archives.

Parallel corpus for PEFT – a set of 50,000 "instruction-response" pairs in the format "legacy code – modern equivalent". The data was prepared by experts who included code for migration to SAP S/4HANA and supplemented with synthetically generated examples using migration rules from official SAP documentation.

A set of advantage data for Alignment – 8,000 triplets (prompt-chosen-rejected), where "chosen" is the option with high IAQS, and "rejected" is the option with lower IAQS. The data was generated automatically using the above-described procedure for forming preference pairs.

For independent evaluation of the models, a test sample was prepared containing 1,000 pairs (obsolete code – reference_modern_code) that were not used during any stage of training. The test examples cover typical migration tasks: improving the efficiency of SELECT queries, replacing outdated functional modules with modern classes, refactoring of outdated constructs.

The IAQS component calculation methodology involves the following steps:

– syntax checking using ABAP Parser (a component of ABAP Development Tools). The M_{syn} metric is calculated as the percentage of fragments without syntax errors;

– validation of functional equivalence by dynamically executing code in the SAP S/4HANA environment. Sets of unit tests were prepared for each pair (old code – new code). Thanks to access to the SAP S/4HANA system, the verification was performed not on mock-ups, but on a real ABAP processor. The metric (pass@1) is calculated as the percentage of generated fragments which are compiled successfully and successfully passed all functional tests on the first attempt [12];

– calculation *of* M_{sem} (CodeBLEU) relative to the reference code using the codebleu library [19, 20]. Given the specifics of the proprietary ABAP language and the lack of publicly available stable grammars (Tree-sitter grammars) for building AST in Python, the calculation was performed in an adapted mode. N-gram match and weighted N-gram match components (with customised weights for ABAP keywords) were used, which made it possible to evaluate the lexical and structural correspondence of the code without the need to export syntax trees from SAP systems.

Each model was evaluated on a full test sample with a fixed seed (random seed=42) to ensure reproducibility of results. The hyperparameters for fine-tuning are presented in Table 2.

Table 2 – Hyperparameters for fine-tuning

Parameter	CPT stage	PEFT stage (QLoRA)	Alignment stage (ORPO)
Learning Rate	1e–5	2e–4	5e–6
Batch Size	8	16	8
Gradient Accumulation Steps	4	2	4
Epochs	1	3	2
LoRA Rank (r)	–	16	16
LoRA Alpha	–	32	32
LoRA Dropout	–	0.05	0.05
Quantization Type	–	4-bit NF4	4-bit NF4
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.01	0.01	0.01
Max Sequence Length	2048	2048	2048
Warmup Steps	500	100	50
ORPO λ	–	–	0.1

DOI 10.15588/1607-3274-2026-1-8

For QLoRA, 4-bit NF4 (Normal Float 4) quantisation with double quantisation was used, which reduced the model's memory footprint.

5 RESULTS

To evaluate the effectiveness of the proposed approach, not only individual metrics were calculated, but also an integral IAQS assessment at each stage of fine-tuning. The results are presented in Table 3.

Table 3 – Results of model evaluation on the test sample

Model	Syntactic correctness (M_{syn}), %	Semantic similarity (M_{sem})	Functional correctness (M_{func}), %	IAQS (w=1/3)
Base Model (zero-shot)	12.4	0.215	3.1	0.117
After CPT	28.7	0.365	11.2	0.255
After PEFT (QLoRA)	78.3	0.724	64.8	0.718
After Alignment (ORPO)	82.1	0.761	68.5	0.756

Figure 1 shows the visualisation of the dynamics of improvement in the IAQS integral assessment and its components at each stage of fine-tuning.

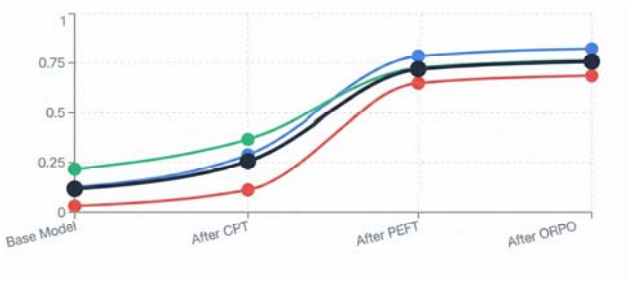


Figure 1 – Dynamics of IAQS growth and its components at the stages of fine-tuning

The graph shows the non-linear nature of quality improvement.

The base LLM shows very low results across all metrics, which is explained by the practical absence of ABAP code in its baseline training data.

The CPT stage provides basic adaptation to the domain, increasing semantic similarity (M_{sem}) from 0.215 to 0.365 (+69.8%).

The largest increase is observed at the PEFT stage, from 0.255 to 0.718 (+181.6%), confirming the critical importance of supervised fine-tuning for achieving functional correctness. The ORPO stage improves (+5.3%) all components simultaneously.

Analysis of the results shows that each stage contributes to a significant and consistent increase in the overall integral quality assessment of ABAP code.

The baseline model demonstrates very low code quality (IAQS = 0.117), which is explained by the lack of ABAP representation in its baseline training data.

The CPT stage increases it more than twice (IAQS = 0.255, an increase of +117.9%).

The PEFT stage provides the largest increase, after which the integral quality assessment metric rises to 0.718 (an increase of +181.6%).

The final alignment stage, controlled by IAQS, brings the integral assessment to 0.756, which is a 546% improvement over the baseline model.

6 DISCUSSION

The results confirm the central thesis of the work: formalising quality assessment in the form of an integral metric and using it as a target function allows for systematic improvement of the results of the ABAP code automated migration process.

A detailed analysis shows that different stages of fine-tuning have different effects on IAQS components.

The continuous pre-training (CPT) stage mainly improves the semantic component M_{sem} (from 0.215 to 0.365, an increase of +69.8%), familiarising the model with idioms and constructs of the ABAP language. This is because CPT adapts the tokenizer and internal representations of the model to the specific syntax of ABAP, making the generated code more "natural" in terms of structure. However, functional correctness increases insignificantly (from 3.1% to 11.2%) because the model has not yet learned the specific task of migration.

The parameter-efficient fine-tuning (PEFT) stage provides the biggest leap in functional correctness M_{func} (from 11.2% to 64.8%, an increase of +478.6%), as at this stage the model learns the specific task of converting obsolete constructs into modern equivalents based on 50,000 examples of pairs (legacy code, modern code). This stage also significantly improves syntactic correctness M_{syn} (from 28.7% to 78.3%, an increase of +172.8%), indicating successful assimilation of the rules for generating valid code. Semantic similarity also increases (from 0.365 to 0.724, +98.4%) as the model learns to generate code that is structurally similar to the reference samples.

The final stage of preference-based alignment (ORPO) provides "fine-tuning" of all components simultaneously. Although the absolute increase is smaller (+5.3% for IAQS, +4.9% for M_{syn}, +5.1% for M_{sem}, +5.7% for M_{func}, this stage is critical for achieving high quality, as it trains the model to distinguish between "good" and "acceptable" code versions. The improvement in semantic-structural similarity (CodeBLEU from 0.724 to 0.761) is particularly noticeable, confirming that the model has learned to generate code that is not only functionally correct but also structurally close to idiomatic ABAP.

It is important to note the synergistic effect between the stages: each subsequent stage builds on the results of the previous one, creating a cumulative improvement in

DOI 10.15588/1607-3274-2026-1-8

quality. CPT lays the foundation for domain knowledge, PEFT trains for a specific task, and ORPO performs targeted model alignment to achieve the maximum value of the IAQS comprehensive quality indicator.

It is necessary to note the limitations of the proposed approach and ways to overcome them.

Firstly, the choice of weights for IAQS is subjective [10]. Although equal weighting (w=1/3) is a reasonable starting point [10], in industrial projects these weights can be calibrated to reflect specific business priorities. For example, in critical financial systems, functional correctness may have a higher weight (w_{func}=0.5, w_{syn}=0.3, w_{sem}=0.2), while in projects with an emphasis on maintainability, semantic similarity (w_{sem}=0.5, w_{func}= 0.3, w_{syn}=0.2). Future research could use data-driven methods (e.g., regression analysis or analytical hierarchy process) to determine optimal weights based on historical migration project data.

Secondly, benchmark-based metrics such as CodeBLEU may be prone to “surface bias”, favouring textual similarity over true functional equivalence [20]. Research [21] has shown that CodeBLEU can give high scores to code that looks similar to the benchmark but has subtle semantic differences. This risk is deliberately mitigated in IAQS by including the pass@1 functional correctness metric, which evaluates the actual behaviour of the code through the execution of unit tests [12]. This combination makes IAQS more robust to evaluation errors than any single metric, as the code must simultaneously pass functional tests (pass@1) and be structurally similar to a high-quality benchmark (CodeBLEU).

Thirdly, the scalability of the approach to very large code bases (>100,000 lines) still needs empirical confirmation. The current study focused on the migration of individual functions and methods (average length 50 lines), whereas industrial projects often require the migration of inter-module dependencies, global variables, and complex integration scenarios. Future work should investigate how IAQS scales up to the level of entire applications and whether additional metric components are needed (e.g., inter-module compatibility assessment).

Fourthly, the current implementation of IAQS does not take into account such important aspects of code quality as performance, security, and energy efficiency. Although the ISO/IEC 25010 [6] standard includes the characteristic “Performance Efficiency”, its integration into IAQS requires the development of automated methods for measuring the performance of generated ABAP code, which is a non-trivial task due to the need to execute the code in a real SAP environment.

The proposed approach paves the way for the development of more reliable and predictable automated code migration systems. The IAQS integral metric can be integrated into SAP projects as an objective indicator of code quality during migration, allowing teams to track progress and automatically identify problematic code fragments. The proposed fine-tuning methodology can be used as a template for adapting LLM to other domain-specific languages and code migration tasks.

The improvement in economic performance is expected to be significant. If automated migration with IAQS 0.756 can replace even 50% of manual work, this potentially reduces the total cost of an SAP S/4HANA migration project by 20–30%, considering that code migration accounts for up to 40% of the total workload [3].

CONCLUSIONS

This paper presented an approach to the automated migration of legacy ABAP custom code based on the principles of intelligent reengineering and objective code quality assessment. An integral quality assessment metric, IAQS, was proposed, which provides a comprehensive view of the quality of the generated code by combining syntactic, functional, and semantic characteristics based on the ISO/IEC 25010 and ISO/IEC 25040 standards.

A three-stage approach to LLM training (CPT → PEFT → ORPO) was also developed and tested, which purposefully aligns the model to achieve the maximum value of the integral quality assessment metric. A key feature of the methodology is the use of the IAQS metric itself to automatically generate preference data at the alignment stage, creating a self-consistent tuning cycle. The research results confirm that this approach allows for significant and controlled improvement in the quality of migrated code. The final model achieved a high IAQS score (0.756), which is a 546% improvement over the baseline LLM.

The scientific novelty lies in the development of the method for ensuring the quality of legacy ABAP custom code, based on the IAQS code quality assessment metric and a three-stage approach to LLM fine-tuning. Unlike existing approaches that use disjointed code quality assessment metrics, IAQS is an integral metric specifically designed for comprehensive assessment of ABAP code migration quality based on ISO/IEC 25040 principles and composite indicator theory.

A distinctive feature of the proposed three-stage approach to LLM alignment is the use of the IAQS integral metric in the automatic generation of preference data for the ORPO algorithm. This creates a self-consistent cycle of target alignment “metric-model”, where the IAQS integral metric directly controls the learning process, ensuring target alignment with respect to the integral quality indicator.

The **practical significance** is that the proposed approach paves the way for the creation of more reliable and predictable automated code migration systems, which will reduce costs in computer system intelligent reengineering projects. The IAQS integral metric can be implemented in SAP projects as an objective indicator of the effectiveness of the migration process of legacy custom code.

Prospects for further research include studying various weighting schemes for IAQS based on expert assessments and real project data, expanding metrics to

DOI 10.15588/1607-3274-2026-1-8

account for performance, security, and energy efficiency aspects, and comparative testing of the proposed approach with existing methods in real SAP S/4HANA migration projects. A separate promising area of research is the deepening of semantic similarity metrics. We plan to develop a mechanism for integration with open static analysis tools such as abaplint or internal SAP parsers. This will allow the generation and export of complete abstract syntax trees (AST) in JSON format directly to the evaluation pipeline, ensuring that deep structural relationships and data flow are taken into account in the CodeBLEU metric.

ACKNOWLEDGEMENTS

This work was performed as part of the research project "Information technologies of intelligent computing" (state registration number 0124U004188) of the Software Tools Department of Zaporizhzhia Polytechnic National University.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors contributions: Oleg Pozdnyakov: data curation, formal analysis, investigation, methodology; validation, visualization, writing – original draft. Anzhelika Parkhomenko: conceptualization, supervision, writing – review & editing.

Data availability: The data will be provided upon reasonable request to the authors by email to oleg.pozdnyakov@zp.edu.ua.

Software availability: The manuscript has no associated software.

Use of artificial intelligence tools: The authors used the Qwen 2.5 Coder 14B model in the "Experiments" section to generate ABAP code. The model was sequentially fine-tuned in three stages, and the results of each stage were controlled based on the calculation of component metrics and the integral IAQS metric. Comparative analysis confirmed the improvement of the integral quality indicator of the generated ABAP code. Additionally, during the preparation of this paper, the authors used DeepL in order to: grammar and spelling check, paraphrase and reword. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

REFERENCES

1. Hardy P. Migrating custom code to SAP S/4HANA. Boston, Rheinwerk Publishing Inc., 2020, 333 p.
2. Ktern AI. SAP S/4HANA 2022: The ultimate custom code migration guide [Electronic resource]. 2025. Access mode: https://ktern.com/article/sap-custom-code-migration-guide-2024/.
3. SAPinsider. Technical guide: using ABAP Test Cockpit for SAP S/4HANA Transition [Electronic resource], 2017. Access mode: https://sapinsider.org/articles/technical-guide-using-abap-test-cockpit-for-sap-s-4hana-transition/.
4. Pozdnyakov O. A., Parkhomenko A. V. Migration of custom code to new versions of complex computer systems using methods and models of intelligent reengineering, *Scientific Works of DonNTU. Series "Informatics, Cybernetics and Computer Engineering"*, 2025, Vol. 2 (41), pp. 86–98 (in Ukranian).
5. Pozdnyakov O. A., Parkhomenko A. V. Research and selection of large learning modes for automation of ABAP-code migration, *Management of the development of complex systems*, 2025, Vol. 63, pp. 191–200 (in Ukrainian).
6. ISO/IEC 25010:2023. Systems and software engineering – systems and software quality requirements and evaluation (SQuaRE). Product quality model. [Electronic resource]. Access mode: https://www.iso.org/ru/standard/78176.html.
7. ISO/IEC 25040:2024. Systems and software engineering – systems and software quality requirements and evaluation (SQuaRE). Evaluation process. [Electronic resource]. Access mode: https://www.iso.org/standard/83467.html.
8. Nardo M., Saisana M., Saltelli A., Tarantola S., Hoffman A., Giovannini E. Handbook on constructing composite indicators: Methodology and user guide. Paris, OECD Publishing, 2008, 162 p.
9. Yurchyshyn V. M. Methodological Approaches to Assessing Software Quality for Oil and Gas Industry Facilities, *Methods and Instruments of Quality Control*, 2020, Vol. 2 (45), pp. 40–57. DOI:10.31471/1993-9981-2020-2(45)-40-57 (in Ukranian).
10. Prokofiev I. Method of Static Analysis of Code Quality Using Machine Learning, *Measuring and Computing Technology in Technological Processes*, 2025, Vol. 3, pp. 126–133. DOI:10.31891/2219-9365-2025-83-17 (in Ukrainian).
11. Qwen Team. Qwen2.5-Coder series: Powerful, Diverse, Practical [Electronic resource]. 2024. Access mode: https://qwenlm.github.io/blog/qwen2.5-coder-family/.
12. Rozière B. , Gehring J., Gloeckle F. et al. Code Llama: Open foundation models for code, *Meta AI Technical Report*, 2023, 38 p. DOI:10.48550/arXiv.2308.12950.
13. Weyssow M., Zhou X., Kim K. et al. Exploring parameter-efficient fine-tuning techniques for code generation with large language models, *ACM Transactions on Software Engineering and Methodology*, 2024, Vol. 33(6), pp. 1–25. DOI:10.1145/3714461.
14. Ray J. Teach an old dog new tricks: LLM continual pre-training (CPT) [Electronic resource], 2025. Access mode: https://medium.com/better-ml/teach-an-old-dog-new-tricks-llm-continual-pre-training-cpt-684cfb931247.
15. Han Z., Gao C., Liu J. et al. Parameter-efficient fine-tuning for large models: a comprehensive survey, *Transactions on Machine Learning Research*, 2024, pp. 1–44. DOI:10.48550/arXiv.2403.14608.
16. Hu E. J., Shen Y., Wallis P. et al. LoRA: Low-rank adaptation of large language models, *International Conference on Learning Representations (ICLR 2022)*, pp. 1–13. DOI:10.48550/arXiv.2106.09685.
17. Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L. QLoRA: Efficient finetuning of quantized LLMs, *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023, pp. 1–28. DOI:10.48550/arXiv.2305.14314.
18. Hong J., Lee N., Thorne J. ORPO: Monolithic preference optimization without reference model, *2024 Conference on*

DOI 10.15588/1607-3274-2026-1-8

Empirical Methods in Natural Language Processing, 2024, pp. 11170–11189. DOI:10.18653/v1/2024.emnlp-main.626.
19. Ren S., Guo D., Lu S. et al. CodeBLEU: a method for automatic evaluation of code synthesis, *Computer Science. Software Engineering*, 2020, pp. 1–8. DOI:10.48550/arXiv.2009.10297.
20. Microsoft. CodeXGLUE: CodeBLEU evaluation metric [Electronic resource]. 2025. Access mode: https://github.com/microsoft/CodeXGLUE/tree/main/Code-Code/code-to-code-trans/evaluator/CodeBLEU.
21. Bhattacharjee A., Dwyer C. Analyzing and mitigating surface bias in code evaluation metrics, *Computer Science. Software Engineering*, 2025, pp. 1–22. DOI:10.48550/arXiv.2509.15397.

Received 09.12.2025.
Accepted 09.01.2026.
Published 27.03.2026.

УДК 004.89

ОЦІНКА ТА ЗАБЕЗПЕЧЕННЯ ЯКОСТІ МІГРОВАНОГО ABAP-КОДУ ЗА ДОПОМОГОЮ ІНТЕГРАЛЬНОЇ МЕТРИКИ ТА МОДЕЛЕЙ ГЕНЕРАТИВНОГО ШТУЧНОГО ІНТЕЛЕКТУ

Поздняков О. А. – аспірант кафедри програмних засобів, Національний університет «Запорізька політехніка», Запоріжжя, Україна. ROR: https://ror.org/03aph1990. ORCID: https://orcid.org/0009-0006-3955-802X.
Пархоменко А. В. – канд. техн. наук, доцент, доцент кафедри програмних засобів, Національний університет «Запорізька політехніка», Запоріжжя, Україна. ROR: https://ror.org/03aph1990. ORCID: https://orcid.org/0000-0002-6008-1610.

АНОТАЦІЯ

Актуальність. Автоматизація процесу міграції успадкованого користувацького коду при переході на нову версію системи S/4HANA за допомогою великих мовних моделей (LLM) є перспективним напрямом. Проте оцінка якості згенерованого коду залишається невирішеною проблемою, оскільки існуючі підходи використовують розрізнені метрики, що не дозволяють комплексно оцінити та забезпечити якість програмного коду для подальшого використання без додаткового доопрацювання.

Мета роботи – підвищення ефективності процесу інтелектуального реінжинірингу комп'ютерної системи на основі методу комплексного оцінювання та забезпечення якості мігрованого користувацького ABAP-коду.

Метод. Розроблений метод базується на двох ключових компонентах. Інтегральна метрика оцінки якості IAQS (Integral ABAP Quality Score) комплексно враховує синтаксичні, функціональні та семантичні характеристики коду та ґрунтується на положеннях міжнародних стандартів якості програмного забезпечення ISO/IEC 25010, ISO/IEC 25040, а також теорії композитних індикаторів. Триетапний підхід до донавчання LLM (Qwen 2.5 Coder 14B) включає безперервне попереднє навчання (CPT), параметро-ефективне донавчання (PEFT) та вирівнювання на основі переваг (Alignment) на основі алгоритму ORPO. При цьому використання розробленої метрики IAQS для формування набору даних переваг на етапі вирівнювання створює механізм керованого вдосконалення, а саме визначає напрямок адаптації LLM.

Результати. Результати експериментальних досліджень демонструють, що реалізація розробленого методу дозволяє покращити як окремі показники якості програмного коду, так і інтегральну метрику оцінку якості IAQS в цілому. Фінальна модель, донавчена на основі запропонованого триетапного підходу, дозволила досягти високого значення IAQS (0.756), що демонструє суттєве підвищення у порівнянні з базовою моделлю (0.117).

Висновки. Дослідження представляє новий проблемно-орієнтований підхід до автоматизованої міграції ABAP-коду при інтелектуальному реінжинірингу комп'ютерних систем. Запропонована інтегральна метрика IAQS є основою для створення формалізованої та об'єктивної системи оцінки якості програмного забезпечення, згенерованого LLM у контексті міграції успадкованого користувацького коду. Продемонстровано, що послідовне донавчання LLM на основі триетапного підходу з використанням IAQS забезпечує суттєве підвищення інтегрованого показника якості згенерованого програмного коду.

КЛЮЧОВІ СЛОВА: якість програмного забезпечення, інтегральна метрика, великі мовні моделі, міграція успадкованого користувацького коду, донавчання LLM.

ЛІТЕРАТУРА

1. Hardy P. Migrating custom code to SAP S/4HANA / P. Hardy. – Boston : Rheinwerk Publishing Inc., 2020. – 333 p.
2. Ktern AI. SAP S/4HANA 2022: The ultimate custom code migration guide [Electronic resource] / AI. Ktern. – 2025. – Access mode: https://ktern.com/article/sap-custom-code-migration-guide-2024/.
3. SAPinsider. Technical guide: using ABAP Test Cockpit for SAP S/4HANA Transition [Electronic resource] / SAPinsider. – 2017. – Access mode: https://sapinsider.org/articles/technical-guide-using-abap-test-cockpit-for-sap-s-4hana-transition/.
4. Поздняков О. А. Міграція користувацького коду на нові версії складних комп'ютерних систем з використанням методів і моделей інтелектуального реінжинірингу / О. А. Поздняков, А. В. Пархоменко // Наукові праці ДонНТУ. Серія «Інформатика, кібернетика та обчислювальна техніка». – 2025. – № 2 (41). – С. 86–98.
5. Поздняков О. А. Дослідження та вибір великих мовних моделей для автоматизації міграції ABAP-коду/ О. А. Поздняков, А. В. Пархоменко // Збірник наукових праць «Управління розвитком складних систем». – 2025. – № 63. – С. 191–200.
6. ISO/IEC 25010:2023. Systems and software engineering – systems and software quality requirements and evaluation (SQuaRE) – Product quality model. [Electronic resource]. – Access mode: https://www.iso.org/ru/standard/78176.html.
7. ISO/IEC 25040:2024. Systems and software engineering – systems and software quality requirements and evaluation (SQuaRE) – Evaluation process. [Electronic resource]. – Access mode: https://www.iso.org/standard/83467.html.
8. Handbook on constructing composite indicators: Methodology and user guide / [M. Nardo, M. Saisana, A. Saltelli et al.]. – Paris : OECD Publishing, 2008. – 162 p.

DOI 10.15588/1607-3274-2026-1-8

9. Юрчишин В. М. Методологічні підходи щодо оцінки якості програмного забезпечення для об'єктів нафтогазового комплексу / В. М. Юрчишин // Методи та прилади контролю якості. – 2020. – № 2 (45). – С. 40–57. DOI:10.31471/1993-9981-2020-2(45)-40-57.
10. Прокоф'єв І. Метод статичного аналізу якості коду з допомогою машинного навчання / І. Прокоф'єв // Вимірювальна та обчислювальна техніка в технологічних процесах. – 2025. – № 3. – С. 126–133. DOI:10.31891/2219-9365-2025-83-17.
11. Qwen Team. Qwen2.5-Coder series: Powerful, Diverse, Practical [Electronic resource] / Qwen Team // QwenLM Blog. – 2024. – Access mode: https://qwenlm.github.io/blog/qwen2.5-coder-family/.
12. Code Llama: Open foundation models for code / [B. Rozière, J. Gehring, F. Gloeckle et al.] // Meta AI Technical Report. – 2023. – 38 p. DOI:10.48550/arXiv.2308.12950.
13. Exploring parameter-efficient fine-tuning techniques for code generation with large language models / [M. Weyssow, X. Zhou, K. Kim et al.] // ACM Transactions on Software Engineering and Methodology. – 2024. – Vol. 33, № 6. – P. 1–25. DOI: 10.1145/3714461.
14. Ray J. Teach an old dog new tricks: LLM continual pre-training (CPT) [Electronic resource] / J. Ray // Medium. – 2025. – Access mode: https://medium.com/better-ml/teach-an-old-dog-new-tricks-llm-continual-pre-training-cpt-684cfb931247.
15. Parameter-efficient fine-tuning for large models: A comprehensive survey / [Z. Han, C. Gao, J. Liu et al.] // Transactions on Machine Learning Research. – 2024. – P. 1–44. DOI:10.48550/arXiv.2403.14608.
16. LoRA: Low-rank adaptation of large language models / [E. J. Hu, Y. Shen, P. Wallis et al.] // International Conference on Learning Representations (ICLR 2022). – P. 1–13. DOI:10.48550/arXiv.2106.09685.
17. QLoRA: Efficient finetuning of quantized LLMs / [T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer] // 37th Conference on Neural Information Processing Systems (NeurIPS 2023). – 2023. – P. 1–28. DOI:10.48550/arXiv.2305.14314.
18. Hong J. ORPO: Monolithic preference optimization without reference model / J. Hong, N. Lee, J. Thorne // 2024 Conference on Empirical Methods in Natural Language Processing. – 2024. – P. 11170–11189. DOI:10.18653/v1/2024.emnlp-main.626.
19. CodeBLEU: A method for automatic evaluation of code synthesis / [S. Ren, D. Guo, S. Lu et al.] // Computer Science. Software Engineering. – 2020. – P. 1–8. DOI:10.48550/arXiv.2009.10297.
20. Microsoft. CodeXGLUE: CodeBLEU evaluation metric [Electronic resource] / Microsoft. – 2025. – Access mode: https://github.com/microsoft/CodeXGLUE/tree/main/Code-Code/code-to-code-trans/evaluator/CodeBLEU.
21. Bhattacharjee A. Analyzing and mitigating surface bias in code evaluation metrics / A. Bhattacharjee, C. Dwyer // Computer Science. Software Engineering. – 2025. – P. 1–22. DOI:10.48550/arXiv.2509.15397.

DOI 10.15588/1607-3274-2026-1-8

ПРОГРЕСИВНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

PROGRESSIVE INFORMATION TECHNOLOGIES

UDC 004.41

A DESIGN PATTERN FOR ENABLING FUNCTIONAL STABILITY IN SOFTWARE SYSTEMS

Bychkov O. S. – Dr. Sc. (Engin.), Professor, Head of the Department of Software Systems and Technologies, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. ROR: https://ror.org/02aaqv166. ORCID: https://orcid.org/0000-0002-9378-9535.

Moroz M. V. – Post-graduate student of the Department of Software Systems and Technologies, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. ROR: https://ror.org/02aaqv166. ORCID: https://orcid.org/0000-0001-6953-683X.

ABSTRACT

Context. Modern software systems operate in dynamic and harsh environments where internal and external failures, unexpected disturbances, direct attacks, and resource constraints challenge the consistent provision of core functionalities. In these contexts, ensuring functional stability – where the quality of each system function remains within a predetermined stable range despite failures or environmental anomalies – is critical, especially for safety-critical and high-availability applications.

Objective. The primary objective of this work is to develop and justify an enabling design pattern that provides the architectural backbone for achieving functional stability in software systems. The main focus is to provide a flexible solution that facilitates dynamic adaptation while maintaining robust system behavior.

Method. We propose a novel pattern that combines the dynamic strategy selection capabilities with the loose coupling between components afforded by an event-driven approach. This enabling pattern decouples system components by enforcing communication solely through standardized event types and allows each module to select an appropriate adaptation strategy based on its current context. The described pattern was used to build a design of a real-life example that aims to implement stable object tracking functionality for autonomous quad-platforms. The proposed design was evaluated using design-level metrics alongside qualitative comparisons with existing adaptive approaches.

Results. Our analysis shows that the enabling pattern achieves significant modularity and adaptability. Key object-oriented metrics indicate minimal interdependencies among modules and a clear separation of concerns. The design proposal demonstrates that the pattern supports dynamic behavior adjustment through flexible strategy selection and serves as an enabler for functional stability by providing a robust architectural backbone for software systems.

Conclusions. The scientific novelty of this work is twofold: firstly, the novel pattern is obtained in our study, providing dynamic adaptation through context-aware strategy selection; secondly, functional stability received further development in the area of software architecture. The proposed pattern offers a robust, scalable, and maintainable architectural solution, with significant practical implications for the design of adaptive, resilient software systems.

KEYWORDS: software design patterns, functional stability, event processing, adaptive behavior, autonomous systems.

NOMENCLATURE

C is a set of software components (modules) that form the system SW;

C_i is a specific component of the system SW, responsible for executing one or more functions;

c_i is an internal state or context of component C_i, which affects how the strategy selection function σ behaves;

E is a situation space representing the full range of environmental and operational conditions in which the system may operate;

$E_{expected}$ is a subset of situation space E that contains situations anticipated by the system designer;

$E \backslash E_{expected}$ is a set of unexpected or unhandled situations that are not explicitly anticipated during system design;

F is a set of system functions provided by the system SW;

f_i is a particular function from the set F, implemented by a specific component C_i;

M is a mapping function that defines the system's behavior;

$O(f)$ is a set of all possible output states of the component that provides function f;

$o(f)$ is an output state of the component that provides function f at time t;

P is an enabling architectural software design pattern;

s is a specific situation from space E that may affect the system's functioning at runtime;

$S_{stable}^{(f)}$ is a set of output states considered stable for function f;

ST is a set of strategies available for adaptation within the system;

st is a specific adaptation strategy from the set ST that a component may use to react to a detected event;

DOI 10.15588/1607-3274-2026-1-9

SW is a target software system under analysis, composed of interacting components that together implement the system's core functionality;

T is a transient interval after which the output state of function *f* is expected to reside within the stable set $S_{stable}^{(f)}$;

σ is a strategy selection function that determines the most suitable strategy for a component based on the current situation and internal context.

INTRODUCTION

Modern software systems are increasingly deployed in environments characterized by rapid changes, complex dependencies, unexpected disturbances, and resource limitations. These conditions – ranging from internal failures and external attacks to unanticipated operational anomalies – pose significant challenges to maintaining consistent, reliable functionality. The concept of functional stability [1], defined as the ability of a system to preserve its core functions within predetermined stable boundaries despite adverse conditions, has traditionally been applied in mechanical, physical, and decentralized systems. However, its application to software systems and software architectures is less explored, creating a gap in both scientific research and practical implementations.

The current state of research in adaptive and resilient software architecture emphasizes the use of classical design patterns alongside approaches for dynamic reconfiguration and self-adaptation. These methods provide valuable mechanisms for enabling systems to adjust their behavior in response to changing operational conditions. However, while they offer important insights into dynamic adaptation and fault tolerance, many of these approaches address only isolated aspects of system resilience. Consequently, a comprehensive architectural solution that unifies dynamic adaptation with long-term functional stability remains lacking. This gap motivates our investigation into an enabling design pattern that not only supports flexible strategy selection and decoupled communication between components but also provides a robust framework for maintaining stable system behavior in complex, dynamic environments.

The object of the study is the adaptive process within software systems that encounter a diverse range of operational situations – including both expected and unforeseen disturbances.

The subject of the study is the enabling design pattern, an architectural backbone that integrates dynamic strategy selection with decoupled, event-driven communication among components. This pattern is intended to provide the structural support necessary for systems to maintain stable functionality over time, even under conditions of stress or failure.

The purpose of the work is to develop and justify a pattern-based solution that provides the necessary structural support for achieving functional stability in complex software systems. To achieve this aim, we have set the following tasks: to analyze existing approaches and identify their limitations with respect to functional stability; to design a pattern that is able to fulfill the mentioned requirements; to evaluate the proposed pattern.

1 PROBLEM STATEMENT

Let *SW* be a software system composed of a set of components $C=\{C_1,C_2,\ldots,C_n\}$ that collectively implement a set of functions $F=\{f_1,f_2,\ldots,f_r\}$. The system *SW* operates in an environment characterized by a situation space *E*, where each situation $s \in E$ is an element of ^{k}. These situations represent various conditions – including internal and external failures, disturbances, and other operational anomalies – that the system may encounter.

The system is modeled by a mapping function M: $E \rightarrow O(f)$, where $O(f)=\{ o_1^{(f)}, o_2^{(f)}, \ldots, o_m^{(f)} \}$ represents the set of output states or responses of a component that provides function *f* after encountering a situation *s*. In response to a situation, a component may either change its state or maintain its current state.

The overall goal is to achieve functional stability for each function $f \in F$ under situation space *E*. Formally, for every function *f*, there exists a subset of stable states $S_{stable}^{(f)}$ such that after a transient interval *T*, the state $o_t^{(f)}$ corresponding to function *f* satisfies formula (1):

$$o_t^{(f)} \in S_{stable}^{(f)}, \forall t > T. \quad (1)$$

To address these challenges, the problem targeted by this article is to develop an enabling software design pattern *P* that provides the architectural backbone for achieving functional stability in dynamic environments. This pattern by its design considerations should support the system *SW* with the ability to satisfy formula (1) despite failures, disturbances, or unanticipated operational conditions. In details, the target pattern *P* should be responsible for next aspects.

– Enabling adaptability. Allowing each component C_i to select a strategy from a set *ST* based on the current situation $s \in E$ and its internal state (context) c_i. This selection is formalized by a function $\sigma:(s,c_i) \rightarrow st \in ST$. By adjusting a strategy component could transition into a stable state, satisfying the formula (1).

– Handling expected situations. It is expected that a system designer defines a subset $E_{expected} \subset E$ that covers the range of anticipated conditions. Recognizing that $E \backslash E_{expected}$ may include unexpected situations, *P* should support default or fallback strategies to mitigate these unhandled cases.

– Supporting extensibility and scalability. Ensuring that the pattern *P* is designed in a modular way to allow easy maintenance, extension, and scaling of both the strategies *ST* and the overall system functions *F*.

2 REVIEW OF THE LITERATURE

Functional stability [1] is defined as the property of an object to preserve the execution of its primary functions over a specified time, within limits set by normative

DOI 10.15588/1607-3274-2026-1-9

requirements, even when exposed to counteractive influences and streams of failures, malfunctions, or errors. This concept is particularly valued in contexts where uninterrupted performance is critical, such as in safety-critical systems or environments prone to frequent disturbances. The methodological approach to achieving functional stability can be divided into four compact stages [2]:

– detection of an abnormal situation associated with degradation in the quality of functioning due to the influence of destabilizing factors;

– identification of an abnormal situation;

– making a decision on restoring the functioning process;

– restoring functioning by redistributing functions and tasks between undamaged elements.

Historically, the concept of functional stability has been applied extensively to mechanical and physical systems – such as for example onboard aircraft systems [3, 4], control and navigation systems [5, 6], and distributed information networks [7, 8] – where reliability under failure conditions is paramount. However, this concept remains poorly defined and underexplored in the domain of software engineering, particularly at the software architecture and design levels. But there are some connected terms that are represented in the area of software architecture: in some academic circles, “functional stability” is a common term, while Western literature often employs related concepts such as self-adaptive, robust, or resilient system design. Further we provide brief definitions of each term and related literature review.

According to the source [9] self-adaptive software is identified as software that possesses the capability to autonomously modify its behavior in response to changes in its operating environment or internal state. Such systems continuously monitor their context and, upon detecting deviations from desired behavior or performance goals, reconfigure themselves automatically to satisfy both functional and non-functional requirements.

The standard [10] defines robustness as the degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions.

The CNSS Glossary [11] determines resilience as the ability to prepare for and adapt to changing conditions and withstand and recover rapidly from disruptions. Resilience includes the ability to withstand and recover from deliberate attacks, accidents, or naturally occurring threats or incidents.

These definitions, while distinct in their emphasis, collectively highlight the system’s capacity to maintain and restore core functionality under adverse conditions and are interrelated in that they contribute to the overall goal of designing systems that can endure, adapt, and recover in harsh dynamic environments.

A first notable foundation for achieving adaptiveness in software design is provided by the work of the GoF [12], which shaped modern software architecture. The GoF introduced a catalog of design patterns that offer proven solutions to recurring problems in software design. Among these, the Observer and Strategy patterns stand out as particularly relevant for enhancing system adaptiveness. The Observer pattern decouples the subject from its observers, allowing components to react in real time to state changes. It enables systems to dynamically adjust their behavior in response to environmental shifts, thereby supporting continuous functional performance. By encapsulating interchangeable algorithms, the Strategy pattern empowers systems to select the most appropriate behavior at runtime. This flexibility is essential for maintaining adaptive responses under varying operational conditions.

The thesis [13] which is summarized in the work [14] presents a catalog of design patterns specifically aimed at enabling software systems to adjust their behavior dynamically at runtime. The work systematically categorizes design patterns that support dynamic adaptiveness, outlining how various patterns address different aspects of runtime adaptation. The patterns are organized into clear categories based on their roles within the adaptive process:

– monitoring and analysis patterns focus on continuously observing system behavior and environmental conditions, providing essential data for triggering adaptation;

– planning and decision patterns encapsulate strategies for selecting among multiple behavioral alternatives; it leverages concepts from control theory and optimization to guide adaptive decision-making;

– execution and reconfiguration patterns responsible for the implementation of adaptive changes at runtime, such as dynamic component replacement and state migration, ensuring that the system can reconfigure itself seamlessly.

The thesis focuses on dynamic adaptiveness without addressing the concept of functional stability, so there is an opportunity to integrate these concepts. The works [15, 16] introduce innovative autonomic design patterns that complement the taxonomy of dynamically adaptive systems presented in the thesis. The work [15] proposes an event-based pattern for web services, triggering adaptive responses via system events, while the study [16] presents an adaptive reconfiguration compliance pattern that enables systems to adjust configurations while meeting compliance standards.

The work [17] introduces a comprehensive taxonomy of design patterns aimed at enabling runtime reconfiguration in software systems. It emphasizes the importance of decoupling configuration logic from core business functionality, thereby allowing components to adapt dynamically without disrupting overall operations. Key ideas include mechanisms for dynamic state transfer, policy-driven reconfiguration, and runtime component replacement – each facilitating seamless adaptations in response to changing environments or internal conditions. Overall, this source provides a structured framework for

DOI 10.15588/1607-3274-2026-1-9

understanding how pattern-based approaches can support software adaptiveness.

In the more recent study [18] authors present a framework for dynamic software adaptation that leverages runtime architectural models to manage both planned and unplanned changes. The paper categorizes adaptation into three types – algorithmic, configuration, and architectural – and introduces state machine-based adaptation patterns that explicitly govern component transitions from active to quiescent states.

The article [19] presents a system designed to automatically detect, diagnose, and repair faults in sensor networks during runtime. The framework is built upon the MAPE-K model [20] – an established autonomic computing paradigm which structures the self-healing process into five distinct phases:

– monitor – the system continuously gathers sensor data and other relevant metrics;

– analyze – collected data are processed to detect anomalies or deviations that may indicate sensor faults;

– plan – upon detecting an issue, the system formulates a remediation strategy;

– execute – the planned corrective actions are applied to restore proper operation;

– knowledge – a shared repository is maintained to store historical data, learned patterns, and decision-making rules for future incidents.

By leveraging this model, the framework not only addresses transient errors in sensor data but also supports dynamic adaptation to evolving fault conditions, thereby enhancing the overall reliability and availability of sensor-based systems. Although the framework addresses self-healing, it does not explicitly integrate the broader notion of functional stability into its design. There is an opportunity to combine the self-healing approach with functional stability principles. The framework is tailored to sensor networks, focusing primarily on data acquisition and fault remediation in that context.

In summary, while the reviewed literature proposes many adaptive and self-healing solutions, these approaches predominantly address how systems can alter their behavior in response to change. However, there remains a gap: the concept of functional stability is underexplored at the software design and architectural levels. This gap underscores the need for novel design patterns and methodologies that explicitly integrate functional stability into software architectures, thereby fostering systems that are not only adaptive but also inherently resilient and robust.

3 MATERIALS AND METHODS

Our approach for ensuring functional stability in software systems is based on the assumption that a software component can achieve functional stability if it is capable of dynamically adapting to changes in its operational environment. Specifically, if each software component C_i can detect events that signify changes in the situation space E and then select an appropriate adaptation strategy from a set ST – using a strategy selection function σ – then software system can maintain its output state within a predetermined stable set $S_{stable}^{(f)}$. In this approach, the presence of multiple strategies, including a fallback strategy for handling unexpected events and critical failures, enables the module to adjust its behavior in real time. Consequently, by dynamically switching among these strategies as counteractions to detected disturbances, a software component can preserve the quality of its primary functions and, in turn, contribute to the overall functional stability of the system.

In this section, we introduce our novel design pattern that aims to enable functional stability in software systems while expanding the adaptive patterns collection. Our pattern is born from the idea of combining the flexibility of the Strategy pattern [12] with the dynamic notification features of the Observer pattern [12]. The Strategy pattern excels at allowing interchangeable algorithms, and the Observer pattern provides the ability to trigger actions in response to incoming events. Below, you'll find the UML diagrams and a detailed description of our pattern. These materials will walk you through how the pattern is structured, how each component interacts, and the rationale behind our design choices.

One key aspect of our approach is evolving from the traditional Observer pattern to a more flexible Publish-Subscribe model [21]. In the classic Observer pattern, subscribers must know the Observer (publisher) directly, which can create tight coupling between components. With Publish-Subscribe subscribers don't need to know the publisher. Instead, they simply subscribe to specific event types. This decoupling enhances flexibility and makes it easier to manage complex, dynamic systems.

We should also mention that event handling is a core element of our pattern, serving as the engine for adaptivity and a one of key contributors to functional stability. By centering our design around event handling, we enable a decoupled, dynamic interaction between components, where events trigger specific responses without requiring direct connections between publishers and subscribers. At the same time, it is important to note that we will not be focusing on how events are generated or detected within the system. We assume that events are produced by various entities through different approaches – whether for example by physical sensors, monitoring tools, or through the specialized "monitoring and analysis patterns" [13, 14, 20]. Our primary concern here is how these events are handled to adapt the behavior of individual components and the overall system.

Our pattern is expected to be applied for software components (modules), and under the term "software component" we understand any self-contained, logically cohesive unit that encapsulates a distinct functionality within a system. Each module should be designed to operate independently, featuring well-defined interfaces that facilitate interaction via our event-driven approach. Such modular structure supports the decoupling and dynamic adaptivity of our design. In addition, our pattern acknowledges that software components can exist in

DOI 10.15588/1607-3274-2026-1-9

hierarchical structures. Some modules may function as submodules of larger modules, forming dependency relationships that need careful handling. Our approach is designed to accommodate these nested configurations, ensuring that even when modules depend on one another, the event-driven mechanism maintains robust decoupling and clear communication channels.

With the foundational concepts in place, let's now dive into the details of our pattern. The pattern is a behavioral design approach that enables software components to dynamically adjust their behavior in response to events that arise in a system. By merging the strategic decision-making capabilities of the Strategy pattern with the loose coupling of the Publish-Subscribe model, this pattern allows components to select the most appropriate operational strategy when specific events occur.

When an event is triggered, a dedicated handler within the component is able to evaluate the event context (the event may contain some valuable information, such as reason, creation time, criticality etc) along with a system state and provide a decision about activating the best-suited strategy from a set of available options. For instance, strategy changes may involve initiating a recovery process, modifying the execution algorithm to maintain the module's core functionality, switching to backup resources, switching to a default strategy that degrades the functionality in case of unexpected and unknown events, adapting to new environmental conditions, or adjusting task processing priorities, among other possibilities.

It's worth mentioning that in our approach, the event handler is not required to change the strategy for every event. Certain events may fall outside a module's ability to resolve on its own. When a module experiences an unknown or critical error that impedes its normal operation, it can notify its dependent modules about the malfunction. These dependent modules are then expected to adjust their strategies accordingly, effectively shifting the responsibility for resolving the event to a higher hierarchical level. Taking this into account, it can be seen that our pattern allows each module to play a dual role – acting both as a publisher and as a subscriber. This means that a module can react to received events while also generating new events in response to its current state. This cascading approach to event processing supports multi-level and hierarchical error management, ultimately enhancing the overall reliability and adaptability of the system. In summary, our design supports adaptation of individual components by allowing them to change their operational strategies (for example, by switching algorithms) and by reconfiguring their submodules.

An example of such hierarchical structure is introduced in Fig. 1. Here, the highest-level module functions as the main component that orchestrates the system, yet its overall functionality depends on the services provided by the lower-level modules. In this view, the downward pointing dependency arrows effectively show that while the high-level module is in a position to steer the configuration and behavior of the lower modules, it remains fundamentally dependent on them to supply the necessary functionality.

Furthermore, the pattern encourages developers to clearly identify and define critical events and to design targeted response strategies. Developers are urged to thoroughly analyze the range of potential events – both anticipated and unforeseen – that might influence system behavior. For each identified event, a corresponding handling strategy should be pedantically crafted. Such careful planning and targeted strategy development ensure that each module can gracefully adapt to changing conditions while remaining isolated from unintended side effects in other parts of the system.

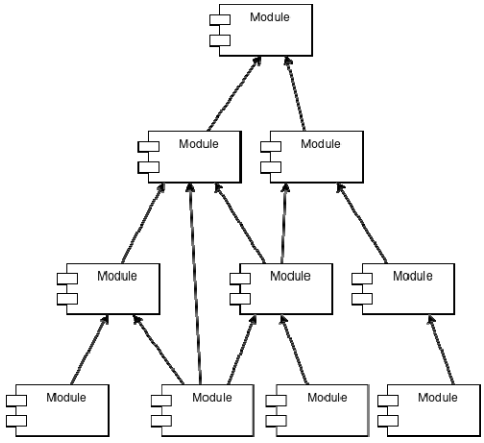


Figure 1 – An example of hierarchical module structure

The UML class diagram of the pattern, which is presented on Fig. 2, illustrates the key components and their relationships within the pattern, further is the description of every item.

– IPublisher. This interface defines the contract for any entity that is responsible for publishing events. It declares the method for event publication, ensuring that any implementing class can deliver events to the system's central coordinator.

– IEventSubscriptionService. This interface outlines the operations required for managing event subscriptions. It includes methods for subscribing and unsubscribing components, ensuring that events are delivered only to those subscribers that have expressed interest.

– IEventSubscriber. The IEventSubscriber interface specifies the method(s) a component must implement to handle incoming events. It standardizes event processing so that every subscriber can react appropriately when notified of an event.

– IStrategy. This interface defines the operational algorithm or behavior that a subscriber may adopt when processing an event. It encapsulates the logic for adapting the module's functionality in response to system or environment changes.

– IStrategySelector. The IStrategySelector interface establishes the contract for selecting an appropriate strategy based on the current context. This context may include for example information provided by an event or

DOI 10.15588/1607-3274-2026-1-9

system's state, allowing the selector to choose the best-fit strategy for the situation.

– Event. An Event represents any change, notification, or trigger that might be significant to system components. Each event is associated with a predefined type (EventType) and can carry various details (such as a timestamp, publisher identity, cause, or description) that subscribers might require for processing the event.

– EventType. Defined as an enumeration, EventType specifies the different categories of events that the system can handle. It serves as a key for mapping events to their corresponding subscribers and facilitates efficient event routing within the system.

– AnEventProducer. This abstract class encapsulates the logic for generating events. AnEventProducer aggregates an instance of the IPublisher interface, which it uses to publish events. By decoupling event generation from event distribution, it ensures that modules remain independent of the specific mechanisms used to deliver events.

– EventManager. Serving as the central coordinator of the pattern, the EventManager implements both the IPublisher and IEventSubscriptionService interfaces. It is responsible for managing subscriptions and for delivering events to the appropriate subscribers. To accomplish this, it maintains an internal mapping (subscribers map) that associates each event type with a list of interested subscribers, ensuring that events are routed correctly.

– ConcreteModule. A ConcreteModule plays the dual role of event subscriber and, optionally, event producer. It implements the IEventSubscriber interface to process incoming events and, in some cases, extends AnEventProducer to generate events as well. Upon receiving an event, the ConcreteModule may leverage a strategy selection mechanism to dynamically choose an appropriate response strategy, thereby adapting its behavior to maintain system stability.

The UML sequence diagram is illustrated on Fig. 3. This diagram conveys the dynamic interplay between the

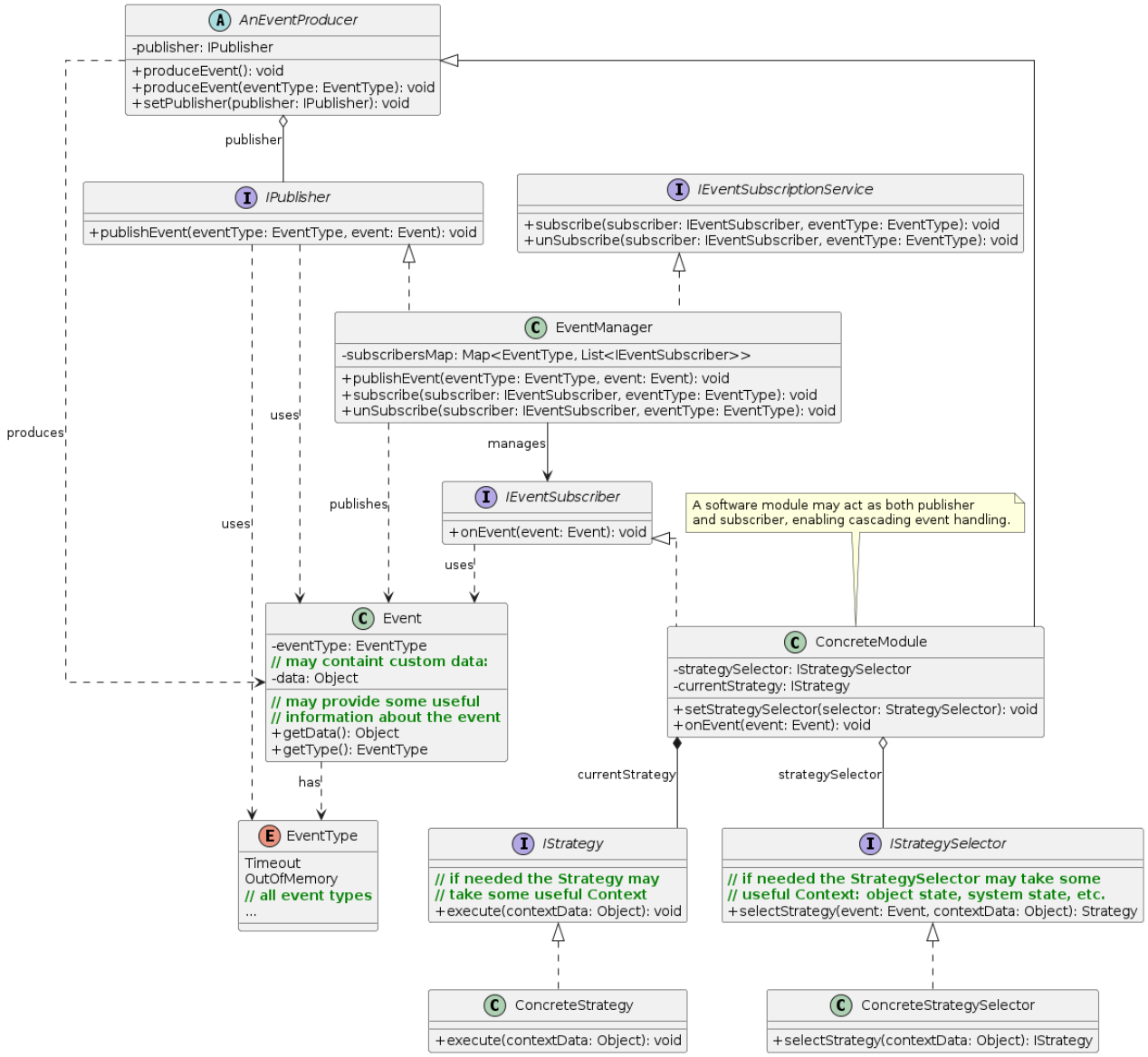


Figure 2 – UML class diagram of the proposed pattern

DOI 10.15588/1607-3274-2026-1-9

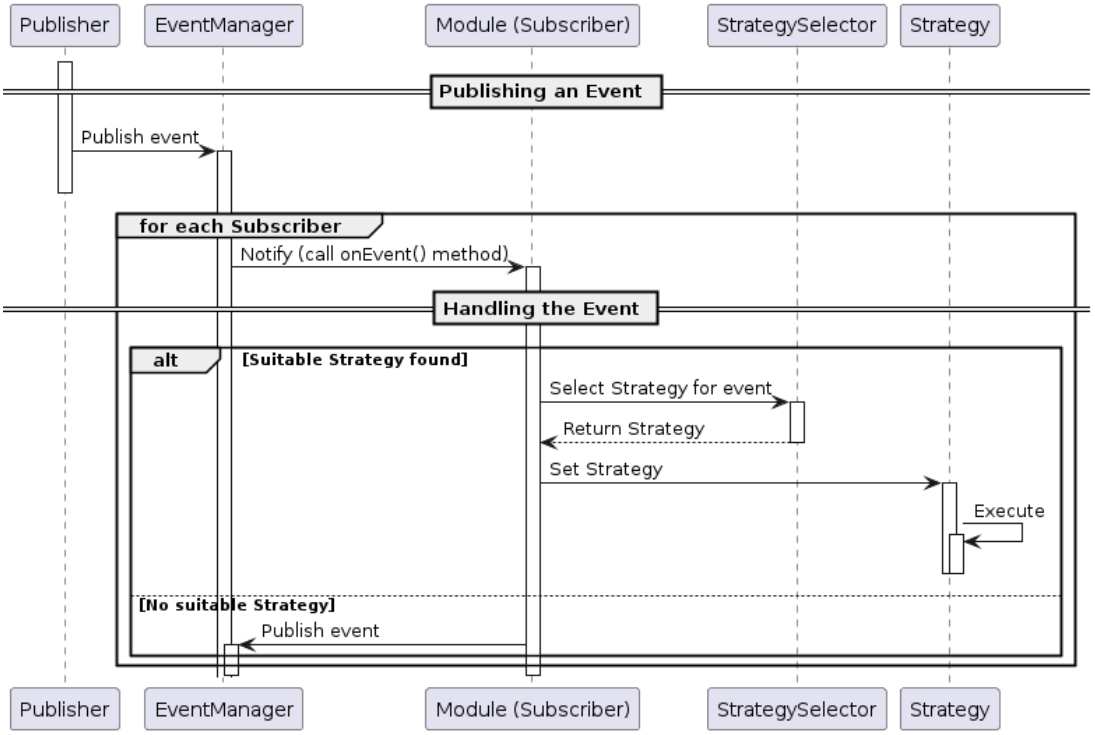


Figure 3 – UML sequence diagram of the proposed pattern

key components in our pattern, showcasing how the system adapts to events through a cascading mechanism. The diagram highlights the inherent flexibility built into the design. Upon receiving the event, each subscriber evaluates it using its StrategySelector. The sequence diagram illustrates that if a subscriber identifies an appropriate response, it will execute the corresponding strategy to adapt its behavior. However, if no suitable strategy is found, the subscriber escalates the situation by publishing a new event, effectively shifting the responsibility to higher-level components. This cascading approach not only underscores the dual role of modules – as both publishers and subscribers – but also reinforces a multi-tiered error management process that enhances the overall reliability and adaptability of the system.

In summary, our proposed pattern fulfills the described approach of providing functional stability for software systems by enabling each software component to maintain multiple strategies – represented via the IStrategy interface – and dynamically react to changes. The pattern ensures that, upon detection of relevant events, a component can select and execute the most appropriate adaptation strategy through its strategy selection interface IStrategySelector. This design not only supports flexible adaptation but also lays the architectural foundation for achieving functional stability across the system.

4 EXPERIMENTS

In our experimental evaluation, we address the challenge of designing a robust software system that can maintain functional stability under dynamic conditions. Rather than focusing on specific code implementations, our experiment demonstrates how the proposed enabling pattern influences the design of an application intended for real-life challenges, such as stable visual object tracking for autonomous platforms.

The solution under design is expected to detect, track, and recover from tracking failures in real time. Initially, the system analyzes the video stream to detect an object based on predefined criteria. Upon detection, an event is generated that triggers the transition from an object detection mode to an object tracking mode. When the target moves an appropriate event occurs and the corresponding module uses a strategy selection mechanism to determine whether to adjust the camera orientation or reposition the platform using its wheels.

The experimental design is considered to support several scenarios.

– Initial detection and transition to tracking. To start the tracking process, the system must first detect the target object. In this initial step, the system employs “find” and “detect” strategies by exploring the camera view and, if necessary, changing its physical location. Once the target is detected, a detection event is published and routed to all interested parties, initiating the tracking process.

– Dynamic tracking with adaptive strategies. As the target moves, a series of events reflecting its motion are generated. The strategy selector evaluates these events and dynamically chooses the appropriate response – adjusting the camera for minor movements or driving the wheels for larger positional changes. This scenario

DOI 10.15588/1607-3274-2026-1-9

demonstrates the pattern's ability to maintain functional stability by adapting to ongoing changes.

– Target loss and recovery. In situations where the target temporarily leaves the field of view or becomes occluded, the system generates a "target lost" event. This event triggers a recovery procedure, wherein the system searches for the target based on the last known tracking data

The pattern's capability to dynamically switch between different strategies ensures that the system can adapt to environmental variability and unexpected disturbances, thereby achieving reliable performance under a wide range of conditions. To present our experimental design we have developed a series of simplified UML class diagrams that detail the architecture of the experimental system. These diagrams illustrate the core infrastructure, along with the domain-specific modules that implement dynamic adaptation. Note that some core components (e.g., EventManager) and interfaces are not present in UML diagrams for better readability, but they are still essential for the design.

Fig. 4 presents the components related to the camera functionality. The CameraModule, responsible for capturing frames and reorienting the camera, extends the common AnEventProducer and implements the IEventSubscriber interface. It interacts with a dedicated CameraStrategySelector that dynamically selects between the CameraTrackingStrategy and CameraSearchStrategy based on incoming events.

The Movement Domain diagram presented in Fig. 5 details the architecture for platform repositioning. The MovementModule, which manages the physical adjustments of the system, similarly extends AnEventProducer and implements IEventSubscriber. It utilizes a MovementStrategySelector to choose between strategies such as WheelTrackingStrategy and WheelSearchStrategy. This enables responsive adjustments to the platform's position in reaction to target movement or loss.

In the diagram on Fig. 6, the Image Analysis Domain is depicted. The ImageAnalysisModule, tasked with processing captured images and analyzing target characteristics, also inherits from AnEventProducer and implements IEventSubscriber. It employs an ImageAnalysisStrategySelector to dynamically select among strategies such as FindObjectStrategy, TrackObjectStrategy, and RecoverFindStrategy. This modular approach facilitates robust analysis and quick recovery when tracking is disrupted.

By focusing on architectural design rather than implementation specifics, our experiment demonstrates that the proposed enabling pattern can serve as a robust architectural backbone for adaptive systems. The pattern's ability to support both dynamic strategy selection and component reconfiguration highlights its potential to enhance system functional stability in real-world applications.

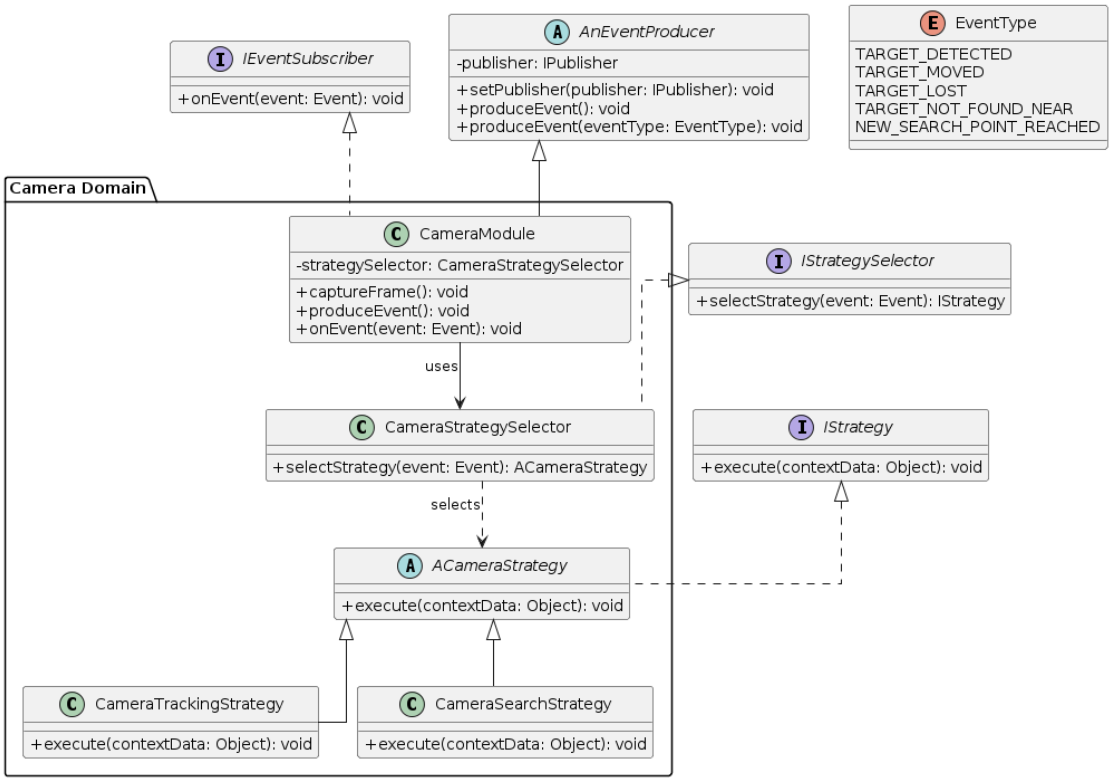


Figure 4 – The UML class diagram for Camera domain

DOI 10.15588/1607-3274-2026-1-9

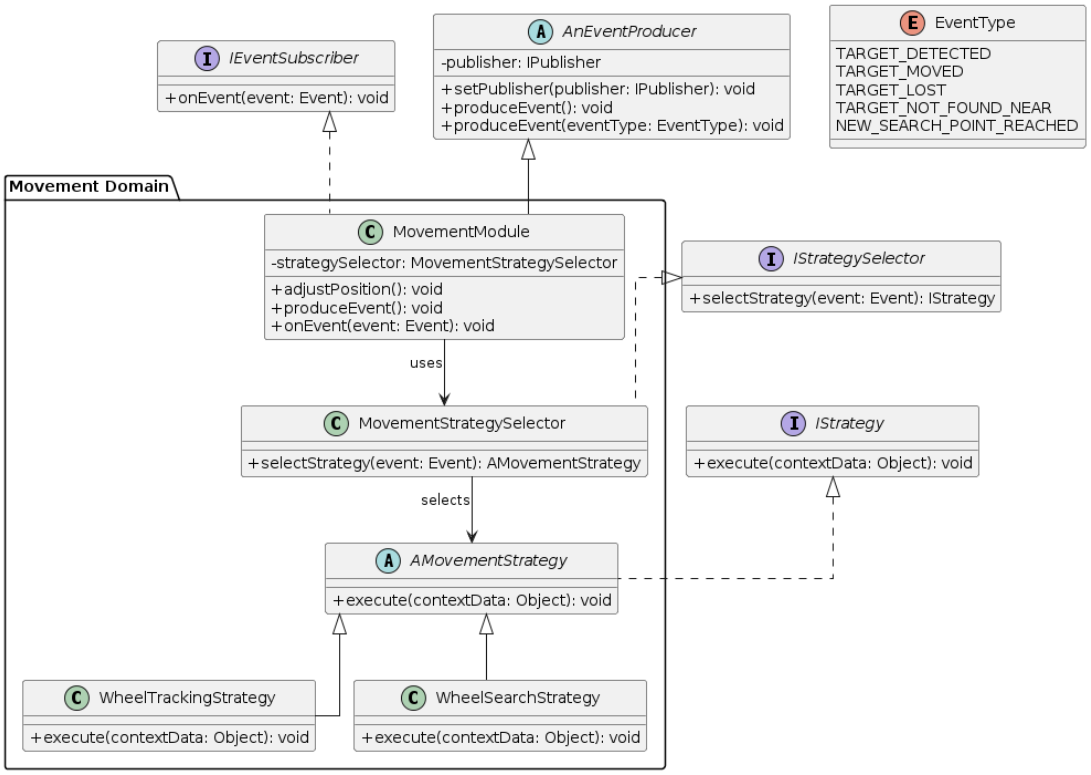


Figure 5 – The UML class diagram for Movement domain

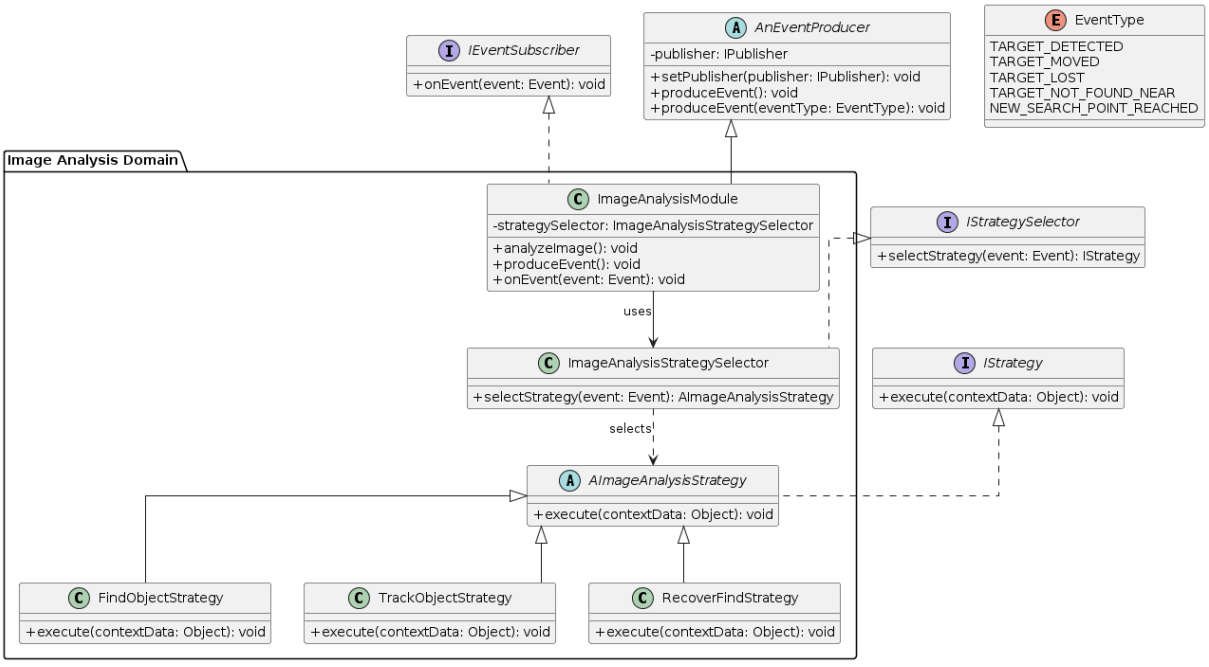


Figure 6 – The UML class diagram for Image Analysis domain

5 RESULTS

The proposed design pattern was evaluated from the perspective of structural quality, modularity, and readiness for adaptation in dynamic software systems. Rather than benchmarking low-level algorithmic performance, the assessment focuses on design-level metrics and scenario-based maintainability analysis.

To quantify structural quality, object-oriented design metrics were calculated based on the UML architecture presented in the Experimental section. Their values are shown in Table 1.

DOI 10.15588/1607-3274-2026-1-9

Table 1 – Object-oriented metrics for the proposed design

Metric	Value	Interpretation
CBO (Coupling Between Objects) [22]	4	Low coupling; components interact solely via event types
LCOM (Lack of Cohesion in Methods, LCOM4) [22]	0.19	High cohesion; modules encapsulate tightly related functionality
RFC (Response for Class) [22]	36	Moderate response surface; limited and testable method complexity
DIT (Depth of Inheritance Tree) [22]	3	Moderate abstraction depth due to use of interfaces and base classes

These metrics confirm the pattern's architectural strength: decoupled modules, focused responsibilities, and extensibility through clearly defined interfaces.

Beyond metric-based evaluation, the design was assessed according to established SOLID [23] principles. In general, our design satisfies the SOLID principles:

– it demonstrates high cohesion and a clear separation of responsibilities (SRP);

– it is open to extension through subclassing while remaining stable (OCP);

– it supports substitutability of components without loss of functionality (LSP);

– it employs client-specific interfaces to prevent unnecessary dependencies (ISP);

– it inverts dependencies so that high-level modules rely on abstractions (DIP).

Qualitatively, the architecture shows:

– high modularity – by decoupling modules via events, each component can be developed and maintained independently, which facilitates scalability and easier integration of new features;

– dynamic adaptability – the use of dedicated strategy selectors within each domain allows the system to respond adaptively to dynamic conditions, thereby enhancing functional stability;

– maintainability – clear separation of core and domain-specific components reduces complexity and makes the system easier to understand and maintain.

These qualities collectively validate the robustness, adaptability, and maintainability of our proposed pattern.

Although the study does not rely on a specific implementation, an approximate estimation of runtime overhead was performed based on architectural assumptions. The event loop involves four stages: event creation, dispatch, strategy selection, and strategy execution. Table 2 outlines estimated average durations for each step, assuming single-process deployment with in-memory event dispatch.

Table 2 – Estimated runtime overhead per event

Operation	Estimated duration (μs)
Event creation and enqueueing	5–10
Dispatch to subscribers	10–15
Strategy selection	3–5
Strategy switch and execution	4–6
Total estimated per event	22–36

Given that typical processing intervals in control or video-tracking applications range from 20 to 50 milliseconds, even handling multiple events per frame results in overhead well under 1–2% of total cycle time. These figures validate that the design is lightweight and suitable for real-time reactive systems.

To further analyse maintainability, several common evolution scenarios were considered. For each, we assessed the required changes and the relative complexity involved in extending the system using the proposed pattern. Table 3 presents these findings.

Table 3 – Change scenarios and their impact on the proposed design

Scenario	Required changes	Relative impact
Adding a new event type	Define a new event type, register subscribers, implement matching strategy	Low
Replacing tracking algorithm	Add new strategy and register it; no changes to other modules	Low
Adding a notification module for alerting	Implement a subscriber, subscribe to target events	Low
Introducing criticality-based prioritisation of events	Extend event metadata, modify dispatch logic or event manager policies	Medium
Changing event routing policy (e.g., filtering, batching)	Modify event manager internals or insert pre-processors	Medium
Adding a dynamic strategy selector that uses ML models	Extend selector logic, integrate model loading, handle decision fallback	Medium

The analysis shows that typical functional extensions involve localized changes and benefit from the modular structure. Scenarios that modify global coordination logic (e.g., routing policies) introduce higher complexity, which is expected due to their system-wide implications.

6 DISCUSSION

Our evaluation confirms that the proposed software design pattern successfully addresses the stated problem, described at the beginning of the article. Specifically, the pattern meets the following requirements:

– enabling adaptability – each component dynamically selects a strategy based on the current situation and internal state;

– handling expected and unforeseen situations – the inclusion of fallback strategies ensures that even when unexpected conditions arise, the system maintains its stability;

– supporting extensibility and scalability – the modular architecture, with clearly decoupled components that communicate solely via event types, facilitates maintenance, extension, and scaling.

These points are supported by our design-level metrics and qualitative assessments, which together demonstrate that the pattern provides a robust architectural backbone for enabling functional stability.

Our proposed pattern distinguishes itself from reviewed solutions in several ways.

DOI 10.15588/1607-3274-2026-1-9

– Integrated Adaptability Through Strategy Selection. Whereas traditional patterns like Observer and Strategy provide individual benefits – event notification and algorithmic interchangeability – our pattern fuses these concepts together. By dynamically selecting among domain-specific strategies, our pattern not only reacts to events but does so in a way that supports continuous functional stability.

– Decoupling and Modularity. Several approaches in the literature [13–16, 19, 20] emphasize reconfiguration or centralized fault handling. Our pattern achieves adaptability through a decoupled, event-driven architecture, ensuring that components interact solely through standardized event types. This decoupling minimizes interdependencies and supports a highly modular design. Such modularity enables incremental extension and easier maintenance.

– Architectural Focus on Functional Stability. While many studies address adaptiveness [13–16, 20] or self-healing [19] – typically measured in terms of recovery time or system throughput – our work explicitly targets functional stability. Our pattern ensures that, despite disturbances, the output state of each system function remains within an acceptable stable range after a transient period. This focus on functional stability fills a gap identified in the literature where adaptive techniques are rarely evaluated against the criteria of long-term functional stability.

– Combination of Strategy and Reconfiguration Approaches. Some literature [13–16] contrasts strategy-based adaptation with dynamic reconfiguration. Our pattern uniquely combines these perspectives: while components may change their operational strategies at runtime, these strategy changes can incorporate subcomponent reconfigurations as needed. This hybrid approach not only provides flexibility but also ensures that the system maintains its essential functionality even as underlying configurations evolve.

Despite its strengths, our design is not without challenges. The event-driven architecture introduces inherent complexity in managing event flows, which can complicate debugging and performance analysis. Additionally, the overhead of dynamic strategy selection may impact performance under extremely high event loads. These trade-offs are considered acceptable given the substantial gains in adaptability and modularity; however, they warrant further empirical investigation to optimize system performance in large-scale deployments.

The proposed pattern is particularly well-suited for applications requiring high adaptability, such as autonomous systems, real-time monitoring, and fault-tolerant computing environments. Its modular nature enables integration with existing systems and supports iterative enhancement.

CONCLUSIONS

This work addresses the scientific problem of ensuring functional stability in dynamic, resource-constrained software systems by developing an enabling design pattern. Our proposed pattern provides an architectural backbone that decouples system components through event-driven interactions and supports dynamic strategy selection.

The scientific novelty of this work is twofold. Firstly, our novel pattern supports dynamic adaptation by allowing each module to select an appropriate strategy based on the current operational context. This novel pattern is firstly obtained in our study and is substantiated by quantitative design-level metrics – such as low coupling and inferred high cohesion – which demonstrate the robustness and maintainability of the design. Secondly, the functional stability received further development by bridging the gap in the software architecture viewpoint. In this regard, our approach ensures that system functions remain within predetermined stable states despite disturbances, effectively uniting theoretical concepts with practical architectural design.

The practical significance of the proposed pattern is that it is applicable to a wide range of adaptive systems, including autonomous platforms, real-time monitoring, and fault-tolerant computing environments. Its modularity and clear separation of concerns facilitate easier integration, maintenance, and future extension. As a recommendation, practitioners can adopt this pattern as a foundational element in designing resilient architectures where adaptability and stability are critical.

Prospects for further research are to focus on advancing the pattern by incorporating additional features, such as prioritization of events and multi-threading support, which could further enhance the pattern's scalability and performance in high-demand environments. In parallel, the development of specialized tools and methodologies for streamlined pattern testing, debugging, and integration is highly recommended.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: Oleksii Bychkov: pattern conceptualization, writing – review & editing; Mykola Moroz: searching and reviewing the literature, pattern design and evaluation, writing – original draft.

Data availability: The manuscript has no associated data.

Software availability: The manuscript has no associated software.

Use of artificial intelligence tools: The authors used artificial intelligence technologies in creating the submitted work: X-GPT-4 model was used in order to perform style, grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

DOI 10.15588/1607-3274-2026-1-9

REFERENCES

1. Barabash O. V., Sobchuk V. V., Musienko A. P. et al. System analysis and method of ensuring functional sustainability of the information system of a critical infrastructure object, *System analysis and artificial intelligence*. Cham, Springer, 2023, Vol. 1107, pp. 177–192. DOI: 10.1007/978-3-031-37450-0_11
2. Barabash O. V., Svynchuk O. V., Salanda I. P. et al. Ensuring the functional stability of the information system of the power plant on the basis of monitoring the parameters of the working condition of computer devices, *Advanced Information Systems*, 2024, Vol. 8, № 2, pp. 107–117. DOI: 10.20998/2522-9052.2024.2.12
3. Kalashnyk G. A., Kalashnyk-Rybalko M. A. Strategy for provision of the functional stability of integrated complexes of modern and advanced aircraft onboard equipment, *Perspective trajectory of scientific research in technical sciences*. Riga, Baltija Publishing, 2021, Section 10, pp. 186–202. DOI: 10.30525/978-9934-26-085-8-10
4. Kalashnyk G. A., Kalashnyk-Rybalko M. A. Methodology for ensuring the functional stability of aircraft integrated modular avionics complex, *Science and Technology of the Air Force of Ukraine*, 2024, № 4 (53), pp. 30–40. DOI: 10.30748/nitps.2023.53.04
5. Firsov S. N., Pishchukhina O. A. Intelligent support of multilevel functional stability of control and navigation systems, *Radio Electronics, Computer Science, Control*, 2018, № 2, pp. 177–183. DOI: 10.15588/1607-3274-2018-2-20
6. Barabash O. V., Tverdenko H. M., Sobchuk V. V.et al. The assessment of the quality of functional stability of the automated control system with hierarchic structure, *Proceedings of the 2nd International Conference on System Analysis & Intelligent Computing*, 2020, pp. 1–4. DOI: 10.1109/SAIC51296.2020.9239122
7. Sobchuk V. V., Barabash O. V., Musienko A. P. et al. Analysis of the main approaches and stages for providing the properties of the functional stability of the information systems of the enterprise, *Sciences of Europe*, 2019, Vol. 1, № 42, pp. 41–44.
8. Barabash O. V., Sobchuk V. V., Musienko A. P. et al. System analysis and method of ensuring functional sustainability of the information system of a critical infrastructure object, *System Analysis and Artificial Intelligence*. Cham, Springer, 2023, Section 11, pp. 177–192. DOI: 10.1007/978-3-031-37450-0_11
9. Salehie M., Tahvildari L. Self-adaptive software, *ACM Transactions on Autonomous and Adaptive Systems*, 2009, Vol. 4, № 2, pp. 1–42. DOI: 10.1145/1516533.1516538
10. IEEE Standard Glossary of Software Engineering Terminology : IEEE Std 610.12-1990. [Effective from 1990-12-31]. New York, IEEE, 1990, 84 p.
11. Committee on National Security Systems Glossary : CNSSI 4009. [Effective from 2022-03]. Fort Meade, CNSS, 2022, 252 p.
12. Gamma E., Helm R., Johnson R., Vlissides J. Design Patterns: Elements of Reusable Object-Oriented Software. Boston, Addison-Wesley, 1995, 395 p.
13. Ramirez A. J. Design patterns for developing dynamically adaptive systems : thesis … master of science in computer science. East Lansing, Michigan State University, 2008, 244 p. DOI: 10.25335/ttfn-qx40
14. Ramirez A. J., Cheng B. H. C. Design patterns for developing dynamically adaptive systems, *Proceedings of the 2010 ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems*, 2010, pp. 49–58. DOI: 10.1145/1808984.1808990
15. Mannava V., Ramesh T. A novel event based autonomic design pattern for management of webservices, *Advances in Computing and Information*, 2011, pp. 142–151. DOI: 10.1007/978-3-642-22555-0_16
16. Mannava V., Ramesh T. A novel adaptive re-configuration compliance design pattern for autonomic computing systems, *Procedia Engineering*, 2012, Vol. 30, pp. 1129–1137. DOI: 10.1016/j.proeng.2012.01.972
17. Gomaa H., Hussein M. Software reconfiguration patterns for dynamic evolution of software architectures, *Proceedings of the Fourth Working IEEE/IFIP Conference on Software Architecture (WICSA 2004)*, 2004, pp. 79–88. DOI: 10.1109/wicsa.2004.1310692
18. Gomaa H., Albassam E. Run-time software architectural models for adaptation, recovery and evolution, *Proceedings of the MODELS 2017 Satellite Event*, 2017, pp. 193–200.
19. Nguyen T. A., Aiello M., Yonezawa T. et al. A self-healing framework for online sensor data, *2015 IEEE International Conference on Autonomic Computing (ICAC)*, 2015, pp. 295–300. DOI: 10.1109/icac.2015.61
20. An architectural blueprint for autonomic computing : white paper, IBM Corporation. Armonk, NY, 2006, 37 p.
21. Schmidt D. C., Buschmann F., Henney K. Pattern-oriented software architecture. Hoboken, NJ, John Wiley & Sons, 2007, pp. 28–29.
22. Chidamber S. R., Kemerer C. F. A metrics suite for object oriented design, *IEEE Transactions on Software Engineering*, 1994, Vol. 20, № 6, pp. 476–493. DOI: 10.1109/32.295895
23. Martin R. Clean architecture: A craftsman's guide to software structure and design. Boston, Prentice Hall, 2018, 420 p.

Received 17.03.2025.
Accepted 14.01.2026.
Published 27.03.2026.

УДК 004.41

ПАТЕРН ПРОЄКТУВАННЯ ДЛЯ ЗАБЕЗПЕЧЕННЯ ФУНКЦІОНАЛЬНОЇ СТІЙКОСТІ В ПРОГРАМНИХ СИСТЕМАХ

Бичков О. С. – д-р техн. наук, професор, завідувач кафедри програмних систем та технологій Київського національного університету імені Тараса Шевченка, Київ, Україна. ROR: https://ror.org/02aaqv166. ORCID: https://orcid.org/0000-0002-9378-9535.

Мороз М. В. – аспірант кафедри програмних систем та технологій Київського національного університету імені Тараса Шевченка, Київ, Україна. ROR: https://ror.org/02aaqv166. ORCID: https://orcid.org/0000-0001-6953-683X.

АНОТАЦІЯ

Актуальність. Сучасні програмні системи працюють у динамічних та жорстких умовах, де внутрішні та зовнішні збої, непередбачувані збурення, прямі атаки та обмеження ресурсів ускладнюють стабільне надання основних функцій. У таких умовах забезпечення функціональної стійкості, коли якість кожної функції системи залишається в межах заздалегідь визначеного стабільного діапазону, незважаючи на збої чи аномалії середовища, є критично важливим, особливо для систем, де безпека та висока доступність мають першорядне значення.

Мета роботи – розробка та обґрунтування патерну проєктування, який забезпечує архітектурну основу для досягнення функціональної стійкості в програмних системах. Головний акцент робиться на наданні гнучкого рішення, яке сприяє динамічній адаптації при збереженні надійної роботи системи.

Метод. Запропоновано новий патерн проєктування, який поєднує можливості динамічного вибору стратегії з перевагами слабкого зв'язку між компонентами, який забезпечується моделлю на основі подій. Цей патерн роз'єднує компоненти

DOI 10.15588/1607-3274-2026-1-9

системи, забезпечуючи взаємодію виключно через стандартизовані типи подій, та дозволяє кожному модулю обирати відповідну стратегію адаптації на основі його поточного контексту. Описаний патерн був використаний для побудови дизайну прикладу, спрямованого на впровадження функціональності стабільного трекінгу об'єктів для автономних квадроплатформ. Запропонований дизайн оцінювався за допомогою дизайн-метрик та якісних порівнянь з існуючими адаптивними підходами.

Результати. Проведений аналіз показав, що патерн забезпечує значну модульність та адаптивність. Ключові об'єктно-орієнтовані метрики свідчать про мінімальну взаємозалежність між модулями та чіткий розподіл обов'язків. Запропонований дизайн демонструє, що патерн підтримує динамічну адаптацію поведінки за рахунок гнучкого вибору стратегії і слугує засобом забезпечення функціональної стійкості, створюючи надійну архітектурну основу для програмних систем.

Висновки. Наукова новизна цієї роботи є подвійною: по-перше, у дослідженні отримано новий патерн, що забезпечує динамічну адаптацію через контекстно-орієнтований вибір стратегії; по-друге, функціональна стійкість отримала подальший розвиток в області програмної архітектури шляхом заповнення прогалини між теоретичними концепціями стійкості та практичним проєктуванням систем. Запропонований патерн пропонує надійне, масштабоване та підтримуване архітектурне рішення з вагомими практичними наслідками для розробки адаптивних, стійких програмних систем.

КЛЮЧОВІ СЛОВА: патерни проєктування програмного забезпечення, функціональна стійкість, обробка подій, адаптивна поведінка, автономні системи.

ЛІТЕРАТУРА

1. System analysis and method of ensuring functional sustainability of the information system of a critical infrastructure object / [O. V. Barabash, V. V. Sobchuk, A. P. Musienko et al.] // System analysis and artificial intelligence. – Cham : Springer, 2023. – Vol. 1107. – P. 177–192. DOI: 10.1007/978-3-031-37450-0_11
2. Ensuring the functional stability of the information system of the power plant on the basis of monitoring the parameters of the working condition of computer devices / [O. V. Barabash, O. V. Svynchuk, I. P. Salanda et al.] // Advanced Information Systems. – 2024. – Vol. 8, № 2. – P. 107–117. DOI: 10.20998/2522-9052.2024.2.12
3. Kalashnyk G. A. Strategy for provision of the functional stability of integrated complexes of modern and advanced aircraft onboard equipment / G. A. Kalashnyk, M. A. Kalashnyk-Rybalko // Perspective trajectory of scientific research in technical sciences. – Riga : Baltija Publishing, 2021. – Section 10. – P. 186–202. DOI: 10.30525/978-9934-26-085-8-10
4. Kalashnyk G. A. Methodology for ensuring the functional stability of aircraft integrated modular avionics complex / G. A. Kalashnyk, M. A. Kalashnyk-Rybalko // Science and Technology of the Air Force of Ukraine. – 2024. – № 4 (53). – P. 30–40. DOI: 10.30748/nitps.2023.53.04
5. Firsov S. N. Intelligent support of multilevel functional stability of control and navigation systems / S. N. Firsov, O. A. Pishchukhina // Radio Electronics, Computer Science, Control. – 2018. – № 2. – P. 177–183. DOI: 10.15588/1607-3274-2018-2-20
6. The assessment of the quality of functional stability of the automated control system with hierarchic structure / [O. V. Barabash, H. M. Tverdenko, V. V. Sobchuk et al.] // Proceedings of the 2nd International Conference on System Analysis & Intelligent Computing. – 2020. – P. 1–4. DOI: 10.1109/SAIC51296.2020.9239122
7. Analysis of the main approaches and stages for providing the properties of the functional stability of the information systems of the enterprise / [V. V. Sobchuk, O. V. Barabash, A. P. Musienko et al.] // Sciences of Europe. – 2019. – Vol. 1, № 42. – P. 41–44.
8. System analysis and method of ensuring functional sustainability of the information system of a critical infrastructure object / [O. V. Barabash, V. V. Sobchuk, A. P. Musienko et al.] // System Analysis and Artificial Intelligence. – Cham : Springer, 2023. – Section 11. – P. 177–192. DOI: 10.1007/978-3-031-37450-0_11
9. Salehie M. Self-adaptive software / M. Salehie, L. Tahvildari // ACM Transactions on Autonomous and Adaptive Systems. – 2009. – Vol. 4, № 2. – P. 1–42. DOI: 10.1145/1516533.1516538
10. IEEE Standard Glossary of Software Engineering Terminology : IEEE Std 610.12-1990. – [Effective from 1990-12-31]. – New York : IEEE, 1990. – 84 p.
11. Committee on National Security Systems Glossary : CNSSI 4009. – [Effective from 2022-03]. – Fort Meade : CNSS, 2022. – 252 p.
12. Design Patterns: Elements of Reusable Object-Oriented Software / [E. Gamma, R. Helm, R. Johnson, J. Vlissides]. – Boston : Addison-Wesley, 1995. – 395 p.
13. Ramirez A. J. Design patterns for developing dynamically adaptive systems : thesis … master of science in computer science / Andres J. Ramirez – East Lansing, Michigan State University, 2008. – 244 p. DOI: 10.25335/ttfn-qx40
14. Ramirez A. J. Design patterns for developing dynamically adaptive systems / A. J. Ramirez, B. H. C. Cheng // Proceedings of the 2010 ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems. – 2010. – P. 49–58. DOI: 10.1145/1808984.1808990
15. Mannava V. A novel event based autonomic design pattern for management of webservices / V. Mannava, T. Ramesh // Advances in Computing and Information. – 2011. – P. 142–151. DOI: 10.1007/978-3-642-22555-0_16
16. Mannava V. A novel adaptive re-configuration compliance design pattern for autonomic computing systems / V. Mannava, T. Ramesh // Procedia Engineering. – 2012. – Vol. 30. – P. 1129–1137. DOI: 10.1016/j.proeng.2012.01.972
17. Gomaa H. Software reconfiguration patterns for dynamic evolution of software architectures / H. Gomaa, M. Hussein // Proceedings of the Fourth Working IEEE/IFIP Conference on Software Architecture (WICSA 2004). – 2004. – P. 79–88. DOI: 10.1109/wicsa.2004.1310692
18. Gomaa H. Run-time software architectural models for adaptation, recovery and evolution / H. Gomaa, E. Albassam // Proceedings of the MODELS 2017 Satellite Event. – 2017. – P. 193–200.
19. A self-healing framework for online sensor data / [T. A. Nguyen, M. Aiello, T. Yonezawa et al.] // 2015 IEEE International Conference on Autonomic Computing (ICAC). – 2015. – P. 295–300. DOI: 10.1109/icac.2015.61
20. An architectural blueprint for autonomic computing : white paper / IBM Corporation. – Armonk, NY, 2006. – 37 p.
21. Schmidt D. C. Pattern-oriented software architecture / D. C. Schmidt, F. Buschmann, K. Henney. – Hoboken, NJ : John Wiley & Sons, 2007. – P. 28–29.
22. Chidamber S. R. A metrics suite for object oriented design / S. R. Chidamber, C. F. Kemerer // IEEE Transactions on Software Engineering. – 1994. – Vol. 20, № 6. – P. 476–493. DOI: 10.1109/32.295895
23. Martin R. Clean architecture: A craftsman's guide to software structure and design / R. Martin. – Boston : Prentice Hall, 2018. – 420 p.

DOI 10.15588/1607-3274-2026-1-9

p-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2026. № 1
e-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2026. № 1

UDC 614.2+574/578+004.38

COMPARISON OF SOFTWARE ARCHITECTURE EVALUATION METHODS APPLICABILITY IN THE CONTEXT OF CQRS WITH EVENT SOURCING ARCHITECTURAL VARIATIONS

Hruzin D. L. – Post-graduate student of the Department of Electronic Computing Machinery, Oles Honchar Dnipro National University, Dnipro, Ukraine. ROR: https://ror.org/00qk1f078. ORCID: 0009-0004-8534-2559.

Lytvynov O. A. – PhD, Associate Professor of the Department of Electronic Computing Machinery, Oles Honchar Dnipro National University, Dnipro, Ukraine. ROR: https://ror.org/00qk1f078. ORCID: 0000-0001-7660-1353.

ABSTRACT

Context. This study is conducted in the context of developing and justifying a methodology for software architecture (SA) evaluation in relation to the Command Query Responsibility Segregation (CQRS) with Event Sourcing (ES) architectural variations.

Objective. This work aims to evaluate and compare the applicability of SA evaluation methods to support the selection of an optimal CQRS with ES architectural variation for real-world projects.

Method. Various SA evaluation methods are applied to enhance objectivity in architectural decisions. However, these methods are not universal; they vary in depth, focus, and required effort. The task considered in this work is the selection among CQRS with ES architectural variations, often structurally similar and thus difficult to distinguish using general-purpose evaluation methods. Comparing architectural variations requires in-depth analysis; however, for most methods, practical implementation is limited by time and resource constraints. The proposed approach identifies the most appropriate SA evaluation method for selecting between CQRS with ES architectural variations. It is based on a validated framework for classifying and comparing SA evaluation methods. In addition to qualitative analysis, the approach introduces a quantitative assessment of applicability to a specific case, allowing for supporting more informed decision-making.

Results. The approach was applied to compare several SA evaluation methods, including Information Technology for Decision-making Support regarding CQRS with ES Architectural Variations (DSAV-CQRSES), a method specifically designed for evaluating variations of the CQRS with ES architecture.

Conclusions. The existing framework of comparing Software Architectures cannot be directly applied to architectural variations (the deviations of the architecture significant for customer). The proposed modifications of the framework are primarily focused on CQRS with ES variations assessment.

KEYWORDS: Software Architecture, Comparison of evaluation methods, CQRS with Event Sourcing, architectural variations.

ABBREVIATIONS

ADL is an Architecture Design Languages;

ALMA is an Architecture-Level Modifiability Analysis;

ATAM is an Architectural Trade-off Analysis Method;

CBAM is a Cost-Benefit Analysis Method;

CMMI is a Capability Maturity Model Integration;

CQRS is a Command Query Responsibility Segregation;

DSAV-CQRSES is an Information Technology for Decision-making Support regarding CQRS with ES Architectural Variations;

ES is an Evant Sourcing;

MCDA is a Multi-Criteria Decision Analysis;

NIMSAD is a Normative Information Model-based System Analysis and Design;

PASA is a Performance Assessment of Software Architecture;

RTP is a Representative Test Project;

SA is a Software Architecture;

SAAM is a Software Architecture Analysis Method;

SQUASH is a Systematic Quantitative Analysis of Scenarios' Heuristics.

NOMENCLATURE

AS_i is an alternative (architectural strategy) being evaluated;

Cont_{ij} is a contribution of AS_i to QA_j;

E is a vector that contains effectiveness values;

E_{asc} is a sorted vector that contains effectiveness values;

e_i is an effectiveness of the i-th method in vector E;

ea_j is an effectiveness of the j-th method in sorted vector E_{asc};

$E_{\max}$ is a sorted vector that contains the highest effectiveness values;

m is a number of criteria;

Max is a maximum possible distance from the etalon;

n is a number of SA evaluation methods considered in the comparison;

p is a vector representing the candidate method;

$QAscore_j$ is a weight assigned to QA_j;

q is a vector representing the reference method;

ref is a vector representing the reference method;

SA is a set of SA evaluation methods;

SA_{opt} is a set of optimal SA evaluation methods;

W_i is a weight of the i-th criterion;

σ is a sorting function by ascending value;

σ' is an inverse function of σ;

DOI 10.15588/1607-3274-2026-1-10

χ is a transformation function from SA to E;

χ' is an inverse function of χ.

INTRODUCTION

Developing large-scale software applications is a complex and resource-intensive process. One of the key aspects of this process is the selection of the most appropriate software architecture (SA). In organizations with low maturity levels, architectural decisions are often made intuitively, primarily based on the prior experience of individual developers. At higher levels of organizational maturity, such as Level 4 of the Capability Maturity Model Integration (CMMI) model [1] (Quantitatively Managed Organization), software development companies are focused on the predictability of quantitative performance improvement objectives and well-justified choices regarding SA solutions.

Within the architectural solutions which represent high level of abstraction SA variations can be seen as deviations which arise from either structural modifications or the application of additional architectural solutions intended to address specific technical challenges while preserving the core principles of SA.

One of the architectures that has multiple variations is the Command Query Responsibility Segregation (CQRS) with Event Sourcing (ES) architecture. This architecture is typically used in software systems with complex structure and business logic. In such systems, even a small architectural change can lead to significant increase of required development effort. For instance, solving the causal event synchronization problem [2] can affect a number of already existing modules. To avoid unexpected expenses during the development of software systems based on CQRS with ES architectural variations, it is necessary to objectify the selection of not only the SA, but also its variation.

A number of methods have been proposed to evaluate and compare SA solutions [3]. One such approach is the Information Technology for Decision-making Support regarding CQRS with ES Architectural Variations (DSAV-CQRSES) method [4]. This raises the issue of identifying the most suitable method for comparison CQRS with ES architectural variations within the boundaries of development team and project requirements and limitations. Several studies [5–6] have attempted to address this issue by providing qualitative comparisons and classification frameworks for SA evaluation methods. Others [7–8] have analysed pairs of methods based on large-scale statistical surveys of their practical application. However, these comparisons are typically generic and do not account for the specificity of CQRS with ES architectural variations and the context of projects or development teams.

Based on practice experience of DBB Software [9] company, it is assumed that the DSAV-CQRSES method is the most appropriate candidate.

Since this study focuses on evaluating the applicability of the DSAV-CQRSES, which enables objective selection of a suitable CQRS with ES architectural variation, a number of scoping restrictions are defined:

– The analysis is limited to the CQRS with ES architecture and its variations.

– Methods for SA evaluation are categorized by application phase as design-time or run-time techniques, with some methods applicable in both phases [3]. This work considers methods that can be applied at the design stage.

– Also, according to [3], evaluation methods may be classified into utility-based, scenario-based, parametric-based, search-based, economics-based, and learning-based categories. This study focuses exclusively on scenario-based methods since the DSAV-CQRSES approach itself belongs to this category and is grounded in use case analysis.

This work provides a brief overview of several SA evaluation methods selected based on a systematic literature review analysis [3, 10–11]. It considers existing approaches for classifying and comparing these methods. It also introduces an SA variation-oriented approach, based on the framework for classifying and comparing SA evaluation methods, which enables the transformation of qualitative assessments into quantitative, thereby facilitating the identification of the most suitable SA variation evaluation method.

The object of study is the process of comparing and assessing the applicability of SA evaluation methods.

The subject of study is methods, approaches, and frameworks for comparing and selecting the most suitable SA evaluation method to assess CQRS with ES architectural variations at the system design stage, as well as to provide an effective strategy for their evolution.

The purpose of the work is to evaluate and compare the applicability of SA evaluation methods to support the selection of an optimal CQRS with ES architectural variation within the context of a specific software company and its real-world projects.

1 PROBLEM STATEMENT

A wide range of architectural choices exists, ranging from structural and organizational aspects, such as choosing between monolithic and microservices architectures, to conceptual paradigms including Event Sourcing, Domain-Driven Design, and Service-Oriented Architecture.

To determine the most suitable solution, various methods are proposed, including Architectural Trade-off Analysis Method (ATAM) [12], Software Architecture Analysis Method (SAAM) [13], and DSAV-CQRSES [4], among others. Given the substantial number of available approaches, the question of selecting the most appropriate evaluation method becomes relevant. Several studies [5, 7, 11] have attempted to compare such methods. However, these comparisons are typically qualitative in nature and not designed to support the evaluation of architectural variations.

DOI 10.15588/1607-3274-2026-1-10

Suppose given a set of scenario-based SA evaluation methods $SA=\{sa_1, sa_2, \ldots, sa_n\}$ consolidated from a comprehensive literature review and primary research sources, as well as project-specific requirements, priorities, and constraints provided by the decision-making team in the form of (i) a vector of attribute weights *w* and (ii) a vector of optimal (target) attribute values *ref*.

Thus, the problem – evaluating the applicability and effectiveness of scenario-based SA evaluation techniques for comparing CQRS with ES architectural variations in the context of a specific projects – consists in determining a subset $SA_{opt} \subseteq SA$ that contains one or more evaluation methods most suitable for the comparison task. Optimality is determined by (i) a qualitative assessment that reveals similarities and differences among the considered methods, and (ii) a quantitative assessment of each method's applicability with respect to *w* and *ref.*

2 REVIEW OF THE LITERATURE

Unfortunately, no techniques were found that specifically address the evaluation of the applicability of SA evaluation methods within the context of CQRS with ES architectural variations. However, several approaches exist for comparing these methods in general.

One such approach is the framework for classifying and comparing software architecture evaluation methods proposed in [5], which was developed by identifying similarities and differences among existing evaluation methods. The framework introduces a set of guiding questions designed to characterize SA evaluation methods. These questions include both general aspects, such as the number of quality attributes (QAs) considered and the estimated man-days required for applying the method, as well as more specific ones, such as whether the method provides support or guidance on non-technical aspects (e.g., social, organizational, managerial, or business issues) involved in the evaluation process. The proposed classification parameters (questions) were validated in [14] through expert surveys.

An extension to this framework was later published in [6], which arranged each element within four components, according to the Normative Information Model-based System Analysis and Design (NIMSAD) [15] evaluation framework. Three new dimensions were also added: Input and Output Management, Application Domain, and Stakeholder Benefits.

The output of applying this framework is a summary table describing the key characteristics of various SA evaluation methods. Although this helps narrow down the selection space and reduces the need to deeply study each method individually, the comparison remains qualitative and does not assess how well a particular method applies to a specific team and project.

An alternative comparison methodology is found in [7–8], which presents a family of experiments comparing Quality-Driven Architecture Derivation and Improvement [16] and ATAM [12]. The comparison was based on six parameters. Two of them are measurable: total time spent applying the method and effectiveness, which indicates how close the participant came to selecting the optimal architecture for the given project. A third parameter, efficiency, is calculated as the ratio of effectiveness to the application time. The remaining three parameters were drawn from the Technology Acceptance Model: Perceived Ease of Use, Perceived Usefulness, and Intention to Use. These were assessed through a post-experiment Likert-scale questionnaire containing a set of closed questions for each variable.

Among these parameters, effectiveness is perhaps the most interesting metric in the context of objectivity. It reflects how accurately a participant with relatively low qualifications can select an appropriate architectural solution using a given evaluation method. It was computed using the Euclidean distance between the n-dimensional vector of Non-Functional Requirements (NFR) values attained by the architecture selected by the participant and the optimal vector of values that could be achieved. Unfortunately, the authors did not provide details on the actual vector's values or how they were derived.

While this approach provides quantitative metrics for comparing two methods, it still does not answer the question of which method is best suited for a particular case. Moreover, the use of unqualified participants and a limited set of projects raise concerns about objectivity. Changes in participants, project characteristics, or NFRs could significantly impact the results of such experiments.

3 MATERIALS AND METHODS

Based on a systematic literature review [10–11], the most cited [3] scenario-based methods for SA design-time evaluation were selected: SAAM, ATAM, Cost-Benefit Analysis Method (CBAM), Architecture-Level Modifiability Analysis (ALMA), Systematic Quantitative Analysis of Scenarios' Heu-ristics (SQUASH) and Performance Assessment of Software Architecture (PASA).

This section is structured as follows:

– A proposal of an SA variation-oriented approach for assessing the applicability of selected methods to compare CQRS with ES architectural variations.

– A short overview of these methods, including the DSAV-CQRSES approach.

As a basis to compare SA variation evaluation methods, an enhanced version of the framework for classifying and comparing software architecture evaluation methods [6], modified concerning the NIMSAD evaluation framework.

However, the original framework is unsuitable for comparing SA evaluation methods in the context of their applicability to architectural variations, specifically, variations of CQRS with ES, for several reasons. First, the original set of questions does not accommodate the particular characteristics required for evaluating architectural variations. Secondly, it relies on qualitative rather than quantitative assessment, which delegates considerable analytical effort to the user.

Let us first define the terms SA and SA variation in order to clarify the distinction between them.

DOI 10.15588/1607-3274-2026-1-10

SA definition. Software architecture refers to the fundamental structure of a software system, encompassing its components, their relationships, and the principles guiding its design. In this context, multiple definitions exist. For example:

Lloyd and Galambos [17] define reference architectures as domain-specific architectural templates that aim to address the architectural concerns for a particular class of problems.

Bass et al. [18]: "The software architecture of a program or computing system is the structure or structures of the system, which comprise software elements, the externally visible properties of those elements, and the relationships among them." Similar to the previous definition, this highlights that architecture can be understood as a metamodel [19].

Clements et al. [20]: "Software architecture is the set of structures needed to reason about a software system and the discipline of creating such structures and systems. Each structure comprises software elements, relations among them, and properties of both elements and relations." This definition conceptualizes architecture at a deeper level of abstraction – the model level.

In [21], software architecture is considered from several perspectives. On one hand, it is viewed as a set of basic concepts and constraints within which application functionality is to be specified and integrated. On the other hand, it serves as a means for addressing technical issues and quality requirements, as well as for assessing application functionality. As a result, Fritz Solms defines SA as the software infrastructure within which software components that address functional requirements of the software system can be specified, deployed, and executed.

In the presentation [22], Dr. Jean-Claude Franchitti provides several definitions, among which the following can be regarded as operating at a higher level of abstraction while also being the most comprehensive: "A set of artifacts (that is: principles, guidelines, policies, models, standards, and processes) and the relationships between these artifacts, that guide the selection, creation, and implementation of solutions aligned with business goals".

Based on the definitions above, SA is defined in the context of comparing SA evaluation methods as follows.

Definition: Software architecture is a structured set of artifacts (that is: principles, guidelines, policies, models, standards, and processes) and the relationships between these artifacts, established to enable a software system to satisfy functional and non-functional business requirements. It is characterized by the extent to which a system instantiated upon it fulfills specified business requirements under defined constraints.

SA variation definition. SA solutions are typically designed for a broad class of software systems and therefore offer the most universally applicable practices. Given the increasing need for flexibility in applications and systems, a trend has emerged toward flexible architectures [23] that can adapt to changes in business requirements and advancing technology stacks. The evolutionary architecture approach [24] assumes that SA must be continuously tested and adapted to produce more effective solutions, while ensuring that such evolution does not compromise key architectural concerns. As SAs are applied to various projects within a company, they give rise to a family of architectural variations that differ in complexity, performance, development time, and the level of developer expertise required. These differences have an impact on the development cost and maintainability of the software application.

In some sense, variations may be considered as deviations from the original SA. However, unlike those discussed in [25], they are aimed at improving the software product and optimizing the development process. Variations emerge as responses to specific technical challenges (e.g., event replay performance issues during aggregate reconstruction, complexity of event versioning, etc.) and as methods for reducing development and maintenance complexity.

SA variations, like architecture itself, are described by a metamodel; however, in the case of variations, the metamodel exhibits a higher degree of precision. From the perspective of Evolutionary Architecture, a set of architectural variations can be regarded as different stages of architectural evolution.

Thus, the following definition of an SA variation can be provided.

Definition: Software Architecture Variation is a purposeful deviation from a reference SA, characterized by the modification of one or more artifacts of this SA (that is: principles, guidelines, policies, models, standards, and processes) or the relationships between these elements. It is strategically implemented to optimize a software system's ability to fulfil both functional and non-functional business requirements with maximum efficiency, as measured by current and projected resource expenditure, development effort, and maintenance complexity.

Typically, SA variations emerge and evolve within bounded organizational contexts (e.g., enterprises or development communities) in response to specific technical challenges, operational constraints, or evolving business imperatives of projects under development. Thus, the essence of a SA variation lies in the modification of one or more artifacts of the architecture (i.e., principles, guidelines, policies, models, standards, and processes) or in a deviation from them. However, such deviations must not involve abandoning the fundamental principles of the architecture. For example, rejecting the segregation of commands and queries within the CQRS approach would constitute a departure from the architectural paradigm itself and therefore cannot be regarded as a variation. Similarly, discarding the event store and event-based operations in Event Sourcing also violates core principles. In contrast, using an event store that no longer serves as the source of truth may be considered a deviation from the original architecture and thus qualify as a variation. Ultimately, the determination of whether a modification constitutes a new architectural solution or a variation of an existing one should remain within the discretion of the modification creator.

DOI 10.15588/1607-3274-2026-1-10

For the majority of SA evaluation methods, the QA set commonly used in SA evaluation includes characteristics such as usability, security, reliability, portability, cost, and others. These attributes are also formalized in ISO/IEC 25002:2024 [26]. While all of these parameters are relevant to SA evaluation, the nature of CQRS with ES architectural variations places primary emphasis on maintainability and performance. Within the context of CQRS with ES, other QAs are largely equivalent across the majority of variations [4] and are therefore not the focus of differentiation.

Taking into account the specific characteristics of architectural variations, the questions proposed by the original framework for classifying and comparing SA evaluation methods should be revised.

It is a good point for the method if it includes not only an SA definition but also a definition of SA variation. That will help the team to understand the specificity of architectural candidates. The next objective of the framework is to clarify whether the method's specific goals include the comparison of architectural variations. QAs should cover those that are relevant to the user's needs for comparing candidates within the context of their application (i.e., SA variation-oriented QAs).

As an output, most methods produce documentation describing the architectural candidates. In terms of output quality, the following aspects are evaluated:

– Whether the documentation is ready for direct use during the implementation phase or further elaboration is required.

– Whether trade-offs between architectural alternatives are explicitly addressed.

– Whether the method highlights a recommended architecture or just provides information to support further decision-making.

Different methods offer various benefits. For architectural variation comparison, the most valuable ones include the identification of the optimal solution and an approximate estimate of its implementation effort.

One of the most challenging aspects of applying SA evaluation methods is involving business stakeholders in evaluating QAs. It is often difficult to explain technical QAs (such as modifiability or portability) to non-technical team members. Therefore, it is a great point if stakeholder participation is minimized, and they are asked to assess only easily understandable criteria (e.g., estimated development time in man-hours, or system response time in milliseconds). The usability of a method is also affected by the required resources, such as team size and the time investment. Spending several man-weeks to evaluate each architectural solution can be unacceptably expensive.

Methods provide different forms of architectural candidate description. Formal modeling languages such as Unified Modeling Language [27] or other Architecture Design Languages (ADL) [28] (e.g., module or logical views) are often recommended. However, the depth of these descriptions also differs: some methods provide only a high-level overview, while others model architecture with greater precision. For distinguishing between structurally similar SA variations, especially when relying on expert judgment, clear, formal, and sufficiently detailed descriptions are essential. This leads to the question: Does the method provide a formal and detailed description of the architectural candidates?

Even when detailed descriptions are used, if a method relies on expert judgment, it introduces a human factor, making the evaluation subject to uncertainty. Therefore, it is important to clearly distinguish which parts of the evaluation are based on expert opinions and which are grounded in experiments or statistical data. Although evaluation approaches that rely on empirical experiments and statistical analysis may offer increased objectivity, they are still subject to uncertainty. Thus, it is essential to consider the extent to which a method is affected by uncertainty and whether it includes mechanisms for uncertainty mitigation.

Finally, regarding the validation of a method, it is necessary to clarify whether the method has been validated specifically in the context of architectural variations or not.

The components and attributes of the framework and the evaluation questions are presented in Table 1.

Table 1 – The components and attributes of the framework and the evaluation questions

Component	Elements	Original explanation	Variation-oriented explanation
Context	SA definition	Does the method explicitly consider a particular definition of SA?	How closely do the method's definitions of SA and SA variation align with those considered in this paper?
	Specific goal	What is the particular goal of the methods?	Is the selection between structurally similar SA alternatives addressed by the method's objectives?
	Quality attributes	How many and which quality attributes are covered by the method?	Are variation-specific QAs (e.g., complexity and performance) covered by the method?
	Applicable stage	Which is the most appropriate development phase to apply the method?	Is the method applicable at the design stage of system development?
	Input & output	What are the inputs required and outputs produced?	To what extent is the resulting technical documentation ready for direct use during implementation? Does the method facilitate trade-off analysis? Does the method identify a recommended architecture, or does it only provide additional information about the alternatives?
	Application domain	What is/are the application domain(s) the method is mostly applied?	Is the method validated or considered applicable in domains similar to the target system?

DOI 10.15588/1607-3274-2026-1-10

Continuation of Table 1

Stakeholders	Benefits	What are the benefits of the method to the stakeholders?	Does the method support the selection of the optimal solution? Does the method provide an approximate estimate of the effort required for system implementation?
	Involved Stakeholders	Which groups of stakeholders are required to participate in the evaluation?	Does the method require involvement of multiple stakeholders? How many actions or inputs are expected from stakeholders?
	Process support	How much support is provided by the method to perform various activities?	How much support is provided by the method to perform various activities?
	Socio-technical issues	How does method handle non-technical (e.g. social, organisational issues)?	Does the method address non-technical issues (e.g., social and organizational factors)?
	Required resources	How many man-days are required? What is the size of evaluation team?	How long does the application of the method take? What is the typical team size required to apply the method? How demanding are the architectural skills required from the users of the method?
Contents	Method's activities	What are the activities to be performed and in which order to achieve the goals?	What are the activities to be performed and in what order to achieve the goals?
	SA description	What form of SA description is recommended (e.g., formal, informal, particular ADL, views etc.)?	Does the method provide a formal and detailed description of the architectural candidates?
	Evaluation approaches	What types of evaluation approaches are used by the method?	To what extent is the method subject to uncertainty? Are there any techniques applied to mitigate uncertainty? How much of the evaluation is grounded in quantitative metrics and experimental data versus expert assessment?
	Tool support	Are there tools or experience repository to support the method and its artefacts?	Does the method offer tools to facilitate or partially automate its use?
Reliability	Maturity of method	What is the level of maturity (inception, development, refinement or dormant)?	What is the level of maturity (inception, development, refinement or dormant)?
	Method's validation	Has the method been validated? How has it been validated?	Has the method been validated? Has the method been applied specifically to architectural variations?

The modified framework highlights key points relevant for comparing architectural variations; however, the comparison remains qualitative. To obtain quantitative indicators of the effectiveness of SA evaluation methods, the following algorithm is proposed:

1) Evaluate each criterion by answering the framework's questions for each method using a Yes, Yes/No, No scale (Y, Y/N, N) for binary-type questions, and a Low, Medium, High scale (L, M, H) for degree-based questions. Note: This point applies to all criteria except "Method's activities". While it is important for qualitative comparison, it does not influence the decision regarding the selection of an SA evaluation method and should not be relied upon in quantitative evaluation.

2) Assign weights to each criterion based on its impact on the effectiveness of the method's application. It is proposed to determine weights through expert judgment.

It is important to note that the results of the first two steps can be reused in subsequent applications of the method.

3) Determine quantitative equivalents for the values H, M, L and Y, Y/N, N.

4) Define a reference (etalon) method, representing the characteristics for the ideal candidate, the team would desire to use, and the weights (priorities) assigned to each question. To do this, the team answers the framework's questions, specifying the responses and setting the priority that are ideal for their context. This process is suggested to be conducted through structured collective discussion and subsequent voting.

5) The formalized responses to the framework's questions constitute a multidimensional vector for each candidate method and the reference method. The values of the vector's elements are replaced with their quantitative equivalents.

$$p = (Y, N, L, H), iff \quad Y = H = 3, N = L = 1$$
$$then \quad p = (3,1,1,3).$$

6) The deviation of each candidate method from the reference method is measured using the Euclidean distance metric:

$$D(p,q) = \sqrt{\sum_{i=1}^{m} w_i \cdot (p_i - q_i)^2}, i \in [1,m],$$

where w_i represents the weight of the i-th criterion, q is the vector representing the reference method, p is the vector representing the candidate method, m is the number of criteria.

7) Based on the computed distances, the effectiveness of each method's applicability is derived using the formula (1):

$$Effectiveness(p) = 1 - \frac{D(p, ref)}{\text{Max}}, \qquad (1)$$

where ref is the vector representing the reference method, p is the vector representing the candidate method, Max is the maximum possible distance from the etalon.

8) As a result, a vector of effectiveness values for the SA evaluation methods is obtained:

$$E = \langle e_1, e_2, ..., e_n \rangle, e_i = \chi(sa_i), sa_i \in SA, i \in [1,n],$$

DOI 10.15588/1607-3274-2026-1-10

where $\chi: SA \to E$ and $\chi': E \to SA$ are transformation functions, *SA* is the set of SA evaluation methods, e_i is the effectiveness of the *i*-th method, *n* is the number of SA evaluation methods considered in the comparison.

9) Next, by sorting this array, we obtain an ordered vector E_{asc}.

$$E_{asc} = \sigma(E) = \langle ea_1, ea_2, ..., ea_n \rangle, \quad (2)$$

where $ea_1 \le ea_2 \le \ldots \le ea_n$, $\sigma: E \to E_{asc}$ is the sorting function by ascending value and $\sigma': E_{asc} \to E$ is an inverse function of σ.

10) Determining a vector $E_{\max}$, consisting of elements whose values are close to the maximum value in vector E_{asc} (ea_n). In most cases, this set will consist of a single element:

$$E_{\max} = \max_j(ea_j) = \langle ea_n \rangle. \quad (3)$$

However, when the difference between ea_n and ea_{n-1} is not significant, the decision-maker should consider several options. In this case, the vector $E_{\max}$ will include multiple elements (Fig. 1).

$$E_{\max} = \{ea \in E_{asc} \mid d = ea_n - ea_{n-k}, d < th, k \in \mathbb{N}\}, \quad (4)$$

where *th* is the value that defines the maximum allowable deviation from the maximum value at which ea_j is still considered as one of the most effective values.

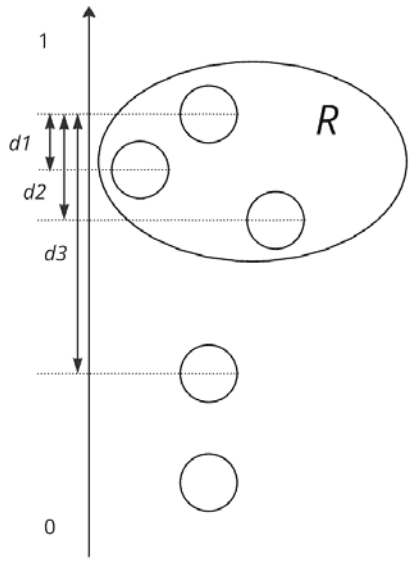


Figure 1 – Set of most suitable results $E_{\max}$. *d*1, *d*2 < *th*, *d*3 > *th*

11) The output of the algorithm is a set $SA_{opt} \subseteq SA$, that contains the most suitable methods deriving from $E_{\max}$:

$$SA_{opt} = \chi'(\sigma'(E_{\max})). \quad (5)$$

The initial step in comparing SA evaluation methods is the description of each method:

Software Architecture Analysis Method [13] is one of the earliest methods proposed for software architecture evaluation. Initially, it focused primarily on maintainability QA. However, subsequent extensions of the method have incorporated additional concerns such as reusability [29], flexibility and complexity [30], and so forth.

Prior to applying the method, the architectural alternatives under evaluation must be described in detail. The SAAM process consists of five main steps: i. characterize a canonical functional partitioning for the domain; ii. map the functional partitioning onto the structural decomposition of the architecture; iii. select a set of QAs for evaluation; iv. identify a set of concrete tasks which test the desired QAs; v. evaluate the degree to which each architecture provides support for each task.

While SAAM is primarily a qualitative evaluation method, one of its distinctive features is the use of small-scale experiments and benchmarks. For example, a benchmark might involve measuring the time or effort required to add a new option to a user's menu bar, reflecting some piece of application functionality. Based on qualitative assessments and experimental benchmarks of application modification, evaluators analyse alternatives and determine whether the candidate provides architectural support in the context of the QA.

The outcome of applying SAAM includes a report that documents the architectural alternatives and their qualitative evaluation of the project's quality requirements, as well as the experimental applications and modification benchmarks developed during the assessment.

The duration of the evaluation depends on the number of alternatives and the complexity of the experimental tasks. On average, conducting a SAAM assessment requires approximately three days for a team of four evaluators, in addition to stakeholder participation and preparatory activities [5].

Limitations of the method include a strong reliance on expert judgment, which introduces a human factor, as well as the lack of quantitative metrics for assessing QAs. Furthermore, SAAM does not explicitly address trade-offs between conflicting QAs.

Architectural Trade-off Analysis Method [12] is a technique for analysing SA that provides a well-defined sequence of evaluation steps, including a schedule and clearly defined roles for each participant. The application of ATAM involves a preparation phase, typically handled by the lead architect or architecture team, and two evaluation phases.

The first phase of ATAM is architecture-centric. It begins with a presentation of the ATAM to the evaluation

DOI 10.15588/1607-3274-2026-1-10

team, followed by the analysis of the problem domain and identification of business goals, specification of requirements for architectural solutions, and the selection of relevant QAs, which can be different (e.g., availability, security, usability, and modifiability). A Quality Attribute Utility Tree is then generated.

Each architectural alternative is then analysed with a focus on four aspects: understanding the approach, identifying well-known weaknesses, recognizing sensitivity points, and finding interactions and trade-offs with other approaches. For each alternative, both approach-specific and quality-attribute-specific questions are discussed, and risks associated with the identified concerns are documented. The outcomes of this phase include: a document with the system's business requirements, a Quality Attribute Utility Tree, and a document outlining the architectural alternatives along with the associated risks and concerns.

The second phase is stakeholder-centric and concentrates on eliciting stakeholder points of view and verifying the results of the first phase. It involves a brainstorming session in which a wide range of scenarios are generated. The scenarios cover both system use cases and possible future modifications. These scenarios are then prioritized via stakeholder voting. Based on these prioritized scenarios, each architectural alternative is analysed again. Any sensitivity or trade-off point is treated as a potential risk.

The final deliverables of ATAM include the documented architectural approaches, the set of scenarios and their prioritization, the collection of attribute-based questions, the utility tree, the list of identified risks and non-risks, as well as the discovered sensitivity and trade-off points. Based on this information, the team selects the architectural approach that demonstrates the lowest level of risk. If none of the evaluated alternatives is deemed sufficiently suitable, the process iterates with the preparation of additional architectural candidates.

Successful application of ATAM requires a multidisciplinary team. A minimum of three evaluators and a representative set of stakeholders is recommended. The core evaluation spans two full working days, besides preparation time. ATAM does not provide formal tool support, although certain steps could potentially be optimized with modern technologies such as large language models. Like most scenario-based approaches, the quality of the evaluation outcome is strongly dependent on the expertise of the participants.

Cost-Benefit Analysis Method [31] is an architecture evaluation technique that extends the ATAM framework by incorporating cost-benefit analysis of architectural design decisions. The prerequisites for applying CBAM include a prior ATAM assessment and the availability of a cost estimation model for implementing architectural alternatives.

The CBAM process involves several steps. Initially, stakeholders assess the benefits of QAs by assigning them relative weights and ranking each alternative based on its contribution (Cont) to each QA on a scale from –1 to 1. This is typically done through stakeholder voting. The benefit of each architectural strategy is then computed using the following formula:

$$Benefit(AS_i) = \sum_j (Cont_{ij} \times QAscore_j).$$

In the next step, the expected cost of each alternative is estimated. If precise data (e.g., statistical data or specific implementation prices) is available, it is used directly; otherwise, it is suggested to estimate the cost using qualitative scales such as High, Medium, or Low. The last step before making a decision is the calculation of the Return on Investment (ROI) metric for each alternative using the following formula:

$$ROI(AS_i) = Benefit(AS_i) / Cost(AS_i).$$

This method heavily depends on expert judgment, particularly due to the abstract nature of some QAs (e.g., reliability or modifiability), which may be interpreted inconsistently across team members, leading to uncertainty in results. Because of this, in the second version of CBAM [32], QAs are replaced with specific scenarios for each QA, and additional evaluation iterations are introduced. As a result, the assessment yields not a single benefit value per architectural alternative, but a set of values. The minimum and maximum values in this set are interpreted as the boundaries of a confidence interval, reflecting the uncertainty inherent in the expert evaluations.

The application of CBAM requires a team comparable in size to that used for ATAM and typically takes approximately two additional working days. The outcomes of this method include the results obtained through ATAM (utility tree, scenarios, risks), ROI estimates for each alternative, and a confidence matrix representing the degree of certainty in the ranking under uncertainty conditions.

Architecture-Level Modifiability Analysis [33] is a method for SA evaluation with a focus on modifiability (e.g., maintenance cost prediction and risk assessment). The evaluation flow consists of five main steps: determine the aim of the analysis, describe the software architecture, find the set of relevant scenarios, determine the effect of the scenarios, and conclude the analysis.

During the ALMA application, the evaluator sets one or more analysis objectives. The first objective is to estimate the cost of maintenance or modification of the application. The second is to identify potential changes for which the architecture is inflexible and to assess the associated risks. The third objective is to support architectural decision-making by comparing multiple SA candidates and selecting the most suitable one.

The method for comparing candidates focuses on finding extreme scenarios that stress the architecture and evaluating how each candidate responds. The scoring process is left to the discretion of the evaluators. This may involve comparing quantitative metrics, if available, or aggregating expert judgment ratings across all scenarios.

DOI 10.15588/1607-3274-2026-1-10

To improve result accuracy, the method recommends conducting the evaluation in several iterations. During each iteration, the team discusses a set of questions (or experiments) for each candidate architecture and assigns ratings for every scenario.

The result of applying the ALMA method is an assessment of the architecture candidates in terms of maintainability and reusability. The data is presented in a table that, for each candidate, indicates which application components need to be modified or added, along with the affected requirements. The method does not provide mechanisms for handling trade-offs between maintainability and other QAs.

Systematic Quantitative Analysis of Scenarios' Heuristics [34] is a method that focuses on the quantitative evaluation of both architectural alternatives and stakeholder-defined objectives to reduce uncertainty. Its application includes identifying stakeholders and objectives, making objectives quantifiable, analysing and aggregating scenarios, improving scenarios, and selecting the optimal options. To enhance precision, the analysis process is conducted in multiple iterations.

In [34], QAs such as usability and performance are used as evaluation criteria. For each QA, a set of objectives is formulated in the form of refined system requirements. Stakeholders assign quantitative values to each objective across five levels (ranging from Minimal to Excellent). Each level is associated with a distinct colour.

Subsequently, SA candidates are analysed in the context of the defined objectives, taking into account detailed technical aspects. Where metric-based evaluation is not feasible, it is suggested to use expert judgment with a Low / Medium / High scale.

The output of the method is presented as a table, where the columns represent SA candidates and the rows represent objectives, grouped by QAs. Each cell contains the evaluation of an SA candidate for the given objective and is color-coded according to the corresponding level.

A key limitation of the method is its complexity in application. When evaluation is tailored to a non-typical task, a considerable number of objectives can lead to high preparation and experimentation effort, which makes the method difficult to apply in practice.

Performance Assessment of Software Architecture [35] is a method focused on achieving performance objectives. Its goal is either to improve the performance of an existing system or to select an SA that meets the performance requirements of a new system.

The PASA application flow consists of 10 steps. The first step is an overview of the reasons for conducting an architectural assessment and the assessment process itself, presented to the entire team, including developers and stakeholders. Then, documentation is prepared for the architecture of the existing or planned application (if not already available), and an overview of the architecture is presented to the team. In the following steps, critical use cases affecting responsiveness and scalability, as well as key performance scenarios, are identified. These scenarios are documented and typically represented using UML.

Next, the team identifies the performance objectives. Based on the presented data, the participants conduct a more detailed discussion of the architecture and its specific features that influence key performance scenarios. If the analysis reveals performance issues, architectural alternatives that meet the performance requirements are proposed. The results and recommendations are then presented to the entire team, and an economic analysis of the costs and benefits of the proposed solution is conducted.

Performance improvement methods include practices such as identifying deviations from the current architectural style, detecting performance-related antipatterns [36], proposing alternative interactions between components, etc. Using software performance engineering techniques [37], evaluators perform performance modeling, including best- and worst-case analyses to address uncertainty.

The assessment process usually takes 1–2 weeks. The outputs of the method include:

– Documentation of the current architecture and main processes (if not already available).

– A set of architectural alternatives that meet performance requirements, with recommendations for selecting among them.

– A rough comparison of the costs of analysis and subsequent improvements versus the potential costs of additional hardware and development effort that would have been required if the problems had not been detected in time.

The main drawbacks of the method are the lack of trade-off analysis with other QAs and the focus only on critical use cases. If a large number of scenarios need to be evaluated, the analysis may become time-consuming.

Information Technology for Decision-making Support regarding CQRS with ES Architectural Variations [4] is a method designed to evaluate evolutionary architectural variations within a single SA style. Its primary focus is on comparing different variants of CQRS with ES architecture. Given the nature of these variations, the central QAs under consideration are maintainability and performance. Other QAs are largely equivalent across the majority of variations and are therefore not the focus of differentiation, at least for CQRS with ES architecture. DSAV-CQRSES takes into account detailed technical aspects, but unlike SQUASH, it introduces a classification of typical use cases, which significantly reduces the effort required to conduct the experiments.

The method supports three distinct evaluation objectives: assessing implementation/modification complexity, supporting architectural decision-making by selecting the most suitable variation, and evaluating migration complexity between variations.

During the preparation phase for selecting an architectural variation, requirements are gathered, a preliminary estimation of the number of use cases of various types is calculated, and the expected proportion of addition new / modification old use cases after the release of the minimum viable product phase is assessed.

DOI 10.15588/1607-3274-2026-1-10

The support architectural decision-making by selecting the most suitable SA variation algorithm looks as follows:

1. Identification of QAs.
2. Informal description of the considered architectural variations.
3. Estimation of implementation/modification complexity for each variation:
    3.1. Formalization of the architectural variations:
        3.1.1. Classification of use case types;
        3.1.2. Identification of processes corresponding to each use case class;
        3.1.3. Formal process modelling using the activity model.
    3.2. Evaluation of the processes.
    3.3. Evaluation of the architectural variation.
4. Identification of metrics requiring a Representative Test Project (RTP):
    4.1. Definition of the RTP.
    4.2. Implementation of the RTP.
    4.3. Metric collection based on the RTP.
    4.4. Aggregation and preparation of results for applying a Multi-Criteria Decision Analysis (MCDA) method (automated).
5. Application of the MCDA method, such as AHP (automated).

The method can be used to achieve each objective independently. Its primary advantage is the minimal reliance on expert judgment. While human input is needed during the process modelling phase, high-level architecture expertise is not required to construct use case and process diagrams. Due to the automation of calculations, stakeholders do not need to predefine QA priorities; instead, they can visually analyse the data to identify the most suitable variation under different conditions and select an appropriate trade-off.

The output includes the visualized applicability metrics of each architectural variation under different QA values (in percentages), as well as documentation describing the processes.

It is important to note that the time required to apply the method increases with the number of use case classes. The method is specifically designed for evaluating CQRS with ES architecture, which typically involves two to three use case classes, depending on stakeholder needs. For applications with a large number of unique processes, the method may become less applicable due to the associated effort.

4 EXPERIMENTS

The experiment involves comparing a set of scenario-based methods for SA design-time evaluation: SAAM, ATAM, CBAM, ALMA, SQUASH, PASA and DSAV-CQRSES. It consists of three stages.

1) The methods are qualitatively compared using the modified comparison framework.

2) The qualitative evaluation is translated into a quantitative form, and criterion weights are applied. As a result, each method is represented by a multidimensional vector corresponding to its evaluation on each criterion.

3) A reference method is then introduced, reflecting the characteristics desired by the DBB Software team for evaluating CQRS with ES architectural variations. The effectiveness of each method is calculated using formula (1), after which the methods are ranked, and the top candidates are selected, using formulas (2–4).

5 RESULTS

The results of each stage are summarized in Tables 2–3.

Table 2 – Results of the qualitative comparison of SA evaluation methods

Criteria	SAAM	ATAM	CBAM	ALMA	SQUASH	PASA	DSAV-CQRSES
How closely do the method's definitions of SA and SA variation align with those considered in this paper?	SA definition is left to users.	SA definition is left to users.	SA definition is left to users.	Not provided.	Not provided.	Not provided.	Specified definition for SA variations.
Is the selection between structurally similar SA alternatives addressed by the method's objectives?	No.	No.	No.	No.	No.	Different solutions applied to the same architecture are considered to address performance issues.	Yes. Method is developed for SA variations comparison.
Are variation-specific QAs (e.g., complexity and performance) covered by the method?	Just modifiability (complexity).	Yes.	Yes.	Just modifiability (complexity).	Yes.	Just performance.	Yes.

DOI 10.15588/1607-3274-2026-1-10

Continuation of Table 2

Is the method applicable at the design stage of system development?	Yes.	Yes.	Yes.	Yes.	Yes.	Yes.	Yes.
To what extent is the resulting technical documentation ready for direct use during implementation?	Informal description of candidates, list of scenarios, experimental docs and benchmarks, if available.	Informal description of candidates, list of scenarios and identified risks.	Informal description of candidates, list of scenarios and identified risks.	Informal description of candidates.	Informal description of candidates, experimental docs, if available.	Informal description of candidates and critical processes.	Informal description of candidates, formal specification of their typical processes.
Does the method facilitate trade-off analysis?	No.	Yes, manual.	Yes, manual.	No.	Preparations for trade-off. Left to users.	Between performance and costs, manual.	Between performance and costs, automatic.
Does the method identify a recommended architecture, or does it only provide additional information about the alternatives?	Evaluates the modifiability parameters of the candidates. Left to users.	Identifies risks across multiple QAs. Left to users.	Identifies risks across multiple QAs. Provides a cost/benefit assessment. Left to users.	Evaluates the modifiability parameters of the candidates. Left to users.	Evaluates multiple parameters of the candidates. Left to users.	The selection of recommended options is performed manually by the evaluator.	Yes, based on performance and complexity parameters.
Is the method validated or considered applicable in domains similar to the target system?	Various domains, including medical.	Various domains, including medical.	Various domains, including medical.	Various domains, including medical.	Verified on medical domain.	Various domains.	Domains compatible with CQRS with ES, including medical.
Does the method support the selection of the optimal solution?	Yes.	Yes.	Yes.	Yes.	Yes.	Yes.	Yes.
Does the method provide an approximate estimate of the effort required for system implementation?	Yes, based on benchmarks.	No.	Not explicitly, based on costs evaluation.	Not explicitly, based on components to be changed.	Not explicitly, based on experiments.	No.	Yes, based on experiments.
Does the method require involvement of multiple stakeholders?	Yes.	Yes.	Yes.	Yes.	No, only for discussion and clarification of requirements.	No, only for discussion of requirements and scenarios.	No, only for discussion of requirements.
How many actions or inputs are expected from stakeholders?	Participate in expert evaluation of scenarios.	Participate in expert evaluation of scenarios.	Participate in expert evaluation of scenarios.	Participate in expert evaluation of scenarios.	Formulate refined requirements.	Formulate requirements and critical scenarios.	Formulate requirements.
How much support is provided by the method to perform various activities?	Not explicitly addressed.	Comprehensively covered.	Comprehensively covered.	Embedded in method description.	Embedded in method description.	Embedded in method description.	Embedded in method description.
Does the method address non-technical issues (e.g., social and organizational factors)?	Such issues briefly mentioned.	Sufficiently provided.	Sufficiently provided.	Not explicitly addressed.	Not explicitly addressed.	Not explicitly addressed.	Not addressed. Minimized, due to reduced communication with stakeholders.
How long does the application of the method take?	Apart from initial & post preparation, 3 days.	Apart from initial & post preparation, 2 days.	Apart from initial & post preparation, 2 days (ATAM) and 3 days.	Not specified.	The method requires a considerable amount of time to collect all the required data.	1–2 weeks.	Apart from initial & post preparation, 2 days.

DOI 10.15588/1607-3274-2026-1-10

Continuation of Table 2

What is the typical team size required to apply the method?	4-person evaluation team & stakeholders.	3-person evaluation team & stakeholders.	3-person evaluation team & stakeholders.	Not specified.	Not specified.	Not specified.	2-person evaluation team & stakeholders.
How demanding are the architectural skills required from the users of the method?	High, for the expert judgment and experiments.	High, for the expert judgment.	High, for the expert judgment.	High, for the expert judgment.	High, for the expert judgment and experiments.	High, for the performance modeling and generating alternatives.	Moderate, for experiments and use case documentation.
What are the activities to be performed and in what order to achieve the goals?	6 activities.	9 activities in 2 phases.	9 activities in 2 phases (ATAM) & 4 activities performed over several iterations.	5 activities carried out sequentially.	4 activities & 3 activities performed over several iterations.	10 activities.	4 activities & 6 activities performed over several iterations.
Does the method provide a formal and detailed description of the architectural candidates?	High-level description using ADL; focuses on identifying maintainability-related risks.	High-level description using ADL; emphasizes risks related to multiple QAs.	High-level description using ADL; focuses on economic trade-offs between QAs.	Supports formal description for complex architectures; emphasizes maintainability-related risks.	High-level description using ADL; experiments partially cover system behavior and processes.	Provides detailed description and performance modeling of critical processes and components.	Offers a formal description of SA processes, using ADL and activity models.
To what extent is the method subject to uncertainty?	Highly, due to its reliance on expert judgment.	Highly, due to its reliance on expert judgment.	Highly, due to its reliance on expert judgment.	Highly, due to its reliance on expert judgment.	Moderate, due to partial reliance on expert judgment and remaining variability in the experiments.	Moderate, due to variability in the modeling.	Moderate, due to possible inaccuracies in the source data (use cases) and remaining variability in the experiments.
Are there any techniques applied to mitigate uncertainty?	Implicit.	Implicit, extra iterations of architecture analysis.	Yes, In V2 extra iterations and value boundaries are added.	Implicit.	Implicit, within experiments.	Implicit, within performance modeling.	Implicit, within experiments.
How much of the evaluation is grounded in quantitative metrics and experimental data versus expert assessment?	Primarily expert judgement.	Primarily expert judgement.	Suggested to use statistics or real prices for costs if possible.	Primarily expert judgement. Left to users.	Primarily metrics and experiments.	Primarily metrics and experiments.	Primarily metrics and experiments.
Does the method offer tools to facilitate or partially automate its use?	Partially available.	Not available.	Not available.	Not available.	Not available.	Not available.	Partially automated.
What is the level of maturity (inception, development, refinement or dormant)?	Dormant.	Dormant.	Dormant.	Dormant.	Refinement.	Refinement.	Development.
Has the method been validated?	Yes.	Yes.	Yes.	Yes.	Found a couple articles with its application.	Found a couple articles with its application.	For a couple of projects.
Has the method been applied specifically to architectural variations?	No.	No.	No.	No.	No.	No.	Yes.

A typical project, derived from the analysis of multiple DBB Software projects employing the CQRS with ES architecture, has been selected as the reference case. At this stage, a specific variation needs to be chosen. The software system under development belongs to the medical domain and requires a high level of maintainability and good performance.

Table 3 presents the answers from Table 2 converted into a more formalized form through expert evaluation, describes the reference case, and calculations of the effectiveness of the methods using formula (1).

DOI 10.15588/1607-3274-2026-1-10

Table 3 – Results of the quantitative comparison of SA evaluation methods

Criteria	Weight	SAAM	ATAM	CBAM	ALMA	SQUASH	PASA	DSAV-CQRSES	Ref	Max
How closely do the method's definitions of SA and SA variation align with those considered in this paper?	3	L	L	L	L	L	L	M	H	L
Is the selection between structurally similar SA alternatives addressed by the method's objectives?	5	N	N	N	N	Y/N	Y	Y	Y	N
Are variation-specific QAs (e.g., complexity and performance) covered by the method?	10	Y/N	Y	Y	Y/N	Y	Y/N	Y	Y	N
Is the method applicable at the design stage of system development?	10	Y	Y	Y	Y	Y	Y	Y	Y	N
To what extent is the resulting technical documentation ready for direct use during implementation?	1	Y/N	Y/N	Y/N	Y/N	N	Y/N	Y/N	Y/N	N
Does the method facilitate trade-off analysis?	7	N	Y	Y	N	Y/N	Y/N	Y/N	Y	N
Does the method identify a recommended architecture, or does it only provide additional information about the alternatives?	2	N	N	N	N	N	N	Y	Y	N
Is the method validated or considered applicable in domains similar to the target system?	8	Y	Y	Y	Y	Y	Y/N	Y	Y	N
Does the method support the selection of the optimal solution?	10	Y	Y	Y	Y	Y	Y	Y	Y	N
Does the method provide an approximate estimate of the effort required for system implementation?	4	Y/N	N	Y/N	Y/N	Y/N	Y/N	Y	Y	N
Does the method require involvement of multiple stakeholders?	4	Y	Y	Y	Y	Y/N	Y/N	N	N	Y
How many actions or inputs are expected from stakeholders?	7	H	H	H	H	M	M	L	L	H
How much support is provided by the method to perform various activities?	5	L	H	H	M	M	M	M	H	L
Does the method address non-technical issues (e.g., social and organizational factors)?	2	Y/N	Y	Y	Y/N	Y/N	Y/N	Y/N	Y/N	N
How long does the application of the method take?	5	M	M	H	M	H	H	M	M	L
What is the typical team size required to apply the method?	5	H	M	M	M	M	H	L	L	H
How demanding are the architectural skills required from the users of the method?	8	H	H	H	H	H	H	M	M	L
Does the method provide a formal and detailed description of the architectural candidates?	4	N	N	N	Y/N	N	Y/N	Y	Y	N
To what extent is the method subject to uncertainty?	4	H	H	H	H	M	M	M	L	H
Are there any techniques applied to mitigate uncertainty?	8	Y/N	Y/N	Y	Y/N	Y/N	Y/N	Y/N	Y	N
How much of the evaluation is grounded in quantitative metrics and experimental data versus expert assessment?	7	L	L	M	L	H	H	H	H	L
Does the method offer tools to facilitate or partially automate its use?	2	Y/N	N	N	N	N	N	Y	Y	N
What is the level of maturity (inception, development, refinement or dormant)?	6	H	H	H	H	M	M	L	H	L
Has the method been validated?	6	Y	Y	Y	Y	Y/N	Y/N	Y/N	Y	N
Has the method been applied specifically to architectural variations?	6	N	N	N	N	N	N	Y	Y	N
Euclidean distances		16.37	14.66	13.38	15.23	11.96	12.57	7.55		22.54
Effectiveness		0.27	0.35	0.41	0.32	0.47	0.44	0.67		

DOI 10.15588/1607-3274-2026-1-10

For simplicity in experimentation, the following numeric values are used in this study: L = 1, M = 2, H = 3; N = 1, Y/N = 2, Y = 3.

The weights for each parameter were determined based on the expert judgment of the DBB Software development team conducting the evaluation. They reflected the extent to which each factor influences the team's convenience in applying the method.

For example, the distance between the reference method and DSAV-CQRSES is calculated as follows:

Reference vector:

$$ref = (H,Y,Y,Y,Y/N,Y,Y,Y,Y,Y,N,L,H,Y/N,M,L,M,Y,L,Y,H,Y,H,Y,Y) =$$
$$= (3,3,3,3,2,3,3,3,3,3,1,1,3,2,2,1,2,3,1,3,3,3,3,3,3).$$

Maximum vector:

$$\max = (L,N,N,N,N,N,N,N,N,N,Y,H,L,N,L,H,L,N,H,N,L,N,L,N,N) =$$
$$= (1,1,1,1,1,1,1,1,1,1,3,3,1,1,1,3,1,1,3,1,1,1,1,1,1).$$

Weights vector:

$$w = (3,5,10,10,1,7,2,8,10,4,4,4,7,5,2,5,5,8,4,4,8,7,2,6,6,6).$$

DSAV-CQRSES vector:

$$p_{DSAV-CQRSES} = (M,Y,Y,Y,Y/N,Y/N,Y,Y,Y,Y,N,L,M,Y/N,M,L,M,Y,M,Y/N,H,Y,L,Y/N,Y) =$$
$$= (2,3,3,3,3,2,2,3,3,3,3,3,1,1,2,2,2,1,2,3,2,2,3,3,3,1,2,3).$$

Maximum distance to the Reference:

$$D(\max, ref) = \sqrt{\sum_i w_i \cdot (\max_i - ref_i)^2} = 22.54, i \in [1,25]$$

DSAV-CQRSES to the Reference distance:

$$D(p_{DSAV-CQRSES}, ref) = \sqrt{\sum_i w_i \cdot (p_{DSAV-CQRSES} - ref_i)^2} = 7.55, i \in [1,25].$$

DSAV-CQRSES effectiveness:

$$Effectiveness(p_{DSAV-CQRSES}) = 1 - \frac{D(p_{DSAV-CQRSES}, ref)}{D(\max, ref)} = 1 - \frac{7.55}{22.54} \approx 0.67.$$

Obtain an ordered vector E_{asc} using formula (2):

$$E_{asc} = \{\chi(sa_{SAAM}), \chi(sa_{ALMA}), \chi(sa_{ATAM}), \chi(sa_{CBAM}), \chi(sa_{PASA}), \chi(sa_{SQUASH}), \chi(sa_{DSAV-CQRSES})\}.$$

Determine the result set using formulas (3–5). The maximum allowable deviation (*th*) in formula (4) let us set to 0.05. With this threshold, the resulting set contains a single element:

$$SA_{opt} = \{sa_{DSAV-CQRSES}\}.$$

6 DISCUSSION

The existing approaches reviewed in the literature do not consider evaluation in the context of SA variations. Due to the specific nature and similarity of SA variations, most traditional SA evaluation methods are not well-suited for their comparison. The modified version of the framework for classifying and comparing SA evaluation methods, adapted to support the comparison of methods for evaluating architectural variations, was applied. The method is relatively easy to apply, and its results are supported by factual evidence about the candidates.

The evaluation results demonstrate that traditional scenario-based methods such as SAAM (0.27) and ALMA (0.32) score relatively low. This can be attributed to their reliance on expert judgment, susceptibility to uncertainty, the extensive involvement of non-technical stakeholders, and the lack of detail in the descriptions of architectural candidates. Additionally, these methods do not explicitly support trade-off analysis across multiple QAs.

ATAM, scoring 0.35, improves upon SAAM and ALMA by introducing trade-off analysis, yet still relies heavily on qualitative assessment. Its extension, CBAM (0.41), shows better handling of uncertainty.

The methods that scored higher – PASA (0.44) and SQUASH (0.47). They shift the focus toward quantitative metrics and experimental data. These characteristics make them more applicable for comparing SA variations in practice. However, they require a long execution time; additionally, PASA demands a large team for application, while SQUASH lacks sufficient architectural detail in candidate descriptions.

DSAV-CQRSES (0.67) outperformed other SA evaluation methods in addressing the task of comparing CQRS with ES architectural variations within the DBB Software team. This finding is supported by real-world application, which can be explained by the fact that DSAV-CQRSES was specifically developed for evaluating variations of the CQRS with ES architecture. It includes mechanisms for formal architectural modeling, trade-off analysis, and partial automation. A significant factor that limited its score was the low maturity and validation of the method, as it has only recently been introduced. Without this factor, the score would have increased to 0.76.

DOI 10.15588/1607-3274-2026-1-10

The results of the experiment can be practically applied by DBB Software team when selecting an evaluation method for real-world projects. Moreover, the qualitative and formalized assessment of the methods can be reused by other companies and teams to calculate the applicability of the considered SA evaluation methods for their specific needs.

To enhance the universality of the proposed methodology in practical applications, it is advisable to expand the set of case studies using different evaluation methods, implement multiple architectural variations for each of them, analyze the corresponding metrics (such as development and maintenance speed, performance, etc.), as well as the issues encountered by developers.

CONCLUSIONS

The scientific novelty. The article proposes a methodology for comparing SA evaluation methods. The framework for classifying and comparing SA evaluation methods has been modified for application within the context of CQRS with ES architectural variations. The proposed adapted approach allows assessing the applicability of each method based on a categorized set of variation-related questions. Each method's evaluation is represented as a multidimensional vector. The approach enables a two-stage qualitative and quantitative assessment of method effectiveness. The quantitative stage computes an applicability score using the Euclidean distance from a reference (etalon) vector *ref* to the vector of each method, considering attribute weights *w*, provided by decision-making team.

The adapted framework was applied to determine which SA evaluation method is best suited for assessing CQRS with ES variations by the DBB Software team. The methods compared were SAAM (0.27), ATAM (0.35), CBAM (0.41), ALMA (0.32), SQUASH (0.47), PASA (0.44) and DSAV-CQRSES (0.67). Based on the results of the quantitative evaluation the optimal subset $SA_{opt} = \{sa_{DSAV-CQRSES}\}$, that aligns with practical experience, as DSAV-CQRSES was specifically developed to evaluate variations of the CQRS with ES architecture.

The practical significance. The adapted framework provides a reproducible, measurable basis for selecting an SA evaluation method under real project constraints and can be instantiated for other teams and contexts by re-specifying *w* and *ref*. This, in turn, supports the selection of more appropriate architectural solutions, with positive impact on development and maintenance..

ACKNOWLEDGEMENTS

We sincerely appreciate DBB Software company [9] for providing their proprietary platform, which served as the foundation for our experiment. This platform offered essential capabilities for our research, ensuring the accuracy and reliability of our experimental results.

We also want to express our deep appreciation to Volodymyr Khandetskyi, Head of Electronic Computing Machinery department, for his valuable comments and suggestions.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: Dmytro Hruzin: the methodology for comparing SA evaluation methods; Oleksandr Lytvynov: reviewed and discussed the proposed method and provided recommendations for improving the methodology and the manuscript's presentation.

Data availability: The manuscript has no associated data.

Software availability: The manuscript has no associated software.

Use of artificial intelligence tools: Artificial intelligence tools were used to search for and correct grammatical and punctuation mistakes in the manuscript and to improve the English translation, in parallel with other tools such as Reverso Context.

REFERENCES

1. Chrissis M. B., Konrad M., Shrum S. CMMI for Development: Guidelines for Process Integration and Product Improvement (SEI Series in Software Engineering) 3rd Edition. Boston, Massachusetts, Addison-Wesley Professional, 2011, 688 p.
2. Lytvynov O. A., Hruzin D. L.Critical causal events in systems based on CQRS with Event Sourcing architecture, *Radio Electronics Computer Science Control*, 2024, Issue 3, pp. 119–143. DOI: 10.15588/1607-3274-2024-3-11.
3. Sobhy D., Bahsoon R., Minku L. et al. Evaluation of Software Architectures under Uncertainty: A Systematic Literature Review, *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 2021, Vol. 30, Issue 4, pp. 1–50. DOI: 10.1145/3464305.
4. Lytvynov O. A., Hruzin D. L. Decision-making on Command Query Responsibility Segregation with Event Sourcing architectural variations, *Technology audit and production reserves*, 2025, Vol. 4, Issue 2(84), pp. 37–59. DOI: 10.15587/2706-5448.2025.337168.
5. Babar M. A., Zhu L., Jeffery R. A framework for classifying and comparing software architecture evaluation methods, *Software Engineering, Australian Conference, Melbourne, Victoria, 13–16 April 2004, proceedings*. Melbourne, Victoria, 2004, P. 309. DOI: 10.1109/ASWEC.2004.1290484.
6. Babar M. A., Gorton I. Comparison of Scenario-Based Software Architecture Evaluation Methods, *Software Engineering, 11th Asia-Pacific Conference, Busan, 30 November – 3 December 2004, proceedings*. Busan, 2004, pp. 600–607. DOI: 10.1109/APSEC.2004.38.
7. Abrahão S., Insfran E. Evaluating Software Architecture Evaluation Methods: An Internal Replication, *Conference on Evaluation and Assessment in Software Engineering (EASE 2017) : 21th International Conference, Karlskrona, 15–16 June 2017 : proceedings*. Karlskrona, 2017, pp. 144–153. DOI: 10.1145/3084226.3084253.
8. González H. J., Pelozo I., Gonzales A. et al. Validating a model-driven software architecture evaluation and improvement method: A family of experiments, Information and Software Technology, 2015, Issue 57, pp. 405–429. DOI: 10.1016/j.infsof.2014.05.018.
9. DBB Software's official company site [Electronic resource]. Access mode: https://dbbsoftware.com/.

DOI 10.15588/1607-3274-2026-1-10

10. Fatima I., Lago P. A Review of Software Architecture Evaluation Methods for Sustainability Assessment, *Software Architecture Companion: 20th International Conference (ICSA-C), L'Aquila, 13–17 March 2023 : proceedings*. Los Alamitos, CA: IEEE Computer Society, 2006, pp. 191–194 DOI: 10.1109/ICSA-C57050.2023.00050.
11. Sahlabadi M., Muniyandi R. C., Shukur Z. et al. Lightweight Software Architecture Evaluation for Industry: A Comprehensive Review, *Sensors*, 2022, Vol. 22, Issue 3. DOI: 10.3390/s22031252.
12. Kazman R., Klein M., Clements P. ATAM: Method for Architecture Evaluation : technical report : CMU/SEI-2000-TR-004, ESC-TR-2000-004, Product Line Systems. Pittsburgh, 2000, 71 p.
13. Kazman R., Bass L., Abowd G. et al. SAAM: a method for analyzing the properties of software architectures, *Software Engineering : 16th international conference, Sorrento, 16–21 May 1994 : proceedings.* Washington, DC, IEEE Computer Society Press, 1994, pp. 81–90. DOI: 10.1109/ICSE.1994.296768.
14. Babar M. A., Kitchenham B. Assessment of a Framework for Comparing Software Architecture Analysis Methods, *Evaluation and Assessment in Software Engineering (EASE) : 11th International Conference, UK, 2–3 April 2007 : proceedings*. Swindon, BCS Learning & Development Ltd., 2007, pp. 12–20. DOI: 10.14236/ewic/EASE2007.2.
15. Jayaratna N. Understanding and Evaluating Methodologies: NIMSAD, a Systematic Framewor. New York, McGraw-Hill, Inc., 1994, 259 p.
16. Gonzalez-Huerta J., Insfran E., Abrahão S. Models in Software Architecture Derivation and Evaluation: Challenges and Opportunities, *Model-Driven Engineering and Software Development : the 2nd International Conference, Lisbon, 7–9 January 2014 : proceedings*. Setubal, SCITEPRESS, 2014. – P. 12–31. DOI: 10.1007/978-3-319-25156-1_2.
17. Lloyd P. T. L., Galambos G. M. Technical reference architectures, *IBM Systems Journal*, 1999, Vol. 38, Issue 1, pp. 51–75. DOI: 10.1147/sj.381.005.
18. Bass L., Clements P., Kazman R. Software Architecture in Practice, Second Edition. United States, Addison-Wesley Longman Publishing Co., Inc., 2003, 528 p.
19. Lakhdissi M., Bounabat B. A New Content Framework and Metamodel for Enterprise Architecture and is Strategic Planning, *International Journal of Computer Science Issues*, 2014, Vol. 9, Issue 2. DOI: 10.1201/b16417-9.
20. Clements P., Bachmann F., Bass L. Documenting Software Architectures: Views and Beyond, Second Edition. Boston, Addison-Wesley, 2010, 592 p. ISBN: 978-0-321-55268-6.
21. Solms F. What is Software Architecture, *South African Institute for Computer Scientists and Information Technologists Conference (SAICSIT '12), Pretoria, 1–3 October 2012 : proceedings*. New York, Association for Computing Machinery, 2012, pp. 363–373. DOI: 10.1145/2389836.2389879.
22. Franchitti J. C. Enterprise Architecture Frameworks (EAFs) & Pattern Driven EAFs [Electronic resource]. Access mode: https://cs.nyu.edu/~jcf/classes/g22.3033-007/slides/session2/g22_3033_011_c23.pdf.
23. Galster M., Avgeriou P., Weyns D. et al.Variability in software architecture: Current practice and challenges, *ACM SIGSOFT Software Engineering Notes*, 2011, Vol. 36, Issue 5, pp. 30–32. DOI: 10.1145/2020976.2020978.
24. Ford N., Parsons R., Sadalage P. et al. Building Evolutionary Architectures: Automated Software Governance 2nd Edition. US, O'Reilly Media, 2022, 262 p. ISBN : 978-1492097549.
25. Dakhli S.B.D. Architectural Deviations and Inconsistencies Management: A Framework Based on Information Systems Urbanization, *Procedia Computer Science*, 2021, Vol. 181, pp. 1122–1130. DOI: 10.1016/j.procs.2021.01.309.
26. Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE). Quality model overview and usage : ISO/IEC 25002:2024. [Effective from 2024-03]. – ISO, 2024, 17 p.
27. Rumpe B. Modeling with UML: Language, Concepts, Methods. Switzerland, Springer Cham, 2016, 281 p. ISBN: 978-3-319-33933-7.
28. Chattopadhyay A., Wang Z., Martin G. E. Architecture Description Languages, Handbook of Computer Architecture. New York, Springer, 2024, Processor Design and Programming Flows, pp. 807–839. DOI: 10.1007/978-981-97-9314-3_18.
29. Lung C. H., Bot S., Kalaichelvan K. et al. An approach to software architecture analysis for evolution and reusability, *1997 conference of the Centre for Advanced Studies on Collaborative Research, Toronto, 10–13 November 1997 : proceedings.* US, IBM Press, 1997, P. 15. DOI: 10.1145/782010.782025.
30. Lassing N., Rijsenbrij D., Vliet H. On Software Architecture Analysis of Flexibility, Complexity of Changes: Size isn't Everything, *2nd Nordic Software Architecture Workshop, Ronneby, 12–13 August 1999, proceedings.* Ronneby, 1999, pp. 1103–1581.
31. Kazman R., Asundi J., Klein M.Quantifying the costs and benefits of architectural decisions, Software Engineering : the 23rd International Conference, Toronto Ontario, 12–19 May 2001, proceedings. NW Washington, *IEEE Computer Society*, 2001, pp. 297–306. DOI: 10.1109/ICSE.2001.919103.
32. Moore M., Kaman R., Klein M. Quantifying the value of architecture design decisions: Lessons from the field, *Software Engineering : 25th International Conference (ICSE03), Portland, 3–10 May 2003 : proceedings*. NW Washington, IEEE Computer Society, 2003, pp. 557–562. DOI: 10.1109/ICSE.2003.1201237.
33. Bergtsson P.O., Lassing N., Bosch J. Architecture-Level Modifiability Analysis (ALMA), *Journal of Systems and Software*, 2004, Vol. 69, pp. 129–147. DOI: 10.1016/S0164-1212(03)00080-3.
34. Ionita M. T., America P., Hammer D. K. et al. A Scenario-Driven Approach for Value, Risk, and Cost Analysis in System Architecting for Innovation, *Software Architecture : 4th Working IEEE / IFIP Conference (WICSA 2004), Oslo, 12–15 June 2004, proceedings*, P. 277. DOI: 10.1109/WICSA.2004.1310709.
35. Williams L. G., Smith C. U. PASA^{SM}: An Architectural Approach to Fixing Software Performance Problems, *28th International Computer Measurement Group Conference, Reno, 8–13 December 2002 : proceedings*, pp. 307–320.
36. Smith C. U., Williams L. G. Software performance antipatterns, *Software and performance : the 2nd international workshop (WOSP00), Ottawa, 1 September 2000 : proceedings.* New York, Association for Computing Machinery, pp. 127–136. DOI: 10.1145/350391.350420
37. Smith C. U., Williams L. G. Performance solutions: a practical guide to creating responsive, scalable software, Redwood City, CA, Addison Wesley Longman Publishing Co., Inc., 2002, 544 p. ISBN: 978-0-201-72229-1.

Received 10.09.2025.
Accepted 08.01.2026.
Published 27.03.2026.

DOI 10.15588/1607-3274-2026-1-10

УДК 614.2+574/578+004.38

ПОРІВНЯННЯ МЕТОДІВ ОЦІНЮВАННЯ АРХІТЕКТУРИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ В КОНТЕКСТІ АРХІТЕКТУРНИХ ВАРІАЦІЙ CQRS З EVENT SOURCING

Грузін Д. Л. – аспірант кафедри електронних обчислювальних машин Дніпровського національного університету імені Олеся Гончара, Дніпро, Україна. ROR: https://ror.org/00qk1f078. ORCID: 0009-0004-8534-2559.

Литвинов О. А. – канд. техн. наук, доцент кафедри електронних обчислювальних машин Дніпровського національного університету імені Олеся Гончара, Дніпро, Україна. ROR: https://ror.org/00qk1f078. ORCID: 0000-0001-7660-1353.

АНОТАЦІЯ

Актуальність. Дослідження проводиться в контексті створення та обґрунтування методології оцінювання архітектури програмного забезпечення в рамках варіацій Command Query Responsibility Segregation (CQRS) з Event Sourcing (ES) архітектури.

Метою дослідження є оцінка та порівняння різних методів оцінювання архітектури програмного забезпечення з метою підтримки вибору оптимальної архітектурної варіації CQRS із ES для реальних проєктів.

Метод. Для підвищення об'єктивності процесу ухвалення архітектурних рішень застосовуються різні методи оцінювання архітектури програмного забезпечення. Проте ці методи не є універсальними, оскільки відрізняються глибиною аналізу, спрямованістю та обсягом необхідних ресурсів. Завдання, що розглядається у цьому дослідженні, полягає у виборі між архітектурними варіаціями CQRS з ES, які часто характеризуються високим ступенем структурної подібності та, відповідно, є складними для розрізнення за допомогою загальнопризначених методів оцінювання. Порівняння архітектурних варіацій потребує поглибленого аналізу, однак для більшості методів його проведення на практиці ускладнюється через часові та ресурсні обмеження. Запропонований підхід орієнтований на визначення найбільш доцільного методу оцінювання архітектури ПЗ для підтримки процесу прийняття рішень щодо вибору між варіаціями CQRS з ES. Він ґрунтується на існуючому фреймворку класифікації та порівняння методів оцінювання архітектури. Окрім якісного аналізу, підхід включає кількісну оцінку ступеня придатності до конкретного проєкту, що забезпечує більш обґрунтоване та раціональне прийняття архітектурних рішень.

Результати. Запропонований підхід було застосовано для порівняння кількох методів оцінювання архітектури ПЗ, зокрема Information Technology for Decision-making Support regarding CQRS with ES Architectural Variations (DSAV-CQRSES) – методу, спеціально розробленого для оцінки варіацій архітектури CQRS з ES.

Висновки. Існуюча система порівняння архітектур програмного забезпечення не може бути безпосередньо застосована до архітектурних варіацій (відхилень архітектури, значущих для замовника). Запропонована модифікація фреймворку орієнтовані насамперед на оцінювання варіацій архітектури CQRS з ES.

КЛЮЧОВІ СЛОВА: архітектура програмного забезпечення, порівняння методів оцінювання, CQRS з Event Sourcing, архітектурні варіації.

ЛІТЕРАТУРА

1. Chrissis M. B. CMMI for Development: Guidelines for Process Integration and Product Improvement (SEI Series in Software Engineering) 3rd Edition / M. B. Chrissis, M. Konrad, S. Shrum. – Boston, Massachusetts : Addison-Wesley Professional, 2011. – 688 p.
2. Lytvynov O. A. Critical causal events in systems based on CQRS with Event Sourcing architecture / O. A. Lytvynov, D. L. Hruzin // Radio Electronics Computer Science Control. – 2024. – Issue 3. – P. 119–143. DOI: 10.15588/1607-3274-2024-3-11.
3. Evaluation of Software Architectures under Uncertainty: A Systematic Literature Review / [D. Sobhy, R. Bahsoon, L. Minku et al.]. // ACM Transactions on Software Engineering and Methodology (TOSEM). – 2021. – Vol. 30, Issue 4. – P. 1–50. DOI: 10.1145/3464305.
4. Lytvynov O. A. Decision-making on Command Query Responsibility Segregation with Event Sourcing architectural variations / O. A. Lytvynov, D. L. Hruzin // Technology audit and production reserves. – 2025. – Vol. 4, Issue 2(84). – P. 37–59. DOI: 10.15587/2706-5448.2025.337168.
5. Babar M. A. A framework for classifying and comparing software architecture evaluation methods / M. A. Babar, L. Zhu, R. Jeffery // Software Engineering : Australian Conference, Melbourne, Victoria, 13–16 April 2004 : proceedings. – Melbourne, Victoria: 2004. – P. 309. DOI: 10.1109/ASWEC.2004.1290484.
6. Babar M.A. Comparison of Scenario-Based Software Architecture Evaluation Methods / M. A. Babar, I. Gorton // Software Engineering : 11th Asia-Pacific Conference, Busan, 30 November – 3 December 2004 : proceedings. – Busan: 2004. – P. 600–607. DOI: 10.1109/APSEC.2004.38.
7. Abrahão S. Evaluating Software Architecture Evaluation Methods: An Internal Replication / S. Abrahão, E. Insfran // Conference on Evaluation and Assessment in Software Engineering (EASE 2017) : 21th International Conference, Karlskrona, 15–16 June 2017 : proceedings. – Karlskrona: 2017. – P. 144–153. DOI: 10.1145/3084226.3084253.
8. Validating a model-driven software architecture evaluation and improvement method: A family of experiments / [H. J. González, I. Pelozo, A. Gonzales et al.] // Information and Software Technology. – 2015. – Issue 57. – P. 405–429. DOI: 10.1016/j.infsof.2014.05.018.
9. DBB Software's official company site [Electronic resource]. – Access mode: https://dbbsoftware.com/.
10. Fatima I. A Review of Software Architecture Evaluation Methods for Sustainability Assessment / I. Fatima, P. Lago // Software Architecture Companion: 20th International Conference (ICSA-C), L'Aquila, 13–17 March 2023 : proceedings. – Los Alamitos, CA: IEEE Computer Society, 2006. – P. 191–194 DOI: 10.1109/ICSA-C57050.2023.00050.
11. Lightweight Software Architecture Evaluation for Industry: A Comprehensive Review / [M. Sahlabadi, R. C. Muniyandi, Z. Shukur et al.] // Sensors. – 2022. – Vol. 22, Issue 3. DOI: 10.3390/s22031252.
12. ATAM: Method for Architecture Evaluation : technical report : CMU/SEI-2000-TR-004, ESC-TR-2000-004 / Prod-

DOI 10.15588/1607-3274-2026-1-10

uct Line Systems; R. Kazman, M. Klein, P. Clements. – Pittsburgh, 2000. – 71 p.
13. SAAM: a method for analyzing the properties of software architectures / [R. Kazman, L. Bass, G. Abowd et al.] // Software Engineering : 16th international conference, Sorrento, 16–21 May 1994 : proceedings. – Washington, DC: IEEE Computer Society Press, 1994. – P. 81–90. DOI: 10.1109/ICSE.1994.296768.
14. Babar M. A. Assessment of a Framework for Comparing Software Architecture Analysis Methods / M. A. Babar, B. Kitchenham // Evaluation and Assessment in Software Engineering (EASE) : 11th International Conference, UK, 2–3 April 2007 : proceedings. – Swindon: BCS Learning & Development Ltd., 2007. – P. 12–20. DOI: 10.14236/ewic/EASE2007.2.
15. Jayaratna N. Understanding and Evaluating Methodologies: NIMSAD, a Systematic Framewor / N. Jayaratna. – New York : McGraw-Hill, Inc., 1994. – 259 p.
16. Gonzalez-Huerta J. Models in Software Architecture Derivation and Evaluation: Challenges and Opportunities / J. Gonzalez-Huerta, E. Insfran, S. Abrahão // Model-Driven Engineering and Software Development : the 2nd International Conference, Lisbon, 7–9 January 2014 : proceedings. – Setubal: SCITEPRESS, 2014. – P. 12–31. DOI: 10.1007/978-3-319-25156-1_2.
17. Lloyd P. T. L. Technical reference architectures / P. T. L. Lloyd, G. M. Galambos // IBM Systems Journal. – 1999. – Vol. 38, Issue 1. – P. 51–75. DOI: 10.1147/sj.381.005.
18. Bass L. Software Architecture in Practice, Second Edition. / L. Bass, P. Clements, R. Kazman. – United States: Addison-Wesley Longman Publishing Co., Inc., 2003. – 528 p.
19. Lakhdissi M. A New Content Framework and Metamodel for Enterprise Architecture and is Strategic Planning / M. Lakhdissi, B. Bounabat // International Journal of Computer Science Issues. – 2014. – Vol. 9, Issue 2. DOI: 10.1201/b16417-9.
20. Clements P. Documenting Software Architectures: Views and Beyond, Second Edition / P. Clements, F. Bachmann, L. Bass. – Boston : Addison-Wesley, 2010. – 592 p. ISBN: 978-0-321-55268-6.
21. Solms F. What is Software Architecture / F. Solms // South African Institute for Computer Scientists and Information Technologists Conference (SAICSIT '12), Pretoria, 1–3 October 2012 : proceedings. – New York : Association for Computing Machinery, 2012. – P. 363–373. DOI: 10.1145/2389836.2389879.
22. Franchitti J. C. Enterprise Architecture Frameworks (EAFs) & Pattern Driven EAFs [Electronic resource] / J. C. Franchitti. – Access mode: https://cs.nyu.edu/~jcf/classes/g22.3033-007/slides/session2/g22_3033_011_c23.pdf.
23. Variability in software architecture: Current practice and challenges / [M. Galster, P. Avgeriou, D. Weyns et al.] // ACM SIGSOFT Software Engineering Notes. – 2011. – Vol. 36, Issue 5. – P. 30–32. DOI: 10.1145/2020976.2020978.
24. Building Evolutionary Architectures: Automated Software Governance 2nd Edition / [N. Ford, R. Parsons, P. Sadalage et al.]. – US: O'Reilly Media, 2022. – 262 p. ISBN : 978-1492097549.
25. Dakhli S. B. D. Architectural Deviations and Inconsistencies Management: A Framework Based on Information Systems Urbanization / S. B. D. Dakhli // Procedia Computer Science. – 2021. – Vol. 181. – P. 1122–1130. DOI: 10.1016/j.procs.2021.01.309.
26. Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality model overview and usage : ISO/IEC 25002:2024. – [Effective from 2024-03]. – ISO, 2024. – 17 p.
27. Rumpe B. Modeling with UML: Language, Concepts, Methods / B. Rumpe. – Switzerland : Springer Cham, 2016. – 281 p. ISBN: 978-3-319-33933-7.
28. Chattopadhyay A. Architecture Description Languages / A. Chattopadhyay, Z. Wang, G. E. Martin // Handbook of Computer Architecture. – New York : Springer, 2024. – Processor Design and Programming Flows. – P. 807–839. DOI: 10.1007/978-981-97-9314-3_18.
29. An approach to software architecture analysis for evolution and reusability / [C. H. Lung, S. Bot, K. Kalaichelvan et al.] // 1997 conference of the Centre for Advanced Studies on Collaborative Research, Toronto, 10–13 November 1997 : proceedings. – US : IBM Press, 1997. – P. 15. DOI: 10.1145/782010.782025.
30. Lassing N. On Software Architecture Analysis of Flexibility, Complexity of Changes: Size isn't Everything / [N. Lassing, D. Rijsenbrij, H. Vliet] // 2nd Nordic Software Architecture Workshop, Ronneby, 12–13 August 1999 : proceedings. – Ronneby, 1999. – P. 1103–1581.
31. Quantifying the costs and benefits of architectural decisions / [R. Kazman, J. Asundi, M. Klein] // Software Engineering : the 23rd International Conference, Toronto Ontario, 12–19 May 2001 : proceedings. – NW Washington: IEEE Computer Society, 2001. – P. 297–306. DOI: 10.1109/ICSE.2001.919103.
32. Moore M. Quantifying the value of architecture design decisions: Lessons from the field / M. Moore, R. Kaman, M. Klein // Software Engineering : 25th International Conference (ICSE03), Portland, 3–10 May 2003 : proceedings. – NW Washington: IEEE Computer Society, 2003. – P. 557–562. DOI: 10.1109/ICSE.2003.1201237.
33. Bergtsson P. O. Architecture-Level Modifiability Analysis (ALMA) / P. O. Bergtsson, N. Lassing, J. Bosch // Journal of Systems and Software. – 2004. – Vol. 69. – P. 129–147. DOI: 10.1016/S0164-1212(03)00080-3.
34. A Scenario-Driven Approach for Value, Risk, and Cost Analysis in System Architecting for Innovation. / [M. T. Ionita, P. America, D. K. Hammer et al.] // Software Architecture : 4th Working IEEE / IFIP Conference (WICSA 2004), Oslo, 12–15 June 2004 : proceedings. – P. 277. DOI: 10.1109/WICSA.2004.1310709.
35. Williams L.G. PASA^{SM}: An Architectural Approach to Fixing Software Performance Problems. / L. G. Williams, C. U. Smith // 28th International Computer Measurement Group Conference, Reno, 8–13 December 2002 : proceedings. – P. 307–320.
36. Smith C. U. Software performance antipatterns / C. U. Smith, L. G. Williams // Software and performance : the 2nd international workshop (WOSP00), Ottawa, 1 September 2000 : proceedings. – New York : Association for Computing Machinery. – P. 127–136. DOI: 10.1145/350391.350420
37. Smith C.U. Performance solutions: a practical guide to creating responsive, scalable software / C. U. Smith, L. G. Williams. – Redwood City, CA: Addison Wesley Longman Publishing Co., Inc., 2002. – 544 p. ISBN: 978-0-201-72229-1.

DOI 10.15588/1607-3274-2026-1-10

UDC 519.87:004.02

ABOUT RATIONAL METHODS FOR FINDING OPTIMAL ROUTES IN FUZZY TRAVELING SALESMAN PROBLEMS

Ivohin E. V. – Dr. Sc., Professor, Professor of the Department of System Analysis and Decision Support Theory, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. ROR: https://ror.org/02aaqv166. ORCID: https://orcid.org/0000-0002-5826-7408.

Gavrylenko V. V. – Dr. Sc., Professor, Professor of the Department of Information Systems and Technologies, National Transport University, Kyiv, Ukraine. ROR: https://ror.org/01akgs808. ORCID: https://orcid.org/ 0000-0001-9682-4204.

Yushtin K. E. – PhD, Post-doctoral Student of the Department of System Analysis and Decision Support Theory, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine. ROR: https://ror.org/02aaqv166. ORCID: https://orcid. org/0009-0001-9881-2343.

Ivohina K. E. – Post-graduate student of the Department of Information Systems and Technologies, National Transport University, Kyiv, Ukraine. ROR: https://ror.org/01akgs808. ORCID: https://orcid.org/ 0000-0001-9940-1178.

ABSTRACT

Context. This paper presents the results of a study on the use of triangular fuzzy numbers for determining time-optimal routes in the traveling salesman problem under fuzzy representations of travel duration in a transportation network. To formalize the uncertainty and imprecision of input data – associated with the subjectivity in estimating the time intervals required to travel between individual cities-triangular fuzzy numbers are employed. Various approaches to solving fuzzy traveling salesman problems are examined.

Objective. The goal of the work is to develop algorithms for solving the fuzzy traveling salesman problem based on the implementation of the Bellman-Zadeh parametric optimization methods, the use of a two-criteria approach with a given weight function and the refinement of the scheme for calculating the center of gravity of the membership function graph for a given curve density.

Method. The article considers methods for solving the fuzzy traveling salesman problem, which is formulated as the problem of finding a route to visit a given number of cities without repetitions with a minimum travel time. The parameters of the problem for formalizing the uncertainty and inaccuracy of input data associated with the influence of subjectivity in assessing the duration of time intervals required to travel between individual cities are presented as fuzzy triangular numbers. Different approaches to solving fuzzy traveling salesman problems are considered. The application of the Bellman-Zadeh method, methods taking into account refinements of defuzzified data, and methods based on a multicriteria approach is formalized. Computational experiments are carried out.

Results. Rational algorithms for solving the fuzzy traveling salesman problem based on the Bellman-Zadeh parametric optimization model, multicriteria approach and methods for refining the results of defuzzification of fuzzy data have been developed. In the conducted numerical experiments on solving the traveling salesman problem, fuzzy input data are based on the method for calculating the center of gravity (CoG), the center of gravity of homogeneous and non-homogeneous curves, which are determined by the membership function and the specified reliability values of subjective data. A comparison of the results obtained based on solving the crisp traveling salesman problem and the results based on defuzzified duration values for the fuzzy traveling salesman problem is carried out, according to the results of which the dependence of the solution on the defuzzification method is confirmed.

Conclusions. The article considers the method of formalizing the algorithm for solving the fuzzy traveling salesman problem with the minimum duration of movement along the route based on the Belman-Zadeh method, methods taking into account the refinements of defuzzified data and methods based on the multicriteria approach. Fuzzy triangular numbers are used to formalize the uncertainty of the input data when assessing the duration of movement between individual towns of the transport network. It was made a conclusion about the feasibility of using fuzzy numbers when solving fuzzy traveling salesman problems in real conditions of logistics transportation.

KEYWORDS: fuzzy traveling salesman problem, fuzzy numbers, subjective perception of duration, uncertainty, solution methods, multicriteria approach, defuzzification.

ABBREVIATIONS

TSP is a traveling salesman problem;
CoG is a center of gravity.

NOMENCLATURE

n is a number of cities;

t_{ij} are the travel time between all pairs of vertices;

T is a matrix of t_{ij};

t_{ij} are the travel time between all pairs of vertices;

X is a binary matrix of transitions between vertices x_{ij};

x_{ij} are the elements of matrix X, which equal to 0 or 1;

i is an index;

j is an index;

$\tilde{A}$ is a fuzzy set;

$\mu_{\tilde{A}}(x)$ is a membership function;

supp $\tilde{A}$ is a support of fuzzy set $\tilde{A}$;

Δt_{ij} are width of intervals of fuzzy travel time;

λ is a parameter of optimization Bellman-Zadeh model;

X^{1*} is an optimal values for first TSP;

DOI 10.15588/1607-3274-2026-1-11

X^{2*} is an optimal values for second TSP;

L_1 is a lower value of first TSP optimal solution;

L_2 is a lower value of second TSP optimal solution;

U_1 is an upper value of first TSP optimal solution;

U_2 is an upper value of second TSP optimal solution;

α_1 is a weight coefficient;

α_2 is a weight coefficient;

$\alpha(s)$ is a weight function;

x_C is a coordinate x of CoG of planar curve;

y_C is a coordinate y of CoG of planar curve;

x_C^{ρ} is a coordinate x of CoG of inhomogeneous curve;

y_C^{ρ} is a coordinate y of CoG of inhomogeneous curve;

$\rho(l)$ is a density function;

L is a length of membership function grafik;

Z is a required reliability level.

INTRODUCTION

The challenges associated with organizing and operating various spheres of the economy and management possess distinct characteristics that result in difficulties when solving diverse optimization problems. Some of these issues can be addressed by managerial staff, while others require analytical and optimization methods for production and organizational operations, including planning, coordination, decision-making, and control over the movement and storage of goods, services, and information. Mathematical methods are widely used to solve such problems, enabling the search for effective solutions while taking into account the limitations of the applied problem and the specifics of the available data. Solving certain optimization problems requires the use of non-standard methods based on well-known optimization algorithms.

However, the formal use of such methods is often impossible due to the imprecision (uncertainty) of the parameters and limitations of the real-world process models under consideration. These characteristics require the development and implementation of appropriate methods for formalizing and updating data that take uncertainty into account. One of the most well-known and effective methods for representing imprecise data is based on the use of Zadeh fuzzy sets [1], the essence of which allows for the mathematical expression of the subjectivity and uncertainty of parameter values in many applied problems. In recent years, numerous studies and papers have emerged based on this methodology and its application to various optimization tasks, decision support methods, and beyond.

One of the key applied optimization problems that requires a solution under conditions of uncertainty is the logistics-based Traveling Salesman Problem (TSP) [2]. A typical problem of salesman is to determine the route of a transport network consisting of n interconnected points (cities). The desired route, along which a traveling salesman must visit all cities in the network, must pass through each city only once and must be optimal in time or length [3].

From a mathematical perspective, the TSP is a combinatorial optimization problem, for which various mathematical programming techniques may be employed.

The object of study is a process of finding the optimal route for the fuzzy traveling salesman problem with minimum travel time in the transportation medium.

The subject of study is development of efficient algorithms for solving the fuzzy traveling salesman problem based on the Bellman-Zadeh parametric optimization model, a multi-criteria approach and methods for refining the results of defuzzification of fuzzy data.

The purpose of the work is to develop algorithms for solving the fuzzy traveling salesman problem based on the implementation of the Bellman-Zadeh parametric optimization method, the use of a two-criteria approach with a given weight function and the refinement of the scheme for calculating the center of gravity of the membership function graph for a given curve density.

1 PROBLEM STATEMENT

The task of finding an optimal cyclic tour in the TSP with given pairwise distances or travel durations t_{ij} between all cities in the network what represented by a matrix $T = \{t_{ij}\}, i, j = \overline{1,n}$ $T=\{t_{ij}\}$, $i,j=\overline{1,n}$, reduced to determining solution $X = \{x_{ij}\}$, $x_{ij} \in \{0,1\}$, $i, j=\overline{1,n}$, of the combinatorial optimization problem [4]

$$E = \sum_{i=1}^{n}\sum_{j=1}^{n} t_{ij}x_{ij} \to \min \qquad (1)$$

subject to the constraints

$$\sum_{j=1}^{n} x_{ij} = 1,\ i = \overline{1,n},\ \sum_{i=1}^{n} x_{ij} = 1,\ j=\overline{1,n}, \qquad (2)$$

which represents the classical formulation of the problem, where, in addition to route length or travel duration between locations, other criteria may also be considered, such as transportation cost, travel efficiency taking into account cargo volume or weight, and so on. A characteristic feature of all such formulations is the presence of a single optimality criterion in the route selection process.

It should be noted that, in addition to known and objective parameters of the traveling salesman's movement (such as distance between individual locations in the network, the maximum speed of the vehicle, etc.), one may also consider sets of factors that have an uncertain nature of influence. For example, the travel time along a particular route may depend on the time of day,

DOI 10.15588/1607-3274-2026-1-11

weather conditions, or even the “congestion level” of a given section of the transport network. Moreover, determining the optimal path requires accounting for their combined influence on the selection of the best route. As an illustration, a specific section of the road may experience different levels of traffic load at different times of the day, implying that the time required to pass through this section varies depending on when it is traversed. On the other hand, weather conditions (fog, rain) are also influential factors affecting travel duration across the considered segment. These situations necessitate solving the problem by incorporating subjective assessments of various motion parameters, leading to the formulation and solution of fuzzy traveling salesman problems (FuzzyTSP) [4].

2 REVIEW OF LITERATURE

Numerous works are devoted to the application of fuzzy set theory and fuzzy numbers to solving various applied optimization problems [2–9]. Among the the most recent publications, devoted to solving the traveling salesman problem, one can highlight the work [2], which proposes a method for solving the FuzzyTSP using various membership functions. In [3], a decision-making concept in a fuzzy environment is presented. In [5], a solution to fuzzy linear programming problems with multiple objective functions is proposed. On the other hand, for the classical TSP as a combinatorial optimization problem, many methods based on greedy and heuristic approaches have been proposed, allowing one to find locally optimal solutions (see, for example, [6,7]). Therefore, the integration of these methodologies with the formalization of uncertainty in optimization problems based on fuzzy numbers is of both theoretical and practical interest [4].

This paper examines approaches to improving the objectivity of input parameters in the FuzzyTSP for determining the shortest travel duration under uncertainty. It is proposed to apply and compare the results obtained using the Bellman-Zadeh method [6, 10], methods incorporating refined defuzzified data [11], and approaches based on multi-criteria decision-making [12, 13].

3 MATERIALS AND METHODS

Traditional set theory can be viewed as a particular case of fuzzy set theory. In classical mathematics, a set is defined as a collection of elements (objects) that share some common property [4].

According to Zadeh’s theory, fuzzy sets are defined as subsets of the universal set X as follows:

Definition 1. [4] A fuzzy set $\tilde{A}$ in a universal set X is a collection of ordered pairs $\tilde{A} = \{(\mu_{\tilde{A}}(x), x)\}$, where $\mu_{\tilde{A}} : X \to [0,1]$ is a mapping o the set into the unit interval [0,1] and is called a membership function of fuzzy set $\tilde{A}$.

The value of the membership function $\mu_{\tilde{A}}(x)$ for element $x \in X$ indicates the degree to which x belongs to the fuzzy set A (see Fig. 1). The interpretation of the membership degree $\mu_{\tilde{A}}(x)$ is a subjective measure of how well the element $x \in X$ corresponds to the concept defined by the fuzzy set $\tilde{A}$.

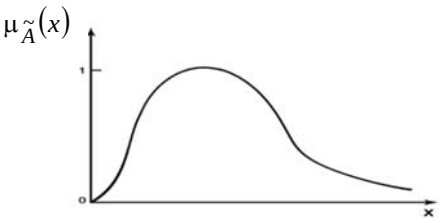


Figure 1 – Example of a membership function

It is noted that for any fuzzy sets $\tilde{A}$ and $\tilde{B}$ $\mu_{\tilde{A}\cup\tilde{B}}(x) = \max(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x))$, $\mu_{\tilde{A}\cap\tilde{B}}(x) = \min(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x))$.

Definition 2. [14] A fuzzy set $\tilde{A}$ is called convex if next inequality satisfied

$$\mu_{\tilde{A}}(\lambda x + (1-\lambda)y) \geq \min(\mu_{\tilde{A}}(x), \mu_{\tilde{A}}(y))$$

for all $x, y \in X, \lambda \in (0,1)$.

A fuzzy set $\tilde{A}$ is called normal if there exists $x \in X$ such that $\mu_{\tilde{A}}(x) = 1$.

Crisp sets $\tilde{A}_\alpha$ are called α-level sets of fuzzy set $\tilde{A}$, where $\alpha \in [0,1]$ is the height of fuzzy set $\tilde{A}$ and supp $\tilde{A} = \{x \in X : \mu_{\tilde{A}}(x) > 0\}$ is the support of fuzzy set $\tilde{A}$.

A fuzzy set A is considered unimodal if there exists only one element $x \in X$ such that $\mu_{\tilde{A}}(x) = h$.

Let us consider the universal set X to be the set of real numbers, i.e., $X = R$.

Definition 3. [4] A fuzzy set A defined over the set of real numbers $\mathbb{R}$ is called a fuzzy number if it satisfies the following properties:

I. the set $\tilde{A}$ is convex in the sense of definition 2;

II. the set $\tilde{A}$ is normal;

III. the membership function $\mu_{\tilde{A}}(x)$ is upper semi-continuous;

IV. the support of the fuzzy number sup $p(\tilde{A})$ is a subset of the universal set R.

Definition 3. [4] A triangular fuzzy number $\tilde{A}$ is an ordered triplet of real numbers (a_1, a_2, a_3), $a_1 \leq a_2 \leq a_3$, with an associated membership function $\mu_{\tilde{A}}(x)$ defined as:

$$\mu_{\tilde{A}}(x) = \frac{x - a_1}{a_2 - a_1}, x \in [a_1, a_2];$$

$$\mu_{\tilde{A}}(x) = \frac{a_3 - x}{a_3 - a_2}, x \in [a_2, a_3]; \quad (3)$$

DOI 10.15588/1607-3274-2026-1-11

$$\mu_{\tilde{A}}(x) = 0, x \notin [a_1, a_3].$$

Clearly, a triangular fuzzy number (a_1, a_2, a_3) is a special case of a unimodal fuzzy set with a height equal to one [15] (see Fig. 2).

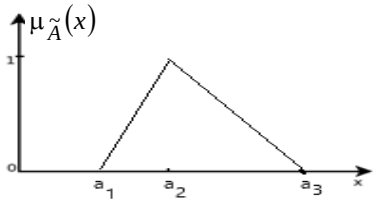


Figure 2 – Example of a triangular fuzzy number

It is well known [4], that triangular fuzzy number (a_1, a_2, a_2) is called a left triangular fuzzy number if its membership function is defined as

$$\mu_{\tilde{A}}(x) = 0, x \le a_1; \quad \mu_{\tilde{A}}(x) = \frac{x - a_1}{a_2 - a_1}, x \in (a_1, a_2);$$

$$\mu_{\tilde{A}}(x) = 0, x \ge a_2, \qquad (4)$$

and a triangular fuzzy number (a_1, a_1, a_2) is called a right triangular fuzzy number with the membership function:

$$\mu_{\tilde{A}}(x) = 0, x \le a_1; \mu_{\tilde{A}}(x) = \frac{a_2 - x}{a_2 - a_1}, x \in (a_1, a_2);$$

$$\mu_{\tilde{A}}(x) = 0, x \ge a_3. \qquad (5)$$

Definition 5. [4] A trapezoidal fuzzy number $\tilde{A}$ is an ordered quadruple of real numbers (a_1, a_2, a_3, a_4), $a_1 \le a_2 \le a_3 \le a_4$, with a membership function $\mu_{\tilde{A}}(x)$ defined as:

$$\mu_{\tilde{A}}(x) = \begin{cases} \dfrac{x - a_1}{a_2 - a_1}, if\ a_1 \le x \le a_2; \\ 1,\ if\ a_2 \le x \le a_3; \\ \dfrac{a_4 - x}{a_4 - a_3}, if\ a_3 \le x \le a_4. \end{cases} \qquad (6)$$

Definition 6. A rectangular-trapezoidal fuzzy number $\tilde{A}$ (see Fig. 3) is a special case of a trapezoidal fuzzy number (a_1, a_1, a_2, a_3), $a_1 \le a_2 \le a_3$, where the membership function $\mu_{\tilde{A}}(x)$ defined as:

$$\mu_{\tilde{A}}(x) = \begin{cases} 0, if\ \ x < a_1\ or\ x > a_3; \\ 1, if\ a_1 \le x \le a_2; \\ \dfrac{a_3 - x}{a_3 - a_2}, if\ a_2 \le x \le a_3. \end{cases} \qquad (6')$$

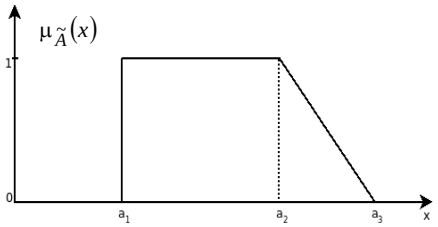


Figure 3 – Trapezoidal fuzzy number

The use of triangular and trapezoidal fuzzy numbers for practical implementation under conditions of uncertainty raises numerous questions regarding their constructiveness and necessitates the development of specialized methods for representing numerical fuzziness. It should be noted that the use of left-sided, right-sided triangular, and rectangular-trapezoidal fuzzy numbers – most commonly employed for formalization – can be considered only within the framework of triangular fuzzy numbers. The inclusion of normalized values from the interval $[a_1, a_2]$ of a rectangular-trapezoidal fuzzy number does not significantly influence the formation of assessments related to the subjectivity of the formalized uncertainty parameters. In this context, the interval-based representation primarily reflects the predetermined variability of the parameters. From the standpoint of fuzzy evaluation, the interval $[a_2, a_3]$, whose corresponding values indicate the degree of subjectivity associated with each estimate, is of critical importance. Therefore, in the future we will limit ourselves to considering only triangular fuzzy numbers for formalizing the uncertainty in travel time in the traveling salesman problem.

Fuzzy Traveling Salesman Problem

Let us consider the fundamentals of optimization theory for solving the FuzzyTSP. In article [6] the authors investigated a classical model for solving optimization problems for decision making in a fuzzy environment, which laid the foundation for the development of most results in the theory of fuzzy decision-making. When analyzing the process of solving optimization problems under uncertainty, it becomes evident that both the objective function and the constraints can be fuzzy, each defined by an appropriate membership function. Since the primary goal of optimization problems is to find solutions where the objective function achieves its optimal value under the given constraints, the solution in a fuzzy context is defined analogously. Finding an optimal solution involves identifying an element in the feasible region that simultaneously delivers the best value of the objective function and satisfies all given constraints. Thus, a "solution" under fuzzy conditions can be interpreted as the intersection of the domains defined by the fuzzy constraints and the fuzzy objective function. An example of such a solution in a fuzzy optimization problem is illustrated in Figure 4.

DOI 10.15588/1607-3274-2026-1-11

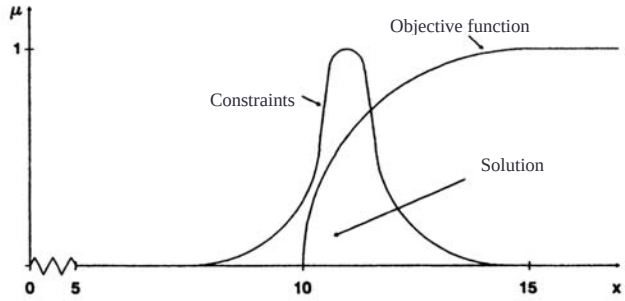


Figure 4 – Example of identifying a solution in a fuzzy optimization problem

It is clear that the "fuzzy linear programming" model cannot be defined unambiguously and that different approaches to implementation are possible depending on the specifics of the real situation that needs to be formalized.

We begin by formulating the basic model of a "fuzzy linear programming problem". For this purpose, let us consider the standard formulation of the classical linear programming problem: find a solution $x \in R^n$ that maximizes the objective function

$$c^T x \to \min, \tag{7}$$

in the feasible solution space defined by the system of constraints

$$Ax \geq b\,,\ x \geq 0\,, \tag{8}$$

where $x = (x_1,...,x_n)^T$ – the vector of decision variables, $c = (c_1,...,c_n)^T$ – the vector of coefficients of the objective function, $A = (a_{ij}), i = \overline{1,n},\ j = \overline{1,m},$ – the constraint coefficient matrix, $b = (b_1,...,b_m)^T$ – the right-hand side vector.

Let us assume that there is a decision maker who determines the required reliability level Z that must be achieved in the objective function in model (7)–(8). Also assume that each constraint in the problem is defined using data from a fuzzy set. In this case, we can formulate the general form of the fuzzy linear programming problem (see, [4]): find a vector such that

$$c^T x \mathrel{\tilde{\leq}} Z, \tag{9}$$

$$Ax \mathrel{\tilde{\geq}} b\,,\ x \geq 0\,. \tag{10}$$

Here, the symbol "$\mathrel{\tilde{\leq}}$" is used to denote the fuzzy variant of the non-strict relation "$\leq$" and can be linguistically interpreted as "essentially less than or equal to." Similarly, the symbol "$\mathrel{\tilde{\geq}}$", respectively, denotes the fuzzified version of the relation "$\geq$", which can be linguistically interpreted as "essentially greater than or equal to". Formally, the fuzzy relations "$\mathrel{\tilde{\leq}}$" and "$\mathrel{\tilde{\geq}}$" can be treated as particular cases of fuzzy relations "non-strictly less" and "non-strictly greater", respectively, over the set $X = R^1$ [16].

In this case, the objective function is defined as the minimization of the linear expression (7), where the value of Z is considered an upper bound of the optimal solution. Let us introduce the notation. $\begin{pmatrix} -c \\ A \end{pmatrix} = D$, $\begin{pmatrix} -Z \\ b \end{pmatrix} = p$, and rewrite the constraints (9)–(10) as follows:

$$Dx \mathrel{\tilde{\geq}} p\,,\ x \geq 0\,. \tag{11}$$

If the value Z and the elements of the vector b are given in the form of fuzzy numbers, then the right-hand side of the system of inequalities (11) will define a vector whose elements can correspond to specific values of the supports of the given fuzzy quantities. Therefore, each of the $(m + 1)$ inequalities in (11) for a given value of the vector p has a set of solutions that are characterized by the value of the corresponding membership function, denoted by $\mu_i(x), i = \overline{1,m+1}$. Without loss of generality, it can be assumed that all membership functions are monotonically increasing. Then, for any $\alpha \in [0,1]$ the values $\mu_i(x) = \alpha$, $i = \overline{1,m+1}$, can be interpreted as degrees of confidence with which the vector x satisfies the crisp inequalities $\sum_{j=1}^{n} d_{ij} x_j \geq \tilde{p}_i(\alpha)$, $i = \overline{1,m+1}$ (where d_{ij} – matrix elements D, and $\tilde{p}_i(\alpha)$ – are the values obtained based on the α-level sets of the fuzzy elements of vector p). If we assume that the decision-maker (DM) is interested not in a fuzzy set of solutions, but in a crisp "optimal" solution of the original problem, then an improvement of the solution (11) can be proposed by solving the following nonlinear programming problem:

$$\mu_{\tilde{P}}(x) = \min_{i=1,m+1} \mu_i(x)\,. \tag{12}$$

If we assume that the decision maker (DM) is interested not in a fuzzy set of solutions, but in a crisp "optimal" solution to the original problem, the solution (12) can be further refined by solving a nonlinear programming problem

$$\max_{x \geq 0} \min_{i=1,m+1} \mu_i(x) = \max_{x \geq 0} \mu_{\tilde{P}}(x)\,. \tag{13}$$

Let us define the type of membership functions $\mu_i(x)$, $i = \overline{1,m+1}$. The values of $\mu_i(x)$, $i = \overline{1,m+1}$, should be equal to 0 if the constraints or objective function are significantly violated, and equal to 1 if they are completely satisfied. In addition, the functions are

DOI 10.15588/1607-3274-2026-1-11

assumed to increase monotonically from 0 to 1 on the interval [0,1], i.e.:

$$\mu_i(x)=\begin{cases}0, & \sum_{j=1}^{n} d_{ij}x_j < p_i,\\ \in[0,1], & p_i \le \sum_{j=1}^{n} d_{ij}x_j \le p_i+q_i,\ i=\overline{1,m+1},\\ 0, & \sum_{j=1}^{n} d_{ij}x_j > p_i+q_i,\end{cases} \quad (14)$$

where q_i, $i=\overline{1,m+1}$, are subjectively defined tolerance thresholds for admissible deviations in the constraints and the objective function.

Under our assumption regarding fuzzy numbers, the membership functions are linearly decreasing over the corresponding "tolerance intervals" $[p_i, p_i+q_i]$, $i=\overline{1,m+1}$:

$$\mu_i(x)=\begin{cases}0, & \sum_{j=1}^{n} d_{ij}x_j < p_i,\\ 1-(\sum_{j=1}^{n} d_{ij}x_j - p_i)/q_i, & p_i \le \sum_{j=1}^{n} d_{ij}x_j \le p_i+q_i,\\ 0, & \sum_{j=1}^{n} d_{ij}x_j > p_i+q_i,\end{cases}$$
$$i=\overline{1,m+1}. \quad (15)$$

Substituting (15) into (13), and after straightforward algebraic manipulation, we obtain the following decision-making criterion for selecting the optimal solution:

$$\max_{x\ge 0} \min_{i=\overline{1,m+1}} \left(1-(\sum_{j=1}^{n} d_{ij}x_j - p_i)/q_i\right). \quad (16)$$

We introduce a new variable $\lambda\in[0,1]$, which corresponds to the minimal membership level of the fuzzy set of "solutions" $\tilde{P}$ defined in (12) for the fuzzy model (11): $1-(\sum_{j=1}^{n} d_{ij}x_j - p_i)/q_i \ge \lambda$.

This yields the Bellman-Zadeh [10] optimization model

$$\max_{x\ge 0} \lambda. \quad (17)$$

under condition

$$\lambda q_i + \sum_{j=1}^{n} d_{ij}x_j \le p_i + q_i,\ \ i=\overline{1,m+1},\ \ x\ge 0. \quad (18)$$

If the optimal solution of problem (17)–(18) is denoted by a vector (λ, x^O), then x^O will be the solution of the maximization problem (13) for the fuzzy optimization model (7)–(8), under the assumption that the membership functions are defined as in (14).

Thus, we come to the conclusion [4], that this optimal solution to the original model (7)–(8) can be found by solving a standard (crisp) linear programming problem with one additional variable and one additional constraint.

Let us apply this approach to solving the fuzzy traveling salesman problem. To solve the FuzzyTSP, it is necessary to take into account the nature of the uncertainty of the problem parameters and develop appropriate methods for finding a route.

In real-world transportation logistics scenarios, it is often impossible to determine precise travel times between cities in a network. When travel times cannot be precisely determined and represented by fuzzy triangular numbers, route finding on a transportation network is formulated as a fuzzy traveling salesman problem.

Let's formulate a mathematical formulation of the FuzzyTSP. The goal is to find a traveling salesman route, represented by a cyclic permutation of the indexes of cities in the transportation network, that minimizes the total travel time. In other words, we aim to minimize the objective function

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{t}_{ij}x_{ij}, \quad (19)$$

where the travel times between nodes are given by a matrix $T=\{t_{ij}\}$, $i,j=\overline{1,n}$, whose elements are triangular fuzzy numbers $\tilde{t}_{ij}=\tilde{t}_{ji}$, $\tilde{t}_{ij}=(t_{ij}, t_{ij}, t_{ij}+\Delta t_{ij})$, $i,j=\overline{1,n}$, and the possible connection routes between cities are specified by the matrix $X=\{x_{ij}\}$, $x_{ij}\in\{0,1\}$, $i,j=\overline{1,n}$ subject to the constraints (2).

To determine the route in the FTSP using the Bellman-Zadeh approach, we solve two crisp TSPs of the form (1), (2), with objective functions

$$Z_l = \min_X \sum_{i=1}^{n}\sum_{j=1}^{n} t_{ij}x_{ij}, \quad (20)$$

$$Z_u = \min_X \sum_{i=1}^{n}\sum_{j=1}^{n} (t_{ij}+\Delta t_{ij})x_{ij}, \quad (21)$$

taking into account the constraints (2), respectively. The solutions yield the optimal values of the objective functions Z_l and Z_u, which correspond to the optimal total durations of the route under lower and upper bounds of fuzzy travel times in the network.

Let us denote the optimal solutions of problems (20), (2) and (21), (2) as

DOI 10.15588/1607-3274-2026-1-11

$$X^{1*} = \arg\min_{X} \sum_{i=1}^{n}\sum_{j=1}^{n} t_{ij}x_{ij},$$

$$X^{2*} = \arg\min_{X} \sum_{i=1}^{n}\sum_{j=1}^{n} (t_{ij}+\Delta t_{ij})x_{ij},$$

respectively. Then we compute

$$L_1 = \min(Z_l, \sum_{i=1}^{n}\sum_{j=1}^{n} t_{ij}x_{ij}^{2*}),\ U_1 = \max(Z_l, \sum_{i=1}^{n}\sum_{j=1}^{n} t_{ij}x_{ij}^{2*}),$$

$$L_2 = \min(Z_u, \sum_{i=1}^{n}\sum_{j=1}^{n} (t_{ij}+\Delta t_{ij})x_{ij}^{1*}),$$

$$U_2 = \max(Z_u, \sum_{i=1}^{n}\sum_{j=1}^{n} (t_{ij}+\Delta t_{ij})x_{ij}^{1*}),$$

which represent, respectively, the lower and upper bounds of the optimal values of the objective functions $\sum_{i=1}^{n}\sum_{j=1}^{n} t_{ij}x_{ij}$ and $\sum_{i=1}^{n}\sum_{j=1}^{n} (t_{ij}+\Delta t_{ij})x_{ij}$.

Then, the solution to the FuzzyTSP (19), (2) is obtained by solving the parametric Bellman-Zadeh optimization problem [10] of the following form:

$$\begin{aligned} &\max_{x} \lambda \\ &\lambda(U_1 - L_1) + \sum_{i=1}^{n}\sum_{j=1}^{n} t_{ij}x_{ij} \le U_1, \\ &\lambda(U_2 - L_2) + \sum_{i=1}^{n}\sum_{j=1}^{n} (t_{ij}+\Delta t_{ij})x_{ij} \le U_2, \\ &\sum_{j=1}^{n} x_{ij} = 1,\ i = \overline{1,n},\ \sum_{i=1}^{n} x_{ij} = 1,\ j = \overline{1,n}, \end{aligned} \tag{22}$$

$x_{ij} = 0$ or 1 for all $i, j = 1,2,...,n$.

Clearly, solving the FuzzyTSP, taking into account its combinatorial nature, via the corresponding optimization formulation requires significant computational and time resources. Therefore, current research in the domain of FuzzyTSP focuses on improving existing methods and/or developing new approaches.

The main results are associated with the use of techniques that transform the triangular fuzzy parameters into a specific representation format [16]. This transformation allows performing arithmetic operations on fuzzy numbers. The implementation of such an approach is among the most commonly used strategies for constructing route sequences in FuzzyTSP, particularly when averaged values of fuzzy input parameters are calculated based on their centers of gravity [17].

Another direction of research in the development and implementation of FTSP solution methods involves the application of a multicriteria approach. Let us consider the fuzzy traveling salesman problem (19), (2) as a bicriteria optimization problem, where the total travel time along the route is to be minimized under fuzzy constraints given by lower and upper bounds. In other words, the classical single-criterion formulation of the TSP is replaced with a formulation containing two criteria:

$$F_1 = \sum_{i=1}^{n}\sum_{j=1}^{n} t_{ij}x_{ij} \to \min, \tag{23}$$

$$F_2 = \sum_{i=1}^{n}\sum_{j=1}^{n} (t_{ij}+\Delta t_{ij})x_{ij} \to \min, \tag{24}$$

where the values t_{ij}, $i,j = \overline{1,n}$, are the elements of the matrix $T = \{t_{ij}\}$, $i,j = \overline{1,n}$, which represent the nominal travel times between all pairs of nodes in the transportation network.

To obtain a compromise solution, we apply the convolution method. Let us introduce weight coefficients $\alpha_1, \alpha_2 > 0$: $\alpha_1 + \alpha_2 = 1$, corresponding to the decision-maker's confidence in the lower and upper bounds of the travel durations. This allows the formation of weighted indicators from the fuzzy input values $\tilde{t}_{ij}$, $i,j = \overline{1,n}$. The solution of problem (8), (9) with two sets of fuzzy parameters t_{ij} and $t_{ij} + \Delta t_{ij}$, $i,j = \overline{1,n}$, respectively, is sought as the optimal solution to a classical TSP of the form (1), (2) with the following weighted criterion:

$$\begin{aligned} F &= \alpha_1 F_1 + \alpha_2 F_2 = \\ &\alpha_1 \sum_{i=1}^{n}\sum_{j=1}^{n} t_{ij}x_{ij} + \alpha_2 \sum_{i=1}^{n}\sum_{j=1}^{n} (t_{ij}+\Delta t_{ij})x_{ij} = \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \{\alpha_1 t_{ij} + \alpha_2 (t_{ij}+\Delta t_{ij})\}x_{ij} \to \min, \end{aligned} \tag{25}$$

subject to the constraint (2). It is important to note that the resulting compromise solution depends on the chosen weight coefficients w_1 and w_2. This approach utilizes only the support intervals of the fuzzy travel durations and does not account for the values of the corresponding membership functions.

Nevertheless, this approach provides a constructive way of solving the fuzzy traveling salesman problem (FTSP) as a bicriteria optimization task for finding a time-optimal travel route based on the criteria defined in (23) and (24), subject to constraints (2).

However, the application of only two criteria–related to the confidence levels of the decision maker in the travel duration bounds – does not fully exploit the convolution method for determining a compromise route in the FTSP defined by (19), (2) as a solution to the associated bicriteria problem.

To generalize the above approach, we introduce confidence indicators through a weight function $\alpha(s) \ge 0$, defined on the interval $s \in [0,1]$, which satisfies the condition:

DOI 10.15588/1607-3274-2026-1-11

$$\int_0^1 \alpha(s)ds = 1. \quad (26)$$

Let us formulate the optimality criterion in the fuzzy traveling salesman problem (19) using the weight function $\alpha(s) \geq 0$, $s \in [0,1]$, as follows:

$$\int_0^1 \alpha(s)\{\sum_{i=1}^{n}\sum_{j=1}^{n}(t_{ij} + s\Delta t_{ij})x_{ij}\}ds = \\ = \sum_{i=1}^{n}\sum_{j=1}^{n}\{\int_0^1 \alpha(s)(t_{ij} + s\Delta t_{ij})ds\}x_{ij} \to \min. \quad (27)$$

Assuming that $\alpha(s) = 0$ for all $s \notin [0,1]$, the integral in (27) evaluates the weighted average of the linear travel time functions $t_{ij} + s\Delta t_{ij}$, $i,j = \overline{1,n}$, with respect to the confidence levels specified by the weight function $\omega(x)$. In this formulation, the entire range of values within $\alpha(s)$, is used to derive defuzzified estimates for fuzzy travel times intervals $t_{ij} + s\Delta t_{ij}$, $s \in [0,1]$, $i,j = \overline{1,n}$, allowing the fuzzy TSP to be reduced to the classical problem (1), (2).

It should be noted that defining an additional confidence function $\omega(x)$ may impose preferences for certain values $t_{ij} + s\Delta t_{ij}$, $s \in [0,1]$, $i,j = \overline{1,n}$, , derived from the support intervals of fuzzy travel durations, without directly accounting for their fuzziness. To incorporate the fuzzy nature of the travel time intervals, corresponding membership functions $\mu_{\tilde{t}_{ij}}(t_{ij} + s\Delta t_{ij})$, $s \in [0,1]$, $i,j = \overline{1,n}$ should be used.

Let us denote $g_{ij}(s) = \mu_{\tilde{t}_{ij}}(t_{ij} + s\Delta t_{ij}) \cdot (t_{ij} + s\Delta t_{ij})$, $s \in [0,1]$, $i,j = \overline{1,n}$. Then for any arbitrary weight function $\alpha(s) \geq 0$, $s \in [0,1]$ $\alpha(s)$, values $\int_0^1 \alpha(s)g_{ij}(s)ds$, will be considered as the weighted average values for each fuzzy triangular number $\tilde{t}_{ij}$, $i,j = \overline{1,n}$. Then, the defuzzified values of the travel durations on the network are computed by integrating both the membership functions and the preference weights defined by $\omega(x)$, and the optimality criterion (12) $\alpha(s)$, $s \in [0,1]$, and the optimality criterion (12) for the fuzzy traveling salesman problem becomes:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\{\int_0^1 \alpha(s)g_{ij}(s)ds\}x_{ij} \to \min. \quad (28)$$

Finally, the fuzzy traveling salesman problem with objective function (19) is reduced to a single-criterion optimization problem with the objective function in the form of (28), by applying a multicriteria approach and a specific linear convolution with an interval-defined weight function. This enables the refinement of route calculations within the fuzzy TSP framework.

Finally, one of the most effective techniques employed in solving fuzzy traveling salesman problems is the defuzzification of fuzzy numbers based on the computation of the center of gravity (CoG) of the fuzzy set [18]. In this case, the "averaged" value of a triangular fuzzy number $\tilde{A} = (a_1, a_2, a_3)$ is considered to be the value computed as the center of gravity of a planar figure bounded by the abscissa axis and the graph of the membership function of the fuzzy set. For a discrete fuzzy set, the formula is given by:

$$CoG = \frac{\sum_{i=1}^{n}\mu(x_i)\cdot x_i}{\sum_{i=1}^{n}\mu(x_i)}, \quad (29)$$

where x_i are values from the universal set, and, $\mu(x_i)$ is the membership degree of each value, $i = \overline{1,n}$, in the case of a continuous representation of the fuzzy number, the formula takes the form:

$$CoG = \frac{\int_{a_1}^{a_3} x \cdot \mu(x)dx}{\int_{a_1}^{a_3} \mu(x)dx}. \quad (30)$$

The quantity CoG obtained by this method represents the abscissa of the center of gravity of a homogeneous planar figure. It is important to note that a fuzzy set (or fuzzy number) is defined by the graph of its membership function over the interval of its support. In the case of a triangular fuzzy number, the membership function is piecewise linear, which allows the CoG value to be replaced by computing the center of gravity of a homogeneous curve (x_c, y_c), using the formulas

$$x_c = \frac{\int_{a_1}^{a_3} x\sqrt{1+[\mu'(x)]^2}\,dx}{\int_{a_1}^{a_3} \sqrt{1+[\mu'(x)]^2}\,dx}, \\ y_c = \frac{\int_{a_1}^{a_3} \mu(x)\sqrt{1+[\mu'(x)]^2}\,dx}{\int_{a_1}^{a_3} \sqrt{1+[\mu'(x)]^2}\,dx}. \quad (30)$$

The center of gravity of a plane curve is defined as a point on the plane at which the static moment about any coordinate axis is equal to the static moment of the curve itself about the same axis. In other words, the center of gravity of the curve corresponds to the "average" value of the triangular fuzzy number.

When applying the above averaging methods based on the CoG concept, it is generally assumed that the density of the membership function curve and the associated plate

DOI 10.15588/1607-3274-2026-1-11

is homogeneous. However, one can also assume that the curve has a variable density function $\rho(l)$, $0 \le l \le L$, where the value of $\rho(l)$ depends on the position of each point on the graph of the fuzzy number's membership function (here L is the arc length of the curve). Without loss of generality, we can assume that the density values lie within the interval $[0,1]$. In this case, the position of the center of gravity of a continuous inhomogeneous curve is given by

$$x_c^{\rho} = \frac{\int_{a_1}^{a_4} \rho(l) x \sqrt{1+[\mu'(x)]^2}\, dx}{\int_{a_1}^{a_4} \sqrt{1+[\mu'(x)]^2}\, dx}, \qquad (32)$$

$$y_c^{\rho} = \frac{\int_{a_1}^{a_4} \rho(l) \mu(x) \sqrt{1+[\mu'(x)]^2}\, dx}{\int_{a_1}^{a_4} \sqrt{1+[\mu'(x)]^2}\, dx}.$$

By interpreting the density function $\rho(x)$ as the credibility (or reliability) function of the membership values, the triangular fuzzy number is thus represented as a type-2 fuzzy number [19]. Accordingly, by expressing the density dependence on the location l on the curve through an arbitrary function $\rho(l)$: $R^1 \to [0,1]$, $0 \le l \le L$, one obtains new "averaged" characteristics of the center of gravity and the corresponding degree of membership (as per equation (31)).

4 EXPERIMENTS

We conduct numerical experiments using various approaches to solve the fuzzy traveling salesman problem in which the travel duration along the network is defined by fuzzy right-triangular numbers. The computations are carried out using a model of a logistic transport network with precisely defined values of travel time for all possible segments of movement [20].

The optimal solution to the classical traveling salesman problem on the specified network corresponds to the route

$$1 \to 2 \to 6 \to 10 \to 11 \to 8 \to 5 \to 9 \to 7 \to 4 \to 3 \to 1, \qquad (33)$$

for which the total duration amounts to 156 units.

To simulate the fuzzy version of the traveling salesman problem, the fuzzy travel time along each path between cities is modeled using the rule:

$$\tilde{t}_{ij} = (t_{ij}, t_{ij}, t_{ij} \cdot 1.8 + 1.5 + (i+j) \cdot 0.75), \qquad (34)$$

where i, j are the indices of the cities and t_{ij} is the crisp (precisely defined) travel times.

5 RESULTS

Applying different defuzzification methods to the fuzzy travel durations results in solutions that fully coincide with the travel sequence given in (33). Table 1 presents the lower and upper bounds of travel durations between cities i and j within the transport network, the weighted travel durations computed using $\alpha_1 = \alpha_2 = 0.5$, the *CoG* values of the corresponding fuzzy durations, as well as the calculated centroids of homogeneous and inhomogeneous graphs of linear membership functions (using $\rho(l) = e^{-l}, 0 \le l \le L$) and the corresponding optimal values of the total travel time for the salesman (see the row labeled "Duration"), which are obtained by the aforementioned methods. The last column reports the results obtained using the Bellman-Zadeh method with parameter $\lambda = 0.15$.

It is evident that the invariance of the route across different solution methods for the fuzzy traveling salesman problem is due to the uniform increase in all fuzzy duration parameters, which is rarely observed under real-world conditions. In this case, the primary objective of the presented results is to visually demonstrate the previously discussed solution techniques for the fuzzy TSP. It is worth noting that the most informative and qualitatively accurate outcomes were achieved in numerical experiments involving fuzzy travel durations between cities, in which the computation of travel time accounts for the density function of the membership graph (see Table 1, column x_c^{ρ}).

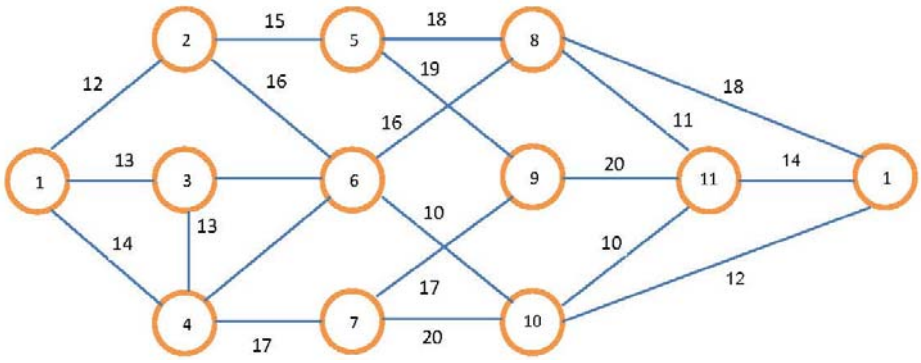


Figure 5 – An example of a transport network in the traveling salesman problem [4]

DOI 10.15588/1607-3274-2026-1-11

Table 1 – Results of Solving the FuzzyTSP with Fuzzy Travel Duration Values in the Network [4]

i	j	t_{ij}	$t_{ij} + \Delta t_{ij}$	$0.5\,t_{ij} + 0.5\,(t_{ij} + \Delta\, t_{ij})$	CoG	x_c	x_c^{ρ}	x^O
1	2	12	25.35	24.675	17.2337	17.103	17.01	14.0025
1	3	13	27.90	26.95	18.8037	18.4807	18.17	15.235
1	4	14	30.45	29.225	20.3742	20.0508	20.016	16.4675
1	8	18	40.65	38.325	26.6586	26.3284	26.241	21.3975
1	10	13	31.35	28.675	19.1806	19.160	19.105	15.7525
1	11	14	35.70	31.85	22.0797	22.0195	22.011	17.255
2	5	15	33.75	31.875	22.1870	22.1852	22.156	17.8125
2	6	16	36.30	34.15	23.7584	23.3543	23.256	19.045
3	4	13	30.15	28.075	19.5295	19.512	19.314	15.5725
3	6	15	35.25	32.625	22.6731	22.1125	22.105	18.0375
4	6	16	37.80	34.9	24.2450	24.108	24.092	19.27
4	7	17	40.35	37.175	25.8170	25.6245	25.602	20.5025
5	8	18	43.65	39.825	27.6335	27.3799	27.367	21.8475
5	9	19	46.20	42.1	29.2058	29.0481	29.031	23.08
6	8	16	40.80	36.4	25.2245	25.1274	25.121	19.72
6	10	10	31.50	25.75	17.7688	17.1875	17.085	13.225
7	9	17	44.10	39.05	27.0429	27.0356	27.026	21.065
7	10	20	50.25	45.125	31.2685	31.2386	31.214	24.5375
8	11	11	35.55	28.775	19.8387	19.1381	19.123	14.6825
9	11	20	52.50	46.25	32.0067	31.9191	31.694	24.875
10	11	10	35.25	27.625	19.0165	18.246	18.228	13.7875
Duration		156	396.3	354.15	245.6485	242.1097	241.212	192.045

6 DISCUSSION

Several remarks should be noted. The procedure for finding a solution to the fuzzy traveling salesman problem based on the Bellman-Zadeh method is based on multiple solutions of a crisp problem for different parameter values. To solve the problem, a genetic algorithm was used [21], which allowed us to obtain a solution relatively quickly. However, the transport network in the problem under consideration is small, and when it increases, finding a solution based on the Bellman-Zadeh method will be significantly limited. Thus, solving practical problems of finding the optimal traveling salesman route in a fuzzy setting is characterized by low performance and significant requirements for computing resources. To speed up obtaining the result, it is proposed to use methods based on defuzzification of the duration of movements between network nodes by calculating the gravity centroids of the graphs of the corresponding membership functions.

CONCLUSIONS

This paper presents the results of a study on the use of triangular fuzzy numbers for determining time-optimal routes in the traveling salesman problem under fuzzy representations of travel duration in a transportation network. To formalize the uncertainty and imprecision of input data – associated with the subjectivity in estimating the time intervals required to travel between individual cities-triangular fuzzy numbers are employed. Various approaches to solving fuzzy traveling salesman problems are examined. The application of the Bellman-Zadeh method, methods incorporating refined defuzzified data, and methods based on multicriteria decision-making are formalized. The interpretation of averaged values for right-sided triangular fuzzy numbers is analyzed. An enhancement of defuzzified values is proposed based on the computation of the center of gravity of the membership function curve and the construction of type-2 fuzzy sets, which allows for improved objectivity of the input parameters and yields better results. In the conducted numerical experiments on solving the traveling salesman problem with fuzzy travel durations, the influence of various defuzzification techniques is demonstrated. These include the use of the center of gravity (CoG), the centroid of homogeneous and inhomogeneous curves defined by membership functions, and the assigned confidence values of subjective data. A comparison is made between the results obtained from solving the crisp version of the traveling salesman problem and those derived from defuzzified values in the fuzzy case. The outcomes confirm the dependence of the solution on the defuzzification method applied. The study concludes that using triangular fuzzy numbers is appropriate and effective for solving fuzzy traveling salesman problems in real-world logistic transportation scenarios.

DOI 10.15588/1607-3274-2026-1-11

ACKNOWLEDGEMENTS

The work is supported by the state budget scientific research project of the National Transport University "System research and information technologies in the transport industry, telecommunications, industry and business" (state registration number 0124U003679).

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: Eugene Ivohin: formulation of the general statement of the problem and research methodology; Valery Gavrylenko: discussion of research directions and methods for solving the problem; Konstantin Yushtin: development of a methodology for processing fuzzy numbers and defuzzification methods, Kateryna Ivohina: experimental study of solution search methods.

Data availability: The manuscript has no associated data.

Software availability: The manuscript has no associated software.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

REFERENCES

1. Zadeh L. A. Fuzzy sets, *Information and Control*, 1965, No. 8, pp. 338–353.
2. Ghiani G., Laporte G., Musmanno R. Introduction to Logistics Systems Planning and Control. John Wiley & Sons, Ltd, 2013, 377 p. DOI:10.1002/9781118492185
3. Davendra D. Traveling salesman problem: theory and applications. Books on Demand, 2010, 338 p.
4. Gavrylenko V. V., Ivohin E. V., Ivohina K. E., Yushtin K. E. Optimization models of transport and network flows in the problems of supporting decision-making in information management systems, *In "Innovative trends in the development of information control systems and technologies"*. Under the general editorship of Doctor of Economics, Professor Ustenko S. V. Kyiv, KNEU named after Vadym Hetman, 2024, pp. 233–255. https://ir.kneu.edu.ua/handle/2010/46976
5. Kumar A., Gupta A. Methods for solving fuzzy assignment problems and fuzzy travelling salesman problems with different membership functions, *Fuzzy Information and Engineering*, 2011, No. 3(1), pp. 3–21.
6. Bellman R. E., Zadeh L. A. Decision making in fuzzy environment, *Management science*, 1970, No. 17, pp. 144–164.
7. Zimmermann H.-J. Fuzzy programming and linear programming with several objective functions, *Fuzzy Sets and Systems*, 1978, No.1, pp. 45–55.
8. Christofides N. Vehicle routing in the traveling salesman problem. Lawler, Lenstra, RinooyKan and Shmoys, John Wiley eds., 1985, pp. 431–448.
9. Fuling Tien. Applying interactive fuzzy multi-objective Linear programming to transportation planning decisions, *Journal of information and optimization sciences*, 2006, No. 27(1), pp. 107–126.
10. Kosheleva O., Kreinovich V. Why Bellman-Zadeh approach to fuzzy optimization, *Applied Mathematical Sciences*, 2018, Vol. 12, No. 11, pp. 517–522.
11. Yushtin K., Ivohin E. About defuzzification methods influence on fuzzy traveling salesman problem's solving, *Artificial Intelligence*, 2024, No. 1 (98), pp. 64–72. DOI: 10.15407/jai2024.01.064
12. Ivohin E., Gavrylenko V., Ivohina K. One approach to solving the fuzzy traveling salesman problem based on a multicriteria approach, *Artificial Intelligence*, 2025, No. 2 (103), pp. 84–94. DOI: 10.15407/jai2025 .02.084
13. Ivohin E., Yushtin K. A method for solving a single fuzzy multicriteria traveling salesman problem, *Artificial Intelligence*, 2024, No. 4 (101), pp. 142–150. DOI: 10.15407/jai2024.04.142
14. Bablu Jana, Tapan Kumar Roy Multi-objective fuzzy linear programming and its application in transportation model, *Tamsui Oxford Journal of Mathematical Sciences*, 2005, No. 21(2), pp. 243–268.
15. Kaufman A., Gupta M. M. Introduction to fuzzy arithmetic: theory and applications. Van Nostrand Reinhold Co. Inc., Workingham, Berkshire, 2003, 351 p.
16. Zimmermann H.-J. Fuzzy set theory and its application. Kluwer, Boston, 1992, 525 p. DOI: 10.1007/ 978-94-010-0646-0
17. Voskoglou M. G. Fuzzy sets, fuzzy logic and their applications, *Mathematics*, 2020, 452 p. DOI: 10.3390/ books978-3-03928-521-1
18. Van Broekhoven E., De Baets B. Fast and accurate center of gravity defuzzification of fuzzy system outputs defined on trape-zoidal fuzzy partitions, *Fuzzy Sets and Systems*, 2006, Vol. 157, No. 7, pp. 904–918.
19. Mendel J., Robert J. Type-2 fuzzy sets made simple, *IEEE Transactions on Fuzzy Systems*, 2002, No. 10 (2), pp. 117–127. DOI: 10.1109/91.995115.
20. Ivohin E., Gavrylenko V., Ivohina K. On the recursive algorithm for solving the traveling salesman problem on the basis of the data flow optimization method, *Radio Electronics, Computer Science, Control*, 2023, No. 3, pp. 141–147. DOI:10.15588/1607-3274-2023-3-14.
21. Ivohin E. V., Yushtin K. E. Solving the fuzzy traveling salesman problem using genetic algorithm with clustering by ward's method, *Conference Intelligent Transport Systems: Ecology, Safety, Quality, Comfort: ITS ESQC-2024: proceedings.* Springer, *2024*, Vol. 1, pp. 199–210. DOI: 10.1007/978-3-031-87376-8

Received 08.09.2025.
Accepted 08.01.2026.
Published 27.03.2026.

DOI 10.15588/1607-3274-2026-1-11

УДК 519.87:004.02

ПРО РАЦІОНАЛЬНІ МЕТОДИ ПОШУКУ ОПТИМАЛЬНИХ МАРШРУТІВ У НЕЧІТКИХ ЗАДАЧАХ КОМІВОЯЖЕРА

Івохін Є. В. – д-р фіз.-мат. наук, професор, професор кафедри системного аналізу та теорії прийняття рішень Київського національного університету імені Тараса Шевченка, Київ, Україна. ROR: https://ror.org/02aaqv166. ORCID: https://orcid.org/0000-0002-5826-7408

Гавриленко В. В. – д-р фіз.-мат. наук, професор, професор кафедри інформаційних систем і технологій Національного транспортного університету, Київ, Україна. ROR: https://ror.org/01akgs808. ORCID: https://orcid.org/ 0000-0001-9682-4204

Юштін К. Е. – канд. фіз.-мат. наук, докторант кафедри системного аналізу та теорії прийняття рішень Київського національного університету імені Тараса Шевченка, Київ, Україна. ROR: https://ror.org/02aaqv166. ORCID: https://orcid. org/0009-0001-9881-2343

Івохіна К. Є. – аспірант кафедри інформаційних систем і технологій Національного транспортного університету, Київ, Україна. ROR: https://ror.org/01akgs808. ORCID: https://orcid.org/ 0000-0001-9940-1178.

АНОТАЦІЯ

Актуальність. Важливою сучасною проблемою є швидке відновлення та оптимізація управління логістикою. В залежності від поставленої задачі існує багато різних математичних методів та підходів до вирішення різних логістичних задач, розв'язування яких набуває широкого практичного впровадження. Його конкретний зміст залежить від характеру проблеми та повноти наявних даних. Іноді для розв'язання відомих задач, однією з яких є задача комівояжера, вдається знайти нетипові методики на основі поєднання декількох обчислювальних схем та методів.

Ціль. Мета роботи – розробити алгоритми розв'язання нечіткої задачі комівояжера на основі реалізації методів параметричної оптимізації Беллмана-Заде, використання двокритеріального підходу із заданою ваговою функцією та уточнення схеми розрахунку центру ваги графіка функції належності для заданої щільності кривої.

Метод. У статті розглядаються методи розв'язування нечіткої задачі комівояжера, що формулюється як задача знаходження маршруту відвідування заданої кількості міст без повторень з мінімальною тривалістю руху. Параметри задачі для формалізація невизначеності та неточності вхідних даних, пов'язаних з впливом суб'єктивності в оцінках тривалості, необхідних для переміщення між окремими містами проміжків часу, подаються у вигляді нечітких трикутних чисел. Розглянуто різні підходи, що дозволяють розв'язувати нечіткі задачі комівояжера. Формалізовано застосування методу Белмана-Заде, методів з урахуванням уточнень дефазифікованих даних та методів на основі багатокритеріального підходу. Проведено обчислювальні експерименти.

Результати. Розроблено раціональні алгоритми розв'язання нечіткої задачі комівояжера на основі параметричної оптимізаційної моделі Беллмана-Заде, багатокритеріального підходу та методів уточнення результатів дефазифікації нечітких даних У проведених чисельних експериментах з розв'язання задачі комівояжера з нечітко заданою тривалістю переміщень продемонстровано вплив різних варіантів дефазифікації нечітких вхідних даних на основі методу розрахунку центра тяжіння (CoG), центру ваги однорідної та неоднорідної кривих, які визначаються функцією належності та заданими величинами надійності суб'єктивних даних. Проведено порівняння результатів, отриманих на основі вирішення чіткої задачі комівояжера, та результатів на основі дефазифікованих значень тривалості для нечіткої задачі комівояжеру, за ітогами якого підтверджено залежність розв'язку від способу дефазифікації.

Висновки. У статті розглянуто метод формалізації алгоритму розв'язання нечіткої задачі комівояжера з мінімальною тривалістю руху за маршрутом на основі методу Белмана-Заде, методів з урахуванням уточнень дефазифікованих даних та методів на основі багатокритеріального підходу. Для формалізації невизначеності вхідних даних при оцінці тривалості переміщення між окремими містами транспортної мережі використовуються нечіткі трикутні числа. Зроблено висновок про доцільність використання нечітких чисел при розв'язанні нечітких задач комівояжера в реальних умовах логістичних перевезень.

КЛЮЧОВІ СЛОВА: нечітка задача комівояжера, нечіткі числа, суб'єктивне сприйняття тривалості, невизначеність, методи розв'язування, багатокритеріальний підхід, дефазифікація.

ЛІТЕРАТУРА

1. Zadeh. L. A. Fuzzy sets / L. A. Zadeh // Information and Control. – 1965. – No. 8. – P. 338–353.
2. Ghiani G. Introduction to Logistics Systems Planning and Control / G. Ghiani, G. Laporte, R. Musmanno. – John Wiley & Sons, Ltd, 2013. – 377 p. DOI:10.1002/9781118492185
3. Davendra D. Traveling salesman problem: theory and applications / D. Davendra. – Books on Demand, 2010. – 338 p.
4. Optimization models of transport and network flows in the problems of supporting decision-making in information management systems / [V. V. Gavrylenko, E. V. Ivohin, K. E. Ivohina, K. E. Yushtin] // In "Innovative trends in the development of information control systems and technologies". – Under the general editorship of Doctor of Economics, Professor Ustenko S.V. – Kyiv, KNEU named after Vadym Hetman, 2024. – P. 233–255. https://ir.kneu.edu.ua/handle/2010/46976
5. Kumar A. Methods for solving fuzzy assignment problems and fuzzy travelling salesman problems with different membership functions / A. Kumar, A. Gupta // Fuzzy Information and Engineering. – 2011. – No. 3 (1). – P. 3–21.

DOI 10.15588/1607-3274-2026-1-11

6. Bellman R. E. Decision making in fuzzy environment / R. E. Bellman, L. A. Zadeh // Management science. – 1970. – No. 17. – P. 144–164.
7. Zimmermann H.-J. Fuzzy programming and linear programming with several objective functions / H.-J. Zimmermann // Fuzzy Sets and Systems. – 1978. – No. 1. – P. 45–55.
8. Christofides N. Vehicle routing in the traveling salesman problem / N. Christofides. – Lawler, Lenstra, RinooyKan and Shmoys, John Wiley eds., 1985. – P. 431–448.
9. Fuling Tien Applying interactive fuzzy multi-objective Linear programming to transportation planning decisions / Tien Fuling // Journal of information and optimization sciences. – 2006. – No. 27(1). – P. 107–126.
10. Kosheleva O. Why Bellman-Zadeh approach to fuzzy optimization / O. Kosheleva, V. Kreinovich // Applied Mathematical Sciences. – 2018. – Vol. 12, No. 11. – P. 517–522.
11. Yushtin K. About defuzzification methods influence on fuzzy traveling salesman problem's solving / K. Yushtin, E. Ivohin // Artificial Intelligence. – 2024. – No. 1 (98). – P. 64–72. DOI: 10.15407/jai2024.01.064
12. Ivohin E. One approach to solving the fuzzy traveling salesman problem based on a multicriteria approach / E. Ivohin, V. Gavrylenko, K. Ivohina // Artificial Intelligence. – 2025. – No. 2 (103). – P. 84–94. DOI: 10.15407/jai2025 .02.084
13. Ivohin E. A method for solving a single fuzzy multicriteria traveling salesman problem / E. Ivohin, K. Yushtin // Artificial Intelligence. – 2024. – No. 4 (101). – P. 142–150. DOI: 10.15407/jai2024.04.142
14. Bablu Jana. Multi-objective fuzzy linear programming and its application in transportation model / Bablu Jana, Tapan Kumar Roy // Tamsui Oxford Journal of Mathematical Sciences. – 2005. – No. 21(2). – P. 243–268.
15. Kaufman A. Introduction to fuzzy arithmetic: theory and applications / A. Kaufman, M. M. Gupta. – Van Nostrand Reinhold Co. Inc., Workingham, Berkshire, 2003. – 351 p.
16. Zimmermann H.-J. Fuzzy set theory and its application / H.-J. Zimmermann. – Kluwer, Boston, 1992. – 525 p. DOI: 10.1007/ 978-94-010-0646-0
17. Voskoglou M. G. Fuzzy sets, fuzzy logic and their applications / M. G. Voskoglou // Mathematics. – 2020. – 452 p. DOI: 10.3390/ books978-3-03928-521-1
18. Van Broekhoven E. Fast and accurate center of gravity defuzzification of fuzzy system outputs defined on trapezoidal fuzzy partitions / E. Van Broekhoven, B. de Baets// Fuzzy Sets and Systems. – 2006. – Vol. 157. – No. 7. – P. 904–918.
19. Mendel J. Type-2 fuzzy sets made simple/ J. Mendel, J. Robert // IEEE Transactions on Fuzzy Systems. – 2002. – No. 10 (2). – P. 117–127. DOI: 10.1109/91.995115.
20. Ivohin E. V. On the recursive algorithm for solving the traveling salesman problem on the basis of the data flow optimization method / E. Ivohin, V. Gavrylenko, K. Ivohina // Radio Electronics, Computer Science, Control. – 2023. – No. 3. – P. 141–147. DOI:10.15588/1607-3274-2023-3-14
21. Ivohin E. V. Solving the fuzzy traveling salesman problem using genetic algorithm with clustering by ward's method / E. V. Ivohin, K. E. Yushtin // Conference Intelligent Transport Systems: Ecology, Safety, Quality, Comfort: ITS ESQC-2024: proceedings, Springer, 2024. – Vol. 1. – P. 199–210. DOI: 10.1007/978-3-031-87376-8

DOI 10.15588/1607-3274-2026-1-11

p-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2026. № 1
e-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2026. № 1

UDC 004.421.2:519.6:004.89

A STUDY OF THE PERFORMANCE OF ANY-ANGLE THETA* ALGORITHMS ON WEIGHTED GRID MAPS FOR ROUTE PLANNING

Kis Y. – Post-graduate student of the Department of Discrete Analysis and Intelligent Systems, Ivan Franko National University of Lviv, Ukraine. ROR: https://ror.org/01s7y5e82. ORCID: https://orcid.org/ 0009-0009-7816-237X.

Shcherbyna Y. M. – PhD, Professor of the Department of Discrete Analysis and Intelligent Systems, Ivan Franko National University of Lviv, Ukraine. ROR: https://ror.org/01s7y5e82. ORCID: https://orcid.org/ 0000-0002-4942-2787.

Kunanets N. E. – Dr. Sc., Professor of the Department of Information Systems and Networks, Lviv Polytechnic National University, Lviv, Ukraine. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0000-0003-3007-2462.

Yarymovych Y. A. – Post-graduate student of the Department of Information Systems and Networks, Lviv Polytechnic National University, Lviv, Ukraine. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0009-0006-1391-3214.

ABSTRACT

Context. The article addresses the study of free-direction pathfinding algorithms, in particular the Theta* algorithm, and evaluates their performance on weighted grid maps in order to determine optimal routes for delivering goods to a firearms store. This research is carried out in the broader context of developing an information system for tracking and managing arms sales and logistics under complex conditions. One of the main motivations is that any-angle methods can produce more realistic and natural-looking paths compared to the classical A* algorithm.

Objective. The purpose of the study is to analyze the performance of three Theta*-based algorithms – Basic Theta*, Lazy Theta*, and Strict Theta*– on both uniform and weighted square grids, with special emphasis on execution time and path cost metrics. The work aims to generalize the applicability of these algorithms to weighted environments and to propose improvements suitable for real-world route planning scenarios.

Method. The principles of A*, the three Theta* variants, and path post-processing smoothing techniques are presented. The research describes the transition from unweighted uniform square grids to weighted grids and highlights the complexity of calculating accurate path costs when applying any-angle approaches. Visual demonstrations of algorithmic behavior were implemented using the Unity game engine. Performance metrics were measured separately for uniform and weighted grids to ensure comparative analysis.

Results. The results include comparative evaluations of Basic Theta*, Lazy Theta*, Strict Theta*, and classical A* algorithms. The analysis identifies conditions under which each algorithm performs effectively, as well as factors that limit their applicability in weighted environments. It is shown that path length and path cost may differ substantially in weighted grids, leading to new considerations for cost-based optimization. Based on the experiments, a generalization of the Basic Theta* algorithm is proposed to enhance its suitability for weighted square grids, and a potential extension of the Strict Theta* algorithm to this context is outlined.

Conclusions. The findings demonstrate that while any-angle algorithms provide smoother and more realistic routes, their effectiveness in weighted environments depends on careful adaptation of cost functions. The research highlights their value not only for simulating complex virtual environments and agent behaviors in games and robotics but also for practical applications in logistics, particularly in the development of an information system for tracking and managing firearms sales. The proposed algorithmic adaptations may contribute to improving delivery planning and supply chain efficiency, including the modeling of weapons delivery routes under wartime conditions.

KEYWORDS: pathfinding, path planning, square grid, any-angle algorithm, path cost, weighted grid, Theta*.

ABBREVIATIONS

AI – Artificial Intelligence;
A* – A-star algorithm;
API – Application Programming Interface;
GIS – Geographic Information System;
SCM – Supply Chain Management;
WGM – Weighted Grid Map;
UML – Unified Modeling Language;
ms – milliseconds.

NOMENCLATURE

s is a node corresponding to the point of departure;
S is a set of grid cells;
N is a number of nodes;
W_r is a road weight coefficients;
$R_{i,j}$ is a risk factor for each cell;
$T_{i,j}$ is a traffic intensity;
$T(P)$ is a travel time;
E is an efficiency and safety indicator;
P is a logistics cost parameters;
g is a node corresponding to the firearms store (destination);
D is an optimal path;
L is a path length;
$L(s,g)$ is a set of all feasible routes between s and g that do not pass through forbidden or hazardous cells;
V is a set of grid cells (nodes);
$v_{i,j}$ is a set of grid cells (nodes) indexed by coordinates (*i*,*j*);
$d_{i,j}$ is a road quality coefficient (road surface condition);
$r_{i,j}$ is a risk factor (e.g., presence of checkpoints, crime level);

DOI 10.15588/1607-3274-2026-1-12

$t_{i,j}$ is an average traffic;

$p_{i,j}$ is a logistics cost (fuel, duties, etc.);

α is a weight for road surface condition;

β is a weight for risk level;

γ is a weight for traffic density;

δ is a weight for surveillance or control measures;

$r_{\max}$ is a maximum allowable risk of a cell;

Ω is a set of cells designated as risk zones or restricted areas;

λ is a congestion impact coefficient;

v_{avg} is a verage vehicle speed;

P is a set of all valid routes between s and g that do not pass through forbidden or dangerous cells;

V is a set of nodes (grid cells) indexed by coordinates (i,j);

v_k is a k-th node (vertex) in the grid representing a specific position or cell;

v_{k+1} is a k+1-th node in the path or in the visibility sequence;

$v_{i,j}$ is a node is assigned a weight (traversal cost);

$C(P)$ is a weighted length (cost) of a path P;

$C(v_k, v_{k+1}))$ is a transition cost between nodes v_k and v_{k+1};

$\bar{\bar{w}}_{k,k+1}$ is an averaged weight of the cells crossed by the segment $[v_k, v_{k+1}]$;

$w(v_k))$ is a weight of node v_k.

INTRODUCTION

Nowadays, pathfinding algorithms are widely used across numerous fields related to computer science, including robotics [1], logistics, navigation systems [2], routing protocols [3, 4], and video games [5]. It is evident that different application domains require different approaches to solving the problem of pathfinding – or, as some sources refer to it, path planning. For example, in logistics, pathfinding may involve traversing graph vertices representing warehouses between which goods are transported. In contrast, in video games, the virtual environment is typically discretized and represented as a square grid or navigation mesh. Another potential variation arises in scenarios involving incomplete information about the graph structure, such as in routing protocols, where routers (treated as graph nodes) operate with limited knowledge and attempt to forward data toward a destinatio – often prioritizing reachability over optimality.

Given the broad spectrum of use cases, pathfinding has been extensively studied. However, this does not imply that there is no room for improvement of existing algorithms. This study focuses on the exploration of any-angle pathfinding algorithms, particularly Theta*, when applied to square grid environments. These algorithms are capable of producing shorter paths in Euclidean space compared to the classical A* algorithm [6], which is constrained to move between adjacent grid nodes.

The primary emphasis of this work lies in investigating the performance of Theta*-based algorithms on weighted square grids, where their any-angle nature provides a substantial advantage over A*. Moreover, paths generated by any-angle methods often appear visually more plausible, as agents are not restricted to movement along fixed angular directions but can adjust their trajectory according to the geometry and structure of the environment.

The relevance of this topic stems from the growing demand for realistic simulations of complex virtual environments [7], as well as the need for intelligent decision-making in agent navigation. One such example might be simulating a video game character crossing a river.

The relevance of this research also lies in adapting the results of studying Theta* algorithms for planning supply routes to firearms stores, particularly under complex logistical conditions. The results of the study are integrated into an information system for managing firearms sales, including an AI module for selecting and re-selecting routes as the situation changes, simulating risk scenarios, and visualizing routes on a logistics dashboard. In addition, this research is relevant for security services, private security companies, military logistics, specialty goods stores, and the public sector, where high accuracy, flexibility, and realism of planned routes are required. All these factors can be converted into a quadratic grid and expressed as the weight of a specific cell, which indicates either the time required to traverse that cell or the probability of failure when passing through it.

A distinguishing feature of this paper is its comparative analysis of Theta* algorithms against the standard A* algorithm and a variant that incorporates path smoothing. The smoothing process partially mimics the any-angle nature of Theta*, but it has limitations – primarily because it only smooths paths generated by A*, which can negatively affect the cost of the resulting path, especially when operating on a weighted grid.

The object of research is the processes of searching for optimal routes in discrete spatial models (grids), taking into account the weighted characteristics of the environment.

The subject of research of the study is any-angle pathfinding algorithms, in particular the variants of the Theta* algorithm (Basic Theta*, Lazy Theta*, Strict Theta*) and their application to weighted square grids for route planning tasks.

The obtained results are distinguished by their **scientific novelty**, which lies in the proposal to use the algorithms lies in the fact that, for the first time, the specific features of applying Any-Angle Theta* algorithms to weighted square grids have been analyzed, with consideration of the difference between path length and path cost, which is critical for optimization tasks on maps with weighted coefficients. The methodology for quantitatively

DOI 10.15588/1607-3274-2026-1-12

evaluating the efficiency of Theta* algorithms has been improved by comprehensively accounting for two metrics-execution time and path cost – in environments with different types of grids (uniform and weighted). Furthermore, the generalization of the Basic Theta* algorithm for application in weighted environments has been further developed, and directions have been identified for extending the Strict Theta* algorithm to weighted maps, opening new opportunities for practical use in logistics systems and complex route management.

1 PROBLEM STATEMENT

In the context of secure logistics, the task of planning safe and efficient delivery routes for high-risk goods, particularly firearms, is becoming increasingly complex due to unstable infrastructure, dynamic security conditions, traffic restrictions, and the need for real-time adaptation. Traditional grid-based algorithms such as A* do not always ensure sufficient accuracy, flexibility, or adaptability, since they rely on fixed movement directions and do not adequately reflect weighted environmental constraints such as risks, terrain conditions, or traffic.

Input Variables: S, N, W_r, $R_{i,j}$, $T_{i,j}$, P, $\alpha, \beta, \gamma, \delta$, s, g.

Desired Outcomes (Output Variables): D, L, $C(P)$, $T(P)$, E.

Dependencies: D depends on spatial constraints S and N, L depends on road and obstacle weights W_r.

$C(P)$ is a function of road quality, risk factor, traffic, and logistics costs:

$$C(P) = \sum_{k=0}^{n-1} \frac{(length(v_k, v_{k+1})}{v_{avg}} + \lambda \cdot t_{i,j}. \quad (1)$$

$T(P)$ depends on average traffic and speed constraints:

$$C(P) = \sum_{k=0}^{n-1} \frac{(length(v_k, v_{k+1})}{v_{avg}} + \lambda \cdot t_{i,j}. \quad (2)$$

$length(v_k, v_{k+1})$ is the Euclidean distance between nodes v_k and v_{k+1}. E depends on minimization of cost and risk under time constraints:

$$T(P) \le T_{\max}, \; r_{i,j} \le r_{\max}. \quad (3)$$

The problem reduces to finding a path P^* such that:

$$P^* = \arg\min_{P \in L(s,g)} C(P).$$

Mathematical modeling of these parameters enables the construction of optimal, safe, and realistic routes on weighted grid maps. This justifies the need to investigate the performance of Any-Angle Theta* algorithms (Basic, Lazy, Strict) as adaptive tools for secure logistics and route planning in high-risk environments.

2 REVIEW OF THE LITERATURE

Previous studies have explored the application of Theta* algorithms on weighted square grids. In the original paper proposing the Basic Theta* algorithm [8], a generalized version was presented and compared against A* and Field D*. In that study, the cost of the Theta* path was calculated along the line connecting two vertices, considering the cumulative weights of all intersected cells. The weighted Theta* paths were, on average, 3% shorter than those produced by A*, but only in cases where the grid contained large contiguous areas of cells with baseline weight values.

An alternative approach was proposed in which the entire map was assigned a weight factor [9], and this factor was incorporated into the heuristic distance function to the goal. Through such pre-analysis of map complexity, algorithms in the Theta* family achieved higher efficiency by reducing execution time during pathfinding.

There also exists a generalization of the Lazy Theta* algorithm adapted for pathfinding in weighted 3D environments [10]. The cost computation approach in that work was similar to that of [8], but the primary focus was on three-dimensional space.

Furthermore, a dynamically hybrid algorithm, Non-uniform-Theta*, was developed for autonomous ground vehicle navigation in environments containing both static and dynamic obstacles [11]. This algorithm enables real-world maneuvering by integrating both path planning and agent-level motion control in dynamic settings.

To date, no dedicated study has been found that evaluates the performance of the Strict Theta* algorithm on weighted square grids. In general, the most promising approach for computing accurate and optimal paths appears to be the line-based cost evaluation method, which considers all intersected cell weights along the line-of-sight between vertices. Given that the primary advantage of Theta* algorithms over A* lies in their ability to produce lower-cost paths, this cost-evaluation method was adopted in the present study.

In contrast to prior works, this paper evaluates the performance of all three Theta* algorithms (Basic, Lazy, and Strict) on a weighted square grid, and compares them against both the standard A* algorithm and A* with path smoothing.

3 MATERIALS AND METHODS

This study adopts a simulation-based experimental methodology, combining concepts from computational geometry, optimization, and artificial intelligence to analyze the performance of directional pathfinding algorithms – specifically Theta* and its modifications – on weighted grid maps. The research aims to evaluate route quality, realism, and risk-aware cost optimization in weapon delivery logistics under uncertain and dynamic conditions.

DOI 10.15588/1607-3274-2026-1-12

The simulation environment is built in Unity to reflect realistic terrain conditions, traffic, and risk zones. This allows reproducible and controlled experiments for testing routing algorithms under varying logistical constraints.

The routing environment is modeled as a Weighted Grid Map (WGM), where the delivery region (e.g., Lviv Oblast or border zones) is discretized into grid cells (e.g., 50×50 meters). Each cell carries a weight $w_{j,i}$, which is computed using a composite function:

$$w_{j,i} = \alpha \cdot d_{j,i} + \beta \cdot r_{j,i} + \gamma t_{j,i} + \delta p_{i,j}, \quad (4)$$

represent road conditions, security risk, traffic intensity, and surveillance density respectively, while $\alpha, \beta, \gamma, \delta$ are tunable coefficients based on scenario-specific priorities (e.g., minimizing risk in high-value cargo delivery). $\alpha, \beta, \gamma, \delta \in R+$ are the weighting coefficients reflecting priorities.

The experimental framework evaluates three variants of the Theta* family:

– Basic Theta*: prioritizes shortest, realistic paths with arbitrary angles;

– Lazy Theta*: optimistically assumes line-of-sight and defers visibility checks;

– Strict Theta*: enforces obstacle-hugging behavior for higher realism in constrained urban settings.

Each algorithm is benchmarked across multiple scenarios, including uniform and weighted maps, short-range and long-haul delivery simulations. Over 2500 runs were conducted, with randomized source-target locations and consistent pseudorandom seeds for reproducibility.

The results of each algorithm were assessed using total route cost (sum of weights along the path), computation time (ms), path realism (visual coherence and feasibility of navigation). For dynamic adaptation, an API connection to real-time traffic data (e.g., simulated Google Traffic) and threat updates was considered in the weighted grid recalculation.

Route visualization and metric logging were performed via Unity and Python, while data analysis and comparisons used Pandas, Matplotlib, and NumPy libraries.

Inspired by techniques in AI-driven navigation and perceptual learning, the weighted environment simulates realistic logistical bottlenecks. While this research does not employ GANs or image-based synthesis, the methodology follows a multi-layer abstraction similar to preprocessing in computer vision. Grid generation, threat encoding, and cost-map construction serve as “preprocessing layers”, and path optimization mimics an inference step, outputting cost-effective, secure routes.This layered modeling allows flexible integration into AI-based decision support systems, e.g., those incorporating reinforcement learning or probabilistic planning for supply chain operations.

In the current conditions, ensuring the logistics of high-risk goods supply, particularly weapons, is becoming an increasingly complex task due to unstable infrastructure, dynamic changes in the security environment, traffic restrictions, and the need for real-time adaptation to external factors. Traditional pathfinding algorithms, such as A*, are not always capable of providing sufficient accuracy, flexibility, and efficiency in route planning under such conditions, as they operate in a discrete grid-based environment with fixed movement directions and limited adaptability to changing weight characteristics of the route. This leads to inefficient routes that fail to account for risks, road conditions, dynamic constraints, and other critical parameters.

The problem lies in the lack of sufficiently universal, flexible, and adaptive tools capable of modeling a complex logistics environment using a weighted spatial model and finding routes that not only minimize distance but also consider critical parameters of the delivery environment.

In this context, it becomes necessary to investigate the efficiency and adaptation of free-direction pathfinding algorithms such as Theta*, particularly their operation on weighted grid maps, as a means of building realistic, safe, and optimal routes in a high-risk environment for delivering goods to firearms stores.

The task is to determine a safe and efficient delivery route for goods to a firearms store, taking into account spatial constraints, potential risk zones, and the need to minimize time and transportation cost. To model the environment, a weighted grid is used, where each cell has its own weight that reflects the complexity or danger of traversing it (for example, the presence of obstacles, surveillance cameras, checkpoints, or other risk factors).

It is necessary to find a route from a starting point (warehouse or logistics center) to a firearms store so that:

– the total cost of the route is minimal;

– the route remains within permitted areas and maximally avoids high-risk zones;

– the trajectory is as close as possible to the optimal (shortest) path, taking into account the possibility of movement not only horizontally and vertically but also at arbitrary angles.

To solve this problem, a free-direction Theta* algorithm is proposed. Unlike classical grid-based pathfinding algorithms (such as A* and D*), Theta* allows the construction of routes that are not limited to grid axes and reduces unnecessary turns, producing paths closer to the straight-line optimum.

The result is a route that simultaneously meets the criteria of safety (avoiding dangerous cells) and efficiency (minimizing both path length and total weighted cost).

Let the delivery environment be discretized as a weighted grid

$$G = (V, E). \quad (5)$$

$V = v_{i,j}$ is the set of grid cells (nodes) indexed by coordinates (*i*,*j*); $E \subseteq V \times V$ is the set of edges connecting adjacent nodes (in Theta*, connections in arbitrary directions are allowed if the conditions of direct visibility

DOI 10.15588/1607-3274-2026-1-12

are satisfied); $C: V \times V \to R+$ is the cost function for moving between nodes, which takes into account the properties of the cells through which the segment passes. Each node $v_{i,j}$ is assigned a weight (traversal cost) $w(v_{i,j}) \geq 0$, determined by the conditions of the terrain and safety. Start and goal nodes $s \in V, g \in V$. For each grid cell $w_{i,j}$ we define its weight as (4).

We define the cost function of a route. Let the path P be a sequence of nodes $P = (v_0, v_1, \ldots v_k)$, $v_0 = s_{start}$ is a start node, with the goal $v_n = v_{goal}$. The weighted length of the path is defined as:

$$C(P) = \sum_{k=0}^{k-1} (\| v_{k+1} - v_k \| \cdot \frac{w(v_k) + w(v_{k+1})}{2}) \quad (6)$$

$\| v_{k+1} - v_k \|$ is the Euclidean distance between the centers of the grid cells (for Theta*, both diagonal and arbitrary directions are taken into account). v_k, v_{k+1} is a any pair of adjacent nodes between which there is direct visibility in the grid.

Objective (optimality criterion) it is required to find a route.

And minimizes the total cost:

$$\frac{\min}{p} \sum_{k=0}^{k-1} C(v_k, v_{k+1}), \quad (7)$$

where the transition cost is

$$C(v_k, v_{k+1}) = length(v_k, v_{k+1}) \cdot \overline{\overline{w}}_{k,k+1}. \quad (8)$$

It is required to find a path P* that satisfies

$$P^* = arg \frac{\min}{P \in L(s,g)} C(P), \quad (9)$$

$L(s,g)$ for which $w(v) = \infty$. A path is considered feasible if:

$$\forall_k, visibility(v_k, v_{k+1}) = true \text{ і } r_{i,j} \leq r_{\max}.$$

For any pair of consecutive nodes (v_k, v_{k+1}), a clear line-of-sight without intersecting obstacles is required. In this case $v_k \notin \Omega, \forall v_k \in P$.

To define the delivery time constraint, let T(P) denote the total travel time along the route (a function of the path length and traffic):

$$T(P) = \sum_{k=0}^{k-1} (\frac{length(v_k, v_{k+1})}{v_{avg}} + \lambda \cdot t_{i,j}). \quad (10)$$

It is required to (3). The generalized formalized problem can be expressed as follows:

$$\frac{\min}{P} \sum_{k=0}^{n-1} C(v_k, v_{k+1}) \quad (11)$$

subject to $T(P) \leq T_{\max}, r_{i,j} \leq r_{\max}$, $visibility(v_k, v_{k+1}) = true$.

Thus, the problem reduces to finding a path of minimal cost, taking into account the cell weights and safety constraints, using the Theta* algorithm, which allows constructing routes with freely chosen directions.

The pathfinding problem assumes the existence of a certain graph structure over which the search process is performed. In this case, the graph is represented by a square grid, where each cell has eight neighbors (including diagonals). Grid cells can exist in one of two states: unblocked or blocked, visualized as white and gray cells, respectively, in Figure 1. Unblocked cells may also carry a numerical weight value, indicating an increased traversal cost when passing through the respective cell. A square grid in which cells contain such weight values is referred to as a weighted square grid.

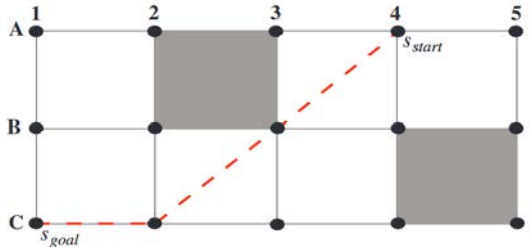


Figure 1 – An illustration of an unweighted square grid; the computed path is shown as a dashed line

In the pathfinding process, movement occurs through the corners of grid cells, which serve as the graph's vertices. The set of all vertices is denoted as *S*, and an individual vertex is denoted as s, where $s \in S$. The goal of the pathfinding problem is to determine an unblocked path from the starting vertex start to the target vertex send. An example of such a path is shown as a dashed line in Figure 1. A path is considered unblocked if every vertex along the path has line-of-sight to the next vertex. Line-of-sight is defined as the condition in which a straight line connecting any two vertices does not pass through the interior of any blocked cell or between two blocked cells along a shared edge.

Key criteria for evaluating pathfinding algorithms include path optimality and computational efficiency. An optimal path is the shortest possible path, while efficiency refers to the minimization of computation time during the search. Due to their any-angle nature, Theta* algorithms are capable of finding Euclidean-optimal paths, often yielding shorter paths than those produced by the classical A* algorithm, which is limited to grid-adjacent move-

DOI 10.15588/1607-3274-2026-1-12

ments. The same advantage is observed when operating on weighted square grids.

Since Theta* algorithms are derived from the A* algorithm [6], the following section outlines the fundamental principles of A*, based on which the key modifications introduced by any-angle algorithms are identified.

A* is conceptually straightforward: starting from the initial vertex sstart, it aims to find a path to the target vertex sgoal that minimizes the total cost. The algorithm operates by constructing a search tree, beginning at sstart and incrementally expanding paths one vertex at a time until the goal is reached.

At each iteration of the main loop, the A* algorithm must determine which vertex to expand next. The vertex with the lowest numerical value of $f(s)$ is selected. This value represents an estimated cost of the shortest possible path from the sstart vertex sstart, passing through s, to the goal vertex sgoal. The function $f(s)$ is defined as the sum of two components: the actual cost $g(s)$, which is the known cost from sstart to s, and the heuristic estimate $h(s)$, which approximates the remaining distance from s to sgoal. The choice of heuristic depends on the specific problem domain. In this study, the heuristic function is defined as the Euclidean distance between s and sgoal.

The value of f is updated at each iteration during the expansion of vertex s. All neighboring vertices $[s_1, s_2, ...s_8]$ compute their respective $f(sn)$ values. If the newly computed value is lower than the previously stored one (at the beginning of the algorithm, the value of f of all vertices is $+\infty$), update this value to the newly calculated one and update the parent vertex with vertex s.

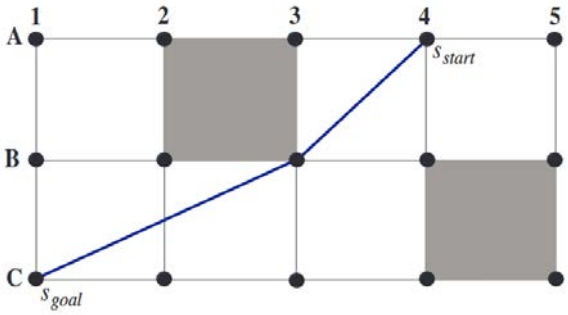


Figure 2 – The true shortest path is shown as a solid line from sstart to sgoal

This process defines the operation of the optimal path-finding algorithm A* [5]. To obtain even shorter paths, the Theta* algorithm [7] was introduced. As mentioned earlier, Theta* searches for the shortest paths in Euclidean 2D space and is not restricted to movement in just eight grid directions. This flexibility allows Theta* to approximate true shortest paths on a square grid more accurately. The any-angle Theta* algorithms discussed in this work do not guarantee perfectly optimal Euclidean paths, but they typically produce paths that are shorter than those generated by A*. An example of an optimal A* path is shown in Figure 1, while Figure 2 illustrates a true shortest path found using an any-angle Theta* algorithm.

The key distinction of Theta* algorithms lies in the fact that the parent of any vertex along the path may be any other vertex on the grid that is reachable via an unblocked line of sight. In contrast, in the A* algorithm, a vertex can only have one of its immediate neighbors as a parent.

In practice, the only functional difference in the Basic Theta* algorithm is that, during the expansion of a vertex s and subsequent evaluation of the *f*-value of a neighboring vertex s', two possible paths are considered. These paths are illustrated in Figure 3.

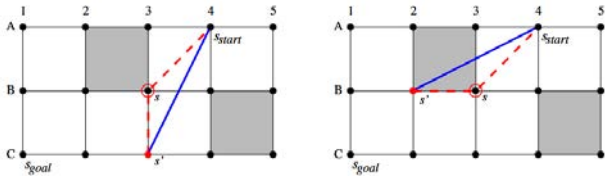


Figure 3 – Path 1 is shown as a dashed line. Path 2 is shown as a solid line

Path 1 corresponds to the standard path evaluated by A*, where the g-value is equal to the sum of the distance from parent(s) to s, and from s to s'. In contrast, Path 2 checks whether a direct connection exists between parent(s) and s', bypassing vertex s.

If a straight, unobstructed line of sight exists between parent(s) and s', this path is selected instead. According to the triangle inequality, Path 2 can never be longer than Path 1, since any side of a triangle is not longer than the sum of the lengths of the other two sides.

The smoothed A* algorithm adopts a similar principle, but on a smaller scale. While Basic Theta* evaluates Path 2 during the expansion phase, the smoothing process applies the same idea after the A* algorithm has already computed a complete path.

Given a sequence of vertices from the resulting path, the smoothing process iterates through them, checking whether a later vertex in the sequence has a direct line of sight to the currently processed vertex. If such a vertex is found, the intermediate vertices can be bypassed.

This post-processing approach is efficient, as it operates on a fixed path. However, it is limited to the set of vertices discovered by A*, and cannot search for truly optimal paths. This behavior is illustrated in Figure 4.

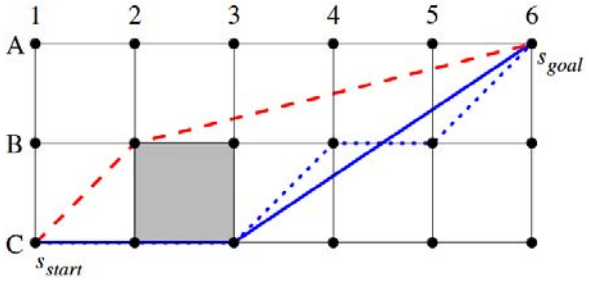


Figure 4 – The true shortest path is depicted by a dashed line, the path found by A* is represented by a dotted line, and the smoothed path is shown as a solid line

DOI 10.15588/1607-3274-2026-1-12

The Lazy Theta* algorithm [12] optimistically assumes that Path 2 is always available. As a result, it significantly reduces the number of computations performed during the vertex expansion phase, since it does not initially verify whether the assumed path truly provides an unobstructed line of sight. Instead, in the next iteration, it re-evaluates whether the assumption was valid and, if not, reverts to Path 1 without changing the current vertex. This strategy is effective in reducing runtime but heavily relies on the triangle inequality, which does not always hold when a square grid is transformed into a weighted grid.

The Strict Theta* algorithm [13] further introduces the concept of a tightly surrounding path. A path is considered tightly surrounding if every change in direction closely wraps around a specific obstacle. On uniform square grids, the optimal path is typically one that is tightly surrounding. However, not all tightly surrounding paths are necessarily optimal. Therefore, while Strict Theta* does not guarantee optimal pathfinding, it increases the likelihood of identifying near-optimal paths, as it leverages this additional spatial constraint – something that Basic Theta* does not account for.

Strict Theta* is implemented largely in the same manner as Basic Theta*, with the exception of an additional check to determine whether the path tightly surrounds an obstacle – this check is performed in constant time. If the path to the current vertex is not tightly surrounding, an additional penalty distance is added to the g-value after the vertex is expanded. The penalty distance is defined as sqrt(2) – 1, which approximately equals 0.42.

The check to determine whether the path $[\, parent(s),\ s,\ s' \,]$ is tightly surrounding requires examining a single grid cell. Among the four cells adjacent to vertex s, only the one that lies within the interior of the angle $\angle parent(s),\ s,\ s')$ less than 180 degrees is evaluated. This segment is considered tightly surrounding if and only if the corresponding cell is blocked. This behavior is illustrated in Figure 5.

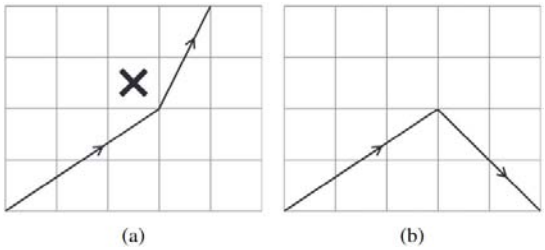


Figure 5 – (a) Since the grid cell within the angle less than 180 degrees is blocked, the path is considered tightly surrounding; (b) the path is not tightly surrounding

The following section describes the methodologies used for computing the path cost on a weighted square grid. Figure 6 provides a visual representation that serves as a reference for verifying the described path cost computation techniques.

Since A* navigates between neighboring vertices, there are three possible cases for assigning cell weights during path cost computation. In these examples, cells are denoted by the name of their top-left vertex. For instance, the cell enclosed by vertices $A1$, $A2$, $B1$, and $B2$ is referred to as $A1$, and its weight is denoted as weight ($A1$). The distance between two vertices is represented as $c(s_1, s_2)$.

In the first case, the path proceeds from the starting vertex ($s_{start} = B2$) to the goal vertex ($s_{goal} = A1$) within the same cell, i.e., diagonally. In this scenario, the path length equals $c(B2, A1)$, while the path cost is computed as the product of the path length and the weight of the cell it passes through. Thus, the path cost is calculated according to Equation (1)

$$c(B2, A1) * weight(A1) .$$

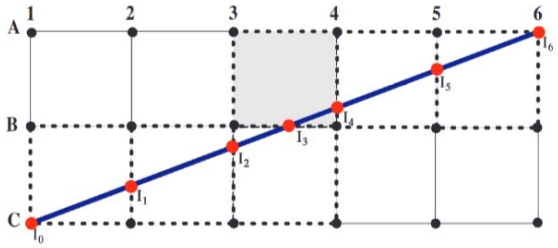


Figure 6 – The line represents the path found by the any-angle algorithm

In the second case, the movement occurs from the starting vertex ($s_{start} = B2$) to the goal vertex ($s_{goal} = A2$) along the edge shared by two unblocked cells, i.e., a horizontal or vertical move. Similarly, the path length is calculated as c(B2, A2); however, in this case, it is multiplied by the average weight of the two adjacent cells. Therefore, the path cost in this scenario is computed according to Equation (2)

$$\frac{c(B2, A2) \bullet (weight(A1) + weight(A2))}{2}.$$

In the third case, the movement proceeds from the starting vertex ($s_{start} = B1$) to the goal vertex ($s_{goal} = A1$) along the edge between one unblocked cell and one blocked cell, which is again a horizontal or vertical move. In this scenario, cell $A0$ does not exist; if it did, it would be considered blocked. The path cost is calculated as the path length multiplied by the weight of the unblocked cell, according to Equation (3)

$$c(B1, A1) \bullet weight(A1) .$$

To some extent, the situations described are appropriate to apply when calculating the path cost during the operation of Theta* algorithms. At the same time, let us analyze another unique case.

In the fourth case, the movement proceeds from the starting vertex ($s_{start} = C1$) to the goal vertex ($s_{goal} = A6$) across multiple grid cells. The straight line from vertex C1 to vertex A6 is divided into several segments at the points where it intersects the cell boundaries ($l_0, l_1, ... l_6$). The length of each segment is denoted as

DOI 10.15588/1607-3274-2026-1-12

$c(l_n, l_{n+1})$, and the cost of each segment is calculated as the segment length multiplied by the weight of the cell through which the segment passes. The total path cost from $C1$ to $A6$ is the sum of the segment costs that make up the path. As an example, the cost of the first segment is computed as follows:

$$c(l_0, l_1) \bullet weight(B1).$$

To compute the lengths of specific path segments, the Fast Voxel Traversal Algorithm [14] is employed. Although originally designed for use in three-dimensional space, this algorithm performs equally effectively in two-dimensional environments.

The following section discusses the modifications related to the operation of Theta* algorithms on weighted grid maps. While working with uniform grid maps, the triangle inequality property held true, since the cost of Path 2 (direct diagonal) was never greater than that of Path 1 (via intermediate vertex), assuming cells had equal weights. Consequently, the implementation of the Basic Theta* algorithm was designed to prefer Path 2 without verifying whether Path 1 might be shorter.

Since the work is now being conducted on a weighted grid map, the triangle inequality no longer holds in many cases. As a result, a more deliberate decision must be made between Path 1 and Path 2. This implies that an explicit cost comparison between both paths is required, and the one that truly yields the shorter total cost should be selected.

The Lazy Theta* algorithm optimistically selects Path 2, which immediately suggests that its performance will degrade significantly, as it was originally designed under the assumption that the triangle inequality always holds.

Similarly, the Strict Theta* algorithm introduced the concept of tightly bounding paths, assuming that changes in the optimal path direction occur only at the edges of obstacles. This assumption becomes invalid with the introduction of weighted grid maps. Nevertheless, the algorithm was also extended to compare the costs of Path 1 and Path 2 before making a final decision.

A weighted grid map was generated with a width of 125 cells and a height of 50 cells. The size of an individual cell was set to 0.4 meters. Every second cell was randomly assigned a weight value. The weight value was randomly selected as a fractional number from 0 to 3 inclusive, and this value was added to the base weight value of – 1. Thus, the weight values will vary from 1 to 4 in every second cell.

The evaluation was conducted by executing 2500 algorithm runs. The start and goal positions were selected randomly. A fixed seed value was set for the pseudorandom number generator to ensure that the paths being searched were identical across all algorithm executions. The final results represent the average path cost and execution time of the pathfinding process. The distance is measured in meters, and the time in milliseconds.

Initially, the execution time and path cost were measured for the algorithms on an unweighted grid map. The results are presented in Table 1

As can be seen from the obtained results, all any-angle algorithms require more time for pathfinding while providing no significant advantages in the resulting path length compared to the smoothed A* algorithm. The smoothed A* executes only 1.13% longer than the original A*, while the cost of the path it finds is only 0.32% higher than that of Strict Theta*, which produces the shortest paths. However, the time required by Strict Theta* is 21.26% greater than that of the original A*.

Although the execution times of all algorithms remain low – within the range of milliseconds – these results highlight the efficiency of the path smoothing operation on a uniform, unweighted grid.

The results of the algorithms on the weighted grid are presented in Table 2.

Table 1 – Execution time and costs [15]

	A*	A* smoothing	Basic Theta*	Lazy Theta*	Strict Theta*
Execution time (ms)	0.7499	0.7584	0.8459	0.8418	0.9093
Execution time (%)	100%	101.13%	112.8%	112.25%	121.26%
Distance (m)	37.031	35.856	35.757	35.764	35.74025
Distance (%)	100%	96.83%	96.56%	96.58%	96.51%

Table 2 – Results of the algorithms [15]

	A*	A* smoothing	Basic Theta*	Lazy Theta*	Strict Theta*
Execution time (ms)	1.3536	1.3756	2.7028	2.996	2.2826
Execution time (%)	100%	101.63%	199.67%	221.34%	168.63%
Distance (m)	50.1497	60.7711	45.92569	60.25556	46.70809
Distance (%)	100%	121.18%	91.58%	120.15%	93.14%

DOI 10.15588/1607-3274-2026-1-12

The execution time of all algorithms increased significantly due to the additional computational overhead required for pathfinding on a weighted grid. Furthermore, the runtime difference between Theta* algorithms and A* became more pronounced.

The smoothed A* algorithm maintained a similar relative runtime compared to the unweighted grid case, which is expected, since the transition to a weighted grid has little effect on the smoothing procedure itself. However, the cost of the resulting path increased by 21%, as the only nodes operated on by the smoothing procedure are those previously identified by the A* algorithm, which does not account for cell weights in the same manner as any-angle pathfinding.

Basic Theta* exhibits a twofold increase in pathfinding time. This is primarily due to the evaluation of the fourth path cost case, which involves applying the fast voxel traversal algorithm to calculate path cost during any-angle traversal across the grid. In return, however, the algorithm achieves an 8.5% reduction in path cost compared to A*, which is a significant improvement – especially when contrasted with the 3% improvement reported in [8]. Considering that the algorithm still operates within a millisecond timescale, this improvement may justify the use of Theta* on a weighted grid.

Lazy Theta*, on the other hand, shows a 20% increase in path cost and a 120% increase in runtime compared to baseline A*. As Lazy Theta* relies on the triangle inequality and optimistically assumes that it always holds, it fails to perform correctly on weighted grids where this assumption no longer holds true. For correct usage, adjustments described in [10] should be considered.

Strict Theta* was expected to behave similarly to Lazy Theta*, since its primary strategy involves identifying tightly wrapping paths – an approach thought to be less relevant in weighted grids, where direction changes are not necessarily constrained to obstacle corners for optimality. However, it appears that the penalty distances enforced by Strict Theta* help guide the search more effectively – allowing it to reach goals faster than Basic Theta*, especially over shorter distances. The trade-off is a slightly higher path cost than that of Basic Theta*.

After conducting additional testing exclusively over long distances (with start and end points placed at opposite edges of a grid twice the original size), the following results were obtained:

Basic Theta*: Execution time = 11.8 ms; Path cost = 115.2 m.

Strict Theta*: Execution time = 17.3 ms; Path cost = 117.1 m.

These results indicate that Strict Theta* can indeed find paths faster than Basic Theta* on a weighted grid – but only when the grid size is relatively small. At a certain threshold, the accumulated penalty distances in the priority queue begin to interfere with pathfinding efficiency.

A promising direction for improving Strict Theta* in the context of weighted grids lies in introducing dynamically adjusted penalty values during direction changes. To achieve this, one must define a penalty computation function tailored to the current map, and then modulate the penalty from 0 up to a predefined value based on the weight of the cell being wrapped. This approach would enable the algorithm to prune more paths that exhibit suboptimal direction changes and yield a more performant implementation – comparable to the results presented in Table 2 – while remaining applicable to grids of arbitrary size.

Figure 7 illustrates paths generated by the A*, A* with smoothing, and Basic Theta* algorithms. It is evident that A* progresses in a “staircase” fashion due to its restriction to eight movement directions between neighboring vertices. A with smoothing* generates more natural-looking paths based solely on A*’s explored vertices, yet this comes at the cost of higher overall path cost. In contrast, Basic Theta* has the flexibility to traverse in any direction between vertices on the grid, which allows it to discover shorter and more visually plausible paths. The advantage of Basic Theta* becomes even more pronounced when weights are overlaid on the grid – clearly showing how the path bypasses high-cost cells (Figure 8).

Given the low absolute execution time (on the order of milliseconds) even for relatively large grid maps, Basic Theta* proves to be a viable solution for large-scale projects. It offers a simple implementation, low-cost paths, and realistic path shapes, making it especially suitable for applications where visual plausibility and path optimality outweigh minimal time savings.

Generalization of Theta* application in weapon supply logistics involves the use of a weighted grid map as the routing space. The delivery region map (e.g., a city or an administrative area) can be divided into a square grid, where each cell is assigned a weight that reflects parameters such as road surface condition (asphalt, cobblestone, dirt), risk level (especially high risk in border zones or areas with increased criminal activity), average traffic, as well as the presence of surveillance cameras, checkpoints, or patrol units, which is particularly relevant under military or special delivery conditions. A detailed description of this procedure implies that the delivery region map can be represented as a Weighted Grid Map, implementing a spatial model of the environment in the form of a two-dimensional grid. In this grid, each cell corresponds to a specific square segment of space (e.g., 10×10 or 50×50 meters – depending on the selected scale).

DOI 10.15588/1607-3274-2026-1-12

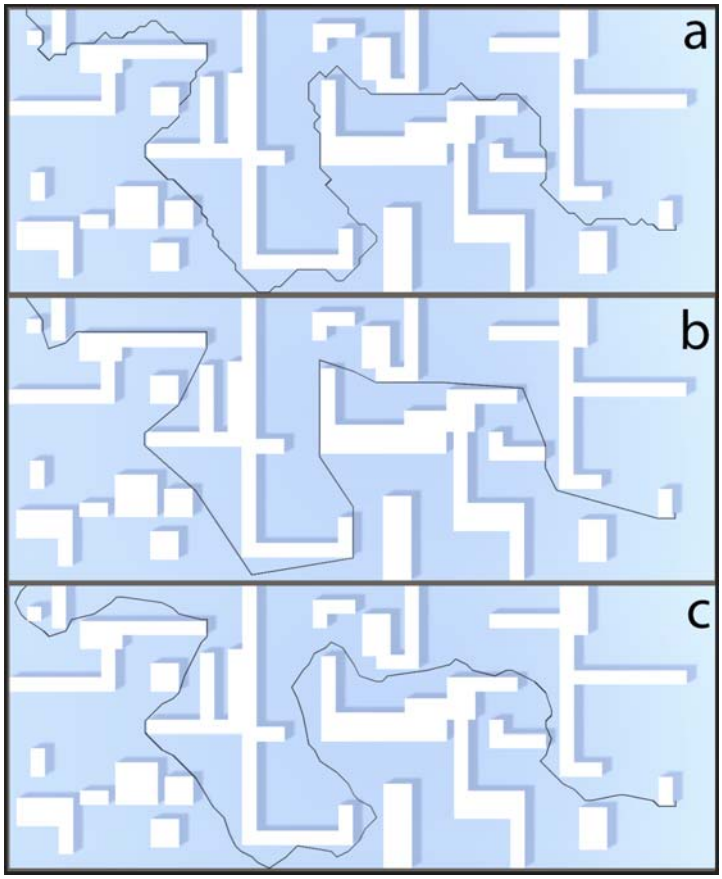


Figure 7 – Paths generated by the algorithms: a – A*, b – A* with smoothing, c – Theta*

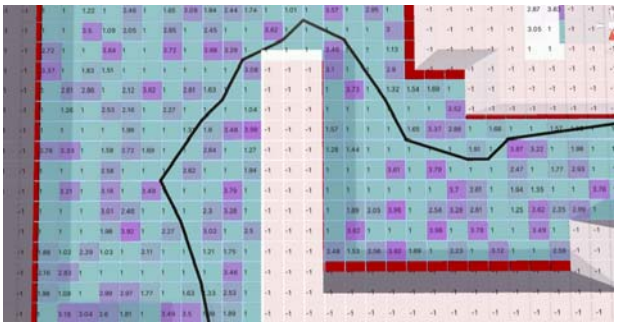


Figure 8 – The path found by the Basic Theta* algorithm is plotted on top of a square grid with weights

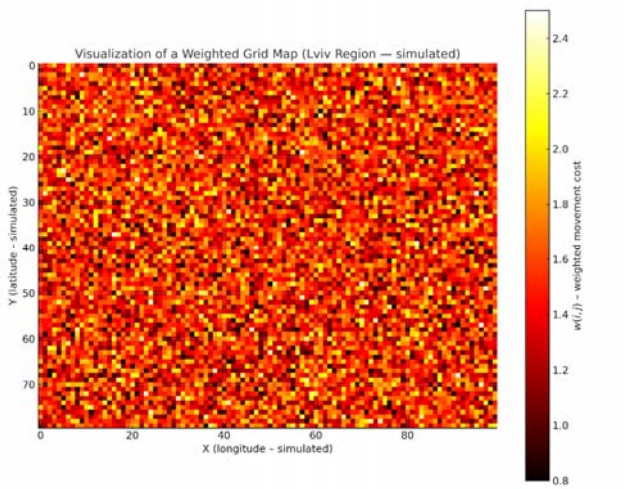


Figure 9 – Visualization of a Weighted Grid Map (Lviv Region – simulated)

5 RESULTS

Each cell is assigned a numerical weight $w(i, j)$, which determines the cost of movement through that cell and directly influences the results of the Theta* pathfinding algorithm. The formula for calculating the cell's weight accounts for the selected factors and may include both road-related and logistics-security characteristics of the environment formula (4).

The weighting coefficients $\alpha, \beta, \gamma, \delta$ in the weighted grid formula are used to determine the relative significance of each factor in computing the total traversal cost of a grid cell (Figure 9). Their selection depends on the routing objectives, application context (e.g., military logistics vs. civilian delivery), and priorities such as safety, time, or cost.

The coefficients $\alpha, \beta, \gamma, \delta$ can be determined using several approaches. The first method is expert-based tuning, in which values are assigned by security specialists, logisticians, or analysts based on the specific nature of the cargo. For example, in the context of weapon delivery, security takes the highest priority, which may result in settings such as $\alpha = 0.5$, since road quality is important but not critical; $\beta = 3.0$, indicating risk is treated as a critical factor; $\gamma = 1.5$, reflecting moderate importance of traffic; and $\delta = 2.0$, emphasizing the impact of surveil-

DOI 10.15588/1607-3274-2026-1-12

lance infrastructure such as cameras and checkpoints on route selection.

The second approach involves normalization and weight summation, where all coefficients are scaled so that $\alpha+\beta+\gamma+\delta=1$. This enables preservation of relative importance while remaining independent of absolute values.

The third method is the Analytic Hierarchy Process (AHP), which involves pairwise comparisons of factors by importance (e.g., security is more important than traffic, thus $\beta>\gamma$), followed by the construction of a priority matrix and computation of a weight vector.

Finally, the fourth approach is data-driven learning. Given access to historical data on routes, risks, and incidents, regression models or neural networks can be applied to derive coefficients that minimize cost, time, or delivery risk. Logistic regression or other machine learning techniques may be used for this purpose.

The coefficients α,β,γ and δ define the routing priorities. They should be adjusted depending on the target scenario: secure delivery, speed, stealth, etc. (Table 4).

The best results are achieved when combining expert analysis, normalization, and – when possible – machine learning based on historical data.

For example, the road surface condition parameter $r_{i,j}$ can be derived from satellite imagery and classified as follows:

– asphalt $r_{i,j}=1.0$;

– cobblestone $r_{i,j}=1.5$;

– dirt road $r_{i,j}=2.5$;

– impassable section $r_{i,j}=\infty$ (blocked).

Table 4 – Example of Weights for Weapon Delivery (High-Risk Scenario)

Parameter	Priority Level	Weight (Expert-Based)
Road Condition r	Medium	α = 0.2
Risk Level t	High	β = 0.4
Traffic s	Low	γ = 0.1
Surveillance v	High	δ = 0.3
Road Condition r	Medium	α = 0.2

The risk level parameter $t_{i,j}$ defines the likelihood of cargo loss or obstacles along the route. For instance:

– low risk (central areas) $t_{i,j}=0.5$;

– medium risk (suburbs, industrial zones) $t_{i,j}=1.0$;

– high risk (border or high-crime zones) $t_{i,j}=2.0-3.0$.

The average traffic parameter $s_{i,j}$ based on Google Traffic data, reflects possible delays caused by congestion or high road occupancy:

– free-flowing traffic $s_{i,j}=0.8$;

– moderate traffic $s_{i,j}=1.2$;

– heavy traffic 1. $s_{i,j}=1.8$.

The surveillance/control parameter $v_{i,j}$, derived from municipal safety maps, indicates the presence of surveillance cameras, checkpoints, or patrol units:

– no surveillance $v_{i,j}=0.5$;

– moderate control $v_{i,j}=1.0$;

– high surveillance (restricted access zones) $v_{i,j}=2.5$.

The optimal delivery route is constructed based on the minimal cumulative path weight, rather than merely the geometric distance. This requires dynamic weight updates – e.g., through APIs such as Google Traffic or public service data. Such spatial modeling allows the integration of logistical risks into the routing process, which is particularly relevant for arms delivery, where security, efficiency, and route realism are critical.

The choice of algorithm depends on the specific needs of the logistics scenario. The Basic Theta* algorithm is appropriate when the priority is to generate a short route that accounts for realistic turns – especially important for delivering heavy or bulky cargo where frequent or sharp maneuvers should be avoided. It is the most versatile option and is well suited for minimizing delivery costs.

Lazy Theta* is not recommended for weighted grid maps due to its lower accuracy in path cost estimation. However, it may be used in real-time applications where rapid path generation is critical.

Strict Theta* is justified in dense urban environments, where the route must accurately bypass obstacles such as buildings, restricted zones, or barriers. This makes it effective for deliveries to central city areas or in situations with frequent movement restrictions.

An important aspect is the customization of metrics and weights: the weight of a cell may incorporate travel cost (fuel expenses, tolls, duties), delay probability (queues, weather), and the risk of loss or confiscation – especially relevant for sensitive cargo such as firearms.

Theta* has several advantages that make it suitable for such use cases: routes appear realistic for navigation; the algorithm adapts well to dynamic risk maps or traffic conditions (Figure 10); it integrates easily with AI-based systems for risk assessment or decision-making.

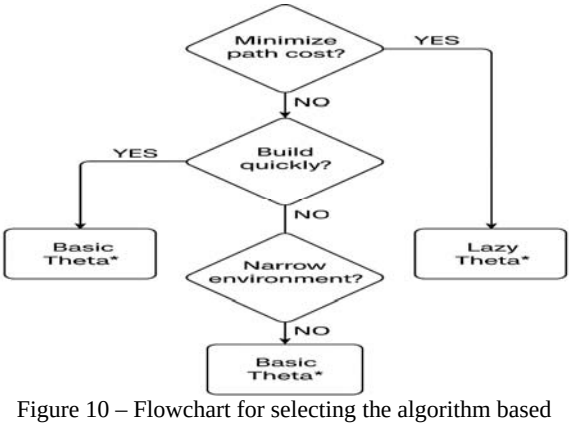


Figure 10 – Flowchart for selecting the algorithm based on operational conditions

DOI 10.15588/1607-3274-2026-1-12

From a technical perspective, the algorithm can be integrated into a Supply Chain Management (SCM) system, enabling: route planning with policy and constraint compliance; integration with GPS tracking; visualization of results in the logistics dashboard; use in simulation-based training for logistics personnel.

The study of Theta* algorithms, particularly their adaptation for weighted grid maps, opens new opportunities for optimizing supply routes – especially for sensitive goods such as weapons, where accuracy, reliability, security, and path realism are of critical importance. When combined with GIS, AI, and risk management systems, these algorithms become the core of modern intelligent logistics.

This UML component diagram illustrates the integration of the routing module into the firearm sales management information system. The main component, WeaponSalesManagementSystem, is responsible for managing the entire order and delivery cycle, as well as for interacting with other modules. The RoutePlanningModule is the central element for delivery route construction and includes three main subsystems (Figure 11):

WeightedGridMapEngine, which generates a weighted grid map of the area where each cell accounts for factors such as road conditions, security risks, traffic congestion, and the presence of cameras or checkpoints;

ThetaStarPathfinder, which implements the pathfinding algorithm (e.g., Basic Theta*) and enables optimal route construction based on direct line-of-sight;

RiskAnalyzer, which evaluates the safety level of the route using data from external sources and integration with the security module.

The SecurityModule verifies whether the proposed route is safe for transporting firearms by analyzing information on restricted zones, checkpoints, high-risk areas, and other constraints. Once the route is validated, it is passed to the LogisticsDashboard, where it is visualized for the logistics operator. This interface allows for comparing alternative routes, viewing satellite imagery, and manually adjusting the route if necessary.

The GISDataProvider is a separate service or API that supplies up-to-date geospatial information (e.g., from Google Maps or government sources), which is required for accurate grid generation and real-time risk evaluation.

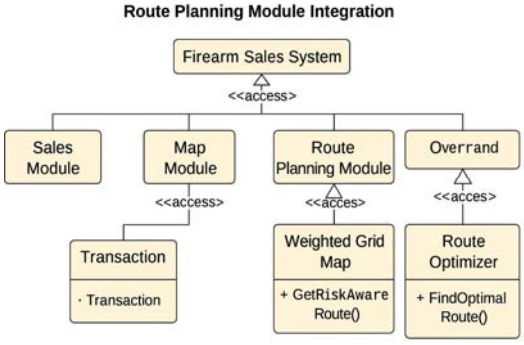


Figure 11 – Route Planning Module Integration

The information system initiates a route planning request [16, 17] and sends it to the RoutePlanningModule, where all processing and analysis are performed. Then, the SecurityModule checks the safety of the route, and finally, the validated route is forwarded to the logistics dashboard. This integration ensures adaptive, secure, and efficient route planning for the delivery of firearms, accounting for real-world risks and environmental constraints.

6 DISCUSSION

The conducted experiments highlighted both the strengths and limitations of applying any-angle Theta* algorithms to weighted grid maps. A key finding is that the Basic Theta* algorithm consistently reduced overall path cost compared to the classical A*, with only a moderate increase in computation time. This confirms earlier research [7] but shows a stronger improvement in weighted environments, which more accurately reflect real-world conditions for safe and efficient route modeling.

By contrast, the Lazy Theta* algorithm proved unsuitable for weighted environments. Its assumption of universal line-of-sight, valid in uniform grids, led to inaccurate cost estimations when weights varied between cells. This caused both runtime and path cost to increase significantly. Lazy Theta* may still be applied in scenarios where rapid computation is critical, but its reliability for cost-aware route planning is limited.

Strict Theta* produced partially unexpected results. While initially thought to offer little benefit in weighted grids, its penalty mechanism improved performance in certain contexts, making its execution time competitive, especially in smaller maps. Although slightly less cost-efficient than Basic Theta*, it offers practical value in constrained or urban-like environments, where precise obstacle avoidance is important.

These findings suggest that none of the Theta* variants is universally optimal across all weighted environments. Instead, algorithm selection should depend on the operational scenario. Basic Theta* is most suitable for general tasks of route modeling where cost minimization and path realism are crucial. Strict Theta* is more relevant in dense environments, while Lazy Theta* is applicable only when rapid path generation outweighs accuracy.

Importantly, the study demonstrates how Theta* algorithms can be integrated into an information system for tracking and managing firearms sales. By representing road conditions, risk levels, and surveillance density as weighted parameters of the grid, the system can model delivery routes that are both safe and efficient. This ensures that the route planning module not only accounts for the shortest distance but also incorporates safety-related constraints – such as avoiding high-risk or restricted areas.

In summary, the presented results advance the theoretical understanding of any-angle algorithms in weighted environments and confirm their practical applicability within intelligent information systems. Specifically, the adaptation of Basic and Strict Theta* algorithms provides a foundation for developing route-planning modules in

DOI 10.15588/1607-3274-2026-1-12

firearm sales management systems, where accuracy, adaptability, and safety of modeled routes are critical.

CONCLUSIONS

This study comprehensively analyzed the efficiency and adaptability of Theta* algorithms (Basic Theta*, Lazy Theta*, and Strict Theta*) when applied to path planning tasks on weighted grid maps. The insights gained not only advance the theoretical understanding of any-angle pathfinding in variable terrain environments but also open up promising avenues for practical implementation – particularly in systems that require high levels of precision, safety, and adaptability.

One of the most promising directions for applying the results of this research is in the development of intelligent route-planning modules within information systems for firearm sales and distribution. In such systems, it is critical to ensure safe, efficient, and legally compliant delivery of weapons across diverse geographic zones, including urban, rural, and high-risk areas.

The integration of the Weighted Grid Map (WGM) model into a firearm logistics information system would enable:

Spatial modeling of delivery routes with real-time adjustments to risk levels, road accessibility, and surveillance density.

Route optimization for safety, using Basic Theta* to avoid high-risk zones, congested areas, or regions with heavy surveillance (e.g., police checkpoints, border patrols).

Dynamic recalculation of routes in case of sudden geopolitical changes or traffic incidents, via integration with real-time data sources such as GIS layers or traffic APIs.

Use of customizable weights and parameters (α, β, γ, δ) for tailoring delivery priorities – e.g., prioritizing stealth in certain operations, or speed and fuel cost in others.

Seamless integration with sales records, enabling automated planning of delivery routes immediately after a weapon transaction is registered in the database.

Furthermore, the results of this study – especially the demonstrated efficiency of the Basic Theta* algorithm on weighted grids – confirm the feasibility of embedding this routing logic into an enterprise-grade information system. This includes systems developed in Java-based frameworks (such as Spring Boot) that already support modules for CRM, logistics, and compliance.

The visual validation via Unity simulations enhances trust in the system's realism and robustness, and such simulations could be used in training environments for logistics personnel, security forces, or system operators.

In conclusion, the Theta* algorithm family – particularly Basic Theta*– can serve as the core routing mechanism in intelligent firearm logistics platforms. When integrated with risk-aware mapping and real-time data, this approach can significantly enhance the security, flexibility, and effectiveness of weapon delivery operations, contributing to national security, regulatory compliance, and operational excellence.

The study highlighted that traditional pathfinding algorithms such as A* may not sufficiently address the complex risk factors and real-time constraints encountered in secure logistics, particularly for the delivery of high-risk goods such as firearms. Therefore, it is recommended that weapon supply chain management systems incorporate weighted grid maps and advanced routing algorithms like Basic Theta*. These approaches better reflect real-world road and threat conditions and provide more realistic and adaptive routing.

Given the findings, it is advisable to train logistics personnel and system developers in the application of spatial modeling, route risk evaluation, and the customization of routing parameters based on regional security and traffic data. Additionally, it is recommended to integrate route planning modules with real-time GIS and security data sources to ensure the adaptability of planned routes to changing field conditions.

Furthermore, as the visual realism and flexibility of Theta* algorithms support decision-making, it is recommended to use these methods not only in operational logistics systems but also in training simulators for logistics officers. This will promote awareness of dynamic routing factors and increase the security of firearm delivery operations.

Finally, regulatory agencies and private logistics providers should collaborate on the standardization of risk-weighted spatial models, enabling broader adoption of intelligent route planning technologies in national and regional weapon distribution systems.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: Yrii Kis contributed to the conceptualization of the study, development and implementation of the Theta* algorithms, simulation experiments in the Unity environment, data collection, and primary drafting of the manuscript; Yrii Shcherbyna supervised the research methodology, contributed to the formalization of mathematical models, analysis of algorithmic performance, and critically reviewed the manuscript for theoretical consistency; Nataliia Kunanets contributed to the system-level design and integration of the routing module into the firearm sales management information system, participated in defining applied use cases, and reviewed the manuscript from the perspective of information systems and applied logistics, and assisted in manuscript editing and technical refinement; Yrii Yarymovych contributed to experimental design, data analysis and visualization, validation of results.

Data availability: The manuscript does not have associated data in a data repository.

DOI 10.15588/1607-3274-2026-1-12

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

REFERENCES

1. Van Den Berg J., Shah R., Huang A., Goldberg K. Anytime nonparametric A, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2011, Vol. 25, № 1, pp. 105–111. Mode of access: https://doi.org/10.1609/aaai.v25i1.7819.
2. Sturtevant N., Geisberger R. A comparison of High-Level approaches for speeding up pathfinding [Electronic resource], *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2010, Vol. 6, № 1, pp. 76–82. Mode of access: https://doi.org/10.1609/aiide.v6i1.12400.
3. Subbotin S. A. Construction of decision trees for the case of low-information features, *Radio Electronics, Computer Science, Control*, 2019, № 1, pp. 121–130.
4. Garcia-Luna-Aceves J., Murthy N. S. A path-finding algorithm for loop-free routing [Electronic resource], *IEEE/ACM Transactions on Networking*, 1997, Vol. 5, № 1, pp. 148–160. Mode of access: https://doi.org/10.1109/90.554729.
5. Lawande S. R., Jasmine G., Anbarasi J. et al. A systematic review and analysis of Intelligence-Based pathfinding algorithms in the field of video games, *Applied Sciences*, 2022, Vol. 12, № 11, P. 5499. Mode of access: https://doi.org/10.3390/app12115499.
6. Hart P., Nilsson N., Raphael B. A formal basis for the heuristic determination of minimum cost paths, *IEEE Transactions on Systems Science and Cybernetics*, 1968, Vol. 4, № 2, pp. 100–107. Mode of access https://doi.org/10.1109/tssc.1968.300136.
7. Fraile-Jurado P., Llovet-Ferrer M., Roig-Munar F. X. Toward Realism: An analysis of coastal environments in Open-World Video Games, *Simulation & Gaming*, 2024, Mode of access: https://doi.org/10.1177/10468781241287900.
8. Daniel K., Nash A., Koenig S., Felner A.Theta: Any-Angle path planning on grids, *Journal of Artificial Intelligence Research*, 2010, Vol. 39, pp. 533–579. Mode of access: https://doi.org/10.1613/jair.2994.
9. Le P. T., Lee K. Weight value and map complexity in Theta* [Electronic resource], *MATEC Web of Conferences*, 2016, Vol. 54, P. 05003. Mode of access: https://doi.org/10.1051/matecconf/20165405003.
10. Rey R., Cobano J. A., Merino L., Caballero F. Generalized Lazy-Theta for 3D path planning considering non-uniform costs, *2022 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2022. Mode of access: https://doi.org/10.1109/icuas54217.2022.9836069.
11. Han S., Wang L., Y. Wang, H. He A dynamically hybrid path planning for unmanned surface vehicles based on non-uniform Theta and improved dynamic windows approach, *Ocean Engineering*, 2022, Vol. 257, P. 111655. Mode of access: https://doi.org/10.1016/j.oceaneng.2022.111655.
12. Nash A., Koenig S., Tovey C. Lazy Theta*: Any-Angle path planning and path length analysis in 3D [Electronic resource], *Proceedings of the AAAI Conference on Artificial Intelligence*, 2010, Vol. 24, № 1, pp. 147–154. Mode of access: https://doi.org/10.1609/aaai.v24i1.7566.
13. Oh S., Leong H. W. Strict Theta*: Shorter motion path planning using taut paths [[Electronic resource], *Proceedings of the International Conference on Automated Planning and Scheduling*, 2016, Vol. 26, pp. 253–257. Mode of access: https://doi.org/10.1609/icaps.v26i1.13744.
14. Amanatides J., Woo A. A Fast Voxel Traversal Algorithm for Ray Tracing. York University, 1987. Mode of access: http://www.cs.yorku.ca/~amana/research/grid.pdf
15. Kis Yu. O. Practicality and efficiency of application of weighted freely directed path finding algorithms Theta*: Qualification (master's) thesis. Lviv, 2024, 54 p.
16. Methods and characteristics of locality-preserving transformations in the problems of computational intelligence, *Radio Electronics, Computer Science, Control*, 2014, № 21, pp. 120–128.
17. Oliinyk A., Subbotin S., Lovkin V. et al. The system of criteria for feature informativeness estimation in pattern recognition, *Radio Electronics, Computer Science, Control*, 2017, № 18, pp. 85–96.

Received 14.10.2025.
Accepted 14.01.2026.
Published 27.03.2026.

УДК 004.421.2:519.6:004.89

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ АЛГОРИТМІВ ANY-ANGLE THETA НА ЗВАЖЕНИХ СІТОЧНИХ КАРТАХ ДЛЯ ПЛАНУВАННЯ МАРШРУТІВ*

Кіс Ю. – аспірант кафедри дискретного аналізу та інтелектуальних систем, Львівський національний університет імені Івана Франка, Україна. ROR: https://ror.org/01s7y5e82. ORCID: https://orcid.org/ 0009-0009-7816-237X.

Щербина Ю. М. – канд. техн. наук, професор кафедри дискретного аналізу та інтелектуальних систем, Львівський національний університет імені Івана Франка, Україна. ROR: https://ror.org/01s7y5e82. ORCID: https://orcid.org/ 0000-0002-4942-2787.

Кунанець Н. Є. – д-р техн. наук, професор кафедри інформаційних систем та мереж, Національний університет «Львівська політехніка», Львів, Україна. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0000-0003-3007-2462.

Яримович Ю. А. – аспірант кафедри інформаційних систем та мереж, Національний університет «Львівська політехніка», Львів, Україна. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0009-0006-1391-3214.

АНОТАЦІЯ

Актуальність. У статті розглядається дослідження алгоритмів пошуку шляхів вільного напряму, зокрема алгоритму Theta*, та оцінюється їхня ефективність на зважених сіточних картах з метою визначення оптимальних маршрутів для доставки товарів до магазину зброї. Це дослідження виконане в ширшому контексті розробки інформаційної системи для відстеження та управління продажем і логістикою зброї в складних умовах. Одним із головних мотивів є те, що методи будь-якого кута (any-angle) здатні генерувати більш реалістичні та природні маршрути порівняно з класичним алгоритмом A*.

Мета. Метою дослідження є аналіз ефективності трьох алгоритмів, заснованих на Theta*: Basic Theta*, Lazy Theta* і Strict Theta*, як на однорідних, так і на зважених квадратних сітках, із особливим акцентом на показниках часу виконання та вартості шляху. Робота спрямована на узагальнення застосовності цих алгоритмів до зважених середовищ і пропозицію удосконалень, придатних для реальних сценаріїв планування маршрутів.

DOI 10.15588/1607-3274-2026-1-12

Метод. Представлено принципи роботи алгоритму A*, трьох варіантів Theta*, а також технік згладжування маршрутів після обчислення. У дослідженні описано перехід від незважених однорідних квадратних сіток до зважених і акцентовано увагу на складності обчислення точних вартостей маршрутів при застосуванні методів будь-якого кута. Візуалізація поведінки алгоритмів була реалізована за допомогою рушія Unity. Показники ефективності вимірювалися окремо для однорідних і зважених сіток, щоб забезпечити порівняльний аналіз.

Результати. Отримані результати включають порівняльну оцінку алгоритмів Basic Theta*, Lazy Theta*, Strict Theta* та класичного A*. Аналіз виявив умови, за яких кожен алгоритм працює ефективно, а також фактори, що обмежують їх застосовність у зважених середовищах. Показано, що довжина маршруту та його вартість можуть суттєво відрізнятися на зважених сітках, що призводить до нових міркувань щодо оптимізації на основі вартості. На основі експериментів запропоновано узагальнення алгоритму Basic Theta* для підвищення його придатності до зважених квадратних сіток, а також окреслено можливе розширення алгоритму Strict Theta* для цього контексту.

Висновки. Результати дослідження показують, що, хоча алгоритми будь-якого кута забезпечують більш плавні та реалістичні маршрути, їх ефективність у зважених середовищах залежить від ретельної адаптації функцій вартості. Дослідження підкреслює їхню цінність не лише для моделювання складних віртуальних середовищ і поведінки агентів у іграх та робототехніці, а й для практичних застосувань у логістиці, зокрема в розробці інформаційної системи для відстеження та управління продажем зброї. Запропоновані алгоритмічні удосконалення можуть сприяти підвищенню ефективності планування доставки та управління ланцюгами постачання, у тому числі моделюванню маршрутів доставки зброї в умовах воєнного часу.

КЛЮЧОВІ СЛОВА: пошук шляху, планування маршруту, квадратна сітка, алгоритм будь-якого кута, вартість шляху, зважена сітка, Theta*.

ЛІТЕРАТУРА

1. Anytime nonparametric A / [J. Van Den Berg, R. Shah, A. Huang, K. Goldberg] // Proceedings of the AAAI Conference on Artificial Intelligence. – 2011. – Vol. 25, № 1. – P. 105–111. – Mode of access: https://doi.org/10.1609/aaai.v25i1.7819.
2. Sturtevant N. A comparison of High-Level approaches for speeding up pathfinding [Електронний ресурс] / N. Sturtevant, R. Geisberger // Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. – 2010. – Vol. 6, № 1. – P. 76–82. – Mode of access: https://doi.org/10.1609/aiide.v6i1.12400.
3. Subbotin S. A. Construction of decision trees for the case of low-information features / S. A. Subbotin // Radio Electronics, Computer Science, Control. – 2019. – № 1. – P. 121–130.
4. Garcia-Luna-Aceves J. A path-finding algorithm for loop-free routing [Електронний ресурс] / J. Garcia-Luna-Aceves, N. S. Murthy // IEEE/ACM Transactions on Networking. – 1997. – Vol. 5, № 1. – P. 148–160. – Mode of access: https://doi.org/10.1109/90.554729.
5. A systematic review and analysis of Intelligence-Based pathfinding algorithms in the field of video games / [S. R. Lawande, G. Jasmine, J. Anbarasi et al.] // Applied Sciences. – 2022. – Vol. 12, № 11. – P. 5499. – Mode of access: https://doi.org/10.3390/app12115499.
6. Hart P. A formal basis for the heuristic determination of minimum cost paths / P. Hart, N. Nilsson, B. Raphael // IEEE Transactions on Systems Science and Cybernetics. – 1968. – Vol. 4, № 2. – P. 100–107. – Mode of access https://doi.org/10.1109/tssc.1968.300136.
7. Fraile-Jurado P. Toward Realism: An analysis of coastal environments in Open-World Video Games / P. Fraile-Jurado, M. Llovet-Ferrer, F. X. Roig-Munar // Simulation & Gaming. – 2024. – Mode of access: https://doi.org/10.1177/10468781241287900.
8. Theta: Any-Angle path planning on grids / [K. Daniel, A. Nash, S. Koenig, A. Felner] // Journal of Artificial Intelligence Research. – 2010. – Vol. 39. – P. 533–579. – Mode of access: https://doi.org/10.1613/jair.2994.
9. Le P. T. Weight value and map complexity in Theta* [Електронний ресурс] / P. T. Le, K. Lee // MATEC Web of Conferences. – 2016. – Vol. 54. – P. 05003. – Mode of access: https://doi.org/10.1051/matecconf/20165405003.
10. Generalized Lazy-Theta for 3D path planning considering non-uniform costs / [R. Rey, J. A. Cobano, L. Merino, F. Caballero] // 2022 International Conference on Unmanned Aircraft Systems (ICUAS). – 2022. – Mode of access: https://doi.org/10.1109/icuas54217.2022.9836069.
11. A dynamically hybrid path planning for unmanned surface vehicles based on non-uniform Theta and improved dynamic windows approach / [S. Han, L. Wang, Y. Wang, H. He] // Ocean Engineering. – 2022. – Vol. 257. – P. 111655. – Mode of access: https://doi.org/10.1016/j.oceaneng.2022.111655.
12. Nash A. Lazy Theta*: Any-Angle path planning and path length analysis in 3D [Електронний ресурс] / A. Nash, S. Koenig, C. Tovey // Proceedings of the AAAI Conference on Artificial Intelligence. – 2010. – Vol. 24, № 1. – P. 147–154. – Mode of access: https://doi.org/10.1609/aaai.v24i1.7566.
13. Oh S. Strict Theta*: Shorter motion path planning using taut paths [Електронний ресурс] / S. Oh, H. W. Leong // Proceedings of the International Conference on Automated Planning and Scheduling. – 2016. – Vol. 26. – P. 253–257. – Mode of access: https://doi.org/10.1609/icaps.v26i1.13744
14. Amanatides J. A Fast Voxel Traversal Algorithm for Ray Tracing / J. Amanatides, A. Woo. – York University, 1987. – Mode of access: http://www.cs.yorku.ca/~amana/research/grid.pdf
15. Kis Yu. O. Practicality and efficiency of application of weighted freely directed path finding algorithms Theta*/ Yu. O. Kis. – Qualification (master's) thesis. – Lviv, 2024. – 54 p.
16. Subbotin S. A. Methods and characteristics of locality-preserving transformations in the problems of computational intelligence / S. A. Subbotin // Radio Electronics, Computer Science, Control. – 2014. – № 21. – P. 120–128.
17. The system of criteria for feature informativeness estimation in pattern recognition / [A. Oliinyk, S. Subbotin, V. Lovkin et al.] // Radio Electronics, Computer Science, Control. – 2017. – № 18. – P. 85–96.

DOI 10.15588/1607-3274-2026-1-12

UDC 004.415.2

DEVELOPMENT OF A CLASS STORAGE REPOSITORY FOR OBJECT-ORIENTED SOFTWARE DEVELOPMENT TECHNOLOGIES

Kungurtsev O. B. – PhD, Professor of the Software Engineering Department, Odessa Polytechnic National University, Odessa, Ukraine. ROR: https://ror.org/05xaz0w84. ORCID: http://orcid.org/0000-0002-3207-7315.

Novikova N. O. – PhD, Associate Professor of the Department of Technical Cybernetics and Information Technologies named after professor R. V. Merct, Odessa National Maritime University, Odessa, Ukraine. ROR: https://ror.org/05qze6v15. ORCID: https://orcid.org /0000-0002-6257-9703.

Buhaeva I. G. – PhD, Associate Professor of the Department of Technical Cybernetics and Information Technologies named after Professor R. V. Merct, Odessa National Maritime University, Odessa, Ukraine. ROR: https://ror.org/05qze6v15**.** ORCID: https://orcid.org /0000-0002-2839-9266.

Vytnova A. I. – Master of the Software Engineering Department, Odessa Polytechnic National University, Odessa, Ukraine. ROR: https://ror.org/05xaz0w84. ORCID: https://orcid.org /0000-0002-4224-3006.

ABSTRACT

Context. Using previously developed code, particularly software classes and class groups related through inheritance, aggregation and composition in object-oriented technologies, significantly reduces software design time.

Objective. Problems arise when it is necessary to store classes for different purposes. In this case, the class name cannot serve as a characteristic for searching. Creating a special structure linking the class name with its purpose will significantly complicate the class repository design. Besides, if relations connect some classes with other classes, there is a need to store their relations along with the class itself, which can significantly complicate both the placement of the class in the repository and its further search. This paper aims to create a special repository of software classes, in which a class is represented by a model that defines its purpose, possible relations with other classes, and role in these relations.

Method. A mathematical model of a software class has been developed, which allows for determining the class designation and possible inter-class relationships. A method of automated placement and search of individual classes and class groups in the class repository is proposed. A storage model is proposed for placing both individual classes and groups of classes connected by inheritance, aggregation and composition relationships. A mechanism has been developed for the automated addition of a separate class or group of classes to the repository, as well as for searching and deleting a separate class or group of classes.

Result. The *ClassCall programme* has been developed to test the proposed solutions. Experiments were conducted to determine the time and quality of placement operations and search of classes and class groups in the repository. The results showed a significant reduction in time for class search compared to known libraries.

Conclusions. The proposed method of automated placement and search of classes based on the class model allows for maintaining class versions, significantly reducing the time to search for the required class and related classes. The method can be used for classes in various object-oriented languages.

KEYWORDS: object-oriented programming, class, class search, composition, inheritance, aggregation, class groups, syntactic analysis.

ABBREVIATIONS

OOP – object-oriented programming.

NOMENCLATURE

abstract is an abstract class
aggreg is a class transformed for aggregation;
attrName is an attribute identification;
attrPurpose is an attribute purpose/designation;
attrType is an attribute type;
cCode is a class code;
cHead is a class header;
cName is a class name;
cNameMain is a main class name;
cNameObject is an object class name;
cNameSource is a source object class name;
cNew is a new introduced class;
compos is a class transformed for composition;
cPurpose is a purpose of class usage;
DictEntry is a dictionary entry;
endAddr is an address of end of method in class code;
fName is a function (method) name;
fPurpose is a function (method) purpose;
fText is a function (method) text;
KeyDict is a key words dictionary;
keyword is a key word;
lang is a programming language;
mC is a set of separate software classes;
mCharacter is a set of necessary characteristics;
mCr is a set of relevant classes;
mGrC is a set of class groups related by some relations;
mGrCr is a set of relevant class groups related by some relations;
mModelC is a model for representing a class;
ordinary is an ordinary class;
queue is a class implementing a queue for aggregation;
Rel is a relation that unites a group;
relation is a relations with other classes;
RepositClasses is a class repository;
rText is a request text;
returnVal is a return value of a function;
Role is a role of class in relation;
sAggred is a set of class names used to implement aggregation;
sArgs is a set of method arguments;

DOI 10.15588/1607-3274-2026-1-13

sAttr is a set of class attributes;

sCharacter is a set of desired characteristics for a searched class;

sClassCode is a set of program class codes;

sClient is a set of client class names;

sCompos is a set of class names used by a given class to implement composition;

sCr is a discovered set of relevant classes;

sGrCr is a discovered set of relevant class groups;

sMClass is a set of class models;

sMeth is a set of functions (methods) of a class;

sResponseClasses1 is a preliminary set of candidate classes for answering a query;

sRKeyword is a set of keywords extracted from a query;

sRsArgs is a set of arguments that return the result of a calculation;

sSynonym is a set of synonyms of a keyword;

startAddr is an address of the start of the method in the class code;

type is a class type;

version is a class version.

INTRODUCTION

When creating object-oriented software, it is possible to significantly reduce design time by using previously developed classes. However, it often turns out that the time to find the required class exceeds the time of its creation. Existing class libraries [1, 2] are usually limited to a specific programming language and a narrow class specialisation (strings, graphical interface, working with files, etc.). Therefore, even within the same design organisation, projects in different subject areas can be carried out, and corresponding class sets can be created. Creating a class for a certain subject area does not mean that it cannot be successfully used in another subject area. Building a multi-subject repository for software classes gives rise to the problem of placing and searching for classes in such a repository.

To solve a certain design problem, interaction relationships of varying degrees of stability are established between classes.

The inheritance relationship (Fig. 1) implies that the derived class always “knows” its parent [3]. When using a class repository, a link from the parent class to the derived class may be needed (for example, to select a class with some additional capabilities), which is not supported by programming languages.

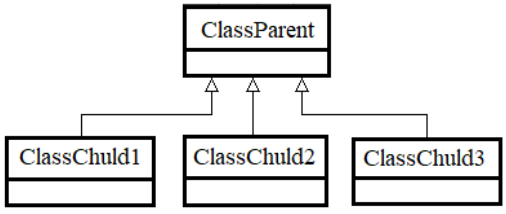


Figure 1 – The inheritance relationship

Moreover, there should be some means of identifying possible inheritance relations among some subsets of repository classes [3]. A composition relation arises if it is necessary to extend the functionality of some class *ClassMain* by using another class *ClassAttr* as an attribute, which *ClassMain* manages alone (Fig. 2a)). In this case, some modification of the attribute class [4] may be required, which is shown in Fig. 2b). Here there may be a need to indicate the availability of this relation in the main class, to place the transformed class-attribute (*ClassAttrConv*) in the repository along with the original class-attribute (*ClassAttr*) and for the latter to specify the class it serves.

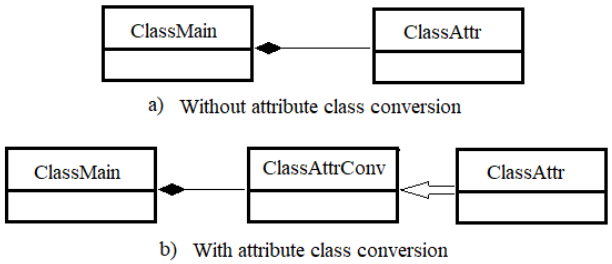


Figure 2 – Composition relationship

The aggregation relation also allows some *ClassMain* to use another class as an attribute (Fig.3a)). However, the *ClassMain* does not create the *ClassAttr* object. This may lead to the use of the *ClassAttr* by multiple *ClassMain* classes. Such a problem is solved by modernising *ClassMain* classes and using a queue class (*ClassQueue*) [5] (Fig. 3b)). Thus, the representation of an aggregation relation in a repository requires placing a number of classes with certain roles in it and fixing the relations between them.

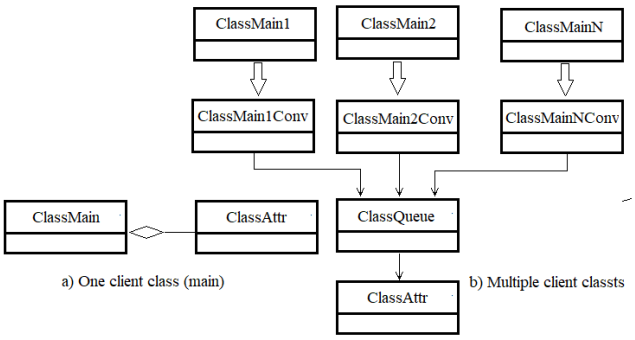


Figure 3 – Aggregation relationship

Given all the above-mentioned, there is a problem of placing and searching in the repository of programme classes intended for solving problems in different subject areas and such class groups connected by certain relations.

The object of study is the processes of defining relationships between software classes, placing and searching for groups of classes connected by relationships in repositories.

DOI 10.15588/1607-3274-2026-1-13

The subject of study is models and repositories of software classes for OOP technologies.

The research aims to reduce the time spent on placement, search, and transformation of classes within the class repository. For this purpose, the following tasks should be solved:

– to create a class model;

– to develop a mechanism for placing a class (class group connected by a certain relation) in the class repository;

– to develop a mechanism for searching a class (class group) by specified characteristics.

1 PROBLEM STATEMENT

Let assume there be a set of separate software classes represented by their codes

$$mC=\{cCod_1, \dots, cCod_n\},$$

as well as a set of class groups linked by some relations

$$mGrC=\{GrC_1, \dots, GrC_k\}.$$

Each group has the form:

$$GrC_i = <Rel, \{<cCod_1, Role>, \dots, < cCod_m, Role>\},$$

where *Rel* is the relation uniting the group,

– *Role* is the role of each class in the relation.

mC and *mGrC* should be placed in a storage repository:

$$(mC, mGrC) \rightarrow Storage.$$

To ensure the placement and search of elements in the storage, it is proposed to use the *ModelC* model for each class, containing its characteristics *sCharacter* . Then the storage can be represented by a tuple

$$Storage = <mModelC, mC, mGrC>.$$

This makes it possible to organize the search for the required class by a request to the repository storage of the form:

$$Request\ (sCharacter) \rightarrow mModelC = sCr,$$

where *sCharacter* is a set of desired characteristics for the sought class,

– *sCr* is a discovered set of relevant classes.

A similar query has a form for finding a group of classes related by the *Rel* relation:

$$Request\ (sCharacter) \rightarrow mModelC = sGrCr,$$

where *sGrCr* is the discovered set of relevant class groups.

2 REVIEW OF THE LITERATURE

Identifying its type is the first task of organising a class repository. Traditionally, program classes are stored in libraries. Such libraries provide ready-made, tested classes for solving everyday tasks like working with text strings, files, graphical interfaces, etc. [2]. They also implement basic functionality that can be used in different projects, and are usually well documented and have a stable API, which makes them easy to use. However, libraries are focused on a specific programming language, so their reuse is limited to a specific platform [1]. In addition, the classes in such libraries are usually highly specialised and cannot always be adapted to new needs, and the relationships between classes in libraries are usually limited to inheritance, which makes it challenging to integrate them into complex systems.

The paper [6] considers the possibility of storing heterogeneous information in a library, but it does not provide for storing groups of related elements with a specific role in the group belonging to the same programming language. The task of evaluating a library from the user's point of view is analysed in [7]. However, among the proposed criteria, there is no evaluation from the point of view of the ability to structure information.

The paper [8] focuses on the issue of creating metadata for storing UML diagrams in a repository. We believe that in the framework of our research, such metadata can be a model of a software class.

Some authors [9, 10, 11] analyse the capabilities of known repositories for software. Undoubtedly, it is possible to organise a repository of classes and class groups within the repository capabilities. However, such a solution seems to be excessively costly.

The second task of building a class repository is the mechanism of placement and search of stored elements (classes, class groups). The efficiency of searching code in natural language using different methods is investigated in the paper [12]. The task can be significantly simplified if, when loading an element into the repository, we use a class model supplemented with natural language text about the purpose of the class and its functions (methods) [13]. This solution can be taken as a basis for searching for the required class in the repository.

However, given the use of classes in different subject areas, the problem of matching the query and class description texts arises. In the study [14], an adaptive text comparison method is proposed, but it is applicable when the degree of overlap between the elements of the compared texts is high. A more universal solution, based on the pre-definition of terms, is proposed in [15]. The study [16] solves the problem by identifying lexically similar words. A solution based on clustering and fuzzy string comparison is proposed in [17]. Such solutions can be adopted if the subject domain is defined and there are enough texts for preliminary analysis. In our case, the set of subject areas is a condition of the problem. In our opinion, a simplified term extraction procedure and an improved method for calculating the Levenshtein distance

DOI 10.15588/1607-3274-2026-1-13

for string comparison, as a universal and quite effective one, should be addressed [18].

When placing and searching classes in the repository, the task of code fragment extraction and comparison may arise. The task of code retrieval has recently become very popular [19]. In [20], an approach to cross-language code search is proposed, which is based on building and scanning a unified abstract syntax tree (UAST) to automate code comparison partially. The method is quite universal, but it seems to be too complicated for our task.

Deep learning methods for determining the semantic properties of programs are analysed in [21]. The methods can give a general characteristic of a program, whereas a class repository needs information about class elements. Comparison of code fragments based on functions, control statements and data types is considered in [22]. A similar task, but in relation to analysing the class structure and selecting its separate elements, was solved in [4], which can be taken as a basis for developing a class repository.

An interesting idea is proposed in [23], where the efficiency of software reuse is determined based on software metrics. However, with respect to the class repository, the metrics-based definition of class inclusion in the repository seems to be too risky.

3 MATERIALS AND METHODS

Repository and class models. The repository model contains a set of class representations (models), a set of terms, and a set of class codes. The class model creates a class representation (class metadata). The purpose of the model is to ensure the class is loaded into the repository and searched for upon request (Fig. 4). The proposed model is a further development of the class model proposed in [13]. A significant development of the proposed model is the inclusion of information about relationships with other classes and the role of the class in relationships. The set of terms ensures a preliminary search for classes in the storage. The extraction of terms (keywords) from the text is discussed in detail in [24].

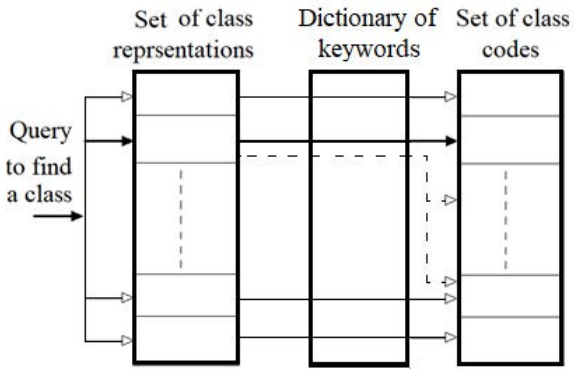


Figure 4 – Class (class group) search scheme)

The class repository is represented as a tuple

$$RepositClasses = <sMClass, KeyDict, sClassCode>, \quad (1)$$

where *sMClass* – set of class models; *KeyDict* – dictionary of keywords; *sClassCode* – set of program class codes.

The *KeyDict* dictionary contains many entries

$$KeyDict = \{DictEntry_1, \dots, DictEntry_k\}.$$

Each entry is represented as a tuple:

$$DictEntry_i = <keyword, sSynonym>,$$

where *keyword* is a key word and *sSynonym* is a set of synonyms for this keyword.

To find the required class, each class is represented by its general purpose of use and the purpose of using each method. Information about the class's purpose, the class type and the relationships will be placed in the class header. Information about the class methods will be placed in the methods view. Class attributes are not planned to be used in the class search procedure, but it makes sense to enter them into the model as additional information for the user when choosing the best class if the search result is a group of classes.

A class is represented by a tuple:

$$c = < cHead , sMeth, sAttr , sKeyword >, \quad (2)$$

where *cHead* – class header; *sMeth* – set of functions (methods) of a class; *sAttr* – set of attributes of a class; *sKeyword* – set of keywords used in the preliminary stage of class search.

Class header. It is proposed to include the general purpose of using the class in the class header as a separate sentence to compare classes.

Thus, the class header is represented as a tuple:

$$cHead=< cName, cPurpose, type, relation, lang, version>, \quad (3)$$

where *cName* – class name; *cPurpose* – purpose of using the class; *type* – class type; *relation* – relationship with other classes: *lang* – programming language; *version* – class version.

The class type can take the following values: *abstract* – abstract class; *ordinary* – ordinary class; *compos* – class transformed for composition; *aggreg* – class transformed for aggregation; *queue* – class implementing a queue for aggregation.

The type of the *relation* element depends on the type of the class.

DOI 10.15588/1607-3274-2026-1-13

For *abstract* and *ordinary* type classes, the relation has the following form:

$$relation =< sParent, sCompos, sAggred>, \quad (4)$$

where *sParent* – set of parent class names for a given class; *sCompos*– set of class names used by a given class to implement composition (object classes can be ordinary classes or classes transformed for composition); *sAggred* – set of class names used by a given class to implement aggregation (object classes can be ordinary classes or classes transformed for aggregation).

For a *compos-type* class, the relation has the following form:

$$relation =< cNameMain, cNameSource>, \quad (5)$$

where *cNameMain* – main class name; *cNameSource* – source object class name

For a *queue-type* class, the relation has the following form:

$$relation =< sClient, cNameObject>, \quad (6)$$

where *sClient* is a set of names of client classes (classes willing to access the object class); *cNameObject* – object class name.

Class attributes. To understand the essence of a class attribute, it is proposed that the concept of the purpose (designation) of the attribute and its type be introduced into the model.

As a result, each attribute from the set *sAttr* will be represented as:

$$Attr=< attrName, attrPurpose, attrType>, \quad (7)$$

where *attrName* – attribute identifier; *attrPurpose* – attribute purpose; *attrType* – attribute type.

Class methods. The main information for searching classes is in the class methods. Therefore, it is proposed that each method's designation be formulated as a short phrase, for example, "calculate the cost of the order". For the method's arguments, the earlier attribute rules should be used.

As a result, each method from the set *sMeth* (2) will take the form as follows:

$$func=< fName, fPurpose, fText, sArgs, returnVal, sRsArgs, startAddr, endAddr>, \quad (8)$$

where *fName* – function (method) name; *fPurpose* – function (method) purpose; *fText* – function (method) text.

sArgs – set of method arguments; each argument is represented by: an identifier *id*, a type *argType*, and a purpose *argPurpose*; *returnVal* = <*retType*, *purpose resp*> – the return value of a function (represented by the value type and target); *sRsArgs* – set of arguments that return the result of a calculation; *startAddr* – address of the start of the method in the class code; *endAddr* – address of the end of the method in the class code.

Repository functions. The following procedures are defined for the classes in the repository:

– Add a class;
– Add a class group;
– Search a class;
– Search a class group;
– Upgrade a class;
– Delete a class;
– Delete a class group.

Add a class. After entering a new class *cNew*, its code is compared with the codes of classes in the repository. If a match is found between the codes cNewCode=cCode$_i$, then the input operation is cancelled. If a match of class names is found cNew.cName=c.cName$_i$, then the specialist must change the name of the loaded class.

The methods *cNew.sMeth* are automatically extracted from the class code.

The specialist must formulate the purpose of the class *cNew.cHead.cPurpose* within one short phrase. The purposes of each method *cNew.func$_i$.fPurpose*, implementing the main purpose of the class, should be formulated in a similar way. Based on the purpose of the class and its methods, the program creates keywords for the class *cNew.sKeyword* and replenishes the dictionary of keywords of the *KeyDict* storage repository.

$$KeyDict.\ sKeyword \cup cNew.sKeyword.$$

Add a class group. The entry operation is cancelled if a class group is entered and all the codes match the group in the repository.

If a class group related by an inheritance relationship is entered, and a match is observed for the code of the parent class, then the derived classes are entered according to the rules for entering a separate class.

If a class group related by a composition relationship is entered, and the codes of the main classes match, but the codes of the attribute classes do not match, then a specialist is called in to correct the error.

If a class group related by an aggregation relationship is entered, and a mismatch in the code is detected for any class in the group, then the entire group is entered under a new name.

If a match in the class names is detected, then the specialist must change the name of the loaded class.

For each class in the group, the operations of describing the class purpose and its methods are performed. The program, based on the purpose of each class and its methods, creates keywords for the class.

Search a class**.** To search for a class, the user must formulate a query/request in the form of text that formulates the purpose of using this class

$$Request = rText,$$

rText– request text.

DOI 10.15588/1607-3274-2026-1-13

The system forms a list of keywords from the query nouns

$$Request \rightarrow < rText, sRKeyword >,$$

where *sRKeyword* – set of keywords extracted from a query.

The system detects the occurrence of query keywords in the repository's *KeyDict* keyword dictionary. If some query keyword is detected in the dictionary,

$$RKeyword_j = DictEntry_i$$

then *sRKeyword* is being updated with synonyms for this keyword.

$$sRKeyword \rightarrow sRKeyword \cup DictEntry_i.sSynonym_i.$$

Based on the comparison of *sRKeyword* with the lists of class keywords, a preliminary set of candidate classes for the response to the request *sResponseClasses1* is formed.

The next step is to compare the text (fuzzy string comparison) of the *rText* request with the texts of the purpose of the $cPurpose_i$ class and its $fPurpose_{i,j}$ methods for each class from *sResponseClasses*1.

As a result, a response to the request is received in the following form:

$$sResponseClasses1 \rightarrow sResponseClasses2 = \{ClassCode_1, \dots, ClassCode_q\},$$

where *q* is set by user.

The resulting list of candidate classes is ranked by the degree of compliance with the query. The user can view the class code from the list.

Additional restrictions can be added to your query, for example, by specifying the class language: *Request* = < *rText, lang* >.

Search a class group. Searching for a class group starts with searching for a single class.

For example, the parent class. Unlike searching for a single class, the relationship of that class to other classes in the group should be added to the query:

$$Request = < rText, relation, type >.$$

The search for a class is performed in the previously specified sequence. Then, at the user's request, it is possible to see the entire group of classes connected by the specified relationship.

4 EXPERIMENTS

In accordance with the proposed model, a program for conducting experiments, *ClassCell*, was developed. Fig. 5 shows the process of loading a class into the repository. The *Composition converter* block [4] was used as a code analyser, and the *TerEx* [24] software product was used to extract keywords.

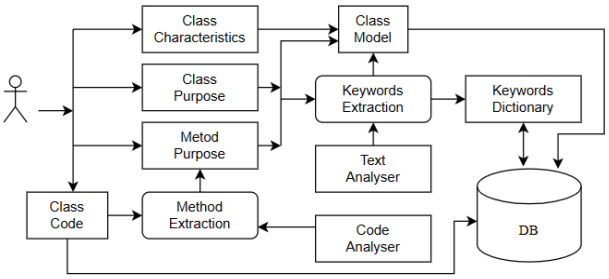


Figure 5 – Scheme of loading a class into a repository

Figure 6 shows the process of searching for a class in the repository. Based on a keyword search, a set of classes, *sClasses1*, is formed at the first stage. At the second stage, as a result of comparing the query texts and class models, the set of proposed classes is reduced to *sClasses2*.

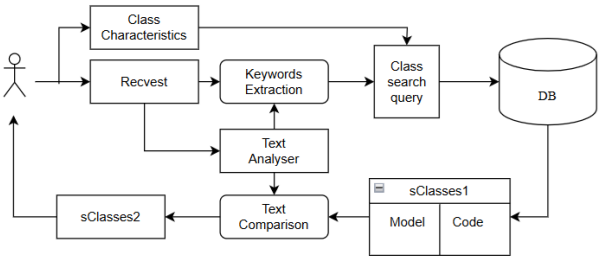


Figure 6 – Searching for classes in the repository

Figure 7 shows the first step of loading a new class into the repository. The user can copy the class text or enter it as a separate file. At the same time, the purpose of the class should be entered.

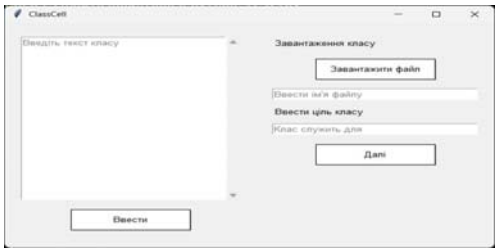


Figure 7 – The first window of the procedure for loading a class into the repository

Fig. 8 shows the window that the user sees after entering a class search query. The program offers a list of classes that were found. If the user wishes, he can see class codes, class methods, and class and method assignments.class codes, class methods, and class and method assignments.

DOI 10.15588/1607-3274-2026-1-13

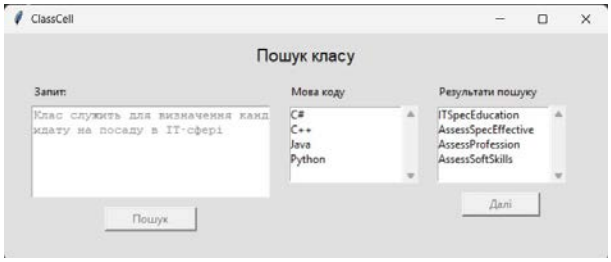


Figure 8 – The first window of the procedure for searching a class in the repository

5 RESULTS

The *ClassCell* program allowed us to conduct a series of experiments to determine the time of loading and searching for classes (100 classes in total). The study's results were the time characteristics and completeness of query responses. Loading a class requires time from 3 to 20 minutes. The average time was 7.9 minutes. At the same time, loading the code was performed in an average of 20.1 seconds. The rest of the time was spent on formulating the purpose of the class and its methods.

The time to search for a class in the repository is from 2 to 7 minutes. Basically, this is the time for composing the query text and viewing the detected classes, since no more than 10 seconds were spent directly searching for a class in the repository.

The experiment showed that increasing the number of classes virtually has no effect on the time needed to search for a class. If we assume that the time for a class search in an unindexed library is about 5 minutes per class, then from the graph (Fig. 9) it follows that even with 10 classes in the repository, the time for a class search in it is 2.5 times less than in the library.

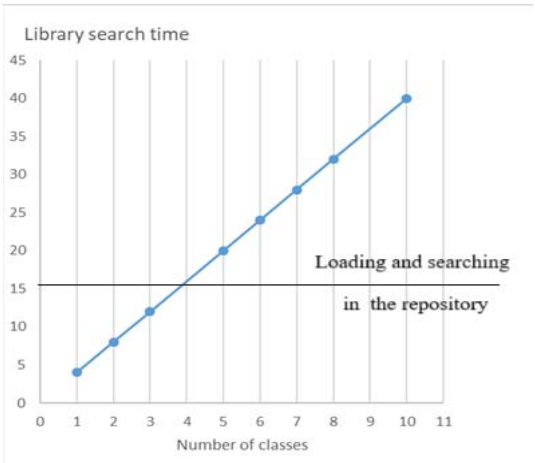


Figure 9 – Time to search for classes in the library and repository

In all cases, the set of classes representing the search result contained all classes relevant to the query subject area, unless a limit was imposed on the number of classes in the query response.

6 DISCUSSION

We assume that the user who uploads a class to the repository is the developer of this class. Then the formation of information about the class designated purpose and methods will not take much time. In addition, we do not limit the volume of this information and do not require precise wording. The more text there is in the class description and in the request for its search (of course, from the subject area under consideration), the more terms will be highlighted, the more likely it is that all the "targeted classes" will be found. At the same time, we are studying the possibility of using artificial intelligence to formulate the purpose and capabilities of a class based on its code.

It should be noted that existing libraries do not offer mechanisms for working with class groups, while they can be partially or entirely reused.

The rationale for limiting the number of classes in a query response remains an open issue.

CONCLUSIONS

A model of a software class has been developed which, unlike existing ones, allows for the organisation of groups of classes linked by inheritance, composition and aggregation relationships.

A mechanism for placing classes in a repository has been developed, providing for automatic replenishment of the dictionary of terms, separate storage of the code and class model with keywords and possible links to other classes involved in the relationship.

A mechanism for class searching based on a natural language query has been developed, providing a two-stage process of searching for relevant classes. At the first stage, it was done by keywords, at the second stage – by comparing the query texts and the class description, the third stage (if necessary) implied determining all classes associated with the desired, specific relationship

The software was created to perform experiments. The experiments showed that, with 10 classes already in the repository, a significant reduction in the time for loading and searching for classes was observed. In all cases, the set of classes representing the search result contained all classes related to the subject area of the query, if no limit was imposed on the number of classes in the response to the query.

The conducted research can serve as a basis for constructing repositories of software classes in design organisations using object-oriented technologies.

ACKNOWLEDGEMENTS

We express our gratitude to the students Bondar Valeriia R. and Gratilova Kateryna O. for their active participation in the development of the Composition Convertor software product, as well as to Master's student Mileiko I.I. for her outstanding contribution to the development of the *TerEx* program. The use of these software solutions significantly reduced the time for creating a program for *ClassCell* experiments.

DOI 10.15588/1607-3274-2026-1-13

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: Oleksii Kungurtsev: development of repository and class models; Nataliia Novikova: developing procedures for adding and deleting classes from the repository; Iryna Buhaieva: developing procedures for searching for a class and a group of classes in a repository; Alina Vytnova: experimental study of methods.

Data availability: All data is provided in the article.

Software availability: Software can be provided upon request.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

REFERENCES

1. Microsoft Corporation. .NET Class Libraries. Microsoft Learn [Electronic resource], 2022. Mode of access: https://learn.microsoft.com/en-us/dotnet/standard/classlibraries. free (date of the application: 01.12.2022). Header from the screen.
2. Krill P. 12 top-notch libraries for C++ programming [Electronic resource]. Electronic text data, InfoWorld, 2022. Mode of access: https://www.infoworld.com/article/2265967/12-top-notch-libraries-for-c-plus-plus-programming.html. free (date of the application: 14.10.2022). Header from the screen.
3. Lee G. Modern Programming: Object Oriented Programming and Best Practices. Packt Publishing, 2019, 266 p.
4. Kungurtsev O. B., Bondar V. R., Gratilova K. O. et al. Method Automated Class Conversion for Composition Implementation, *Radio Electronics, Computer Science, Control*, 2024, №2, pp. 142–149. DOI: 10.15588/1607-3274-2024-2-14
5. Kungurtsev O., Komleva N. Implementation of class interaction under aggregation conditions, *Eastern-European Journal of Enterprise Technologies*, 2024, №2 (128), pp. 20–30. DOI: 10.15587/1729-4061.2024.301011
6. Perhac P., Simonak S. Algorithms and data structure libraries for JAVA, *Acta Electrotechnica et Informatica*, 2020, Vol. 20, No. 1, pp. 39–48. DOI: 10.15546/aeei-2020-0006
7. Tanzil M. H., Uddin G., Barcomb A. "How do people decide?": A Model for Software Library Selection, 1*7th International Conference on Cooperative and Human Aspects of Software Engineering, June 2024*. DOI:10.1145/3641822.3641865
8. Di Felice P. Paolone G., Paesani R. et al. Design and Implementation of a Metadata Repository about UML Class Diagrams. A Software Tool Supporting the Automatic Feeding of the Repository, *Electronics*, 2022, № 11, P. 201. DOI: 10.3390/electronics11020201
9. Dobrzyński B., Sosnowski J. Text Mining Studies of Software Repository Contents, *18th International Conference on Evaluation of Novel Approaches to Software Engineering*, 2023, pp. 562–569. DOI: 10.5220/0011970100003464 ISBN: 978-989-758-647-7; ISSN: 2184-4895
10. Polaczek J., Sosnowski J. Exploring the software repositories of embedded systems: An industrial experience, *Information and Software Technology*, 2021, Vol. 131, P. 106489. DOI: 10.1016/j.infsof.2020.106489
11. Moya J. Repository Software in Software Development [Electronic resource]. Electronic text data. Mode of access: https://www.wearecapicua.com/blog/software-repository-engineering-development. free (date of the application: 25.10.2024). – Header from the screen.
12. Yan S., Yu H., Chen Y. et al. Are the Code Snippets What We Are Searching for? A Benchmark and an Empirical Study on Code Search with Natural-Language Queries, *27th International Conference on Software Analysis, Evolution and Reengineering (SANER)*. Singapore Management University, Singapore, 2020, pp. 344–354. DOI: 10.1109/SANER48275.2020.9054840
13. Kungurtsev O. B., Vytnova A. I. Determination of Inheritance Relations and Restructuring of Software Class Model in the Process of Developing Information Systems, *Radio Electronics, Computer Science, Control*, 2022, № 4, pp. 98–107. DOI: 10.15588/1607-3274-2022-4-8.
14. Kaufman A. R., Klevs A. Adaptive Fuzzy String Matching: How to Merge Datasets with Only One (Messy) Identifying Field, *Published online by Cambridge Universit. Political Analysis*, 2022, Vol. 30 , Issue 4, pp. 590–596.
15. Smirnov A., Shilov N., Evers K. et al. Free Text Customer Requests Analysis: Information Extraction Based on Fuzzy String Comparison, *17th IFIP International Conference on Product Lifecycle Management (PLM)*. Rapperswil, Switzerland, Jul 2020, pp. 193–202. DOI:10.1007/978-3-030-62807-9_16
16. Kalyanathaya K. P., Akila D., Suseendran G. A Fuzzy Approach to Approximate String Matching for Text Retrieval in NLP, *Journal of Computational Information Systems*, 2019, Vol. 15(3), pp. 26–32.
17. Kostanyan A., Harmandayan A. Fuzzy Segmentations of a String [Electronic resource], Electronic text data, Cornell University, 2022. Mode of access: https://doi.org/10.48550/arXiv.2201.13427. free (date of the application: 31.01.2022). Header from the screen.
18. Pikies M., Ali J. Analysis and safety engineering of fuzzy string matching algorithms, *ISA Transactions*, 2020, Vol. 113, P. 1. DOI:10.1016/j.isatra.2020.10.014.
19. Liu C., Xia X., Lo D. et al. Opportunities and Challenges in Code Search Tools [Electronic resource]. Electronic text data, 2020. Mode of access: https://arxiv.org/abs/2011.02297. free (date of the application: 4.11.2020). – Header from the screen.
20. Gorchakov A. V., Demidova L. A. Methods and Algorithms for Cross-Language Search of Source Code Fragments, *2024 International Conference on Information Technologies (InfoTech)*. Sofia, Bulgaria, 11–12 September 2024. DOI: 10.1109/InfoTech63258.2024.10701403
21. Han S., Wang D., Li W., Lu Xuesong et al. A Comparison of Code Embeddings and Beyond [Electronic resource]. Electronic text data. Cornell University,2021. Mode of access: https://doi.org/10.48550/arXiv.2109.07173. free (date of the application: 15.09.2021). Header from the screen.
22. Sudhamani M., Rangarajan L. Code similarity detection through control statement and program features, *Expert Systems with Applications*, 2019, Vol. 132, pp. 63–75. –DOI: 10.1016/j.eswa.2019.04.045
23. Shatnawi R. A Classification of Software Modules into Library and Application Components in the Open-Source Field, *International Journal of Software Engineering and Its Applications*, 2016, Vol. 10(3), pp. 179–190. DOI: 10.14257/ijseia.2016.10.3.16

DOI 10.15588/1607-3274-2026-1-13

24. Kungurtsev O. B., Mileiko I. I., Novikova N. O. Technology for Automated Construction of Domain Dictionaries with Special Processing of Short Documents, *Radio Electronics, Computer Science, Control*, 2024, № 4, P. 148. DOI: 10.15588/1607-3274-2023-4-14

Received 09.07.2025.
Accepted 15.01.2026.
Published 27.03.2026.

УДК 004.415.2

РОЗРОБКА СХОВИЩА КЛАСІВ ДЛЯ ОБ'ЄКТНО-ОРІЄНТОВАНИХ ТЕХНОЛОГІЙ СТВОРЕННЯ ПРОГРАМНИХ ПРОДУКТІВ

Кунгурцев О. Б. – канд. техн. наук, професор кафедри інженерії програмного забезпечення Національного університету «Одеська політехніка», Одеса, Україна. ROR: https://ror.org/05xaz0w84. ORCID: http://orcid.org/0000-0002-3207-7315.

Новікова Н. О. – канд. техн. наук, доцент кафедри Технічна кібернетика й інформаційні технології ім. професора Р. В. Меркта Одеського національного морського університету, Одеса, Україна. ROR: https://ror.org/05qze6v15. ORCID: https://orcid.org /0000-0002-6257-9703.

Бугаєва І. Г. – канд. техн. наук, доцент кафедри Технічна кібернетика й інформаційні технології ім. професора Р. В. Меркта Одеського національного морського університету, Одеса, Україна. ROR: https://ror.org/05qze6v15. ORCID: https://orcid.org /0000-0002-2839-9266.

Витнова А. І. – магістр кафедри інженерії програмного забезпечення Національного університету «Одеська політехніка», Одеса, Україна. ROR: https://ror.org/05xaz0w84. ORCID: https://orcid.org /0000-0002-4224-3006.

АНОТАЦІЯ

Актуальність. Використання раніше розробленого коду, зокрема, програмних класів та груп класів, пов'язаних відносинами спадкування, агрегації та композиції в об'єктно-орієнтованих технологіях, істотно скорочує час проектування програмного забезпечення.

Мета роботи. Проблеми виникають при необхідності зберігання класів різного призначення. У цьому випадку найменування класу не може бути його характеристикою для пошуку. Створення спеціальної структури, що пов'язує ім'я класу з його призначенням, істотно ускладнить сховище класів. Крім цього, якщо деякий клас пов'язаний відносинами з іншими класами, то необхідно зберігати поряд з класом його зв'язки, що може суттєво ускладнити як розміщення класу у сховищі, так і надалі його пошук. Метою роботи є створення спеціального сховища програмних класів, в якому клас представлений моделлю, що дозволяє визначити його призначення, можливі зв'язки з іншими класами та роль у цих зв'язків.

Метод. Розроблено математичну модель програмного класу, яка дозволила визначити призначення класу та можливі зв'язки з іншими класами. Запропоновано метод автоматизованого розміщення та пошуку окремих класів, а також груп класів у сховищі класів. Запропоновано модель сховища для розміщення як окремих класів, так і груп класів, пов'язаних відносинами спадкування, агрегації та композиції. Розроблено механізм автоматизованого додавання окремого класу або групи класів до сховища, а також пошуку та видалення окремого класу чи групи класів.

Результати. Для апробації запропонованих рішень розроблено програму *ClassCell*. Проведено експерименти для визначення часу та якості виконання операцій розміщення та пошуку класів та груп класів у сховищі. Результати показали суттєве скорочення часу на пошук класів порівняно з відомими бібліотеками.

Висновки. Запропонований метод автоматизованого розміщення та пошуку класів на основі моделі класу дозволяє підтримувати версії класів, суттєво скоротити час на пошук потрібного класу та пов'язаних з ним класів. Метод може бути використаний для класів на різних об'єктно-орієнтованих мовах.

КЛЮЧОВІ СЛОВА: об'єктно-орієнтоване програмування, клас, пошук класу, композиція, успадкування, агрегація, групи класів, синтаксичний аналіз.

ЛІТЕРАТУРА

1. Microsoft Corporation. .NET Class Libraries. Microsoft Learn [Electronic resource]. – 2022. – Mode of access: https://learn.microsoft.com/en-us/dotnet/standard/classlibraries. free (date of the application: 01.12.2022). – Header from the screen.
2. Krill P. 12 top-notch libraries for C++ programming [Electronic resource] / Paul. Krill. – Electronic text data. – InfoWorld, 2022. – Mode of access: https://www.infoworld.com/article/2265967/12-top-notch-libraries-for-c-plus-plus-programming.html. free (date of the application: 14.10.2022). – Header from the screen.
3. Lee G. Modern Programming: Object Oriented Programming and Best Practices / G. Lee. – Packt Publishing, 2019. – 266 p.
4. Method Automated Class Conversion for Composition Implementation / [O. B. Kungurtsev, V. R. Bondar, K. O. Gratilova, et al.] // Radio Electronics, Computer Science, Control. – 2024. – № 2. – P. 142–149. DOI: 10.15588/1607-3274-2024-2-14
5. Kungurtsev O. Implementation of class interaction under aggregation conditions / O. Kungurtsev, N. Komleva // Eastern-European Journal of Enterprise Technologies. – 2024. – №2 (128). – P. 20–30. DOI: 10.15587/1729-4061.2024.301011
6. Perhac P. Algorithms and data structure libraries for JAVA / P. Perhac, S. Simonak // Acta Electrotechnica et Informatica. – 2020. – Vol. 20, No. 1. – P. 39–48. DOI: 10.15546/aeei-2020-0006
7. Tanzil M. H. "How do people decide?": A Model for Software Library Selection / M. H. Tanzil, G. Uddin, A. Barcomb // 17th International Conference on Cooperative and Human Aspects of Software Engineering, June 2024. DOI:10.1145/3641822.3641865
8. Di Felice P. Design and Implementation of a Metadata Repository about UML Class Diagrams. A Software Tool Supporting the Automatic Feeding of the Repository / [P. Di Fe-

DOI 10.15588/1607-3274-2026-1-13

lice, G. Paolone, R. Paesani et al.] // Electronics. – 2022. – № 11. – P. 201. –DOI: 10.3390/electronics11020201
9. Dobrzyński B. Text Mining Studies of Software Repository Contents / B. Dobrzyński, J Sosnowski // 18th International Conference on Evaluation of Novel Approaches to Software Engineering. – 2023. – P. 562–569. DOI: 10.5220/0011970100003464 ISBN: 978-989-758-647-7; ISSN: 2184-4895
10. Polaczek J. Exploring the software repositories of embedded systems: An industrial experience / J. Polaczek, J. Sosnowski // Information and Software Technology. – 2021. – Vol. 131. – P. 106489. DOI: 10.1016/j.infsof.2020.106489
11. Moya J. Repository Software in Software Development [Electronic resource] / J. Moya. – Electronic text data. – Mode of access: https://www.wearecapicua.com/blog/software-repository-engineering-development. free (date of the application: 25.10.2024). – Header from the screen.
12. Are the Code Snippets What We Are Searching for? A Benchmark and an Empirical Study on Code Search with Natural-Language Queries / [S. Yan, H. Yu, Y. Chen et al.] // 27th International Conference on Software Analysis, Evolution and Reengineering (SANER). – Singapore Management University, Singapore, 2020. – P. 344–354. DOI: 10.1109/SANER48275.2020.9054840
13. Kungurtsev O. B. Determination of Inheritance Relations and Restructuring of Software Class Model in the Process of Developing Information Systems / O. B. Kungurtsev, A. I. Vytnova // Radio Electronics, Computer Science, Control. – 2022. – № 4. – P. 98–107. DOI: 10.15588/1607-3274-2022-4-8.
14. Kaufman A. R. Adaptive Fuzzy String Matching: How to Merge Datasets with Only One (Messy) Identifying Field / A. R. Kaufman, A. Klevs // Published online by Cambridge Universit. Political Analysis . – 2022. – Vol. 30, Issue 4. – P. 590 – 596.
15. Free Text Customer Requests Analysis: Information Extraction Based on Fuzzy String Comparison / [A. Smirnov, N. Shilov, K. Evers et al.] // 17th IFIP International Conference on Product Lifecycle Management (PLM). – Rapperswil, Switzerland, Jul 2020. – P. 193–202. DOI: 10.1007/978-3-030-62807-9_16
16. Kalyanathaya K. P. A Fuzzy Approach to Approximate String Matching for Text Retrieval in NLP / K. P. Kalyanathaya, D. Akila, G. Suseendran // Journal of Computational Information Systems. – 2019. – Vol 15(3). – P. 26–32.
17. Kostanyan A. Fuzzy Segmentations of a String [Electronic resource] / A. Kostanyan, A. Harmandayan – Electronic text data. – Cornell University, 2022. – Mode of access: https://doi.org/10.48550/arXiv.2201.13427. free (date of the application: 31.01.2022). – Header from the screen.
18. Pikies M. Analysis and safety engineering of fuzzy string matching algorithms / M. Pikies, J. Ali // ISA Transactions. – 2020. – Vol. 113. – P. 1. DOI:10.1016/j.isatra.2020.10.014.
19. Opportunities and Challenges in Code Search Tools [Electronic resource] / [C. Liu, X. Xia, D. Lo et al.] // – Electronic text data.– 2020. – Mode of access: https://arxiv.org/abs/2011.02297. free (date of the application: 4.11.2020). – Header from the screen.
20. Gorchakov A. V. Methods and Algorithms for Cross-Language Search of Source Code Fragments/ A. V. Gorchakov, L. A. Demidova // 2024 International Conference on Information Technologies (InfoTech), Sofia, Bulgaria, 11–12 September 2024. DOI: 10.1109/InfoTech63258.2024.10701403
21. A Comparison of Code Embeddings and Beyond [Electronic resource] / [S. Han, D. Wang, W. Li, Xuesong Lu et al]. – Electronic text data. – Cornell University,2021. – Mode of access: https://doi.org/10.48550/arXiv.2109.07173. free (date of the application: 15.09.2021). – Header from the screen.
22. Sudhamani M. Code similarity detection through control statement and program features/ M. Sudhamani, L. Rangarajan // Expert Systems with Applications. – 2019. – Vol. 132. – P. 63–75. DOI: 10.1016/j.eswa.2019.04.045
23. Shatnawi R. A Classification of Software Modules into Library and Application Components in the Open-Source Field / R. Shatnawi // International Journal of Software Engineering and Its Applications. – 2016. – Vol. 10(3). – P. 179–190. DOI: 10.14257/ijseia.2016.10.3.16
24. Kungurtsev O. B. Technology for Automated Construction of Domain Dictionaries with Special Processing of Short Documents / O. B. Kungurtsev, I. I. Mileiko, N. O. Novikova // Radio Electronics, Computer Science, Control. – 2024. – № 4. – P. 148. DOI: 10.15588/1607-3274-2023-4-14

DOI 10.15588/1607-3274-2026-1-13

UDC 614.2+574/578+004.38

ESTIMATION OF EFFORT OF MIGRATION AMONG DOMAIN-DRIVEN DESIGN ARCHITECTURAL VARIATIONS

Lytvynov O. A. – PhD, Associate Professor, Faculty of Physics, Electronics and Computer Systems, Oles Honchar Dnipro National University, Dnipro, Ukraine. ROR: https://ror.org/00qk1f078. ORCID: 0000-0001-7660-1353.

Khandetskyi V. S. – Dr. Sc., Professor, Head of the Department of Electronic Computing Machinery, Oles Honchar Dnipro National University, Dnipro, Ukraine. ROR: https://ror.org/00qk1f078. ORCID: 0000-0002-6386-4637.

Lytvynov M. O. – Master, Post-graduate student, Faculty of Physics, Electronics and Computer Systems, Oles Honchar Dnipro National University, Dnipro, Ukraine. ROR: https://ror.org/00qk1f078. ORCID: 0009-0000-9765-1501.

ABSTRACT

Context. The article addresses the issue of effort estimation of migration among variations of DDD architecture using a method based on specifications of requirements to increase the predictability of the software migration process.

Objective. The goal of the work is to propose an effective method of effort estimation based on Use Case analysis.

Method. First, a set of rules for rigorous Use Case description adapted for software effort estimation needs are provided. Second, the modified Use Case metamodel and the method of use cases classification based on frame-based knowledge representation model are also suggested. The rigorous description allows to make the estimation of the use cases more precise, using FUSP method, and to build individual predictors for each class of use cases. Thirdly, the method uses historical data taken from previous iterations of the same project, and is based on three trends (optimistic, pessimistic and mean based trend).

Results. The result is the collection of functions used to predict the effort required for the next iteration (measured in person-hours) for each class of use cases.

Conclusions. FUSP method was adapted for task of gaining greater prediction accuracy of effort estimation for migration among variations of DDD architecture using a methodology based on specifications of requirements. The set of conditions to form the Use Case description rules adapted for software migration effort estimation needs is developed. The modified Use Case metamodel and the method of use cases classification based on frame-based knowledge representation model are suggested. It was proposed algorithm for building the individual predictors of each class and for corresponding effort estimation. The coefficient of FUSP person-hours transformation is based on three trends achieved and updated considering the results from previous iterations: the most pessimistic prediction is based on the upper bound, the lower bound predictor plays the role of the optimistic predictor, and the main trend is the meaning among these. The coefficients are used to predict the effort in person-hours required for the next iteration for each class of use cases. The results of experiment, conducted using the test RTP project for this class of software, showed that MMRE for the proposal method is 0.0343, and for the standard method – 0.1094. The obtained results evidence that the classification of use cases along with their rigorous description according to provided rules, and modification of the method by separating prediction logic in accordance with the use case classes makes the prediction more accurate and can be effectively used for effort estimation for DDD architectural variations migration.

KEYWORDS: Use case point, Software Effort Estimation, Fuzzy Use Case Size Point, Domain-Driven Design.

ABBREVIATIONS

DDD is a Domain Driven Design;
FUSP is a fuzzy use case size points;
SDE is a software development effort;
CMMI is a capability maturity models integration;
LOC is a lines of code metric;
FP is a functional points metric;
UCP is an Use-case point;
USP is an Use-case Size Point;
TAF is a Technical Adjustment Factor;
EAF is an Environmental Adjustment Factor;
NL is a natural language;
DTO is a data transfer object;
RTP is a Representative Test Project;
MVP is a minimum valuable product;
CRUD is a set of create, read, update and delete operations;
API is an Application Programming Interface.

NOMENCLATURE

MRE is a magnitude of relative error;
$\hat{y}_i$ is a predicted value;
y_i is a known real value;
i is an iteration;
n is a number of iterations;
x is a value which is considered to be 0.25;
$MMRE$ is a mean magnitude of relative error;
$PRED(x)$ is a prediction criteria within x%;
TPA is a total complexity of actors;
$TPPrC$ is a total complexity of the preconditions;
PCP is a complexity of the main scenario;
$TPCA$ is a total complexity of the alternative scenarios;
TPE is a total exceptions complexity;
$TPPoC$ is a total complexity of postconditions;
$UUSP$ is a numer of Unadjusted Use-case size points;
w_i is a weight of the i-th Technical / Environmental Adjustment Factor;
v_i is a degree of domination of the i-th TAF / EAF;
$\mu_A(x)$ is a degree of membership of element x in fuzzy set A, which in our case is defined by a trapezoidal membership function;

DOI 10.15588/1607-3274-2026-1-14

$FUSP$ is a number of fuzzy use-case size points;

E is an effort in person-hours;

k is an empirical coefficient used to translate FUSP to person-hours;

t is an index of the current iteration;

$t-1$ is an index of the previous iteration;

k_R^t is an actual ratio of effort E_R^t (person-hour) to FUSP^t for t-th iteration;

$E_{\max}^{t+1}$ is a predicted maximum effort needed to realize FUSP^{t+1} in the next iteration;

$E_{\min}^{t+1}$ is a predicted minimum effort needed to realize FUSP^{t+1} in the next iteration;

E_{avg}^{t+1} is a basic prediction.

INTRODUCTION

Today information systems have become an integral part of modern business, the increasing complexity and agility of which require advanced system flexibility, functionality, scalability to provide complete business processes automation and maintainability. One of the basic approaches focused on tackling business domain complexity is DDD [1] approach which encompasses a wide range of architectural variations. Some of the variations are well known [1–3], some appear as a result of their adaptation to pursue specific tasks and undertake certain projects. The situation is complicated by the fact that the requirements changes can cause architectural evolution or even migration of the project to new architecture. According to [4] architecture should be ready to react to changing requirements and end-users and developers feedback by the guided, controllable evolution in the ways to implement more effective solutions and disallow it evolving in a way that harms architectural concerns. In the case of DDD architectural variations it means that managers and developers need to ensure that the architectural modifications are objectively necessary and to estimate the efforts needed to perform the evolution. Accurate estimation of SDE has a strong impact on cost, schedule, functionality and quality of the software [5] and this work is devoted to estimating an effort for DDD architectural variations migration on the example of a project based on the application service-oriented variation of DDD to the Use Case focused variation.

The object of study is evolutionary architecture and architectural migration of information systems.

The subject of study is the problem of the effort estimation of the migration among DDD variations.

The purpose of the work is to provide a method of migration effort estimation of DDD architectural variations based on modified FUSP method.

1 PROBLEM STATEMENT

This work is devoted to the development of methodology for estimation of the complexity of migration among DDD architectural variations using modified FUSP method. In particular, we consider a migration of project based on the application service-oriented variation to the Use Case oriented variation).

The research questions answered by this study are as follows:

RQ1: Which factors can improve effort estimation accuracy using FUSP method based on specifications of requirements in the case of software migration assessment?

RQ2: What strategy of estimation tuning should be used in case of iterative development process?

For the effort estimation accuracy we used the *MMRE* and Percentage of Prediction within x% [6,7], which are the two most common metrics used in software engineering. Both parameters are based on the quantity called *MRE*, which is described by Equation 1.

MMRE and *PRED*(x) are shown in Equation 2 and Equation 3 – respectively.

$$MRE_i = \frac{\left|\hat{y}_i - y_i\right|}{y_i}, \tag{1}$$

$$MMRE = \frac{\sum_{i=1}^{n} MRE_i}{n}, \tag{2}$$

$$PRED(x) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}1, \text{if } MRE_i \le x;\\ 0, \text{otherwise.}\end{cases} \tag{3}$$

According to [8], an accurate effort prediction model should have an $MMRE \le 0.25$ (to ensure that the mean estimation error should be less than 25%) and $PRED(0.25) \ge 0.75$ (meaning no less than 75% of the predicted values with *MRE* are lower than 0.25. Some authors [9] also take into consideration $PRED(0.3)$.

2 REVIEW OF THE LITERATURE

The goal of software effort estimation is to make software development and maintaining process more manageable, and predictable, which is necessary to increase the maturity of the company up to 4-th level of CMMI [10].

According to [11] the software effort is directly proportional to a project size and complexity and inversely proportional to team productivity (4):

$$\text{Effort} = \frac{\text{Size} \cdot \text{Complexity}}{\text{Productivity}}. \tag{4}$$

The question of measuring software size is not trivial, and different metrics can be used, e.g. LOC, FP, Halstead Metrics, McCabes cyclomatic complexity etc.

Among several categories of methods used to estimate the effort (e.g. expert judgement, top-down, bottom-up, design to cost, parametric models), methods of estimation based on specifications of requirements stand out with

DOI 10.15588/1607-3274-2026-1-14

their relative simplicity (they don't require providing software components and algorithms specifications as input data) and ability to estimate software size and effort at early phases of software development.

Evolutionary migration among DDD variations can be regarded as a software development process which allows us to take Requirements Specifications as the main source of information and use appropriate software effort estimation methods.

The basic method of estimation based on specifications of requirements is UCP [12] which has a number of modifications. For example, USP [13], FUSP [14], fuzzy-UCP [15], and ANFUSP [16]. The complete list of methods can be found in [17, 6].

The reason for such variety of modifications relates to four main problems [6] connected to the UCP method foundation.

PRB1. Defining of the complexity in the Use Case models.

PRB2. Adjustments of the Technical Complexity Factors and Environmental Complexity Factors.

PRB3. Classification of Use Cases. It is complicated to define the metrics, e.g., many variations of use case specification style and formality measure the length of a use case. For example, in [10] author argued that defining the UCP model's elements (actors and use cases), and assigning weights to them are the main problems to obtaining a reasonably accurate estimate.

PRB4. The fourth is how to make the right choice in a productivity factor whose units are person-hours for the estimated effort from UCP size measurements – especially when the historical dataset is not available.

In [18] all UCP based Effort Estimation models are classified into three main groups of approaches that focus on UCP factors: adding more complexity levels for the use case/actor weight; discretizing existing complexity levels into more detail options; empirically calibrating complexity weights to the different complexity levels.

FUSP is the extension of UCP which combines two approaches. It is focused on the details of the use-case structure: measures the functionality by counting the number and weight of scenarios, actors, precondition and post conditions [13]. To determine the USP, the sections of the use-case are analyzed, and the following steps are made to calculate USP [19]:

Step 1. *TPA*, *TPPrC*, *PCP*, *TPCA*, *TPE* and *TPPoC* are calculated using Table 1. *UUSP* value is calculated using formula (5)

$$UUSP = TPA + TPPrC + PCP + TPCA + TPE + TPPoC \quad . \qquad (5)$$

Table 1 – Unadjusted Use-case Size Point components classifications

Actor complexity (TPA)		
Complexity	Data (provided to or received from the Use Case)	USP
Simple	≤ 5	2
Average	[6;10]	4
Complex	≥ 10	6
Precondition complexity (TPPrC)		
Complexity	Expressions tested by the condition in precondition	USP
Simple	≤ 1	1
Average	[2;3]	2
Complex	≥ 4	3
Main and alternative scenarios (PCP, TPCA)		
Complexity	Entities + Steps	USP
Very simple	≤ 5	4
Simple	[6;10]	6
Average	[11;15]	8
Complex	[16;20]	12
Very complex	≥ 21	16
Exceptions (TPE)		
Complexity	Tested expressions	USP
Simple	≤ 1	1
Average	[2;3]	2
Compex	≥ 4	3
Postconditions		
Complexity	Entities related with postcondition	USP
Simple	≤ 3	1
Average	[4;5]	2
Complex	≥ 6	3

DOI 10.15588/1607-3274-2026-1-14

Step 2. *TAF* can be calculated using Table 2 by Equation (6)

$$TAF = 0.65 + \left(0.01 \cdot \sum_{i=1}^{14} w_i \cdot v_i \right). \quad (6)$$

Step 3. *EAF* can be calculated using Table 3 by Equation (7)

$$EAF = 0.01 \cdot \sum_{i=1}^{5} w_i \cdot v_i . \quad (7)$$

Step 4. The final value for a use-case is given by Equation (8):

$$USP = UUSP \cdot (TAF - EAF). \quad (8)$$

The use of the classification tables does not allow a gradual change from one complexity category to another. The issue can be resolved by Fuzzification which allows us to solve this problem of classification which is made through the generation of a trapezoidal fuzzy number to each complexity category found on the classification tables [19, 20].

A fuzzy set is represented by a membership function (formula (9)). Each element will have a grade of membership that represents the degree to which a specific element belongs to the set (Fig. 1):

$$Fz[x \in A] = \mu_A(x) : \mathbb{R} \to [0;1], \quad (9)$$

Table 2 – Technical factors

Factor	Requirement	Influence	Weight
F1	Data communication	I1	2
F2	Distributed processing	I2	1
F3	Performance	I3	1
F4	Equipment utilization	I4	1
F5	Transaction Capacity	I5	1
F6	On-line input of data	I6	0.5
F7	User efficiency	I7	0.5
F8	On-line update	I8	2
F9	Code reuse	I9	1
F10	Complex processing	I10	1
F11	Easiness of deploy	I11	1
F12	Easiness operation	I12	1
F13	Many places	I13	1

Table 3 – Environmental factors

Factors	Description	Influence	Weight
E1	Formal development process existence	I1	1.5
E2	Experience with the application being developed	I2	0.5
E3	Experience of the team with the used Technologies	I3	1
E4	Presence on an experienced analyst	I4	0.5
E5	Stable requirements	I5	1
E6	Part time workers	16	2
E7	Motivation	17	–1
E8	Difficulty in programming language	18	–1

DOI 10.15588/1607-3274-2026-1-14

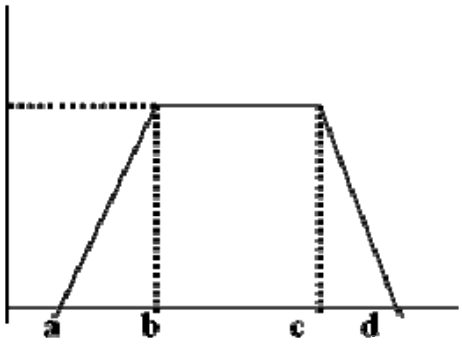


Figure 1 – Trapezoidal Membership function

$$\mu_A(x)=\begin{cases}0, x\le a,\\ 0, x\ge d,\\ 1, b<x\le c,\\ \dfrac{x-a}{b-a}, a<x\le b,\\ \dfrac{d-x}{d-c}, c<x\le d.\end{cases}, \quad (10)$$

Then the effort can be calculated using formula (11)

$$E = FUSP \cdot k\ . \quad (11)$$

In our case of migration among architectural variations of DDD, the method won't be effectively applied to the selected task. The reasons are as follows.

1. To predict the effort, we need to know k, which cannot be based on the historical data of the previous projects, e.g. we cannot use the coefficient applied to realized projects for the estimation of migration purposes, considering that the relationship between effort and size cannot be described by linear function [21].

2. Some factors such as Actor Complexity, Precondition and Postcondition Complexity don't affect the business logic layer and can be omitted. They are required to consider mainly during the development process, not for migration.

3. Technical factors and most environmental factors are oriented to whole project realization tasks and do not cover the specific of business layer and tasks of migration.

Thus, this paper is focused on adaptation of the FUSP method [14, 13] for estimation the efforts of the project migration among DDD variations. The main attention is given to the problems PRB1, PRB3, and PRB4 described above.

3 MATERIALS AND METHODS

The first problem (PRB1) and partially the third one (PRB3) are connected to Use Case definition rules.

The use case model documents the majority of software and system requirements and serves as a contract between stakeholders about the envisioned system behavior [22]. This model consists of two parts: a diagram which represents the relationships among the use cases and use cases descriptions represented as a structured document written in NL. The usage of NL can lead to poor descriptions such as ambiguous, inconsistent and/or incomplete descriptions which lead to missing requirements and eliciting incorrect requirements as well as less comprehensiveness of produced use case models [23]. In works [23, 24] authors proposed a catalog of symptoms of poor descriptions. However, this method cannot solve the problem of effort estimating the requirements development maxim states that requirements shall not consider design aspects.

In [25, 26] noted that to build automated verification of use cases the textual description is not sufficient. In [25] authors provide seven different types of steps: atomic, choice, concurrent, goto, include, success, failure. In [26] use-cases are instrumented by annotations, which are represented by tags (semantical markers) appended to a particular use-case step.

In [10] author asks the main question connected with UCP method issue: How should actors and use cases be defined if the use case model will be used for estimation? Answering the question, the author argued that we cannot ignore the existing architecture and some aspects of the project if we want to obtain accurate enough estimates.

Based on [10] proposition and considering the approaches and strategies mentioned above we provide the following set of rules to the Use Case definition.

Single responsibility of steps. Each step of the Use Case should be defined in accordance with Single Responsibility principle, i.e. responsible for atomic actor or system activity.

Rigorous steps identification. The architectural solution involves use case definition by introducing some additional steps to reflect the specifics of system's behavior. E.g. if the system uses event-driven notification mechanism the steps connected with notification preparation, publishing and handling cannot be omitted or absorbed by other steps.

Alternative flows definition. All the alternatives are represented at least by two steps (condition check and exception handling in case of <abort>).

Use cases decomposition. Some use cases should be split in order to omit repeated logic. However, before the estimation the steps connected to other use cases invocation should be represented as a sequence of atomic steps.

Entity consideration. We provide fixing the relationship between steps and entities which correspond to basic FUSP method.

Implicit use cases. Architecture can involve the requirements by introducing additional use cases. For example, when bounded contexts interact using events mechanism, it causes the introduction of integrated events handling use cases.

Steps and Rules connection. The step responsible for checking business rules should be accompanied by only one alternative. If it is connected to several alternatives, it should be split.

In result informal Use Case definition is transformed into rigorously defined Use Case considering architectural specifics. Modified metamodel of Use Case presented in [27] is shown in Fig. 2.

DOI 10.15588/1607-3274-2026-1-14

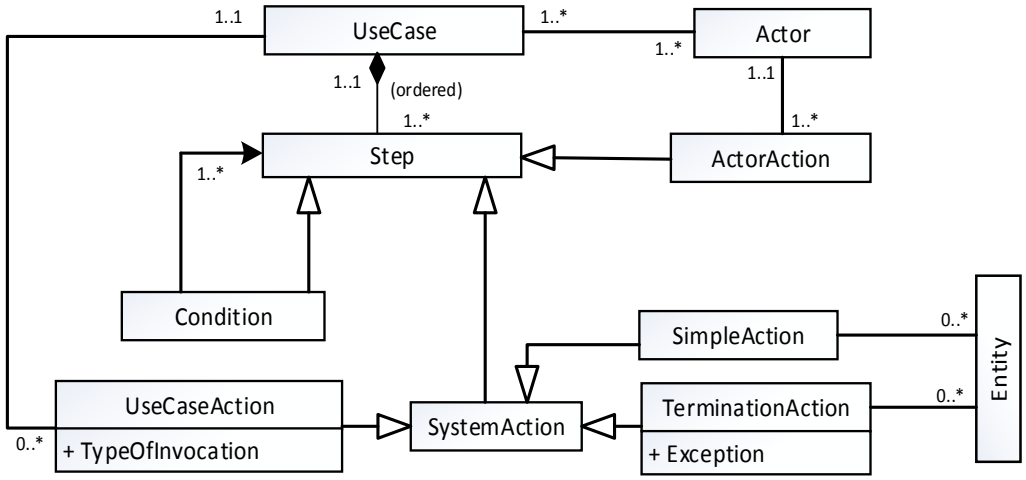


Figure 2 – Metamodel of Use Case

TerminationAction defines the termination of the actor request handling (in the most of cases it means exception handling). Alternatives are resolved using Conditional Steps which are often connected with business rules testing and termination actions (the case when before the termination some actions should be executed is also considered by the model). The step may trigger conditional (extension relationship case) and unconditional (include relationship case) invocation of another use case (see UseCaseAction). Some actions can be connected with entities.

Then we provide to classify the use cases and define the frames (knowledge description structures) connected with the classes of use cases. Each frame is comprised of several slots which represent types of steps, some of which are required, and some are optional which can be represented by the capacity of the slot. Required steps have the capacity equal to 1 or 1..* (one or many). Optional steps have the capacity equaled to 0 or 0..*. For example, the frame which describes a set of use cases of Modify type considering DDD architecture is shown in Table 4.

Table 4 – Modify Use Case Frame

Baseline Step	Alternative		Capacity	Number of steps	Entity
	Condition	Action			
1.Command validation	1a: Validation failed	1a1: Notify	1	3	0
2.Fetch data	2a: Object doesn't exist	2a1: Notify	1	3	1..*
3.Check timestamp	3a: Timestamp is not valid	3a1: Notify	1	3	0
4. Check simple rules	4a: Check simple rules failed	4a1: Notify	0..*	0..*	0
5. Check average rules	5a: Check average rules failed	5a1: Notify	0..*	0..*	0..*
6. Check complex rules	6a: Check complex rules failed	6a1: Notify	0..*	0..*	1..*
7. Check data integrity	7a: Check data integrity failed	7a1: Notify	0..*	0..*	1..*
8. Modify data			1..*	1..*	1..*
9. Prepare events			0..1	0..1	0
10. Publish events			0..1	0..1	0
11.Return response			1	1	0

It means that all the instances of Modify Use Case Type have at least 5 slots which will include at least 11 steps. Some authors [15] proposed to count the number of transactions, omitting the Condition, i.e. an alternative represented as condition-actions would only be counted by the number of actions. Of course, the number of use case types may not be restricted by the create, update, delete and select use cases.

Command validation step addresses input command validation logic performed by the request handler. Fetch data step means entity or several entities retrieval from persistence storage or from cache. Check timestamp step is connected to checking the version of the object to prevent race conditions issues. Checking simple rules means checking simple business rules in the scopes of the fetched data. For example, check the required fields are filled out, validate the data, checking the data on consistency. Checking average rules means checking business rules, involving additional data sources to perform the task. For example, checking the uniqueness of the modified object implies the additional request to the data source or checking the collection of entities. Checking complex rules means using a third-party service to perform checking. For example, checking a credit card balance using bank service. Checking data integrity means checking if the requested operation may violate data integrity. For example, referential integrity. It is worth noting that each rule should be represented as a separate step with the appropriate alternative. For example, when we have several rules connected with data checking, they should not be represented within one step, because it violates single responsibility principle. In our opinion, the main driver for such separation is the exceptions. Modify data slot is responsible for data modification that is the

DOI 10.15588/1607-3274-2026-1-14

core logic of the Use Case. In some cases, it could be connected to transactional modification of several entities, which may reside in several data sources and even perform distributed transactions. Prepare events slot means the preparation of the integration events to publishing (in some cases it may be extraction of domain events from an aggregate after performing the modification, publishing them and translating into integration events; in some cases, creating integration events at the application service level). Publish events slot means delivering events to subscribers (of course, here we primarily mean actors). Return response means creating and returning a response DTO to customer.

Example of fuzzification is shown in Fig. 3. The linguistic variables are not restricted by the 5 basic classes presented in Fig. 3 and could be extended to make the prediction more precise.

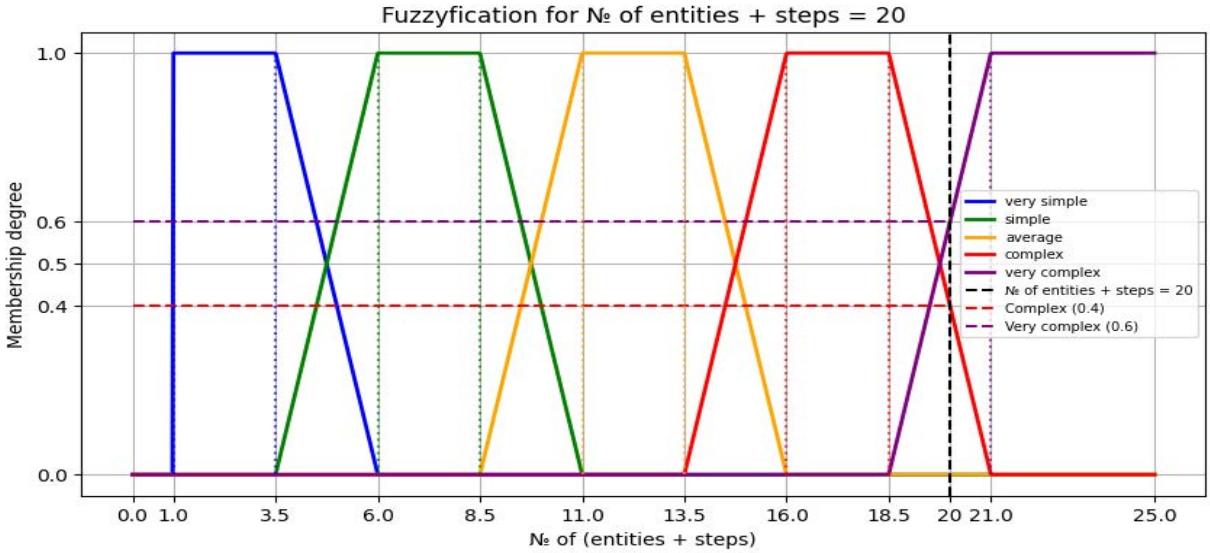


Figure 3 – Example of fuzzification

We propose a methodology which is purposed to predict the effort needed to perform every next iteration of the migration process.

The effort prediction is based on three coefficients used to transform *FUSP* to person-hours: minimal effort estimation, maximum effort estimation, basic effort estimation. While the minimum and maximum effort estimation functions establish bounds of the prediction, the basic effort estimation considers the history of the estimation (in our case we use mean of the values). The definition of the coefficients is represented by the formulas (12)–(15)

$$k_R^t = \frac{E_R^t}{FUSP^t}, \tag{12}$$

$$k_{\max}^t = \begin{cases} k_{\max}^{t-1}, k_R^t \le k_{\max}^{t-1}; \\ k_R^t, \quad k_R^t > k_{\max}^{t-1}. \end{cases} \tag{13}$$

$$k_{\min}^t = \begin{cases} k_{\min}^{t-1}, k_R^t \ge k_{\min}^{t-1}; \\ k_R^t, \quad k_R^t < k_{\min}^{t-1}, \end{cases} \tag{14}$$

$$k_{\text{avg}}^t = \frac{1}{t}\sum_{i=0}^{t-1} k_R^{t-i}. \tag{15}$$

The composition of indices R and t indicates real effort obtained at t-th iteration while indices min, max, avg in combination with t show that the coefficients are used to predict the effort (optimistic, pessimistic, average) for t-th iteration.

The prediction of the effort needed to perform the next iteration is based on the above coefficients as follows (16)–(18)

$$E_{\max}^{t+1} = k_{\max}^t \cdot FUSP^{t+1}, \tag{16}$$

$$E_{\min}^{t+1} = k_{\min}^t \cdot FUSP^{t+1}, \tag{17}$$

$$E_{\text{avg}}^{t+1} = k_{\text{avg}}^t \cdot FUSP^{t+1}. \tag{18}$$

4 EXPERIMENTS

The experiment is based on RTP [28], which is technical station management system, derived from real projects, considering preferable, basic scenarios occurred in real projects of DBB software company [29]. The method of obtaining RTP is as follows. Software projects of the company are classified by architectural solutions and technologies. Each project within the class can be represented by MVP [30] which is used to validate developers' hypotheses about customer needs serving the fastest way to get through the Build-Measure-Learn feedback loop with the minimum amount of effort and the least amount of development time. MVP is regarded as the core of the specific project which can be evaluated by the end-user. MVP projects of the same class, in turn, can be developed on the base of the same prototype. In accordance with [31] RTP relates to evolutionary incremental prototype.

DOI 10.15588/1607-3274-2026-1-14

RTP is not valuable to the customer, because it is primarily focused on typical task resolution in order to help produce MVP as quickly as it is possible, serves as a foundation needed for quick start. In daily practice this concept is known as labs, test projects or sample projects. Figuratively speaking, RTP is a seed from which MVP projects can grow. The connection of RTP and MVPs of several projects of class A is schematically shown in Fig. 4.

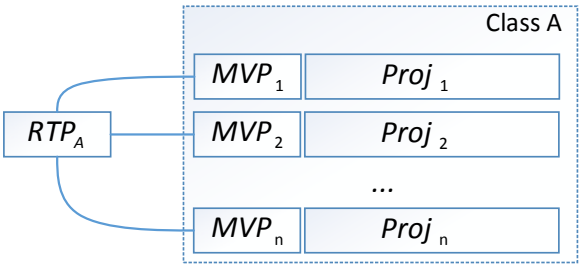


Figure 4 – Representative test project definition

We believe that there is not only one way to extract RTP from a set of projects, but the way we propose is based on use case classifications with mapping the use case classes onto architectural components. In result we get a collection of frames (the example is shown in Table 4). Then we examine the entities and the specifics of their interrelation trying to find a pattern which covers other variants. The domain used for RTP does not matter; the entities may not have all fields the end-use is interested in. The main purpose is to obtain the structural and behavioral pattern of the project.

The metamodels [32] of source and target architectural DDD variations are presented in Fig. 5 and Fig. 6.

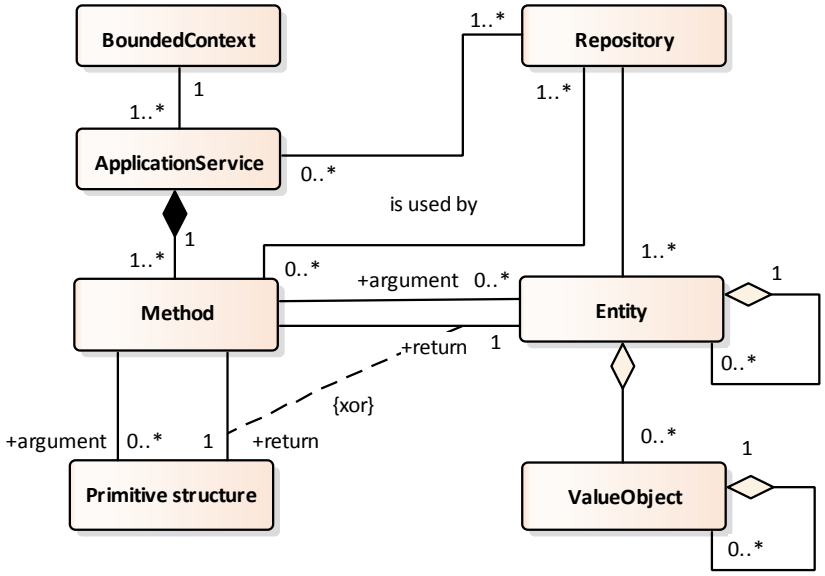


Figure 5 – Application service oriented architectural variations

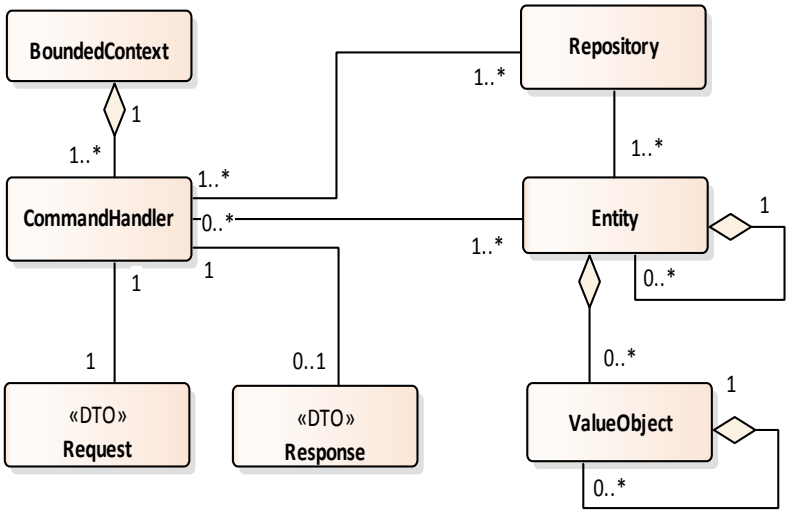


Figure 6 – Use case oriented architectural variations

DOI 10.15588/1607-3274-2026-1-14

The specifics of the source architectural solution are as follows. First, CRUD oriented application services are traced to multiple use cases connected with a certain entity. CRUD operations form four fundamental classes to project information management. According to [28] many use cases start their life as simple CRUD use cases, which later might grow into a full-fledged business use case with rich behavior, but significant part of use cases might remain simple. CRUD orientedness does not contradict DDD, which allows using CRUD for bounded contexts with low complexity [33]. Second, anemic domain model entities can be passed to the application services methods as arguments. These entities obtained by the mapping of DTO at the API controller side. The main reason for that is excluding the introduction of an additional layer of DTOs (which practically double the anemic entities), responsible only for passing the request to the application service methods from the higher system layer. Thirdly, the services are injected into API controllers or hubs using dependency injection mechanism. Controllers are responsible for input, which is represented as an API level DTO, validation, transforming the DTO into domain objects and passing them to appropriate application services. It is worth noting that API level DTO cannot be used as objects to business logic layer, because of several reasons, e.g. supporting several different API with different object representations and naming conventions.

The advantages of the above variation are as follows. Because of the excluding business logic level DTO objects the number of classes is reduced, consequently saving the time and efforts of the developers. The methods of Application services can share the same logic, e.g. validation, check integrity logic etc. In the case of CRUD operations, the testing infrastructure is simple, and the main logic of Test Fixtures may be represented as a generic class. In addition, we can say that the application services and anemic domain entities can be effectively generated using code generator with structure-oriented description model [34–35].

The main disadvantage of the solution is a certain distance from the use case definition (referring to the requirements, business rules) and their code implementation. Use case, which describes the interaction between the actor and the system, is triggered by an actor's request, includes different scenarios-paths comprised of several steps and finishes with a kind of response. The request can be represented as a DTO object or as an event (in case of event-driven architecture using pub-sub pattern) and these DTO structures could not be mapped directly to Entities. The response can be provided as a DTO object and several events used to notify the internal and external (i.e. actors) system components (for example using Event Bus).

Considering the agility of the business which causes the requirements changes, the problems connected with the completeness of the requirements, developers face the challenge of finding an effective process and mechanisms able to respond to functional requirements uncertainty. One of such mechanisms is to make software realization closer to use case definition, mapping the use case to separated class. This idea is implemented using several approaches (e.g. transaction script pattern, CQRS command handlers), including DDD variations, e.g. [12]. The metamodel of the target architectural DDD variation is presented in Fig. 6.

CommandHandler is a separate class which is responsible for task realization connected with a certain use case. The realization of the CommandHandler can be based on Command Bus pattern which provides a mechanism for executing commands in an application in a decoupled and extensible way. By the Use Case we mean the projection of the conventional Use Case onto the Business Logic. It is worth noting that the logic of the methods is comprised of definite blocks of operations which can be classified, and which can be thought of as slots of the frame which represents Method, in case of the first variation, or the CommandHandler in the case of the second one. Some of these slots can be required for definite types of use cases, some are optional. In most cases these slots are connected to the appropriate use case steps or transactions which can also be classified. At this point we face the necessity of the normalization of the use cases in order to make them more rigorously defined.

We restrict the RTP to only one Bounded Context because the logic connected to Bounded Contexts interaction remains the same. We don't consider any technological and environmental changes. In our case, we have 5 entities (Fig. 7) and four basic Use Cases (get, add, modify, remove) connected to them. Totally we have 20 Use Cases. Why can the above combination of entities be considered as a pattern of entity relationships for other projects?

The core of the schema is Worker-Work-Order relationship. Customer and Car entities represent an infrastructure which at first glance may be omitted. But the Customer-Car and Car-Order relationships cannot be covered by the Worker-Work-Order relationship. It seems that Car-Order relationship is similar to Worker-Work relationship, but they are not: Work entity depends on the Order as well.

The combination of entities is common for multiple Bounded Contexts: there could be a number of 1–0..* Car-Order like and several Order-Work-Worker like relationships. Of course, there could be variations of relationships, for example, an entity which depends on more than two entities, but it can be considered as an extension of the Order-Work-Worker relationship type. In Fig. 8 represented relationships examples: a) is covered by the extended Order-Work-Worker relationship; b) taken from [2] is covered by three Customer-Car-Order relationships.

DOI 10.15588/1607-3274-2026-1-14

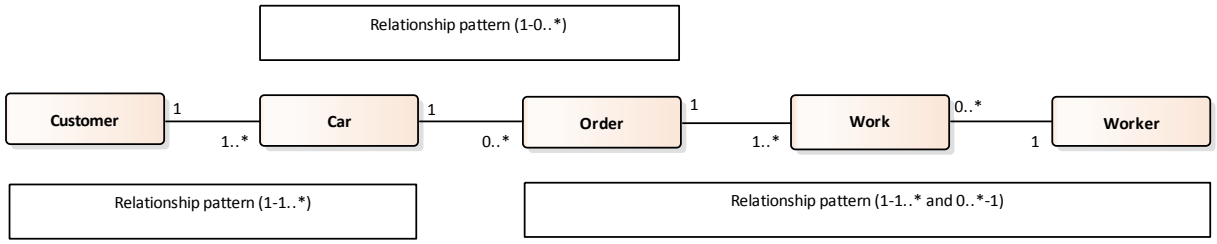


Figure 7 – Main entities of RTP and their relationships

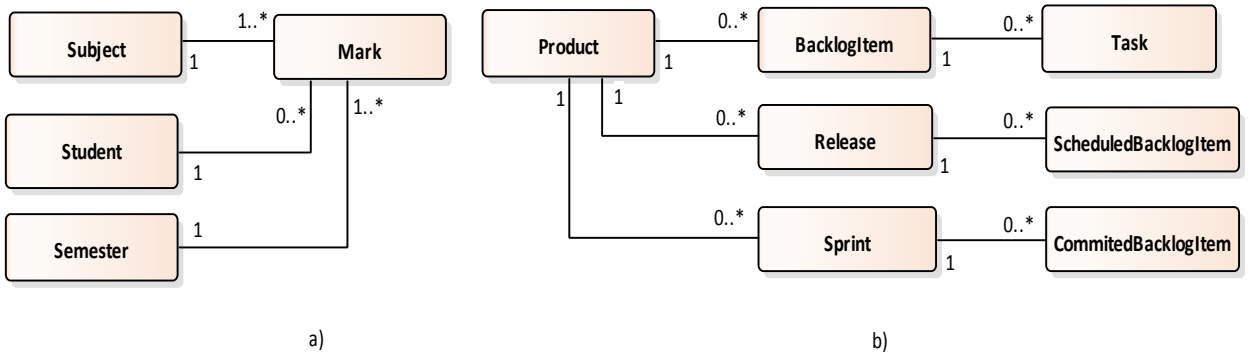


Figure 8 – The relationships covered by the Order-Work-Worker relationship

5 RESULTS

To estimate the effort required to migrate from source architectural variation into the destination one we used the modification of use case size points method [13] without using *TCF–ECF* correction. To check our hypothesis, we calculated *TAF* and *EAF* for DDD architecture as follows. The technical factors (Table 2) in case of selected projects class are: On-line update = 2; Data-communication = 2; Code reuse = 1; Easiness to deploy = 1. Environmental factors are: Stable requirements = 1; Experience with the technologies = 1; Formal development process = 1.5; Experience with the application being developed = 0.5

$$TAF = 0.65 + \left(0.01 \cdot \sum_{i=1}^{14} w_i \cdot v_i \right) = 0.65 + 0.01 \cdot (2+2+1+1) = 0.71,$$

$$EAF = 0.01 \cdot \sum_{i=1}^{5} w_i \cdot v_i = 0.01 \cdot (1+1+1.5+0.5) = 0.03.$$

Thus, *TAF – EAF* adjustment is equaled to 0.68.

The provided modification of FUSP method is as follows. Firstly, we provide a method of Use Case classification and normalization based on single responsibility pattern applied to steps classifying them in accordance with the slot types of the operation frame (e.g. Method, CommandHandler). In result we acquire 4 basic Use Case patterns. Secondly, we separate effort estimations into four different predictors with their own coefficients used to translate obtained use case size points into manhour. Thirdly, we propose an incremental method of the coefficients correction which considers project realization dynamic. The strategy of coefficients correction is based on different bounds estimation: upper, average, minimal bounds. In tables and charts bellow we can see the application of the method to 20 use cases connected to 5 entities.

Let's see the specific of the evaluation on the example of Modify Use Case type. The evaluation of the modify type use cases in FUSP for 5 entities is shown in Table 5. The structure of the modify use cases is defined by the frame presented in Table 4: all required slots are defined, all capacity rules are satisfied.

Let us consider the above description in more detail. Command validation for all the entities is 3-step slot, because the validation is connected to one alternative scenario (condition and notification). Fetching data, checking timestamp and simple rules (testing required fields and fields values conditions) are also estimated as 3 step slots, because checking the conditions implies the existence of an alternative. In case of Customer entity checking average rules slot filled out with the logic of testing the request against the rule "There could not be two customers with the same name in the system" which requires querying the Customer entities collection using repository. In case of Car entity, it is also necessary to check the rule of entity uniqueness which states "There could not be two different cars with the same number". The situation with Worker entity is the same as with Customer entity. There is no checking complex rules logic which implies the connection with third-party services (e.g. banking service), using API etc. Check data integrity means checking referential integrity. In our case Car entity modification requires checking CustomerId reference, whether it is valid or not (in the case when one customer sells the car to another customer it changes). Order has a reference to the Car (CarId field), and Work entity has two references to check: WorkerId (who performs the work), OrderId

DOI 10.15588/1607-3274-2026-1-14

(what order the work belonged to). All slots starting from Modify data and finishing with Publish events one has only one step. The number of entities depends on Fetch data and Check data integrity slots. Thus, in case of Car entity to check the validity of CustomerId the handler requires to interact with the additional Customer entity collection. The same situation with Order (additional Car entity collection) and Work (additional Worker, Order entity collections) entities.

Fuzzification process for Customer Modify Use Case is shown in Fig. 9. The continuous classification is presented in Table 6, considering [19], where p_i – lower value of the linguistic ter T_i in the classification (Table 1, section Main and alternative scenarios); $n_i = (p_i + p_{i+1})/2$; $a_i = n_{i-1}$; $b_i = p_{i+1}$.

Table 5 – Evaluation for Modify type use cases

	Entity				
Step	Customer	Car	Order	Worker	Work
Command validation	3	3	3	3	3
Fetch data	3	3	3	3	3
Check timestamp	3	3	3	3	3
Check Simple rules	3	3	3	3	3
Check Average rules	3	3	0	3	0
Check Complex rules	0	0	0	0	0
Check data integrity	0	3	3	0	6
Modify data	1	1	1	1	1
Prepare Integration Events	1	1	1	1	1
Return response	1	1	1	1	1
Publish events	1	1	1	1	1
Entities	1	2	2	1	3
Summary (UUSP)	20	24	21	20	25
FUSP (unadjusted)	14.4	16	16	14.4	16
USP (UUSP*0.68)	13.6	16.3	14.3	13.6	17
FUSP (adjusted)	8.2	8.5	9.2	8.2	9.6

Table 6 – Main and alternative scenarios classification (see Table 1, section Main and alternative scenarios)

Complexity	Entities + Steps	p	n	a	b	USP
Very simple	$[1;5]$	1	3.5	–	6	4
Simple	$[6;10]$	6	8.5	3.5	11	6
Average	$[11;15]$	11	13.5	8.5	16	8
Complex	$[16;20]$	16	18.5	13.5	21	12
Very complex	$[21;\infty)$	21	23.5	18.5	–	16

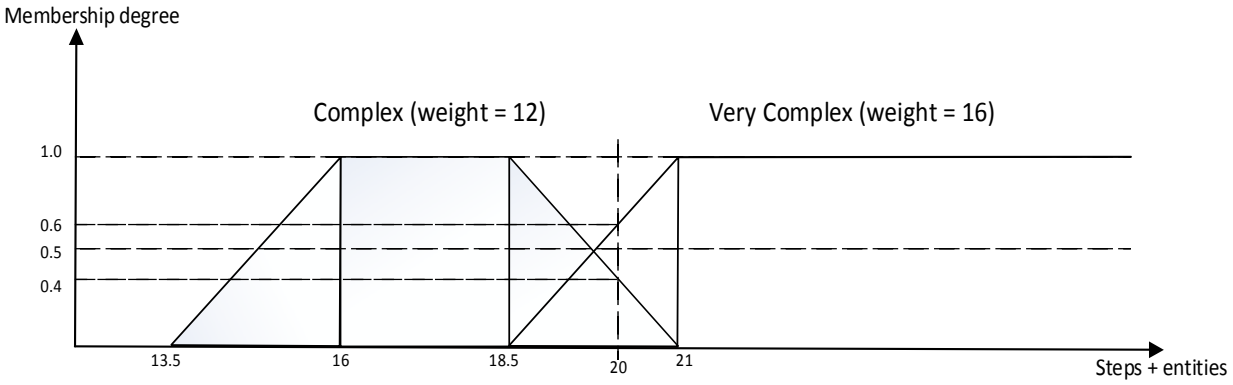


Figure 9 – Example of fuzzification for Customer Modify Use Case

According to FUSP method Modify Customer Use Case (20 steps) belongs to two classes of use cases Complex (weight = 12) and Very Complex (weight = 16) (Table 6). Thus, *FUSP* for this use-case will be evaluated using Formula (10) as follows.

$$FUSP(x) = \frac{b_i - x}{b_i - n_i} \cdot w_i + \frac{x - n_i}{b_i - n_i} \cdot w_{i+1}$$

$$= \frac{21-20}{21-18.5} \cdot 12 + \frac{20-18.5}{21-18.5} \cdot 16 = 14.4.$$

DOI 10.15588/1607-3274-2026-1-14

As we can see from Table 1, in accordance with the core FUSP Method the highest complexity of the use case is 21 steps and more (weight = 16).

The prediction subsequence shown in Table 7 represents 5 iterations. First iteration is devoted to getting initial data which will be used for next iteration prediction. For example, at the initial phase for Modify Use Case type 14.4 FUSP had been realized in 6 manhours, which sets the coefficient (manhours per FUSP) to 0.417. This coefficient is used to predict effort for the next iteration. According to the prediction the next iteration in case of Modify Use Case type would take 6.67 $(16 \cdot 0.417)$ manhours. But in fact, it took 7 manhours, which shifted the max effort coefficient to 0.438 and affected the average FUSP/manhours conversion coefficient: $(0.417 + 0.438)/2 = 0.4275$.

If we use *TAF–EAF* correction the prediction becomes worse (see Table 8). Using the MMRE criteria to estimate the accuracy of prediction we get: $MMRE = 0.0322$ and $MMRE_{(TAF-EAF)} = 0.0831$.

The estimations of the remaining three use case classes are presented in Tables 9–11. The structure of the tables includes three additional columns which indicate basic slot features: capacity, steps and connection to entity (similarly to Modify Use Case Frame description presented in Table 4). Unfortunately, because of the page limits we cannot provide full description of all the use case classes and calculation details. However, the description and calculations were performed similarly to Modify Use Case class.

Resulting comparison of the predictions for proposed and standard methods is shown in Table 12.

Table 7 – Calculations example for Modify Use Cases

FUSP	Total FUSP	Fact (manhours)	Prediction (manhours)		Coefficients (manhours per FUSP) for next iteration effort estimation			
			min	max	avg	min	max	avg
14.4	14.4	6	–	–	–	0.417	0.417	0.417
16	30.4	7 (>max)	6.67	6.67	6.67	0.417	0.438	0.427
16	46.4	6.67	6.67	7.01	6.83	0.417	0.438	0.424
14.4	60.8	5.8(<min)	6	6.31	6.10	0.403	0.438	0.418
16	76.8	6.67	6.44	7.01	6.7	0.403	0.438	0.418

Table 8 – Calculations for Modify Use Cases using TAF–EAF correction

FUSP with TAF, EAF	Total FUSP	Fact (manhours)	Prediction (manhours)		Coefficients (manhours per FUSP) for next iteration effort estimation			
			min	max	avg	min	max	avg
8.2	8.2	6	–	–	–	0.732	0.732	0.732
8.5	16.7	7 (>max)	6.22	6.22	6.22	0.732	0.824	0.778
9.2	25.9	6.67 (<min)	6.73	7.5	7.15	0.725	0.824	0.760
8.2	34.1	5.8(<min)	5.95	6.76	6.23	0.707	0.824	0.747
9.6	43.7	6.67(<min)	6.78	7.91	7.17	0.695	0.824	0.737

Table 9 – Evaluation of Get type use cases

	Entity					Frame		
Step	Customer	Car	Order	Worker	Work	Capacity	Steps	Entity
Command validation	3	3	3	3	3	1	3	0
Fetch data	3	3	3	3	3	1	1	1
Check Simple rules	0	0	0	0	0	0..*	3	0
Check Average rules	0	0	0	0	0	0..*	3	0..*
Return response	1	1	1	1	1	1	1	0
Entities	1	1	1	1	1	–	–	–
UUSP	8	8	8	8	8	–	–	–
FUSP	6	6	6	6	6	–	–	–
USP (UUSP*0.68)	5.44	5.44	5.44	5.44	5.44			
FUSP (adjusted)	4.2	4.2	4.2	4.2	4.2			

DOI 10.15588/1607-3274-2026-1-14

Table 10 – Evaluation of Add type use cases

	Entity					Frame		
Step	Customer	Car	Order	Worker	Work	Capacity	Steps	Entity
Command validation	3	3	3	3	3	1	3	0
Create data	1	1	1	1	1	1	1	1
Check Simple rules	3	3	3	3	3	0..*	3	0
Check Average rules	3	3	0	3	0	0..*	3	0..*
Check Complex rules	0	0	0	0	0	0..*	3	0..*
Save data	1	1	1	1	1	1	1	1..*
Fetch data	1	1	1	1	1	1	1	1
Prepare Integration Events	1	1	1	1	1	1	1	0
Return response	1	1	1	1	1	1	1	0
Publish events	1	1	1	1	1	1	1	0
Entites	1	2	2	1	3	–	–	–
UUSP	16	17	14	16	15	–	–	–
FUSP	12	12	8.8	12	10.4	–	–	–
USP (UUSP*0.68)	10.88	11.56	9.52	10.88	10.2	–	–	–
FUSP (adjusted)	7.9	4.9	6.8	7.9	7.4	–	–	–

Table 11 – Evaluation of Remove type use cases

	Entity					Frame		
Step	Customer	Car	Order	Worker	Work	Capacity	Steps	Entity
Command validation	3	3	3	3	3	1	3	0
Fetch data	3	3	3	3	3	1	1	1
Check Simple rules	0	0	0	0	0	0..*	3	0
Check Average rules	0	0	0	0	0	0..*	3	0..*
Check Complex rules	0	0	0	0	0	0..*	3	0..*
Check data integrity	3	3	3	3	0	0..*	3	0..*
Remove data	1	1	1	1	1	1	1	1
Prepare Integration Events	1	1	1	1	1	1	1	0
Publish events	1	1	1	1	1	1	1	0
Entities	2	2	2	2	1	–	–	–
UUSP	14	14	14	14	10	–	–	–
FUSP	8.8	8.8	8.8	8.8	7.2	–	–	–
USP (UUSP*0.68)	9.52	9.52	9.52	9.52	6.8	–	–	–
FUSP (adjusted)	6.8	6.8	6.8	6.8	4.6	–	–	–

Table 12 – Comparison of Effort/FUSP prediction for proposed and standard methods

Iteration	Proposed method with classification without adjustment		Standard method without classification with adjustment		Fact Effort (manhours)
	FUSP (unadjusted)	Prediction (manhours) Classified	FUSP (adjusted)	Prediction (manhours) Unclassified	
1 (Customer)	41.2	0	27.1	0	17.84
2 (Car)	42.8	18.51	24.4	16.06	18.34
3 (Order)	39.6	17	27	19.04	16.2
4 (Worker)	41.2	17.47	27.1	18.16	16.5
5 (Work)	39.6	16.61	25.8	16.9	16.29

Thus, as a result *MMRE* for the proposal method is 0.0343, and for the standard method – 0.109. The parameter *PRED* for both methods is equal to 1.

6 DISCUSSION

Based on the approaches reviewed, the set of conditions to form rigorous Use Case description rules adapted for software effort estimation needs is developed. The modified Use Case metamodel and the method of use cases classification based on frame-based knowledge representation model are suggested. It was proposed to build individual predictors for each class of use cases using corresponding formulas for effort estimation.

DOI 10.15588/1607-3274-2026-1-14

The experiment is based on RTP derived from real projects of corresponding direction using the MVP methodology.

The initial values for estimation (i.e. FUSP to effort ratio) are acquired during the initial warm-up iteration, i.e. the prediction is driven by the RTP migration iterations and does not rely on statistics from other projects. The result is the collection of functions which are used to predict the effort required for the next iteration (measured in person-hours) for each class of use cases. The coefficient of FUSP person-hours transformation is based on three trends achieved and updated considering the results from previous iterations: the most pessimistic prediction is based on the upper bound, the lower bound predictor plays the role of the optimistic predictor, and the main trend is the meaning among these.

The experiment was conducted for migration among Application service and Use case oriented DDD architectural variations. We take 20 basic use cases related to 5 entities divided into 4 classes (Add, Modify, Get, Remove) and used two variants of prediction: with and without use case classification. We also used filtering described by the formulas (10)–(12): we excluded the old values from the calculation when the deviation of Effort/FUSP ratio was greater than the threshold (in our case it was 0.08). To compare the methods, we used MMRE and PRED criteria.

The results are as follows. Without filtering *MMRE* for the proposal method is 0.0343, and for the standard method – 0.1094. The parameter *PRED* for both methods is equal to 1. In case of filtering the proposal method showed better result: *MMRE* is 0.0309 and for standard method *MMRE* remains the same.

Thus, the obtained results allow us to say that the classification of use cases along with their rigorous description according to provided rules, and modification of the method by separating prediction logic in accordance with the use case classes makes the prediction more accurate and can be effectively used for effort estimation for DDD architectural variations migration.

An experiment was conducted, demonstrating how the proposed method can be applied in practice to describe the use cases, evaluate them, to plan the effort and compare different methods using MMRE and PRED criteria, enabling the project managers to prognose the effort more accurately and developers to determine the development issues. The results obtained can be also reused as initial and historical data when planning similar architectural variations migration in real-world projects.

CONCLUSIONS

The work is focused on providing a theoretical and experimental platform applicable to effort estimation of domain-driven architectural variations migration.

The scientific novelty. FUSP method was adapted for task of gaining greater prediction accuracy of effort estimation for migration among variations of DDD architecture using a methodology based on specifications of requirements.

The set of conditions to form the Use Case description rules adapted for software migration effort estimation needs is developed. The modified Use Case metamodel and the method of use cases classification based on frame-based knowledge representation model are suggested. It was proposed algorithm for building the individual predictors of each class and for corresponding effort estimation. The coefficient of FUSP person-hours transformation is based on three trends achieved and updated considering the results from previous iterations: the most pessimistic prediction is based on the upper bound, the lower bound predictor plays the role of the optimistic predictor, and the main trend is the meaning among these. The coefficients are used to predict the effort in person-hours required for the next iteration for each class of use cases.

The results of experiment, conducted using the test RTP project for this class of software, showed that *MMRE* for the proposal method is 0.0343, and for the standard method – 0.1094. The obtained results evidence that the classification of use cases along with their rigorous description according to provided rules, and modification of the method by separating prediction logic in accordance with the use case classes makes the prediction more accurate and can be effectively used for effort estimation for DDD architectural variations migration.

The practical significance. The experimental part of the article presents the methodology for applying the developed method to describe the use cases, evaluate them, to plan the effort and compare different methods using MMRE and PRED criteria. The results obtained can be also used as initial and historical data when planning similar architectural variations migration in real-world projects.

ACKNOWLEDGEMENTS

We sincerely appreciate DBB Software company [29] for providing their proprietary platform, which served as the foundation for our experiment. This platform offered essential capabilities for our research, ensuring the accuracy and reliability of our experimental results.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: Oleksandr Lytvynov: method of use cases classification based on frame-based knowledge representation model; Volodymyr Khandetskyi: the main idea behind improving estimation accuracy; Mykhailo Lytvynov: FUSP method.

Data availability: The data will be made available on reasonable request from authors by email litvinovma0@gmail.com.

Software availability: The manuscript has associated software in a repository

DOI 10.15588/1607-3274-2026-1-14

https://github.com/MykhayloLytvynov/TechnicalStation.ADMCommandBus;
https://github.com/MykhayloLytvynov/TechnicalStation.Services.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

REFERENCES

1. Evans E. Domain-Driven Design: Tackling Complexity in the Heart of Software. Boston, Addison Wesley Professional, 2003, 560 p.
2. Vernon V. Implementing Domain-Driven Design. Boston, Addison Wesley, 2013, 656 p.
3. Martin R. C. Clean Architecture: A Craftsman's Guide to Software Structure and Design. Hong Kong, Pearson Education Asia, 2017, 432 p.
4. Ford N., Parsons R., Kua P., Sadalage P. Building Evolutionary Architectures, 2nd ed. Sebastopol, O'Reilly Media, 2022, 256 p.
5. Sharma S., Kushwaha D. S. Estimation of Software Development Effort from Requirements Based Complexity, *Procedia Technology*, 2012, Vol. 4, pp. 716–722. DOI: 10.1016/j.protcy.2012.05.116
6. Nhung H. L. T. K., Hoc H. T., Hai V. Van A Review of Use Case-Based Development Effort Estimation Methods in the System Development Context, *Software Engineering and Computer Systems : 2019 International Conference, Kuantan, 24–26 June 2019 : proceedings*. Cham, Springer, 2019, pp. 484–499. (Communications in Computer and Information Science, Vol. 1010). DOI: 10.1007/978-3-030-30329-7_44
7. Hoc H. T., Hai V. Van, Nhung H. L. T. K. AdamOptimizer for the Optimisation of Use Case Points Estimation, *Software Engineering and Computer Systems : 2020 International Conference, Kuantan, 24–26 June 2020 : proceedings*. Cham, Springer, 2020, pp. 747–756.
8. Dolado J. J. On the Problem of the Software Cost Function, *Information and Software Technology*, 2001, Vol. 43, № 1, pp. 61–72.
9. Foss T., Stensrud E., Kitchenham B., Myrtveit I. A Simulation Study of the Model Evaluation Criterion MMRE, *IEEE Transactions on Software Engineering*, 2003, Vol. 29, № 11, pp. 985–995.
10. Diev S. Use Cases Modeling and Software Estimation: Applying Use Case Points, *ACM Software Engineering Notes*, 2006, Vol. 31, № 6, pp. 1–5.
11. Galorath D. D., Evans M. W. Software Sizing, Estimation and Risk Management. Boston, Auerbach Publications, 2006, 573 p.
12. Karner G. Resource Estimation for Objectory Projects. Stockholm, Objective Systems SF AB, 1993, 20 p.
13. Braz M. R., Vergilio S. R. Software Effort Estimation Based on Use Cases, *Proceedings of the 30th Annual International Computer Software and Applications Conference (COMPSAC 2006), Chicago, 17–21 September 2006 : proceedings*. Los Alamitos, IEEE Computer Society, 2006, pp. 221–228. DOI: 10.1109/COMPSAC.2006.77
14. Braz M., Vergilio S. Using Fuzzy Theory for Effort Estimation of Object-Oriented Software, *16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004), Boca Raton, 15–17 November 2004 : proceedings*. Los Alamitos, IEEE, 2004, pp. 196–201.
15. Nassif A. B., Capretz L. F., Ho D. Enhancing Use Case Points Estimation Method Using Soft Computing Techniques, *Global Research in Computer Science*, 2010, Vol. 1, № 4, pp. 12–20. DOI: 10.48550/arXiv.1612.01078
16. Iraji M. S., Motameni H. Object Oriented Software Effort Estimate with Adaptive Neuro Fuzzy Use Case Size Point (ANFUSP), *International Journal of Intelligent Systems and Applications*, 2012, Vol. 4, № 6, pp. 14–24. DOI: 10.5815/ijisa.2012.06.02
17. Wen J., Li S., Lin Z., Hu Y., Huang C. Systematic Literature Review of Machine Learning Based Software Development Effort Estimation Models, *Information and Software Technology*, 2012, Vol. 54, № 1, pp. 41–59.
18. Qi K., Hira A. Calibrating Use Case Points Using Bayesian Analysis, *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM 2018), Oulu, 11–12 October 2018 : proceedings*. New York, ACM, 2018, Article 22. DOI: 10.1145/3239235.3239247
19. Arthi B., Selvarani A. G. Simplified Software Effort Estimation Using Fuzzy Set Theory, *Australian Journal of Basic and Applied Sciences*, 2015, Vol. 9, № 23, pp. 347–353.
20. Azzeh M., Neagu D., Cowling P. I. Analogy-Based Software Effort Estimation Using Fuzzy Numbers, *Journal of Systems and Software*, 2011, Vol. 84, № 2, pp. 270–284. DOI: 10.1016/j.jss.2010.09.028
21. Nassif A. B., Ho D., Capretz L. F. Regression Model for Software Effort Estimation Based on the Use Case Point Method, *2011 International Conference on Computer and Software Modeling, Singapore, 14–16 October 2011 : proceedings*. Singapore, IACSIT Press, 2011, pp. 117–121.
22. Cockburn A. Writing Effective Use Cases. Boston, Addison-Wesley, 2001, 304 p.
23. Seki Y., Hayashi S., Saeki M. Detecting Bad Smells in Use Case Descriptions, *Proceedings of the 27th IEEE International Requirements Engineering Conference (RE'19), Jeju, 23–27 September 2019 : proceedings*. Los Alamitos, IEEE, 2019, pp. 98–108.
24. Seki Y., Hayashi S., Saeki M. Cataloging Bad Smells in Use Case Descriptions and Automating Their Detection, *IEICE Transactions on Information and Systems*, 2022, Vol. 105–D, № 5, P. 849–863.
25. Sinnig D., Khendek F., Chalin P. Partial Order Semantics for Use Case and Task Models, *Formal Aspects of Computing*, 2010, Vol. 23, № 3, pp. 307–332.
26. Simko V., Hauzar D., Hnetynka P., Bures T., Plasil F. Formal Verification of Annotated Textual Use-Cases, *The Computer Journal*, 2015, Vol. 58, № 7, pp. 1495–1529.
27. Durán A., Bernárdez B., Genero M., Piattini M. Empirically Driven Use Case Metamodel Evolution, *UML 2004 : 7th International Conference, Lisbon, 11–15 October 2004 : proceedings*. Berlin, Springer, 2004, pp. 1–11. (Lecture Notes in Computer Science, Vol. 3273).
28. Hombergs T. Get Your Hands Dirty on Clean Architecture 2nd ed. Birmingham, Packt Publishing, 2023, 278 p.
29. DBB Software's Official Company Site [Electronic resource]. Access mode: https://dbbsoftware.com/
30. Ries E. The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses. New York, Crown Business, 2011, 336 p.
31. Floyd C. A Systematic Look at Prototyping, *Approaches to Prototyping*. Berlin, Springer, 1984, pp. 1–18. DOI: 10.1007/978-3-642-69796-8_1

DOI 10.15588/1607-3274-2026-1-14

32. Da Silva A. R. Model-Driven Engineering: A Survey Supported by the Unified Conceptual Model, *Computer Languages, Systems & Structures*, 2015, Vol. 43, pp. 139–155.
33. Milet S. Patterns, Principles, and Practices of Domain-Driven Design. Indianapolis, Wrox, 2015, 800 p.
34. Lytvynov O., Lytvynov M. On Application of the Frame-Based Modeling Language for Information System Development, *System Technologies*, 2023, Vol. 1, № 144, pp. 83–98. DOI: 10.34185/1562-9945-1-144-2023-11
35. Lytvynov O., Khandetskyi V., Lytvynov M. On Use of the Frame-Based Modeling Language for Information System Development, *VIII Vseukrainska naukovo-praktychna konferentsiia "Perspektyvni napriamky suchasnoi elektroniky, informatsiinykh i komp'iuternykh system" (MEICS-2023), Dnipro, 22–24 lystopada 2023 : proceedings.* Dnipro, DNU im. O. Honchara, 2023, pp. 11–12.

Received 03.08.2025.
Accepted 08.01.2026.
Published 27.03.2026.

УДК 614.2+574/578+004.38

ОЦІНКА ЗУСИЛЬ ДЛЯ МІГРАЦІЇ МІЖ АРХІТЕКТУРНИМИ ВАРІАНТАМИ DOMAIN-DRIVEN DESIGN

Литвинов О. А. – канд. техн. наук, доцент, Факультет фізики, електроніки та комп'ютерних систем, Дніпровський національний університет імені Олеся Гончара, просп. м. Дніпро, Україна. ROR: https://ror.org/00qk1f078. ORCID: 0000-0001-7660-1353.

Хандецький В. С. – д-р техн. наук, професор, Завідувач кафедри електронних обчислювальних машин, Дніпровський національний університет імені Олеся Гончара, м. Дніпро, Україна. ROR: https://ror.org/00qk1f078. ORCID: 0000-0002-6386-4637.

Литвинов М. О. – магістр, аспірант, Факультет фізики, електроніки та комп'ютерних систем, Дніпровський національний університет імені Олеся Гончара, м. Дніпро, Україна. ROR: https://ror.org/00qk1f078. ORCID: 0009-0000-9765-1501.

АНОТАЦІЯ

Актуальність. У статті розглядається проблема оцінки трудовитрат при міграції між варіаціями архітектури DDD із використанням методу, що базується на специфікаціях вимог, з метою підвищення передбачуваності процесу міграції програмного забезпечення.

Мета роботи – запропонувати ефективний метод оцінки трудовитрат на основі аналізу Use Case.

Метод. По-перше, запропоновано набір правил для строгої формалізації Use Case, адаптованих під потреби оцінки трудовитрат у розробці програмного забезпечення. По-друге, представлено метамодель модифікованого Use Case і метод класифікації Use Case на основі фреймової моделі подання знань. Такий строгий опис дозволяє точніше оцінювати Use Case за допомогою методу FUSP та створювати окремі предиктори для кожного класу Use Case. По-третє, метод використовує історичні дані з попередніх ітерацій того самого проєкту та ґрунтується на трьох трендах: оптимістичному, песимістичному та середньому.

Результати. Результатом є набір функцій, що використовуються для прогнозування трудовитрат (у людино-годинах), необхідних для наступної ітерації, окремо для кожного класу Use Case.

Висновки. Метод FUSP було адаптовано для підвищення точності прогнозування трудовитрат при міграції між варіаціями архітектури DDD, із застосуванням підходу, заснованого на специфікаціях вимог. Розроблено набір умов для формування правил опису Use Case, адаптованих до задач оцінки трудомісткості міграції програмного забезпечення. Запропоновано метамодель модифікованого Use Case та метод класифікації Use Case на основі фреймової моделі подання знань. Сформульовано алгоритм побудови індивідуальних предикторів для кожного класу Use Case та відповідної оцінки трудовитрат. Коефіцієнт перетворення FUSP у людино-години базується на трьох трендах, що формуються і оновлюються за результатами попередніх ітерацій: найпесимістичніше передбачення визначається верхньою межею, найоптимістичніше – нижньою, а основна оцінка – це середнє між ними. Ці коефіцієнти використовуються для прогнозування трудовитрат у людино-годинах, необхідних для наступної ітерації для кожного класу Use Case. Результати експерименту, проведеного на тестовому проєкті RTP цього класу ПЗ, показали, що середня відносна похибка запропонованого методу становить 0,0343, а стандартного – 0,1094. Отримані результати свідчать про те, що класифікація Use Case разом із їхнім строгим описом за запропонованими правилами, а також модифікація методу шляхом розділення логіки прогнозування відповідно до класів Use Case дозволяє досягти більшої точності та може ефективно використовуватись для оцінки трудовитрат при міграції архітектурних варіацій DDD.

КЛЮЧОВІ СЛОВА: Use Case Point, оцінка трудовитрат програмного забезпечення, Fuzzy Use Case Size Point, Domain-Driven Design.

ЛІТЕРАТУРА

1. Evans E. Domain-Driven Design: Tackling Complexity in the Heart of Software / E. Evans. – Boston : Addison Wesley Professional, 2003. – 560 p.
2. Vernon V. Implementing Domain-Driven Design / V. Vernon. – Boston : Addison Wesley, 2013. – 656 p.
3. Martin R. C. Clean Architecture: A Craftsman's Guide to Software Structure and Design / R. C. Martin. – Hong Kong : Pearson Education Asia, 2017. – 432 p.
4. Building Evolutionary Architectures / [N. Ford, R. Parsons, P. Kua, P. Sadalage]. – 2nd ed. – Sebastopol : O'Reilly Media, 2022. – 256 p.
5. Sharma S. Estimation of Software Development Effort from Requirements Based Complexity / S. Sharma, D. S. Kushwaha // Procedia Technology. – 2012. – Vol. 4. – P. 716–722. DOI: 10.1016/j.protcy.2012.05.116
6. Nhung H. L. T. K. A Review of Use Case-Based Development Effort Estimation Methods in the System Development Context / H. L. T. K. Nhung, H. T. Hoc, V. Van Hai // Soft-

DOI 10.15588/1607-3274-2026-1-14

ware Engineering and Computer Systems : 2019 International Conference, Kuantan, 24–26 June 2019 : proceedings. – Cham : Springer, 2019. – P. 484–499. – (Communications in Computer and Information Science, Vol. 1010). DOI: 10.1007/978-3-030-30329-7_44
7. Hoc H. T. AdamOptimizer for the Optimisation of Use Case Points Estimation / H. T. Hoc, V. Van Hai, H. L. T. K. Nhung // Software Engineering and Computer Systems : 2020 International Conference, Kuantan, 24–26 June 2020 : proceedings. – Cham : Springer, 2020. – P. 747–756.
8. Dolado J. J. On the Problem of the Software Cost Function / J. J. Dolado // Information and Software Technology. – 2001. – Vol. 43, № 1. – P. 61–72.
9. A Simulation Study of the Model Evaluation Criterion MMRE / [T. Foss, E. Stensrud, B. Kitchenham, I. Myrtveit] // IEEE Transactions on Software Engineering. – 2003. – Vol. 29, № 11. – P. 985–995.
10. Diev S. Use Cases Modeling and Software Estimation: Applying Use Case Points / S. Diev // ACM Software Engineering Notes. – 2006. – Vol. 31, № 6. – P. 1–5.
11. Galorath D. D. Software Sizing, Estimation and Risk Management / D. D. Galorath, M. W. Evans. – Boston : Auerbach Publications, 2006. – 573 p.
12. Karner G. Resource Estimation for Objectory Projects / G. Karner. – Stockholm : Objective Systems SF AB, 1993. – 20 p.
13. Braz M. R. Software Effort Estimation Based on Use Cases / M. R. Braz, S. R. Vergilio // Proceedings of the 30th Annual International Computer Software and Applications Conference (COMPSAC 2006), Chicago, 17–21 September 2006 : proceedings. – Los Alamitos : IEEE Computer Society, 2006. – P. 221–228. DOI: 10.1109/COMPSAC.2006.77
14. Braz M. Using Fuzzy Theory for Effort Estimation of Object-Oriented Software / M. Braz, S. Vergilio // 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004), Boca Raton, 15–17 November 2004 : proceedings. – Los Alamitos : IEEE, 2004. – P. 196–201.
15. Nassif A. B. Enhancing Use Case Points Estimation Method Using Soft Computing Techniques / A. B. Nassif, L. F. Capretz, D. Ho // Global Research in Computer Science. – 2010. – Vol. 1, № 4. – P. 12–20. DOI: 10.48550/arXiv.1612.01078
16. Iraji M. S. Object Oriented Software Effort Estimate with Adaptive Neuro Fuzzy Use Case Size Point (ANFUSP) / M. S. Iraji, H. Motameni // International Journal of Intelligent Systems and Applications. – 2012. – Vol. 4, № 6. – P. 14–24. DOI: 10.5815/ijisa.2012.06.02
17. Systematic Literature Review of Machine Learning Based Software Development Effort Estimation Models / [J. Wen, S. Li, Z. Lin et al.] // Information and Software Technology. – 2012. – Vol. 54, № 1. – P. 41–59.
18. Qi K. Calibrating Use Case Points Using Bayesian Analysis / K. Qi, A. Hira // Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM 2018), Oulu, 11–12 October 2018 : proceedings. – New York : ACM, 2018. – Article 22. DOI: 10.1145/3239235.3239247
19. Arthi B. Simplified Software Effort Estimation Using Fuzzy Set Theory / B. Arthi, A. G. Selvarani // Australian Journal of Basic and Applied Sciences. – 2015. – Vol. 9, № 23. – P. 347–353.
20. Azzeh M. Analogy-Based Software Effort Estimation Using Fuzzy Numbers / M. Azzeh, D. Neagu, P. I. Cowling // Journal of Systems and Software. – 2011. – Vol. 84, № 2. – P. 270–284. DOI: 10.1016/j.jss.2010.09.028
21. Nassif A. B. Regression Model for Software Effort Estimation Based on the Use Case Point Method / A. B. Nassif, D. Ho, L. F. Capretz // 2011 International Conference on Computer and Software Modeling, Singapore, 14–16 October 2011 : proceedings. – Singapore : IACSIT Press, 2011. – P. 117–121.
22. Cockburn A. Writing Effective Use Cases / A. Cockburn. – Boston : Addison-Wesley, 2001. – 304 p.
23. Seki Y. Detecting Bad Smells in Use Case Descriptions / Y. Seki, S. Hayashi, M. Saeki // Proceedings of the 27th IEEE International Requirements Engineering Conference (RE'19), Jeju, 23–27 September 2019 : proceedings. – Los Alamitos : IEEE, 2019. – P. 98–108.
24. Seki Y. Cataloging Bad Smells in Use Case Descriptions and Automating Their Detection / Y. Seki, S. Hayashi, M. Saeki // IEICE Transactions on Information and Systems. – 2022. – Vol. 105-D, № 5. – P. 849–863.
25. Sinnig D. Partial Order Semantics for Use Case and Task Models / D. Sinnig, F. Khendek, P. Chalin // Formal Aspects of Computing. – 2010. – Vol. 23, № 3. – P. 307–332.
26. Formal Verification of Annotated Textual Use-Cases / [V. Simko, D. Hauzar, P. Hnetynka et al.] // The Computer Journal. – 2015. – Vol. 58, № 7. – P. 1495–1529.
27. Empirically Driven Use Case Metamodel Evolution / [A. Durán, B. Bernárdez, M. Genero, M. Piattini] // UML 2004 : 7th International Conference, Lisbon, 11–15 October 2004 : proceedings. – Berlin : Springer, 2004. – P. 1–11. – (Lecture Notes in Computer Science, Vol. 3273).
28. Hombergs T. Get Your Hands Dirty on Clean Architecture / T. Hombergs. – 2nd ed. – Birmingham : Packt Publishing, 2023. – 278 p.
29. DBB Software's Official Company Site [Electronic resource]. – Access mode: https://dbbsoftware.com/
30. Ries E. The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses / E. Ries. – New York : Crown Business, 2011. – 336 p.
31. Floyd C. A Systematic Look at Prototyping / C. Floyd // Approaches to Prototyping. – Berlin : Springer, 1984. – P. 1–18. DOI: 10.1007/978-3-642-69796-8_1
32. da Silva A. R. Model-Driven Engineering: A Survey Supported by the Unified Conceptual Model / A. R. da Silva // Computer Languages, Systems & Structures. – 2015. – Vol. 43. – P. 139–155.
33. Milet S. Patterns, Principles, and Practices of Domain-Driven Design / S. Milet. – Indianapolis : Wrox, 2015. – 800 p.
34. Lytvynov O. On Application of the Frame-Based Modeling Language for Information System Development / O. Lytvynov, M. Lytvynov // System Technologies. – 2023. – Vol. 1, № 144. – P. 83–98. DOI: 10.34185/1562-9945-1-144-2023-11
35. Lytvynov O. On Use of the Frame-Based Modeling Language for Information System Development / O. Lytvynov, V. Khandetskyi, M. Lytvynov // VIII Всеукраїнська науково-практична конференція «Перспективні напрямки сучасної електроніки, інформаційних і комп'ютерних систем» (MEICS-2023), Дніпро, 22–24 листопада 2023 : proceedings. – Дніпро : ДНУ ім. О. Гончара, 2023. – P. 11–12.

DOI 10.15588/1607-3274-2026-1-14

UDC 004.93

A FRAMEWORK FOR THE REMOTE MONITORING OF PATIENTS IN THE HEALTHCARE SYSTEM

Mafraq H. I. – Post-graduate student, Department of Information Systems, Faculty of Computing and Information Technology, Mecca, King Abdulaziz University, Jeddah, Saudi Arabia; Lecturer, Department of Information Systems, King Khalid University, Abha, Saudi Arabia. ROR: https://ror.org/052kwzs30. ORCID: https://orcid.org/0000-0003-0331-3021.

Almagrabi A. O. – Assistant Professor, Department of Information Systems, Faculty of Computing and Information Technology, Mecca, King Abdulaziz University, Jeddah, Saudi Arabia. ROR: https://ror.org/02ma4wv74. ORCID: https://orcid.org/0000-0002-4858-9366.

Almagrabi H. – Associate Professor, Department of Information Systems, Faculty of Computing and Information Technology, Mecca, King Abdulaziz University, Jeddah, Saudi Arabia. ROR: https://ror.org/02ma4wv74. ORCID: https://orcid.org/0000-0001-5497-6461.

ABSTRACT

Context. Remote patient monitoring (RPM) technology plays a vital role in developing healthcare services. The medical team can continuously monitor a patient's health status, even outside of hospitals. It is considered one of the most important digital health services, as it facilitates patient care and reduces the spread of disease.

Objective. This paper aims to review current remote patient monitoring (RPM) systems for various diseases. Then proposes a new platform architecture to increase the effectiveness and quality of remote patient care.

Method. The paper analyzes systems for remote monitoring, focusing on the most common systems of several diseases such as diabetes, epilepsy, headache, cardiovascular and heart failure diseases, COVID-19, chronic kidney failure, fainting and unconsciousness, and cancer. Additionally, it provides an overview of the systems with contact and contact-less features, addressing them according to the system type, architectures, technology used, and services they provide.

Results. After analyzing remote patient monitoring (RPM) applications for a variety of diseases, the results highlighted the strengths and weaknesses of existing systems. We then demonstrated how the proposed architecture addresses these shortcomings and develops a scalable and effective solution.

Conclusions. This paper validates the effectiveness of RPM for healthcare development, offering an innovative ontology-based platform that improves service delivery and patient outcomes. This work offers valuable insights for healthcare providers, developers, and policymakers who are advancing remote care solutions.

KEYWORDS: biomedical telemetry, diseases, framework, medical information systems, Telemedicine.

ABBREVIATIONS

DSS is a decision support system;
RNS is a responsive neurostimulation;
EEG is an electroencephalogram;
RELAXaHEAD is an app for Migraine;
WHO is a world health organization;
CKD is a chronic Kidney Disease;
CDSS is a clinical decision support system;
DS is a Diabeo System;
RPM is a remote patient monitoring;
ECG is an electrocardiogram;
ICU is an Intensive Care Unit;
COVID19 is a Corona Virus Disease of 2019;
SWRL is a semantic web rule language;
HTTP is a hypertext transfer protocol;
ICT is an Information & communication technology;
UTAUT2 is a unified theory of acceptance and use of technology;
TMIS is a telemedicine information system.

NOMENCLATURE

H_t is a Health Status at time *t*;

S_t is a Vector of patient vital signs at time *t*;

C_t is a Vector of patient context parameters at time *t*;

O is a Ontology-based Context;

A_t is a New assertions incorporated into the ontology at time *t*;

R is a Set of SWRL rules applied for reasoning;

r_t is an Aggregated risk score derived from patient data;

z_t is a Notification zone classification;

τ_Y, τ_R is a Thresholds for Yellow and Red alert zones;

H^* is a Nearest hospital to the patient;

Ho is a platform's hospitals;

S is a specific hospital within the platform's hospitals;

φ_P, λ_P is a Patient's latitude and longitude;

φ_S, λ_S is a Hospital's latitude and longitude;

$d(\cdot,\cdot)$ is Geographic distance function ;

$f(.)$ is a Processing Function.

INTRODUCTION

Remote patient monitoring is a set of technologies and techniques that empower healthcare representatives, making them integral to the process of tracking patients' health data in real time, monitoring their health condition remotely, and utilizing related information in their treatment plan. The rapid growth of technology has brought many changes in the 21st century and modern society, and healthcare professionals are at the forefront of this transformation [1]. Their role is becoming more essential and

DOI 10.15588/1607-3274-2026-1-15

influential in people's daily lives. In addition, technology affects many aspects of human life, such as the education field, commerce, politics, the work environment, health, and so on. Therefore, healthcare institutions have turned to using new information and communication technologies to improve the quality of their services and their productivity. Furthermore, emerging technologies have enabled healthcare providers to share patient information and monitor them remotely [2].

The emergence of modern technologies and the internet has also contributed to the advancement of telemedicine. These advancements are evident in several areas, including consultations, medical record management, radiology, surgical procedures, remote patient monitoring, patient education, and medication reminders [3]. Remote patient monitoring is considered one of the most important telemedicine services, enhancing immediate access to healthcare [4]. Lin [5] defines telemedicine as the use of communication technologies to address medical concerns. These services promote high-quality healthcare and contribute to cost savings for governments and patients.

Furthermore, modern technologies and telemedicine have enabled patients to access healthcare services anytime, anywhere, eliminating the need for extensive travel to find the best specialists, especially for individuals residing in remote areas [6]. This focus on technology benefits not only patients but also healthcare professionals and policymakers. Studies have proven the effectiveness of patient monitoring in many areas, such as mental health, immunodeficiency, patients with chronic diseases, such as heart patients, and monitoring patients' compliance to medicines [7]. In addition to primary data, the system can collect additional information such as sleep patterns, activity levels, and patient weight. Some systems extend their capabilities to include postoperative management and monitoring of patient wounds [1] as shown in Fig. 1.

Figure 1 – Features of Remote Patient Monitoring System

The object of study is examined remote patient monitoring systems and highlights their shortcomings and limitations. In addition, it examines the technologies, methods used, and system architectures discussed in recent literature. It categorizes remote patient monitoring systems into contact-based and contact-free types. Additionally, it investigates communication channels between doctors, patients, caregivers, and families. Communication is a critical area that needs immediate attention and enhancement. Improving communication between these parties will ensure that families are promptly informed of any progress in the patient's health condition and needs, especially in critical conditions that require immediate attention.

The subject of study focuses on remote patient monitoring systems managing different diseases, like diabetes, epilepsy, headache, cardiovascular and heart failure diseases, COVID-19, chronic kidney failure, fainting and unconsciousness, and finally, cancer. This paper aims to pave the way for future improvements in remote patient monitoring systems, offering a hopeful outlook for the potential advancements in healthcare technology.

The purpose of the work is to offer a comprehensive outline of the proposed platform and a detailed plan. This detailed plan is designed to avoid the most significant limitations and incorporate the most compelling features of previous systems, thereby reinforcing confidence in the study's recommendations.

1 PROBLEM STATEMENT

The problem of remote patient monitoring can be formally defined as:

– A time series of heterogeneous patient vital signs, St.

– A set of patient context parameters at time *t*, Ct, includes medical history, location, and activity.

– Knowledge Base: Represented as an ontology, O, with medical concepts and relationships.

– A set of inference rules, **R**, defined in Semantic Web Rule Language (SWRL).

– Predefined thresholds for risk zones τY (Yellow) and τR (Red).

– A set of hospitals, Ho, with their geographic coordinates $\phi S, \lambda S$.

Now, the problem to find as presented as follows:

– The patient's inferred health status at time *t*, Ht.

– A quantitative risk score, rt.

A notification zone classification, $zt \in \{Green, Yellow, Red\}$, based on comparison of rt with τY and τR.

In the case of a critical alert (zt=Red), get the nearest hospital H^* to the patient's location (ϕP, λP).

The Objective is to minimize the time between the onset of a critical health event and the delivery of an appropriate intervention, through timely and accurate determination of Ht, rt, zt and H^*, subject to constraints of data privacy, system interoperability, and real-time processing requirements. Therefore, a deficiency in current systems is the weak use of semantic compatibility. Ontologies

DOI 10.15588/1607-3274-2026-1-15

provide a structured and interoperable framework across different fields. They unify knowledge, promote its reuse, and simplify problem-solving. This article, therefore, explores the concept of remote patient monitoring and its potential for improvement by identifying limitations in existing systems. The article's main contribution can be summarized as follows:

– To develop a patient-contextual ontology to support the semantic consistency of patient information. This ontology mitigates the impact of technology adoption resistance, enabling patients to benefit from its potential advantages.

– This paper describes an improved platform with an adaptable architecture for different diseases. This means the possibility of creating a universal framework for remote patient monitoring.

– This paper discusses the security and privacy of sensors, and their accuracy in sensing the patient's body. Furthermore, it explains the extent of its ability to protect sensitive patient information.

– The analysis of the communication system within the patient and healthcare institutions is of utmost importance. It plays a critical role in ensuring immediate response and prompt information about any health condition, thereby underlining its urgency and importance.

– The system's engineering for compatibility with multiple operating systems is a testament to its scalability. This, in turn, demonstrates the system's adaptability and reassures users of its acceptance in the technology landscape.

Generally, the problem can be summarized as determining a patient's health status H_t as a function of heterogeneous vital signs S_t and ontology-based contextual information O:

$$f(S_t, O) = H_t.$$

2 REVIEWS OF THE LITERATURE

This research focuses on some remote monitoring systems for specific diseases. It focuses on the most common systems of these diseases: diabetes, epilepsy, headaches, cardiovascular diseases, heart failure, COVID-19, chronic kidney failure, instances of fainting and loss of consciousness, and cancer disease. This section conducts a comprehensive analysis of existing research on remote monitoring for patients with specific diseases. Within each category, two additional classifications distinguish between contact-based and contact-free patient monitoring systems. The investigation aims to determine whether these systems incorporate functionalities for sending alerts to patients and their kin or medical personnel. Additionally, the systems will be assessed based on several elements: the type of application, the architecture utilized, the technology, and the services offered.

Remote monitoring systems for diabetes: Despite the availability of advanced treatment options for diabetes, many patients still struggle to achieve optimal control. The main obstacles to control are non-adherence to medications and dosage adjustments prescribed by the doctor, and difficulties in determining the appropriate insulin dose [8]. These patient concerns can be effectively addressed through remote monitoring and communication [9]. Many technological solutions have been successfully employed in diabetes management. We highlight some illustrative examples of these solutions, including the Diabeo (DS) system, which provides an alert message within the application interface, and the RPM system for pregnant patients, a specialized system designed by Kantorowska et al. [10] for pregnant patients with diabetes.

Monitoring systems for epilepsy patients: Epilepsy, a widespread neurological disorder, affects an estimated 65 million people worldwide. Moreover, seizures can manifest in various ways [11]. The development of systems for recording the number and characteristics of seizures represents a significant advancement for individuals with epilepsy [12]. Remote monitoring systems are capable of accelerating the diagnosis of the type of epilepsy and ensuring immediate medical intervention for patients [13]. Illustrative examples of these systems include monitoring system for epileptic patients using IoT [14], Nelli hybrid system enhancements from [15], EEG at home was designed by Biondi et al., [16], and the RNS designed by Skarpaas et al., [17]. The first system stands out for its unique feature of sending patient notifications.

Monitoring systems for headache patients: Headaches affect more than one billion people worldwide. They primarily affect individuals under 50 years of age, necessitating attention and the development of medical and technological solutions to alleviate their effects [18]. Studies indicate similar satisfaction rates and outcomes between telemedicine visits for headache patients and traditional in-person consultations, confirming the effectiveness of telemedicine [19].Here we review some examples of remote monitoring systems used to manage headache patients. The Leiden Headache Center has devised a web-based electronic diary on a time-bound schedule [20]. Conversely, Minen et al., [21] developed a program called RELAXaHEAD, a smartphone-based electronic diary (e-diary). H- diary application aims to monitor chronic headache patients from a distance [22].

Monitoring systems for cardiovascular disease and heart failure: Emerging technologies have opened up significant possibilities for improving healthcare support for older adults living in their own homes or in nursing homes. These technological advancements can be particularly beneficial in providing electrocardiogram (ECG) monitoring services to a wide range of individuals, including the elderly, athletes, and the general public. Providing these technologies in patients' homes reduces the cost of medical equipment and minimizes reliance on additional resources for caregivers [23].

Remote electronic monitoring of cardiac patients is becoming increasingly prevalent. This method, which involves taking the patient home, alerts, and routine interrogations at fixed intervals, offers a level of convenience that can be reassuring. It allows for increased comfort, faster identification of serious arrhythmias or organ dysfunction, and timely responses from doctors. Moreover,

DOI 10.15588/1607-3274-2026-1-15

remote monitoring screening may reduce the need for stressful in-person visits, particularly for patients with long travel periods or difficulty accessing personal care [24]. This section presents some examples of remote cardiac patient monitoring systems. For instance, remote clinical monitoring of heart patients is described by [15]. Amrita Spandanam was designed to monitor heart patients remotely. A model in Gontarska et al. [2] study estimates the degree of risk based on the vital parameters of a remote patient. The 'ECG Android App' is a mobile application that allows users to visualize their Electrocardiogram (ECG) waves [25].

Monitoring systems for COVID-19 patients: COVID-19, abbreviated from “coronavirus disease 2019”. It is an infectious respiratory disease. It swiftly spread worldwide, prompting the World Health Organization (WHO) to declare a pandemic in 2020 [26]. The global COVID-19 pandemic, with its immediate and widespread impact, has prompted healthcare systems to enhance their utilization of remote patient monitoring (RPM) tools for patient assessment and prioritization from a distance. The surge in COVID-19 cases worldwide has strained healthcare systems, exposing vulnerabilities and jeopardizing patient well-being [27]. This section provides examples of remote monitoring systems employed for managing COVID-19 patients.

Additionally, this section reviews two types of remote monitoring systems: contact-based and non-contact-based. As the name suggests, contact-based systems require physical contact with the patient, such as through wearable devices or sensors. On the other hand, non-contact-based systems can monitor patients from a distance, often using technologies like cameras or remote sensors. Paganelli et al. [28] established an Internet of Things-based framework for monitoring and examining COVID-19 patients in the hospital or home and issuing early warnings. An electronic platform in Sharma et al. study [29] was designed to monitor COVID-19 patients remotely using IoT devices, aiming to contain the spread of the disease.

Monitoring systems for chronic kidney disease patients: chronic kidney disease is a progressive decline in kidney function [30]. Dialysis patients are individuals with significant frailty. Home dialysis is a good solution for enabling these patients to effectively reduce their exposure to the hospital setting [31]. Remote monitoring and online tools provide enhanced convenience and access to care for these patients. These tools facilitate remote consultations between patients and healthcare professionals from the comfort of the patient’s home and bring relief and comfort, knowing that their health is being monitored closely. Remote monitoring systems for kidney failure patients have provided numerous benefits, including reduced hospital visits and improved access to healthcare providers. By utilizing these technologies, patients can receive timely care and support while minimizing disruptions to their daily lives. This section examines contact-based and non-contact-based systems. Markossian et al. [32] designed an app that primarily aims to facilitate self-management for individuals with CKD who do not require dialysis. Scarpioni et al. [31] developed a system to monitor and assist dialysis patients in reducing hospitalization risks during the COVID-19 pandemic.

Remote monitoring for fainting and loss of consciousness: Most of the unconscious patients are in the intensive care unit (ICU). These patients often require multiple life-sustaining devices [33]. However, with technological advancements, healthcare providers can remotely monitor pain, identify potential issues, and take preventive measures. This proactive approach enhances patients’ ability to detect problems early on and also plays a crucial role in reducing complications that may result in hospitalization. Moreover, technology and remote pain monitoring have significantly mitigated barriers to continuous care. This section provides examples of patients in intensive care units who are being monitored. Unlike the other systems in this category, the first two systems possess the communication feature. The system designed by Lee et al. [33] aims to introduce an innovative solution: a remote monitoring system specifically tailored for agitated patients. Emuoyibofarhe et al. [34] designed a remote monitoring system for preterm infants in neonatal ICU incubators. The system utilizes fuzzy rules for modeling and simulation. Garelli et al. [35] have pioneered the development of a groundbreaking platform for remote glucose monitoring, specifically designed for COVID patients in the ICU. A wearable system equipped with a mask contains sensors that capture vital signs has been proposed by Yang [15].

Remote Monitoring for Cancer Patients: Cancer treatment protocols encompass a wide range of procedures, including cancer diagnosis and various interventions. These treatments, such as chemotherapy, radiation therapy, and others, often lead to the development of pain. Digital health tools and technologies help support families, monitor disease symptoms, and remotely determine patients’ pain levels. This side overs some examples of these technologies to monitor cancer patients remotely. Bernier Carney et al. [36] developed a game-based innovative mobile application specifically designed for children aged 6–12 with cancer. Pavic et al. [37] created an application capable of early detection and prevention of health deterioration among cancer patients. ASyMS is a mobile application that enables remote monitoring of cancer patients [38]. Mayo Clinic has introduced a remote monitoring system for cancer patients involving a diverse team of healthcare professionals [39]. To overcome the limitations of these systems, as detailed in the work contributions section, the proposed platform aims to build on the semantic web rules and build an ontology that enables semantic interoperability of patient information. This ontology will play a crucial role in addressing the issue of interoperability between different devices. Furthermore, it provides an architecture that is adapted to various diseases. In addition, the proposed platform is designed to significantly enhance communication between the doctor, patient, patient’s family, and the healthcare giver, fostering a sense of connection and engagement. The commu-

DOI 10.15588/1607-3274-2026-1-15

nication provides notifications that contain developments of the patient's health condition and needs or an alert if there is a critical condition that requires dealing with it. Table 1 provides an overview of the studies we reviewed. It shows the systems that were classified based on diseases, the technology used, their architecture, the method used, limitations, and contact-based systems.

3 MATERIALS AND METHODS

This paper proposes a novel ontology-based framework for remote patient monitoring. The core methodological contribution is the integration of a dynamic patient-contextual ontology with a rule-based inference engine to enable semantic interoperability and real-time, context-aware clinical decision support. The complete architecture of the proposed platform, named Reayah, is illustrated in Figure 2 and consists of three primary components: the actor side, the Reayah management unit, and the database. The actor side encompasses all users of the platform, including patients, doctors, and family members. In addition, it is responsible for collecting data from the Actors. The information that should be entered contains all information about all users of the platform. For instance, the patient possesses the following data: vital signs, location, medical record, medical history, etc. On the other hand, the doctor can view all the information about his patients. Finally, the health provider manages all the system's users, hospitals, appointments, etc. So, this side works as an acquisition operator that captures inputs for the actors without regard to the data processing. The core processing is handled by the Reayah management unit, which is responsible for semantic inference processing. It maps input data to ontology-consistent individuals. It uses ontology rules alongside an inference reasoner to deduce the correct notification. Based on the patient's information and context, the action manager decides on the most appropriate alert (e.g., critical, medium, low). Following this decision, the designated notification will be forwarded to the notification center.

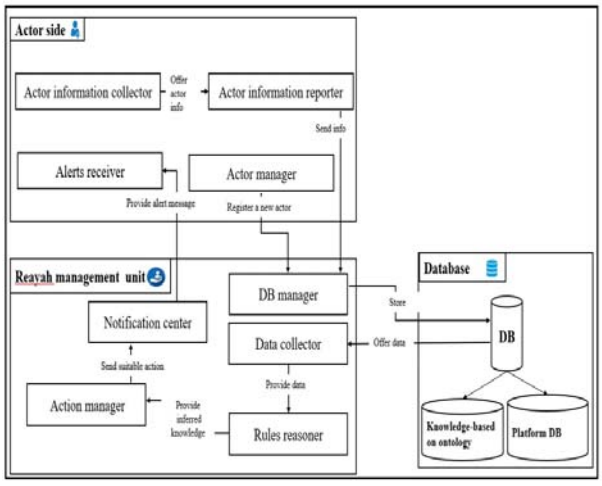


Figure 2 – Proposed Framework

Data persistence is managed by a dual-part database, which includes the Reayah database and the knowledge-based database. The database contains all data of actors, which in this context refers to patients and healthcare providers. For instance, the actor's data includes name, age, date of birth, identity, file number, height, weight, medical record, etc. The healthcare institution database contains all the data of the institution, such as hospitals, doctors, etc. In contrast, the knowledge-based ontology contains predefined rules. SWRL is used to deduce insights that are used within the platform.

This process can be formulated as a generalizable computational method rather than a theoretical description of the platform. In general, after a patient enters their medical information C_t such as [age, medical history, health record, location, activity, and vital signs S_t through the Reayah app, the underlying function F transforms the information into new assertions A_t that are integrated into the ontology in real time. The rule-based inference engine in the ontology determines the estimated health status H_t, risk score r_t, and notification zone z_t

$$f:(S_t,C_t)\to A_t,\quad O=O\_\{t-1\}\cup A_t,$$
$$(H_t,r_t,z_t)=f(O,R),\qquad z_t\in\{Green,Yellow,\mathrm{Re}\,d\},$$

Zone mapping: $z_t=\mathrm{Re}\,d$ if $r_t\geq\tau R$; Yellow if $\tau Y\leq r_t<\tau R$; Green if $r_t<\tau Y$.

The proposed ontology is implemented in Web Ontology Language (OWL 2), and the patient's condition is compared and inferred through the Pellet/Jena interpreter, which uses the Semantic Web Rule Language (SWRL) to infer the abnormal states and the current state of the patient.

Fig. 3 illustrates the workings of the Reayah platform. After a patient submits their vital signs S_t, the patient's context C_t via the mobile app then the A_t is determined. The SWRL-enabled model evaluates the rules to infer abnormal status and calculate a risk score (r_t). The notification zone then assigned to one of three zones

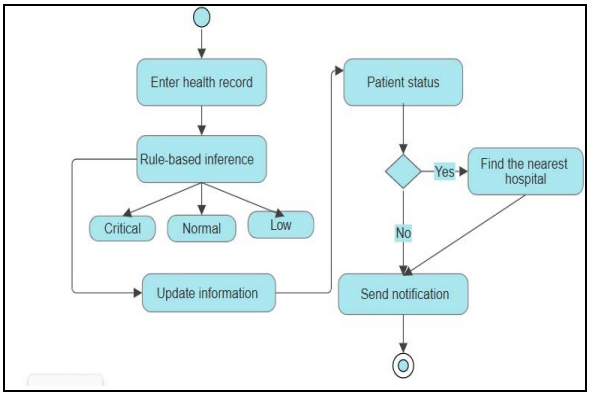


Figure 3 – Activity diagram of Reayah platform

DOI 10.15588/1607-3274-2026-1-15

(green/yellow/red) based on thresholds $(\tau Y, \tau R)$. The knowledge base is continuously updated, and notifications are sent to doctor and family members. In red cases, the nearest hospital (H^*) is determined

$$H^* = \arg\_\min_{S \in Ho} d((\varphi\_P, \lambda\_P), (\varphi\_S, \lambda\_S)).$$

Therefore, the methodology adopted in this study is a formal approach to remote patient monitoring, relying on semantic web rules to give accurate conclusions. It contributes to improving the efficiency, consistency, and accuracy of medical decisions within the proposed platform. Furthermore, it utilizes context-aware messaging to enhance and minimize errors in alerts. Alert messages vary and depend on the patient's condition. They may be considered normal or emergency. All information about patients, doctors, and the correct medical decisions is stored in the proposed platform database. The proposed platform supports many services. It supports messaging services that generate alerts for the patient, the doctor, and the patient's kin automatically at the time of an emergency by the server. Fig. 4 illustrates the data that has been exchanged between the proposed platform units during the platform's operation. First of all, the patient and doctor have to register on the server using his/her own information, such as email, name, and location. The following scenarios explain messaging between patients and doctors in normal situations. For example, the patient enters his daily vital signs as a medical report and sends a consultation to a doctor. The server sends a notification to the concerned doctor. The doctor checks the medical report of the patient and the patient's medical history. After that, he sends medical advice and the appropriate drug dosage to the patient.

On the other hand, we assume that the second patient has an emergency. The patient's blood pressure is high. He enters his vital signs using the Reayah application. Then the data will be analyzed and processed, and a warning alert will be sent to a certain doctor and the family of the patient. The doctor will decide the appropriate medical procedure and send it to the patient. The platform automatically sends medical advice and alerts to the patient's relatives.

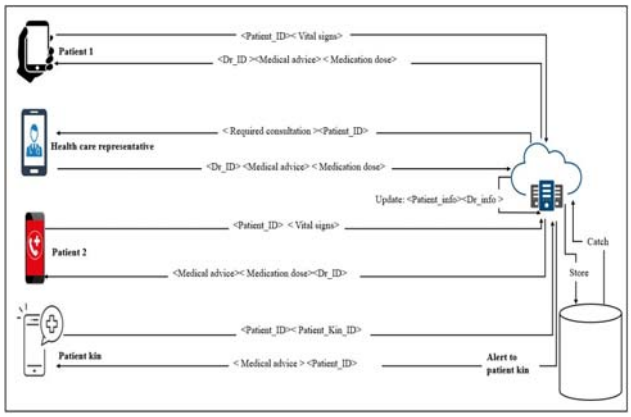


Figure 4 – Data flow of the proposed platform

4 EXPERIMENTS

Remote patient monitoring improves treatment adherence, a crucial aspect of patient care. These approaches have mainly been applied to chronic health conditions. Studies have shown that hospital treatment costs can be significant. However, by delivering health services at home, remote patient monitoring can help reduce time and cost, as patients no longer need to travel to seek medical attention, providing them with more comfort and less inconvenience. This article focuses on exploring the concept of remote patient monitoring, analyzing the associated systems, and highlighting the restrictions in the current systems.

Fig. 5 offers a comprehensive overview of the classification of systems based on specific diseases and the feature of communication between members of the clinical team, patients, and their kin. In this representation, the contact-based patient monitoring systems are represented by blue-filled shapes. Consequently, Remote Patient Monitoring (RPM) for pregnant patients, the Diabeo System, and a prototype for monitoring diabetes patients represent remote monitoring systems for diabetes. The Diablo System involves contact-based monitoring, distinguishing it from the other systems in this category. On the other hand, a Nelli hybrid system, a monitoring system for epileptic patients using IoT, EEG@HOME, and the RNS System, are examples of epilepsy disease remote monitoring. The system for epileptic patients using IoT sends notifications to the patients, providing a unique level of reassurance. Conversely, Leiden Headache Clinic

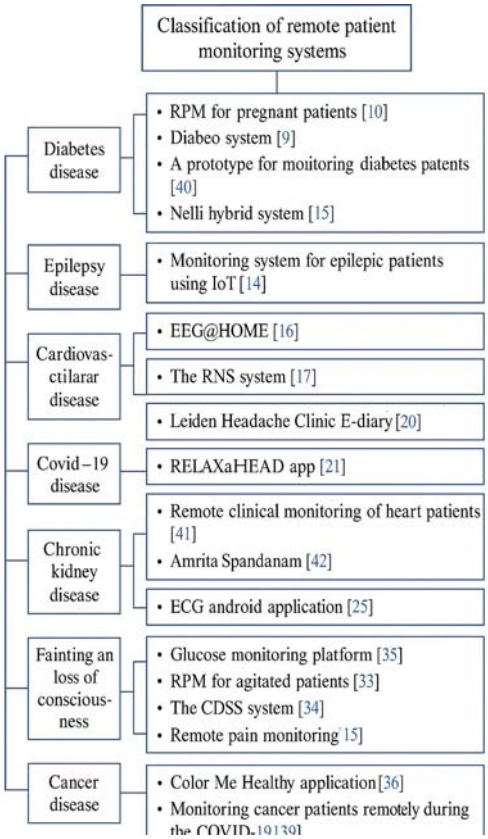


Figure 5 – Classification of remote patient monitoring systems

DOI 10.15588/1607-3274-2026-1-15

E-diary, RELAXaHEAD app, and H-Diary are standard systems for remote patient monitoring with Headache disease. Notably, the H-Diary offers contact-free functionalities compared to the other systems, which may intrigue and interest the audience.

Furthermore, remote clinical monitoring of heart patients, DSS, Amrita Spandanam, and ECG android applications represent remote patient monitoring for heart failure disease. Remote clinical monitoring of heart patients and Amrita Spandanam are contact-based systems, distinguishing them from the other systems in this category. Moreover, an early-warning system for remote monitoring of COVID-19 patients and an electronic platform to monitor COVID-19 patients are examples of remote patient monitoring of COVID-19 disease. The former operates as a contact-based system, while the latter functions as a non-contact-based system. In addition, a self-management mobile app for chronic kidney disease and an emote dialysis monitoring system are examples of RPM for chronic kidney disease. The former operates as a contact-based system, while the latter functions as a non-contact-based system. However, the glucose monitoring platform, RPM for agitated patients, the CDSS system, and remote pain monitoring are examples of RPM for fainting and loss of consciousness. The RPM for agitated patients and the CDSS system are contact-based systems, distinguishing them from the other systems in this category. Finally, the Color Me Healthy application monitors cancer patients remotely during COVID-19, and the Activity Monitoring application and the ASyMS application are examples of RPM for cancer disease. Monitoring cancer patients remotely during COVID-19 is a contact-free feature in contrast with other systems in this category. Context-aware technologies in healthcare offer tangible benefits that can be measured through the results of the studies and systems discussed. Previous systems have shown a reduction in hospital admissions for chronic patients by enabling alerts, adjusting treatment plans, and adhering to medications remotely. Additionally, the systems discussed have shown a decrease in emergency room visits due to timely interventions. This proves that systems can improve patient care by predicting and preventing health crises before they escalate.

5 RESULTS

Although technology has advanced over the years, the systems mentioned have some significant areas for improvement. As a result, we seek to provide a platform for remote patient monitoring, which includes context-aware technologies. Context-aware applications are increasingly being used in healthcare due to their potential to increase efficiency by providing real-time information on patient's health conditions. Context-aware refers to systems that can understand and interact with their physical and digital context [15].

Additionally, our context-aware platform can alert medical staff and patients of critical conditions (see Fig. 6), providing a sense of reassurance and confidence.

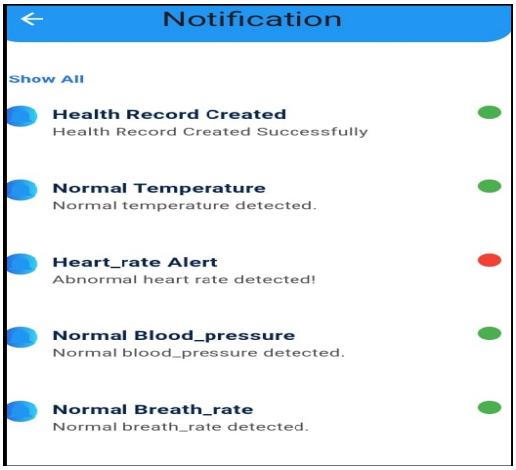


Figure 6 – Reayah platform notifications

Our platform is designed to be context-aware of the patient, focusing on determining the patient's location, health condition, physical activity, and more (see Fig. 7).

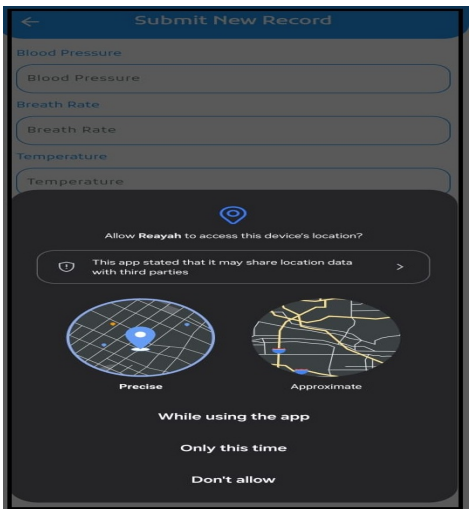


Figure 7 – Determine the location of the patient and their health status

It will use context-aware technology to deliver personalized notifications, analyze individual health data, and send customized messages to patients, their families, or healthcare providers based on specific parameters (see Fig. 8). It is important to recognize that the platform can modify care plans to meet the changing needs of patients, which enhances confidence in our platform's ability to adapt.

Furthermore, the platform can identify the nearest hospital to the patient by utilizing the patient's location. This feature helps ensure faster patient arrival, especially in emergencies. It thus allows healthcare providers to intervene proactively to save the patient's life. Finally, the contextual features of our platform significantly improve remote consultations by providing doctors with real-time data, enabling them to make more informed decisions.

DOI 10.15588/1607-3274-2026-1-15

Figure 8 – Communication between patient and doctor

In addition, deficiencies in communication between patients and doctor were observed in some analyzed systems [17, 29, 31, 39], emphasizing the need to enhance communication between these two parties. This ensures that patients are informed about any developments in their health condition and needs and are alerted in critical conditions that require immediate attention. To enhance alerts' reliability, efficiency, and effectiveness and expand them to include sending the patient's family, we intend to use context-aware messaging in alerts on our proposed platform. The involvement of the patient's family in the messaging system is crucial for providing additional support and care, and it can be instrumental in mitigating false notifications and increasing communication efficiency between the medical staff and the patient's family.

6 DISCUSSIONS

This paper has conducted a comprehensive analysis of remote patient monitoring systems documented in current literature, presenting a wide range of disease categories covered by the existing systems. It specifically focuses on the most common systems of these diseases: diabetes, epilepsy, headache, cardiovascular and heart failure diseases, COVID-19, chronic kidney failure, fainting and unconsciousness, and cancer. After that, the researchers proposed an ontology-based framework consisting of both with-contact and contact-free features by developing a remote monitoring system. In addition, the paper demonstrated the data flow model and comprehensively analyzed the different systems presented by different researchers in the literature. This analysis highlights the significant potential of the evolving healthcare technology field to greatly enhance patient care. The presented generic platform architecture is a pathway for the developers to build a patient monitoring ontology-based system based on context-aware category. Table 1 summarizes a comparison between previous studies and the current study. It illustrates the systems classified based on diseases, the technology used, their structure, the method used, limitations, and the communication-based system. It demonstrates that the current study differs from previous studies in that it relies on an ontology and semantic web rules and is distinguished by its classification of alerts into three zones. Furthermore, the researchers are excited to design a platform for remote patient monitoring that offers enhanced electronic healthcare services through the use of telemedicine information systems (TMIS) and cloud computing platforms. The plan, an ontology-based system, will not only facilitate the semantic interoperability of patient data but also leverage patient context to provide more efficient and effective patient care services.

Table 1 – Comparison between previous studies and the current study

System	Disease Type	Technology Used	Architecture of the system	Method used	Limitations	Contact-based system
RPM for pregnant patients [10]	Diabetics disease.	Bluetooth, MyChart app, and Electronic Health Record.	3-tiered	Replaced the conventional paper-based method of monitoring blood glucose with an Electronic Health Record	The scope of the study was limited to pregnant women.	No.
Diabeo System [9]	Diabetics disease.	Mobile application on Android and iOS platforms and a web portal.	3-tiered.	Machine learning algorithms.	The accuracy rate is not available.	Yes.
A prototype for monitoring diabetes patients [40]	Diabetics disease.	Clinical Decision Support System, knowledge base, and HER.	4-tiered.	Clinical Decision Support Systems (CDSS) and Electronic Health Records (EHR).	The architecture in this study does not encompass a real system of remote patient monitoring.	No.

DOI 10.15588/1607-3274-2026-1-15

Continuation of Table 1

Nelli hybrid system [15]	Epilepsy disease.	Video camera and microphones.	1-tiered.	Machine learning techniques.	Data were collected only from one recording, and the study targeted a group less prone to seizures. Finally, patients with non-motor seizures were not evaluated.	No.
Monitoring system for epileptic patients using IoT [14]	Epilepsy disease.	MATLAB and IoT devices.	1-tiered.	Fuzzy logic.	Insufficient accuracy of the sensors employed for identifying epileptic seizures.	Yes.
EEG@HOME [16]	Epilepsy disease.	Wearable sensor device, EEG recording, and mobile app (Seer app; Seer Medical).	2-tiered.	ANT Neurowas is used to record EEG, self-report self-reporting sensors, and the app collects data related to seizing occurrence app.	Fewer number of participants.	No.
The RNS System [17]	Epilepsy disease.	Tablet, Patient Data Management System (PDMS).	3-tiered.	The physician utilizes a tablet to configure detection and stimulation settings, as well as access and review data from the neurostimulator. The data monitor for the patient's home.	Lack of security.	No.
Leiden Headache Clinic E-diary[20]	Headache disease.	Electronic diary and tablet.	1-tiered.	A web-based survey was sent to the patient through email.	There is an absence of a reliable method to assess patient acceptance.	Yes.
RELAXaHEAD app [21]	Headache disease.	RELAXaHEAD app, electronic diary, and smartphone.	1-tiered.	Self-reported specific details about patients' headaches, sleep-related questions, and medications	The limited sample size and absence of a reliable method to assess patient acceptance.	Yes.
H-Diary [22]	Headache disease.	Web server, Oracle, JAVA, PHP5, JavaScript, HTML, and CSS.	3-tiered.	The patient enters data through daily diaries that contain a questionnaire consisting of yes and no questions.	The absence of a mechanism to assess patient acceptance.	No.
Remote clinical monitoring of heart patients [41]	Cardiova scular disease.	De novo pacemakers, implantable cardiac defibrillators, and follow-up device care.	1-tiered.	Scheduled and unscheduled in-person interrogation before discharge and remote interrogation post-discharge	The limited sample size and the absence of a mechanism to assess the utility of interrogations.	Yes.
DSS for remote patient monitoring of heart disease [2]	Cardiova scular disease.	Deep neural network models and the rule-based model.	1-tiered.	The database was split into three sets: train, validation, and test, with a distribution ratio of 4:1:1.	The capacity for the model was reduced.	No
Amrita Spandanam [42]	Cardiova scular disease.	IoT devices, Wi-Fi and cellular data, mobile phones, and the Cloud.	5-tiered.	Sensors collect data and then analyze it. The severity is measured using Consensus Abnormality Motif technology and other algorithms, and the results are sent to the medical team to take the correct action.	The limited sample size.	Yes.
ECG android application [25]	Cardiova scular disease.	SQL,Bluetooth, IOIO Microcontroller, MATLAB.	3- tiered.	The app leverages microcontroller technology, signal processing algorithms for ECG wave analysis, and communication protocols to ensure secure and private data transfer.	The app was run on Android only.	No.

DOI 10.15588/1607-3274-2026-1-15

Continuation of Table 1

An early-warning system for remote monitoring of COVID-19 patients [28]	COVID-19 disease.	SQL and NoSQL, data mining. Machine learning models and blockchain.	3- tiered.	Data acquired from sensors are analyzed on cloud servers.	The accuracy of the sensor data was not examined.	Yes.
An electronic platform to monitor Covid-19 patients [29]	COVID-19 disease.	Cooza simulator, IoT, artificial intelligence techniques, and Wi-Fi.	3- tiered.	The collected data was analyzed using CAF and KMCCA methods. It was then classified using SVM and KNN.	The model needs more energy.	No.
Self-management mobile app for chronic kidney disease [32]	Chronic kidney disease.	NVivo software and smartphones.	1-tiered.	Patient self-management, recommendations for adherence to medication regimens, avoidance of further nephrotoxic insults, and maintenance of diet.	The limited sample size.	Yes.
Remote dialysis monitoring system [31]	Chronic kidney disease.	Video camera, monitor, microphone, and technology communication.	1-tiered.	Home dialysis network.	The limited sample size.	No.
Glucose Monitoring Platform [35]	Fainting and loss of conscious ness	Different continuous glucose monitor devices and Bluetooth.	1-tiered.	Observed a detailed view of each patient's glucose evolution and other met-rics, automatically up-loaded daily to the plat-form.	The limited sample size.	No.
RPM for agitated patients [33]	Fainting and loss of conscious ness	Microsoft Kinect, computer game graphics, and IBM SPSS.	1-tiered.	Designed a detection system to identify the position of the patient in three-dimensional space.	The accuracy of the sensor data was not examined.	Yes.
The CDSS system [34]	Fainting and loss of conscious ness.	LabVIEW and MATLAB.	3- tiered.	The system utilizes fuzzy rules for modeling and simulation.	There is an absence of a mechanism to test and assess the system's accu-racy.	Yes.
Remote pain monitoring [15]	Fainting and loss of conscious ness.	Sensing devices, Wi-Fi, cloud server, computer, or a smart device	4- tiered.	This device utilizes a facial surface electromyogram (sEMG) to monitor a pa-tient's pain intensity.	The accuracy of the sensor data was not examined.	No.
Color Me Healthy application [36]	Cancer disease.	Game-based application, JavaScript, and tablet.	1-tiered.	Self-report, a checklist of general symptoms, and children express their pain experiences.	The accuracy of the sensor data was not examined.	Yes.
Monitoring cancer patients remotely during the COVID-19 [39]	Cancer disease.	Cellular-enabled tablet, Resideo Life Care Solutions software, Bluetooth-enabled devices	1-tiered.	Patients measure their vital signs regularly; this data is integrated with electronic health records.	The system is implemented for a small number of patients in one healthcare system.	No
Activity Monitoring application [37]	Cancer disease.	Smartphone Galaxy S5 mini, SIM card, the bracelet Everion,	1-tiered.	The patient is at home filling out a daily symptom questionnaire.	The limited sample size and the absence of a mechanism to test and assess the acceptance of wearable devices.	Yes.
The ASyMS application [38]	Cancer disease.	Android mobile phone	1-tiered.	Electronic symptom questionnaires to assess the presence, severity, and distress levels associated with various symptoms.	There is an absence of a mechanism to test and assess the acceptance of the application.	Yes.
Current study	General	Flutter, Laravel and Protégé	1-tiered.	The patient enters his vital data daily, compared with the semantic web rules, and medical advice is sent.	The limited sample size.	Yes

DOI 10.15588/1607-3274-2026-1-15

CONCLUSIONS

This study provides a detailed review of several remote patient monitoring (RPM) systems, focusing on common diseases such as diabetes, epilepsy, cardiovascular disease, chronic kidney disease, and cancer. The analysis highlights significant differences in technological approaches and identifies critical limitations, including issues of semantic consistency and contextual awareness.

The scientific novelty of the study is that researchers propose an innovative ontology-based framework for remote patient monitoring systems, integrating contact-based monitoring methods. This framework utilizes ontology, semantic web rules, and cloud computing to enable the delivery of scalable and efficient healthcare services.

The practical significance of the findings demonstrates the potential of a context-aware and semantically enriching platform to revolutionize telehealth services. By facilitating alignment and intelligent decision-making, the proposed system lays the foundation for remote patient monitoring. A data flow model is included to illustrate how patient context is integrated into context-aware messaging processes for monitoring patient health status. The practical significance of this research lies in its real-world applicability, providing developers and systems engineers with a clear blueprint for designing intelligent and adaptive remote patient monitoring platforms. This framework paves the way for improved patient care and reduced hospital visits.

Prospects for further research are focuses on implementing the proposed system in diverse healthcare settings and larger areas to assess its user acceptance. The study also calls for continued research in the fields of ontology engineering and remote monitoring.

ACKNOWLEDGEMENTS

The authors would like to thank the editorial board, especially Prof. Sergey Subbotin, for his valuable comments that helped enhance the quality of this study.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this study, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: Alaa Almagrabi: methodology of the study; Halimah Mafraq: collect data and interpret the results; Hana Almagrabi: technical side of the study.

Data availability: the data of the study are available to the corresponding author.

Software availability: The paper has no associated software.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies in creating this work.

REFERENCES

1. Malasinghe L. P., Ramzan N., Dahal K. Remote patient monitoring: a comprehensive study, *J Ambient Intell Human Comput*, 2019, No. 1, pp. 57–76. DOI: 10.1007/s12652-017-0598-x.
2. Gontarska K. et al. edited by Tucker A. et al. Predicting Medical Interventions from Vital Parameters: Towards a Decision Support System for Remote Patient Monitoring, *Artificial Intelligence in Medicine*. Cham, Springer International Publishing, 2021, pp. 293–297. DOI: 10.1007/978-3-030-77211-6_33.
3. Alomi Y. A., Aljudaibi S. M. National Survey of Total Parenteral Nutrition Practice in Saudi Arabia: Drug Monitoring and Patient Education at MOH Hospitals, *EC Nutr*, 2016, pp. 784–792. Available at: https://www.researchgate.net/publication/314231507
4. Qureshi B., Tounsi M. A Bluetooth enabled mobile intelligent remote healthcare monitoring system in Saudi Arabia: Analysis and design issues, *18th national computer conference. Citeseer*, 2006. Available at: https://www.researchgate.net/publication/244924911
5. Lin J. C. Applying telecommunication technology to healthcare delivery, *IEEE Engineering in Medicine and Biology Magazine*. IEEE, 1999, No. 4, pp. 28–31. DOI: 10.1109/51.775486.
6. Mainanwal V., Gupta M., Upadhayay S. K. A survey on wireless body area network: Security technology and its design methodology issue, *2015 international conference on innovations in information, embedded and communication systems (ICIIECS)*. IEEE, 2015, pp. 1–5. DOI: 10.1109/ICIIECS.2015.7193088.
7. Leal Filho W., Wall T., Azul A. M., Brandli L., Özuyar P. G. (eds) Remote patient monitoring (RPM), in: Good Health and Well-Being, *Encyclopedia of the UN Sustainable Development Goals*. Cham, Springer, 2020, pp. 583–583. DOI: 10.1007/978-3-319-95681-7_300115.
8. Miller K. M. et al. Current state of type 1 diabetes treatment in the US: updated data from the T1D Exchange clinic registry, *Diabetes care. Am Diabetes Assoc*, 2015, No. 6, pp. 971–978. DOI: 10.2337/dc15-0078.
9. Joubert M. et al. Remote monitoring of diabetes: a cloud-connected digital system for individuals with diabetes and their health care providers, *Journal of diabetes science and technology. SAGE Publications Sage CA*. Los Angeles, CA, 2019, No. 6, pp. 1161–1168. DOI: 10.1177/1932296819834054.
10. Kantorowska A. et al. Remote patient monitoring for management of diabetes mellitus in pregnancy is associated with improved maternal and neonatal outcomes, *American Journal of Obstetrics and Gynecology*, 2023. DOI: 10.1016/j.ajog.2023.02.015.
11. Moshé S. L. et al. Epilepsy: new advances, *The Lancet. Elsevier*, 2015, No. 9971, pp. 884–898. DOI: 10.1016/S0140-6736(14)60456-6.
12. Ricci L. et al. Clinical utility of home videos for diagnosing epileptic seizures: a systematic review and practical recommendations for optimal and safe recording, *Neurological Sciences*. Springer, 2021, pp. 1301–1309. DOI: 10.1007/s10072-021-05040-5.
13. Amin U. et al. Value of smartphone videos for diagnosis of seizures: everyone owns half an epilepsy monitoring unit, *Epilepsia. Wiley Online Library*, 2021, No. 9, pp. e135–e139. DOI: 10.1111/epi.17001.
14. Hassan S., Mwangi E., Kihato P. K. IoT based monitoring system for epileptic patients, *Heliyon*, 2022, No. 6, P. e09618. DOI: 10.1016/j.heliyon.2022.e09618.
15. Yang G. et al. IoT-Based Remote Pain Monitoring System: From Device to Cloud Platform, *IEEE J. Biomed. Health In-*

DOI 10.15588/1607-3274-2026-1-15

form, 2018, No. 6, pp. 1711–1719. DOI: 10.1109/JBHI.2017.2776351.
16. Biondi A. et al. Remote and Long-Term Self-Monitoring of Electroencephalographic and Noninvasive Measurable Variables at Home in Patients With Epilepsy (EEG@HOME): Protocol for an Observational Study, *JMIR Research Protocols*, 2021, No. 3, P. e25309. DOI: 10.2196/25309.
17. Skarpaas T. L., Jarosiewicz B., Morrell M. J. Brain-responsive neurostimulation for epilepsy (RNS® System), *Epilepsy Research*, 2019, pp. 68–70. DOI: 10.1016/j.eplepsyres.2019.02.003.
18. Burch R. C., Buse D. C., Lipton R. B. Migraine: epidemiology, burden, and comorbidity, *Neurologic clinics. Elsevier*, 2019, No. 4, pp. 631–649. DOI: 10.1016/j.ncl.2019.06.001.
19. Bonavita S. et al. Digital triage for people with multiple sclerosis in the age of COVID-19 pandemic, *Neurological sciences*. Springer, 2020, pp. 1007–1009. DOI: 10.1007/s10072-020-04391-9.
20. Ede van E. S. et al. Continuous remote monitoring in post-bariatric surgery patients: development of an early warning protocol, *Surgery for Obesity and Related Diseases*. Elsevier, 2022, No. 11, pp. 1298–1303. DOI: 10.1016/j.soard.2022.06.018.
21. Minen M.T. et al. User Design and Experience Preferences in a Novel Smartphone Application for Migraine Management: A Think Aloud Study of the RELAXaHEAD Application, *Pain Medicine*, 2019, No. 2, pp. 369–377. DOI: 10.1093/pm/pny080.
22. Aljaaf A.J. et al. H-Diary: Mobile Application for Headache Diary and Remote Patient Monitoring, *2018 11th International Conference on Developments in eSystems Engineering (DeSE)*. Cambridge, United Kingdom, IEEE, 2018, pp. 18–22. DOI: 10.1109/DeSE.2018.00010.
23. Berrouiguet S. et al. Fundamentals for future mobile-health (mHealth): a systematic review of mobile phone and web-based text messaging in mental health, *Journal of medical Internet research. JMIR Publications Toronto*. Canada, 2016, No. 6, P. e135. DOI: 10.2196/jmir.5066.
24. Cronin E. M. et al. Remote monitoring of cardiovascular devices: a time and activity analysis, *Heart Rhythm. Elsevier*, 2012, No. 12, pp. 1947–1951.
25. Mohammed J. et al. Internet of Things: Remote Patient Monitoring Using Web Services and Cloud Computing, *2014 IEEE International Conference on Internet of Things(iThings), and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom)*. Taipei, Taiwan, IEEE, 2014, pp. 256–263. DOI: 10.1109/iThings.2014.45.
26. Gunderson M., Melton G. B. Ambulatory Remote Patient Monitoring Beyond COVID-19: Engagement and Sustainment Considerations, *Mayo Clinic Proceedings*. Elsevier, 2022, No. 12, pp. 2184–2185. DOI: 10.1016/j.mayocp.2022.10.016.
27. Alboksmaty A. et al. Effectiveness and safety of pulse oximetry in remote patient monitoring of patients with COVID-19: a systematic review, *The Lancet Digital Health*, 2022, No. 4, pp. e279–e289. DOI: 10.1016/S2589-7500(21)00276-4.
28. Paganelli A. I. et al. A conceptual IoT-based early-warning architecture for remote monitoring of COVID-19 patients in wards and at home, *Internet of Things*, 2022, P. 100399. DOI: 10.1016/j.iot.2021.100399.
29. Sharma N. et al. A smart ontology-based IoT framework for remote patient monitoring, *Biomedical Signal Processing and Control*, 2021, P. 102717. DOI: 10.1016/j.bspc.2021.102717.
30. Webster A.C. et al. Chronic kidney disease, *The lancet*. Elsevier, 2017, Vol. 389, № 10075, pp. 1238–1252. DOI: 10.1016/S0140-6736(16)32064-5.
31. Scarpioni R., Manini A., Chiappini P. Remote patient monitoring in peritoneal dialysis helps reduce risk of hospitalization during Covid-19 pandemic, *J Nephrol*, 2020, No. 6, pp. 1123–1124. DOI: 10.1007/s40620-020-00822-0.
32. Markossian T. W. et al. A Mobile App to Support Self-management of Chronic Kidney Disease: Development Study, *JMIR Human Factors*, 2021, No. 4, P. e29197. DOI: 10.2196/29197.
33. Lee Y.-L. et al. Validation of Agitated Patient Remote Monitoring Alarm System in the Intensive Care Unit, *Studies in Health Technology and Informatics*, 2021, pp. 365–366. DOI: 10.3233/SHTI210747.
34. Emuoyibofarhe J. O. et al. A fuzzy rule-based model for remote monitoring of preterm in the intensive Care Unit of Hospitals, *International Journal of Medical Research & Health Sciences. International Journal of Medical Research & Health Sciences*, 2019, No. 5, pp. 33–3. Available at: https://www.researchgate.net/publication/333508315.
35. Garelli F. et al. Remote Glucose Monitoring Platform for Multiple Simultaneous Patients at Coronavirus Disease 2019 Intensive Care Units: Case Report Including Adults and Children, *Diabetes Technology & Therapeutics*, 2021, pp. 1–3. DOI: 10.1089/dia.2020.0556.
36. Bernier Carney K. M. et al. Communication of pain by school-age children with cancer using a game-based symptom assessment app: A secondary analysis, *European Journal of Oncology Nursing*, 2021, P. 101949. DOI: 10.1016/j.ejon.2021.101949.
37. Pavic M. et al. Feasibility and Usability Aspects of Continuous Remote Monitoring of Health Status in Palliative Cancer Patients Using Wearables, *Oncology*, 2020, No. 6, pp. 386–395. DOI: 10.1159/000501433.
38. Moradian S. et al. Usability Evaluation of a Mobile Phone–Based System for Remote Monitoring and Management of Chemotherapy-Related Side Effects in Cancer Patients: Mixed-Methods Study, *JMIR Cancer*, 2018, No. 2, P. e10932. DOI: 10.2196/10932.
39. Pritchett J. C. et al. Association of a remote patient monitoring (RPM) program with reduced hospitalizations in cancer patients with COVID-19, *JCO Oncology Practice. Wolters Kluwer Health*, 2021, No. 9, pp. e1293–e1302. DOI: 10.1200/OP.21.00307.
40. Ahmad B.I. et al. Remote patient monitoring system architecture for diabetes management, *2017 International Conference on Computing, Engineering, and Design (ICCED)*. Kuala Lumpur, IEEE, 2017, pp. 1–6. DOI: 10.1109/CED.2017.8308120.
41. Yang S. et al. Clinical utility of remote monitoring for patients with cardiac implantable electrical devices, *J Interv Card Electrophysiol*, 2023, No. 4, pp. 961–969. DOI: 10.1007/s10840-022-01406-7.
42. Pathinarupothi R. K., Ramesh M. V., Rangan E. Multi-layer architectures for remote health monitoring, *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2016, pp. 1–6. DOI: 10.1109/HealthCom.2016.7749443

Received 08.10.2025.
Accepted 14.01.2026.
Published 27.03.2026.

DOI 10.15588/1607-3274-2026-1-15

УДК 004.93

СТРУКТУРА ДЛЯ ДИСТАНЦІЙНОГО МОНІТОРИНГУ ПАЦІЄНТІВ У СИСТЕМІ ОХОРОНИ ЗДОРОВ'Я

Мафрак Х. І. – аспірант, кафедра інформаційних систем, факультет обчислювальної техніки та інформаційних технологій, Мекка, Університет короля Абдулазіза, Джидда, Саудівська Аравія; викладач, кафедра інформаційних систем, Університет короля Халіда, Абха, Саудівська Аравія. ROR: https://ror.org/052kwzs30. ORCID: https://orcid.org/0000-0003-0331-3021.

Альмаграбі А. О. – доцент, кафедра інформаційних систем, факультет обчислювальної техніки та інформаційних технологій, Мекка, Університет короля Абдулазіза, Джидда, Саудівська Аравія. ROR: https://ror.org/02ma4wv74. ORCID: https://orcid.org/0000-0002-4858-9366.

Альмаграбі Х. – доцент, кафедра інформаційних систем, факультет обчислювальної техніки та інформаційних технологій, Мекка, Університет короля Абдулазіза, Джидда, Саудівська Аравія. ROR: https://ror.org/02ma4wv74. ORCID: https://orcid.org/0000-0001-5497-6461.

АНОТАЦІЯ

Актуальність. Дистанційний моніторинг пацієнтів (RPM) відіграє ключову роль у трансформації охорони здоров'я, забезпечуючи безперервне відстеження здоров'я в режимі реального часу поза межами традиційних клінічних середовищ. Як наріжний камінь цифрових медичних послуг, RPM сприяє проактивним та профілактичним підходам до догляду.

Мета роботи. Ця стаття має на меті дослідити концепцію RPM, переглянути існуючі системи та запропонувати нову архітектуру платформи для підвищення ефективності, доступності та якості надання медичної допомоги.

Метод. Використовуючи якісний аналітичний метод, дослідження розглядає системи RPM, адаптовані до конкретних умов. Воно класифікує ці системи за режимом роботи, контактним чи безконтактним, та оцінює їхні технології, архітектури та пропоновані послуги. Крім того, воно представляє запропоновану онтологічну платформу RPM, що включає міркування на основі правил для посилення прийняття клінічних рішень.

Результати. Аналіз охоплює застосування RPM для таких станів, як діабет, епілепсія, головний біль, серцево-судинні захворювання, серцева недостатність, COVID-19, хронічна хвороба нирок, рак та непритомність. Він визначає сильні та недоліки існуючих систем та ілюструє, як запропонована архітектура вирішує ці проблеми, надаючи персоналізовані, масштабовані та ефективні рішення для моніторингу.

Висновки. Дослідження підкреслює зростаючу важливість RPM в охороні здоров'я та представляє інноваційну, онтологічно орієнтовану платформу для покращення надання послуг та результатів лікування пацієнтів. Подальші зусилля будуть зосереджені на клінічній валідації та оцінці ефективності в реальних умовах. Ця робота надає цінну інформацію для медичних працівників, розробників та політиків, які вдосконалюють рішення для дистанційної допомоги.

КЛЮЧОВІ СЛОВА: біомедична телеметрія, хвороби, фреймворк, медичні інформаційні системи, телемедицина.

ЛІТЕРАТУРА

1. Malasinghe L. P. Remote patient monitoring: a comprehensive study / L. P. Malasinghe, N. Ramzan, K. Dahal // J Ambient Intell Human Comput. – 2019. – No. 1. – P. 57–76. DOI: 10.1007/s12652-017-0598-x.
2. Predicting Medical Interventions from Vital Parameters: Towards a Decision Support System for Remote Patient Monitoring / K. Gontarska et al. edited by. A. Tucker et al. // Artificial Intelligence in Medicine. – Cham : Springer International Publishing. – 2021. – P. 293–297. DOI: 10.1007/978-3-030-77211-6_33.
3. Alomi Y. A. National Survey of Total Parenteral Nutrition Practice in Saudi Arabia: Drug Monitoring and Patient Education at MOH Hospitals / Y. A. Alomi, S. M. Aljudaibi // EC Nutr. – 2016. – P. 784–792. Available at: https://www.researchgate.net/publication/314231507
4. Qureshi B. A Bluetooth enabled mobile intelligent remote healthcare monitoring system in Saudi Arabia: Analysis and design issues / B. Qureshi, M. Tounsi // 18th national computer conference. Citeseer. – 2006. Available at: https://www.researchgate.net/publication/244924911
5. Lin J. C. Applying telecommunication technology to healthcare delivery / J. C. Lin // IEEE Engineering in Medicine and Biology Magazine. IEEE. – 1999. – No. 4. – P. 28–31. DOI: 10.1109/51.775486.
6. Mainanwal V. A survey on wireless body area network: Security technology and its design methodology issue / V. Mainanwal, M. Gupta, S. K. Upadhayay // 2015 international conference on innovations in information, embedded and communication systems (ICIIECS). IEEE. – 2015. – P. 1–5. DOI: 10.1109/ICIIECS.2015.7193088.
7. Remote patient monitoring (RPM), in: Good Health and Well-Being / [W. Leal Filho, T. Wall, A. M. Azul, L. Brandli, P. G. Özuyar (eds)] // Encyclopedia of the UN Sustainable Development Goals, Cham: Springer. – 2020. – P. 583–583. DOI: 10.1007/978-3-319-95681-7_300115.
8. Current state of type 1 diabetes treatment in the US: updated data from the T1D Exchange clinic registry / K. M. Miller et al. // Diabetes care. Am Diabetes Assoc. – 2015. – No. 6. – P. 971–978. DOI: 10.2337/dc15-0078.
9. Remote monitoring of diabetes: a cloud-connected digital system for individuals with diabetes and their health care providers / M. Joubert et al. // Journal of diabetes science and technology. SAGE Publications Sage CA: Los Angeles, CA. – 2019. – No. 6. – P. 1161–1168. DOI: 10.1177/1932296819834054.
10. Remote patient monitoring for management of diabetes mellitus in pregnancy is associated with improved maternal and neonatal outcomes / A. Kantorowska et al. // American Journal of Obstetrics and Gynecology. – 2023. DOI: 10.1016/j.ajog.2023.02.015.
11. Epilepsy: new advances / S. L. Moshé et al. // The Lancet. Elsevier. – 2015. – No. 9971. – P. 884–898. DOI: 10.1016/S0140-6736(14)60456-6.
12. Clinical utility of home videos for diagnosing epileptic seizures: a systematic review and practical recommendations for optimal and safe recording / L. Ricci et al. // Neurological Sciences. Springer. – 2021. – P. 1301–1309. DOI: 10.1007/s10072-021-05040-5.
13. Value of smartphone videos for diagnosis of seizures: everyone owns half an epilepsy monitoring unit / U. Amin et al. // Epilepsia. Wiley Online Library. – 2021. – No. 9. – P. e135–e139. DOI: 10.1111/epi.17001.
14. Hassan S. IoT based monitoring system for epileptic patients / S. Hassan, E. Mwangi, P. K. Kihato // Heliyon. – 2022. – No. 6. – P. e09618. DOI: 10.1016/j.heliyon.2022.e09618.
15. IoT-Based Remote Pain Monitoring System: From Device to Cloud Platform / G. Yang et al. // IEEE J. Biomed. Health

DOI 10.15588/1607-3274-2026-1-15

Inform. – 2018. – No. 6. – P. 1711–1719. DOI: 10.1109/JBHI.2017.2776351.
16. Remote and Long-Term Self-Monitoring of Electroencephalographic and Noninvasive Measurable Variables at Home in Patients With Epilepsy (EEG@HOME): Protocol for an Observational Study / A. Biondi et al. // JMIR Research Protocols. – 2021. – No. 3. – P. e25309. DOI: 10.2196/25309.
17. Skarpaas T. L. Brain-responsive neurostimulation for epilepsy (RNS® System) / T. L. Skarpaas, B. Jarosiewicz, M. J. Morrell // Epilepsy Research. – 2019. – P. 68–70. DOI: 10.1016/j.eplepsyres.2019.02.003.
18. Burch R. C. Migraine: epidemiology, burden, and comorbidity / R. C. Burch, D. C. Buse, R. B. Lipton // Neurologic clinics. Elsevier. – 2019. – No. 4. – P. 631–649. DOI: 10.1016/j.ncl.2019.06.001.
19. Digital triage for people with multiple sclerosis in the age of COVID-19 pandemic / S. Bonavita et al. // Neurological sciences. – Springer, 2020. – P. 1007–1009. DOI: 10.1007/s10072-020-04391-9.
20. Continuous remote monitoring in post-bariatric surgery patients: development of an early warning protocol / Ede van E. S. et al. // Surgery for Obesity and Related Diseases. – Elsevier. – 2022. – No. 11. – P. 1298–1303. DOI: 10.1016/j.soard.2022.06.018.
21. User Design and Experience Preferences in a Novel Smartphone Application for Migraine Management: A Think Aloud Study of the RELAXaHEAD Application / M. T. Minen et al. // Pain Medicine. – 2019. – No. 2. – P. 369–377. DOI: 10.1093/pm/pny080.
22. H-Diary: Mobile Application for Headache Diary and Remote Patient Monitoring / A. J. Aljaaf et al. // 2018 11th International Conference on Developments in eSystems Engineering (DeSE). Cambridge, United Kingdom: IEEE. – 2018. – P. 18–22. DOI: 10.1109/DeSE.2018.00010.
23. Fundamentals for future mobile-health (mHealth): a systematic review of mobile phone and web-based text messaging in mental health / S. Berrouiguet et al. // Journal of medical Internet research. – JMIR Publications Toronto, Canada. – 2016. – No. 6. – P. e135. DOI: 10.2196/jmir.5066.
24. Remote monitoring of cardiovascular devices: a time and activity analysis / E. M. Cronin et al. // Heart Rhythm. – Elsevier. – 2012. – No. 12. – P. 1947–1951.
25. Internet of Things: Remote Patient Monitoring Using Web Services and Cloud Computing / J. Mohammed et al. // 2014 IEEE International Conference on Internet of Things(iThings), and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom). – Taipei, Taiwan : IEEE. – 2014. – P. 256–263. DOI: 10.1109/iThings.2014.45.
26. Gunderson M. Ambulatory Remote Patient Monitoring Beyond COVID-19: Engagement and Sustainment Considerations / M. Gunderson, G. B. Melton // Mayo Clinic Proceedings. Elsevier. – 2022. – No. 12. – P. 2184–2185. DOI: 10.1016/j.mayocp.2022.10.016.
27. Effectiveness and safety of pulse oximetry in remote patient monitoring of patients with COVID-19: a systematic review / A. Alboksmaty et al. // The Lancet Digital Health. – 2022. – No. 4. – P. e279–e289. DOI: 10.1016/S2589-7500(21)00276-4.
28. A conceptual IoT-based early-warning architecture for remote monitoring of COVID-19 patients in wards and at home / A. I. Paganelli et al. // Internet of Things. – 2022. – P. 100399. DOI: 10.1016/j.iot.2021.100399.
29. A smart ontology-based IoT framework for remote patient monitoring / N. Sharma et al. // Biomedical Signal Processing and Control. – 2021. – P. 102717. DOI: 10.1016/j.bspc.2021.102717.
30. Chronic kidney disease / A. C. Webster et al. // The lancet. Elsevier. – 2017. – Vol. 389, № 10075. – P. 1238–1252. DOI: 10.1016/S0140-6736(16)32064-5.
31. Scarpioni R. Remote patient monitoring in peritoneal dialysis helps reduce risk of hospitalization during Covid-19 pandemic / R. Scarpioni, A. Manini, P. Chiappini // J Nephrol. – 2020. – No. 6. – P. 1123–1124. DOI: 10.1007/s40620-020-00822-0.
32. A Mobile App to Support Self-management of Chronic Kidney Disease: Development Study / T. W. Markossian et al. // JMIR Human Factors. – 2021. – No. 4. – P. e29197. DOI: 10.2196/29197.
33. Validation of Agitated Patient Remote Monitoring Alarm System in the Intensive Care Unit / Y.-L. Lee et al. // Studies in Health Technology and Informatics. – 2021. – P. 365–366. DOI: 10.3233/SHTI210747.
34. A fuzzy rule-based model for remote monitoring of preterm in the intensive Care Unit of Hospitals / J. O. Emuoyibofarhe et al. // International Journal of Medical Research & Health Sciences. International Journal of Medical Research & Health Sciences. – 2019. – No. 5. – P. 33–3. Available at: https://www.researchgate.net/publication/333508315.
35. Remote Glucose Monitoring Platform for Multiple Simultaneous Patients at Coronavirus Disease 2019 Intensive Care Units: Case Report Including Adults and Children / F. Garelli et al. // Diabetes Technology & Therapeutics. – 2021. – P. 1–3. DOI: 10.1089/dia.2020.0556.
36. Communication of pain by school-age children with cancer using a game-based symptom assessment app: A secondary analysis / K. M. Bernier Carney et al. // European Journal of Oncology Nursing. – 2021. – P. 101949. DOI: 10.1016/j.ejon.2021.101949.
37. Feasibility and Usability Aspects of Continuous Remote Monitoring of Health Status in Palliative Cancer Patients Using Wearables / M. Pavic et al. // Oncology. – 2020. – No. 6. – P. 386–395. DOI: 10.1159/000501433.
38. Usability Evaluation of a Mobile Phone-Based System for Remote Monitoring and Management of Chemotherapy-Related Side Effects in Cancer Patients: Mixed-Methods Study / S. Moradian et al. // JMIR Cancer. – 2018. – No. 2. – P. e10932. DOI: 10.2196/10932.
39. Association of a remote patient monitoring (RPM) program with reduced hospitalizations in cancer patients with COVID-19 / J. C. Pritchett et al. // JCO Oncology Practice. Wolters Kluwer Health. – 2021. – No. 9. – P. e1293–e1302. DOI: 10.1200/OP.21.00307.
40. Remote patient monitoring system architecture for diabetes management / Ahmad B. I. et al. // 2017 International Conference on Computing, Engineering, and Design (ICCED). Kuala Lumpur : IEEE. – 2017. – P. 1–6. DOI: 10.1109/CED.2017.8308120.
41. Clinical utility of remote monitoring for patients with cardiac implantable electrical devices / S. Yang et al. // J Interv Card Electrophysiol. – 2023. – No. 4. – P. 961–969. DOI: 10.1007/s10840-022-01406-7.
42. Pathinarupothi R. K. Multi-layer architectures for remote health monitoring / R. K. Pathinarupothi, M. V. Ramesh, E. Rangan // 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom). IEEE. – 2016. – P. 1–6. DOI: 10.1109/HealthCom.2016.7749443.

DOI 10.15588/1607-3274-2026-1-15

UDC 004.056.55

MODIFIED BIOMETRIC TEMPLATE PROTECTION METHOD WITH NONLINEAR TRANSFORMATIONS

Onai M. V. – PhD, Associate Professor of the Department of Computer Systems Software, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine. ROR: https://ror.org/00syn5v21. ORCID: https://orcid.org/0000-0002-4938-8355.

Kosenko O. V. – Bachelor student of the Department of Computer Systems Software, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine. ROR: https://ror.org/00syn5v21. ORCID: https://orcid.org/0009-0001-2115-7918.

ABSTRACT

Context. Biometric data is a common option for authentication or identification. However, it is vulnerable and not replaceable in case of stealing. Several methods for constructing protected biometric templates are proposed in literature, one of them is biohashing. However, linearity of biohashing may be a vulnerability. MLP-hash is similar, but adds nonlinearity. It is modified in this work.

Objective. The goal of this work is to develop a modification of MLP-hash which is faster and allows better separation of users by their templates.

Method. This work focuses on modifying MLP-hash, a biohashing variation with nonlinear transformations. One of the proposed changes is the usage of normalization before applying nonlinear transformation in each layer of MLP-hash. Different normalization methods are investigated and compared. The other proposed change is the simplification of the pseudorandom matrices used in each layer of MLP-hash. Each such matrix is replaced by a block matrix in which blocks that are laying on the diagonal are orthonormal matrices and all other blocks are filled with zeros. Each nonzero block is generated from the user's secret token. In order to make the effect of each nonzero block less localized, a pseudorandom permutation is added before each matrix multiplication and also after all layers. Pseudorandom permutations are also generated with the user's secret token as seed. The proposed method can be used in a similar way to how original MLP-hash and biohashing methods are used: it takes the user's secret token and biometric vector of fixed length and outputs a binary vector of fixed length with the same or smaller dimensionality. MLP-hash with block matrices is compared to the original while applying different normalization techniques and different nonlinear transformations.

Results. The proposed modifications, original MLP-hash and biohashing have been implemented in code. Speed and accuracy of user separation with the usage of these methods have been compared on feature vectors extracted from fingerprints with the usage of Gabor filters.

Conclusions. The conducted experiments have shown an increase of speed and ability to separate user templates from the substitution of proposed block matrices and an increase of ability to separate user templates from the usage of normalization. Comparison of different normalizations and nonlinear transformation has also been conducted. The practical usefulness of the developed method is that it is faster and can be used in applications when users expect no delays while still being difficult to invert. The prospects for further research include testing this method with other biometric modalities, other nonlinear transformations and normalization techniques and an analysis of inversion difficulty of the developed method in comparison to MLP-hash and biohashing.

KEYWORDS: biometrics, biometric template, biometric template protection, one-way transformation, biohashing, elliptic curve cryptography, finite fields.

ABBREVIATIONS

AAD is an average absolute deviation;
EER is an equal error rate;
FAR is a false acceptance rate;
FRR is a false rejection rate;
MLP is a multilayer perceptron;
RNG is a random number generator;
ROI is a region of interest.

NOMENCLATURE

B is a pseudorandom orthonormal matrix used in biohashing;

$\mathbf{B}_{\text{block } i}$ is an i-th diagonal block of block matrix **B**;

$\mathbf{B}_i$ is a pseudorandom orthonormal matrix of i-th layer of MLP-hash or its modifications;

b is a number of nonzero blocks;

b is a binary vector resulting from the usage of biohashing, MLP-hash or its modifications;

b_i is an i-th element of **b**;

$b_{\text{layer } i}$ is a number of nonzero blocks in i-th layer;

$F_{i,\theta}(x, y)$ is a value of pixel at position (x, y) of i-th region of ROI tiling after applying Gabor filter with angle θ;

$\mathbf{F}_x$ is a Sobel filter for estimating gradient along x-axis;

$\mathbf{F}_y$ is a Sobel filter for estimating gradient along x-axis;

f() is a nonlinear transformation;

$f_{\text{custom}}()$ is a custom nonlinear transformation defined in this work;

G is a Gabor filter;

$\mathbf{G}_x$ is a gradient estimation along x-axis;

$\mathbf{G}_y$ is a gradient estimation gradient along y-axis;

H(**a**, **b**) is a hamming distance between binary vectors **a**, **b**;

l is an amount of layers of MLP-hash or its modification;

$M(i, j)$ is a gradient magnitude at position (i, j);

m is a dimensionality of an output of biohashing or its modifications;

DOI 10.15588/1607-3274-2026-1-16

m_i is a dimensionality of an output of i-th layer of MLP-hash or its modifications;

N_{fa} is a number of false acceptances;

N_{fr} is a number of false rejections;

N_{ta} is a number of correct authentications;

N_{tr} is a number of correct rejections;

n is a dimensionality of biometrics feature vector;

n_c is a number of circles breaking the ROI;

n_d is a number of directions breaking the ROI;

n_g is a number of Gabor filters used;

n_i is a number of pixels in the i-th region of ROI tiling;

n_p is a number of pseudorandom projections in one layer;

n_T is a number of threshold values for FRR and FAR calculations;

O() is a worst-case algorithm complexity;

P(i, j) is a Poincare index at position (i, j);

$\mathbf{P}_i$ is a pseudorandom permutation used in i-th layer;

$\mathbf{P}_{i,1}$ is a pseudorandom permutation used before projection of i-th layer;

$\mathbf{P}_{i,2}$ is a pseudorandom permutation used after projection of i-th layer;

$P_{i,\theta}$ is an average of pixel values in the i-th region of ROI tiling after applying Gabor filter with angle θ;

p_i is a number of pseudorandom projections used in parallel in i-th layer;

$\mathbb{R}$ is the set of all real numbers;

r_{inner} is a radius of an innermost circle in ROI;

r_{ROI} is a radius of ROI;

S(**a**, **b**) is a cosine similarity of vectors **a**, **b**;

T is a threshold of a classifier;

t is a secret token;

$V_{i,\theta}$ is an AAD in the i-th region of ROI tiling after applying Gabor filter with angle θ;

$v_x(i, j)$,$v_y(i, j)$ – components of an orientation field;

w is a window size of an orientation field;

x is a biometric feature vector;

y is a result of applying pseudorandom projection to feature vector;

$\mathbf{y}_i$ is an output of i-th layer of MLP-hash or its modification;

$\mathbf{y}_{i,\text{block } j}$ is a part of vector $\mathbf{y}_i$ corresponding to the j-th block;

γ is an eccentricity of a Gabor filter;

$\Delta_k(i, j)$ is a k-th orientation difference at position (i, j) restricted by range $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right]$;

$\delta_k(i, j)$ is a k-th orientation difference at position (i, j);

θ is an angle of a Gabor filter;

θ(i, j) is an orientation at position (i, j);

$\theta_k(i, j)$ is an orientation at k-th neighbour of position (i, j);

λ is a wavelength of a Gabor filter;

$\mu_{\mathbf{x}}$ is a standard deviation of elements within **x**;

σ is a standard deviation of a Gabor filter;

$\sigma_{\mathbf{x}}$ is a standard deviation of elements within **x**;

τ is a binarization threshold;

ψ is a phase shift of a Gabor filter.

INTRODUCTION

Authentication and identification based on biometric data has widespread usage, including device access control, biometric ID-cards, area access control etc. During the process of identification or authentication two biometric templates formed from features of users' biometric data are compared to decide if they belong to the same person. Because biometric samples of the same person slightly vary this comparison is not a full match but rather similarity calculation.

Biometric data is private and its usage and protection is heavily regulated by laws. Biometric templates are usually constructed from the most prominent features, which means that its privacy is also required. There are different methods of biometric template protection, which transform biometric data in a way that prevents inversion but turns similar biometric templates into similar protected templates.

The object of study is the process of biometric template protection.

The process of biometric template protection transforms biometric features to a representation that conceals them while allowing user identification or authentication by measuring similarity between templates. This process should be constructed with regard to variations in biometric samples.

The subject of study is biohashing, a biometric template protection method based on pseudorandom projections and modifications of this method.

This work focuses on a modification of biohashing called MLP-hash, which adds nonlinear transformations to make it harder to invert.

The purpose of the work is to increase the speed of MLP-hash and to increase accuracy of user separation by classifiers based on protected templates generated by it.

1 PROBLEM STATEMENT

Required properties of biometric template protection methods include [1]:

1) non-reversibility: it should not be possible to get original biometric data from protected template;

2) accuracy: biometric system should not lose accuracy from the transformation used for protection;

3) diversity: users should be able to create different templates that are not linkable;

4) revocability: ability to replace template in case it gets stolen.

This work focuses on modifying a biometric template protection method based on biohashing [2]. The resulting method must have all properties listed above, and should be usable in the same way as biohashing: taking as input biometric feature vector $\mathbf{x} \in \mathbb{R}^n$ and secret token t and producing $\mathbf{b} \in \{0, 1\}^m$, $m \leq n$. Besides that the developed modification must be faster than the original MLP-hash.

DOI 10.15588/1607-3274-2026-1-16

2 REVIEW OF THE LITERATURE

There are several biometric template protection methods described in literature. These include projection-based methods like biohashing [2], bloom filters [3], index-of-max hashing [4] etc. These methods apply one-way transformation to biometric features to conceal them.

Biohashing is based on pseudorandom projection. This projection depends on the user's secret token *t*.

Algorithm 1 – Biohashing

Input: $\mathbf{x} \in \mathbb{R}^n$, t
Output: $\mathbf{b} \in \{0, 1\}^m$
1. Generate pseudorandom matrix **B** using *t* as seed
2. Apply Gram-Schmidt process to **B**
3. $\mathbf{y} \leftarrow \mathbf{B} \cdot \mathbf{x}$
4. for $i = 1$ to m do
  4.1. if $y_i < \tau$ then $b_i \leftarrow 0$
  4.2. else $b_i \leftarrow 1$
  4.3. end if
5. end for
6. return **b**

Several improvements to base biohashing are proposed in [5]. These include:

1) normalization of feature vectors before applying biohashing;

2) usage of several projection spaces to increase result dimensionality;

3) usage of several feature permutations to increase result dimensionality.

Binarization and dimension reduction of the feature vector after projection make this process a one-way transformation. However, some authors raise concerns about the linear nature of this transformation saying that this transformation may be partially reversed. In [6] there is a demonstration of reversal for projection-based methods. Although results are demonstrated for lower dimensions than those usually used in practice, this demonstration shows that linear nature is in fact a vulnerability of projection-based methods.

Although biohashing creates a protected template that does not fully reveal original biometric in case of being stolen, the created template should still be stored and transmitted securely and should be renewed immediately upon a suspicion of being compromised, else an attacker may use it for impersonation and unauthorized access may be gained. While being sent through an unsafe communication channel this template may be additionally encrypted with the use of a symmetric cipher such as AES or an asymmetric one, for example with the use of elliptic curve cryptography.

MLP-hash [7] is a modification of biohashing which adds nonlinear transformations to it. It contains l pseudorandom projections using matrices $\mathbf{B}_i$ of size $m_i \times m_{i-1}$, $m_0 = n$, $m_l = m$.

Algorithm 2 – MLP-hash

Input: $\mathbf{x} \in \mathbb{R}^n$, t
Output: $\mathbf{b} \in \{0, 1\}^m$
1. $\mathbf{y}_0 \leftarrow \mathbf{x}$
2. for $i = 1$ to l do
  2.1. Generate pseudorandom matrix $\mathbf{B}_i$ using *t* as seed
  2.2. Apply Gram-Schmidt process to $\mathbf{B}_i$
  2.3. $\mathbf{y}_i \leftarrow f(\mathbf{B}_i \cdot \mathbf{y}_{i-1})$
3. end for
4. for $i = 1$ to m do
  4.1. if $y_i < \tau$ then $b_i \leftarrow 0$
  4.2. else $b_i \leftarrow 1$
  4.3. end if
5. end for
6. return **b**

Nonlinear transformation makes it more difficult to invert a protected template, especially when this transformation is a many-to-one function, which is in general irreversible. In [7], a ReLU (1) is used as a nonlinear transformation, which is a many-to-one function.

$$\mathrm{ReLU}(x) = \begin{cases} 0 \text{ if } x < 0; \\ x \text{ if } x \geq 0. \end{cases} \quad (1)$$

3 MATERIALS AND METHODS

One of the proposed modifications of MLP-hash is the generalization of improvement proposed in [5], namely, the usage of normalization, for the multilayer structure of MLP-hash by applying normalization before each nonlinear transformation. Experimental results from [5] show that usage of normalization before biohashing improves EER in comparison to base biohashing, therefore it is expected that the usage of normalization in MLP-hash will also improve EER in comparison to base MLP-hash, however, the conclusion can only be driven from experimental evidence. Different normalization methods are tried:

1) normalization by range of values (so that all elements of a normalized vector are within the range [− 1, 1]):

$$2 \cdot \frac{\mathbf{x} - \min\limits_{i=1,\ldots,n} x_i}{\max\limits_{i=1,\ldots,n} x_i - \min\limits_{i=1,\ldots,n} x_i} - 1;$$

2) l_2-normalization (so that the length of a normalized vector equals 1):

$$\frac{\mathbf{x}}{|\mathbf{x}|};$$

3) statistical normalization (so that mean of elements in vector is 0 and standard deviation is 1):

$$\frac{\mathbf{x} - \mu_{\mathbf{x}}}{\sigma_{\mathbf{x}}},$$

where $\mu_{\mathbf{x}}$ is the mean of elements of **x** and $\sigma_{\mathbf{x}}$ is the standard deviation of elements in **x**. It should be noted

DOI 10.15588/1607-3274-2026-1-16

that μ and σ are calculated for elements within one vector, not across vectors, so this normalization should be viewed as a soft normalization by range that fits most but not all elements into range [− 1, 1] rather than a statistical tool.

Another proposed modification is the simplification of pseudorandom matrices used in MLP-hash layers by replacing them with block matrices of the following form:

$$\mathbf{B}_i = \begin{pmatrix} \mathbf{B}_{i,\text{ block }1} & 0 & \cdots & 0 \\ 0 & \mathbf{B}_{i,\text{ block }2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{B}_{i,\text{ block }b} \end{pmatrix},$$

where $\mathbf{B}_{i,\text{ block }j}$ is a pseudorandom orthonormal matrix of size $\frac{m_i}{b_{\text{layer }i}} \times \frac{m_{i-1}}{b_{\text{layer }i}}$, $m_0 = n$, $m_l = m$. It should be noted that both m_i and m_{i-1} must be divisible by $b_{\text{layer }i}$. If this modification is intended to be used with layer sizes not divisible by block number, it should be extended to work with blocks of different sizes or with intersecting blocks. Whole matrix $\mathbf{B}_i$ is also orthonormal: scalar product of rows intersecting the same block is 0 because extending block rows by zeros does not change it, scalar product of rows intersecting different blocks is 0 because there is no position at which both vectors contain nonzero element, and the norm of each row is still 1, since it is only affected by nonzero elements which are contained in diagonal blocks and these blocks are orthonormal. Thus simplified matrix still defines a pseudorandom projection. It can be thought of as breaking a feature vector into smaller vectors and projecting each using nonzero blocks.

The values in a biometric feature vector may be correlated, often this correlation is stronger in nearby positions. This means that even projected fragments of vector may form patterns similar to those in the original. In order to break locality of nonzero blocks a pseudorandom permutation is introduced. Although such permutations can be inserted before and after each projection in a multilayer structure, forming layers of the following structure:

$$\mathbf{y}_i = \mathrm{f}(\mathbf{P}_{i,2} \cdot \mathbf{B}_i \cdot \mathbf{P}_{i,1} \cdot \mathbf{y}_{i-1}),$$

some of them are redundant if nonlinear transformation can be defined as applying a function to each element of vector that depends only on selected element and an unordered collection of other elements (examples include just applying any nonlinear function element-wise, normalization by methods listed above or a combination of the two), since permutation that happens before such transformation can be brought outside and can be combined with next layers first permutation:

$$\begin{aligned} &\mathbf{P}_{i+1,1} \cdot \mathrm{f}(\mathbf{P}_{i,2} \cdot \mathbf{B}_i \cdot \mathbf{P}_{i,1} \cdot \mathbf{y}_{i-1}) = \\ &= \mathbf{P}_{i+1,1} \cdot \mathbf{P}_{i,2} \cdot \mathrm{f}(\mathbf{B}_i \cdot \mathbf{P}_{i,1} \cdot \mathbf{y}_{i-1}) = \\ &= \mathbf{P}_{i+1,\text{ combined}} \cdot \mathrm{f}(\mathbf{B}_i \cdot \mathbf{P}_{i,1} \cdot \mathbf{y}_{i-1}), \end{aligned}$$

therefore only one of the two permutations is necessary for each layer except the last, after which a pseudorandom permutation is used. Layer structure is then defined as follows:

$$\mathbf{y}_i = \mathrm{f}(\mathbf{B}_i \cdot \mathbf{P}_i \cdot \mathbf{y}_{i-1}).$$

Note that although a pseudorandom permutation is represented here as matrix multiplication for shorter notation, it does not need to be a matrix multiplication in an actual implementation.

Multiplying matrix of size $m \times n$ by vector of size n consists of $m \times n$ products and $m \times (n - 1)$ sums, so it has time complexity $O(mn)$. In comparison, multiplication of the simplified matrix of size $m \times n$ with b blocks (where m and n are divisible by b) by vector of size n can be decomposed into b multiplications of matrix of size $\frac{m}{b} \times \frac{n}{b}$ by vector of size $\frac{n}{b}$, which means that it has time complexity $O\left(\frac{mn}{b}\right)$. Random permutation in a collection of size n requires at most n copying operations and has a time complexity of $O(n)$. Therefore, a modified projection has a time complexity of $O\left(\frac{mn}{b} + n\right)$. To generate an orthonormal matrix of size $m \times n$, $m \times n$ random numbers need to be generated and then Gram-Schmidt process needs to be applied. Gram-Schmidt process for matrix of size $m \times n$ consists of subtracting projections of rows on rows above them. For a k-th row, $k - 1$ projections must be calculated, each consisting of a dot product of vectors of size n (consisting of n multiplications and $n - 1$ additions) and of product of resulting scalar and vector of size n (n multiplications). Subtraction of a projection consists of n operations. To make the resulting matrix orthonormal, each row is normalized after subtracting projections (n multiplications and $n - 1$ additions and one square root computation). Therefore, applying Gram-Schmidt process to matrix of size $m \times n$ requires

$$\begin{aligned} &\sum_{k=1}^{m}((k-1)(4n-1)+2n) = \\ &= \frac{m(m-1)}{2}(4n-1)+2mn \end{aligned}$$

arithmetic operations and therefore has a time complexity $O(m^2 n)$, ignoring the difference between addition, subtraction and multiplication. For a simplified matrix with b blocks only $\frac{mn}{b}$ numbers must be generated, and

DOI 10.15588/1607-3274-2026-1-16

Gram-Schmidt process on one matrix of size $\frac{m}{b}\times\frac{n}{b}$ has $O\left(\frac{m^2 n}{b^3}\right)$ time complexity. Generation of a random permutation has $O(n)$ time complexity. In summary, generation and usage of the original pseudorandom projection has time complexity $O(m^2 n)$, while generation and usage of simplified projection has time complexity $O\left(\frac{m^2 n}{b^2}+\frac{mn}{b}+n\right)$.

The simplification of matrices reduces not only time complexity of the method, but also space complexity. The projection matrix of size $m \times n$ requires storage for mn floating-point numbers. Gram-Schmidt process (if it is done in-place) requires additional storage for n numbers for projections it creates (at most one projection needs to be stored at the same time) plus a constant amount of numerical variables. Multiplying matrix of size $m \times n$ by vector of size n produces vector of size n. For a modified method, only $\frac{mn}{b}$ numbers are required to represent a projection matrix and Gram-Schmidt process needs to store only $\frac{n}{b}$ plus a constant amount of number variables at once. A pseudorandom permutation of a vector of size n used in modified layers requires $O(n\ln n)$ storage space (considering that each number stored must be an index of an array of size n). In summary, while generation and usage of pseudorandom projection with projection matrix of size $m \times n$ has space complexity $O(mn)$, generation and usage of simplified projection has space complexity $O\left(\frac{mn}{b}+m+n\ln n\right)$.

It should be taken into consideration that while proposed modification of projection matrices significantly reduces computational complexity, it also makes the process easier to partially invert, which should be taken into consideration while choosing the value of parameter b.

A method combining both of the proposed modifications is defined by the following algorithm.

Algorithm 3 – Modified MLP-hash

Input: $\mathbf{x} \in \mathbb{R}^n$, t
Output: $\mathbf{b} \in \{0, 1\}^m$
1. $\mathbf{y}_0 \leftarrow \mathbf{x}$
2. for $i = 1$ to l do
  2.1. Initialize $\mathbf{y}_i$
  2.2. Apply a pseudorandom permutation to $\mathbf{y}_{i-1}$ with seed t
  2.3. for $j = 1$ to $b_{\text{layer } i}$ do
    2.3.1. Generate pseudorandom matrix $\mathbf{B}_{i,\text{ block } j}$ using t as seed
    2.3.2. Apply Gram-Schmidt process to $\mathbf{B}_{i,\text{ block } j}$
    2.3.3. $\mathbf{y}_{i,\text{ block } j} \leftarrow \mathbf{B}_{i,\text{ block } j} \cdot \mathbf{y}_{i-1,\text{ block } j}$
  2.4. end for
  2.5. normalize $\mathbf{y}_i$
  2.6. $\mathbf{y}_i = f(\mathbf{y}_i)$
3. end for
4. Apply a pseudorandom permutation to $\mathbf{y}_l$ with RNG seed t
5. for $i = 1$ to m do
  5.1. if $y_i < \tau$ then $b_i \leftarrow 0$
  5.2. else $b_i \leftarrow 1$
  5.3. end if
6. end for
7. return $\mathbf{b}$

The proposed modification is tested with several different nonlinear transformations, including ReLU, Leaky ReLU, tanh and a custom function with sine component (2), which is denoted as f_{custom} in this work.

$$\text{LeakyReLU}(x)=\begin{cases} ax & \text{if} \quad x<0; \\ x & \text{if} \quad x\geq 0, \end{cases}$$
$$\tanh(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}},$$
$$f_{\text{custom}}(x)=ax+b\,|\,x\,|\sin x. \qquad (2)$$

An additional modification is added to MLP-hash as a generalization of spaces augmentation from [5], which increases dimensionality of biohashing results by using several projection spaces for one input vector and concatenating projected vectors. In this work this principle is used for each layer by replacing layer matrices of size $m_i \times m_{i-1}$ with p_i orthonormal matrices of size $m_i \times m_{i-1}p_{i-1}$ (note that dimensionality of $\mathbf{y}_{i-1}$ increases as well), multiplying $\mathbf{y}_{i-1}$ by each of them and concatenating results into vector of size $m_i \times p_i$. Note that with this modification being implemented spaces augmentation can still be removed from some layers if unnecessary by just setting the corresponding p_i to 1.This modification can also be extended to simplified projections proposed in this work by generating $b_i p_i$ blocks of size $\frac{m_i}{b_i}\times\frac{m_{i-1}p_i}{b_i}$ per each layer matrix instead of b_i, multiplying each $\mathbf{y}_{i-1,\text{ block } j}$ by p_i of these blocks and concatenating resulting vectors. In this work this modification is used only to increase layer sizes for computational speed testing.

4 EXPERIMENTS

The developed method is tested on feature vectors extracted from fingerprint images from the FVC2000 dataset [8]. This dataset contains fingerprints of 10 people, 8 images per fingerprint. These images are grayscale with brightness in range [0,255] and have 300×300 pixels.

Feature vectors are extracted from fingerprint images using a bank of Gabor filters by method similar to the one proposed in [9].

First of all, a pivot point is chosen. In this work it is a singularity point. Singularity point detection is based on

DOI 10.15588/1607-3274-2026-1-16

an orientation field [10], which indicates ridge directions. Before the orientation field computation image is blurred using a Gaussian filter to reduce the effect of noise. Then, image gradient is estimated using Sobel filters [11] of the following form:

$$\mathbf{F}_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}, \quad \mathbf{F}_y = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix},$$

producing gradient estimations $\mathbf{G}_x$ and $\mathbf{G}_y$ along the x and y axes respectively. These are used for computing components of orientation field within non-overlapping windows of size $w \times w$:

$$\mathrm{v}_x(i,j) = 2\sum_{u=iw}^{iw+w-1}\sum_{v=jw}^{jw+w-1}\mathbf{G}_x(u,v)\cdot\mathbf{G}_y(u,v),$$

$$\mathrm{v}_y(i,j) = \sum_{u=iw}^{iw+w-1}\sum_{v=jw}^{jw+w-1}(\mathbf{G}_x(u,v)^2 - \mathbf{G}_y(u,v)^2).$$

These components are then blurred using a Gaussian filter and then orientation is computed as:

$$\theta(i,j) = \frac{1}{2}\arctan\left(\frac{\mathrm{v}_x(i,j)}{\mathrm{v}_y(i,j)}\right).$$

Another value computed from image gradient is the gradient magnitude:

$$\mathrm{M}(i,j) = \sum_{u=iw}^{iw+w-1}\sum_{v=jw}^{jw+w-1}(\mathbf{G}_x(u,v)^2 + \mathbf{G}_y(u,v)^2).$$

It can be used as an estimation of image quality.

Singular points are detected with the use of Poincare index calculated on orientation field. To get its value at some position, 8 positions around it are numbered counterclockwise as $\theta_k(i,j)$, $k = 1, 2, ..., 8$. Differences between them are calculated as $\delta_k(i,j) = \theta_{k+1}(i,j) - \theta_k(i,j)$ for $k = 1, ..., 7$, $\delta_8(i,j) = \theta_1(i,j) - \theta_8(i,j)$, and changed to choose the smaller angle between directions (since directions of fingerprint ridges are ambiguous and may be changed by $\pm\pi$ without changing their meaning):

$$\Delta_k(i,j) = \begin{cases} \delta_k(i,j) & \text{if} \quad |\delta_k(i,j)| < \frac{\pi}{2}; \\ \pi + \delta_k(i,j) & \text{if} \quad \delta_k(i,j) \le -\frac{\pi}{2}; \\ \pi - \delta_k(i,j) & \text{otherwise.} \end{cases}$$

Poincare index is computed from these differences as:

$$\mathrm{P}(i,j) = \frac{1}{2}\sum_{k=1}^{8}\Delta_k(i,j).$$

For closed curves (and closed 8-connected neighbourhood) Poincare index takes one of these values: $-\frac{1}{2}, 0, \frac{1}{2}, 1$, with values different from 0 corresponding to singularity points.

If several singularity points are detected, the pivot point is determined by their average, weighted by gradient magnitude, so that singular points detected in regions with worse quality (which have a higher chance of being wrong) contribute less to the resulting point. If no singular points are detected, the image is discarded.

After the pivot point is chosen, a region of interest is selected. In this work ROI is similar to that from [9] and is bounded by circle of radius r_{ROI} centered at the pivot point and separated into sectors by n_c concentric circles having radiuses $r_{\mathrm{inner}}, r_{\mathrm{inner}} + \frac{r_{\mathrm{ROI}} - r_{\mathrm{inner}}}{n_c - 1}, ..., r_{\mathrm{ROI}}$ (region bound by r_{inner} is not used because it is highly sensitive to changes in pivot position) and by n_d directions defined by angles $0, \frac{2\pi}{n_d}, ..., \frac{2\pi(n_d-1)}{n_d}$, partitioning ROI into $(n_c - 1)n_d$ ring sectors in total.

Gabor filters are matrix filters and they are used to extract texture features from images. Filters used in this work are the real parts of Gabor filters and have the form:

$$\mathbf{G}(x,y;\lambda,\theta,\sigma,\gamma,\psi) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos\left(\frac{2\pi x'}{\lambda}+\psi\right),$$

where $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$. In this work, $\gamma = 1$ (so that gaussian component is radially symmetric), $\psi = 0$ (so that cosine wave peaks at line crossing the filter center), $\sigma = 5$ (chosen empirically), $\lambda = 10$ (chosen empirically to roughly match waves formed by ridges) and θ takes n_g different values to form a Gabor filter bank: $0, \frac{\pi}{n_g}, ..., \frac{\pi(n_g-1)}{n_g}$.

Before Gabor filters are applied an image is normalized so that the mean pixel value is 0 and standard deviation is 1. Then, filters are applied one at a time. For each filter direction θ and each ring sector of ROI an average absolute deviation from mean is calculated as:

$$V_{i,\theta} = \frac{1}{n_i}\left(\sum_{n_i}\left|\mathrm{F}_{i,\theta}(x,y) - P_{i,\theta}\right|\right).$$

DOI 10.15588/1607-3274-2026-1-16

AAD is considered 0 if the sector is located outside of image. The feature vector of a fingerprint consists of these AAD values and have length $(n_c - 1)n_d n_g$.

In this work, $n_c = 5$, $n_d = 8$ and $n_g = 8$, therefore the feature vector has length 256.

MLP-hash with normalizations and MLP-hash with both normalizations and simplified projection matrices are implemented and tested with different normalization methods (min-max, l_2, statistical and without normalization) and different nonlinear transformations (ReLU, leaky ReLU, tanh, f_{custom}). Original MLP-hash is implicitly included as MLP-hash without normalization, without matrix simplification and with ReLU. For comparison, original biohashing is also implemented.

All biohashing variations are tested with a classifier based on Hamming distance. Hamming distance H(**a**, **b**) is the amount of differing elements. Classifier is built such that if H(**a**, **b**) $\le T$, where T is a threshold, **a** and **b** are considered as those from the same user, else they are considered coming from different users. To see how much distinction between templates is lost from the data transformation, another classifier based on cosine similarity S is implemented.

$$\mathrm{S}(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{|\mathbf{a}|\cdot|\mathbf{b}|}.$$

If S(**a**, **b**) $\ge T$, where T is a threshold, vectors **a**, **b** are considered belonging to the same user, else – to different. This classifier is used with unprotected feature vectors.

The metrics being used in this work are FRR, FAR and EER. FRR, or false rejection rate, is the fraction of genuine user authentication attempts that are rejected:

$$\mathrm{FRR} = \frac{N_{fr}}{N_{fr} + N_{ta}}.$$

FAR, or false acceptance rate, is the fraction of imposter authentication attempts that are considered genuine by classifier:

$$\mathrm{FAR} = \frac{N_{fa}}{N_{fa} + N_{tr}}.$$

EER, or equal error rate, is considered equal to FAR when FAR is equal to FRR.

Because there may be no point at which FRR precisely equals FAR for a finite dataset, EER is approximated based on the closest FRR and FAR values. Suppose that n_T parametrisations of the classifier are used, for example, different threshold values T. For these values FRR and FAR pairs are calculated and sorted by descending FRR, forming a sorted list of pairs $(\mathrm{FRR}_i, \mathrm{FAR}_i)$, $i = 1, ..., n_T$. Let i_c be the smallest index at which FRR < FAR. Then EER is approximated as:

$$\mathrm{EER} = \frac{\mathrm{FRR}_{i_c-1} + \mathrm{FAR}_{i_c-1} + \mathrm{FRR}_{i_c} + \mathrm{FAR}_{i_c}}{4}.$$

Biohashing and its modifications are probed in two scenarios: base scenario and stolen token scenario. Base scenario implies that every user has a different secret token and keeps its secrecy, while stolen token scenario implies that users token has lost its secrecy and is used by everyone.

5 RESULTS

FAR, FRR and EER are calculated for the following parameters of MLP-hash modifications: $l = 3$, $m_i = 128$ for all layers. For the modification with matrix simplification, $b = 8$.

Features extracted with the use of Gabor filters are used in the classifier based on cosine similarity. The plot of FRR against FAR for this classifier is presented on Fig. 1. The EER of this classifier is 0.1685.

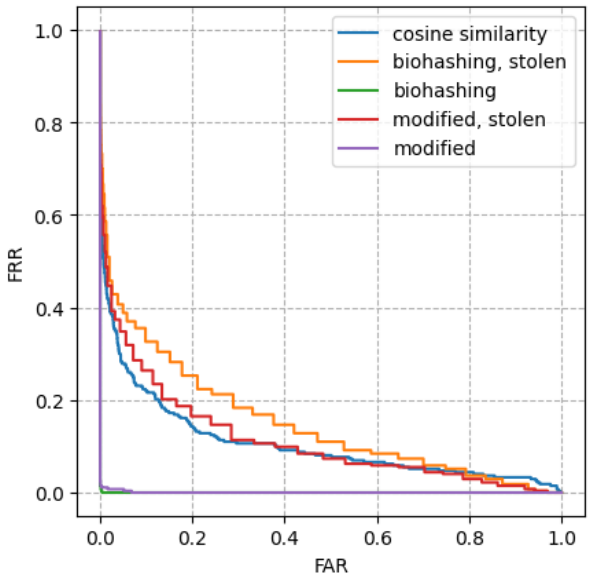


Figure 1 – Plot of FRR versus FAR of classifiers based on unprotected biometric templates (by cosine similarity) and of classifiers based on templates protected by biohashing and its modification proposed in this work, where “stolen” refers to stolen token scenario

The results of applying base biohashing are used in a classifier based on Hamming distance. Plots of FRR against FAR for this classifier are presented on Fig. 1 for comparison with these for unprotected templates. The EER of this classifier is 0.0037 in base scenario and 0.2234 in stolen token scenario.

The results of applying MLP-hash with inclusion of normalization are used in a classifier based on Hamming distance. EER values for base scenario are listed in Table 1, for stolen token scenario – in Table 2.

DOI 10.15588/1607-3274-2026-1-16

Table 1 – EER values of classifier based on Hamming distance with MLP-hash with normalizations in base scenario

	ReLU	leaky ReLU	tanh	f_{custom}
None	0.001450	0.001272	0.010925	0.007313
min-max	0.000089	0.001272	0.001450	0.004884
l_2	0.001450	0.001272	0.001450	0.001272
statistical	0.128294	0.078907	0.089744	0.073222

Table 2 – EER values of classifier based on Hamming distance with MLP-hash with normalizations in stolen token scenario

	ReLU	leaky ReLU	tanh	f_{custom}
None	0.260111	0.208829	0.275921	0.223456
min-max	0.305886	0.295050	0.283323	0.310211
l_2	0.260111	0.208829	0.209834	0.204212
statistical	0.223456	0.213624	0.219844	0.196276

The results of applying MLP-hash with inclusion of normalization and with simplified projections are used in a classifier based on Hamming distance. EER values for base scenario are listed in Table 3, for stolen token scenario – in Table 4. Plots of FRR and FAR for a Hamming distance classifier with the use of the modified method with f_{custom} and statistical normalization are presented on Fig. 1 for comparison with unprotected templates and base biohashing.

Table 3 – EER values of classifier based on Hamming distance with MLP-hash with simplified projections in stolen token scenario

	ReLU	leaky ReLU	tanh	f_{custom}
None	0.000089	0.015898	0.014716	0.014716
min-max	0.054983	0.029342	0.021940	0.016433
l_2	0.000089	0.015898	0.004706	0.010925
statistical	0.001628	0.012999	0.018328	0.010925

Table 4 – EER values of classifier based on Hamming distance with MLP-hash with simplified projections in stolen token scenario

	ReLU	leaky ReLU	tanh	f_{custom}
None	0.276455	0.208829	0.204034	0.211259
min-max	0.378307	0.296767	0.271126	0.241695
l_2	0.276455	0.208829	0.190413	0.190413
statistical	0.254960	0.212441	0.183188	0.186800

For a modified method with the usage of f_{custom} as a nonlinear transformation and statistical normalization, distances between templates of genuine users each having his own token, of different users each having his own token, of genuine users but each time with new token and of different users who use a new token each time are calculated. Histogram of these distances is presented on Fig. 2.

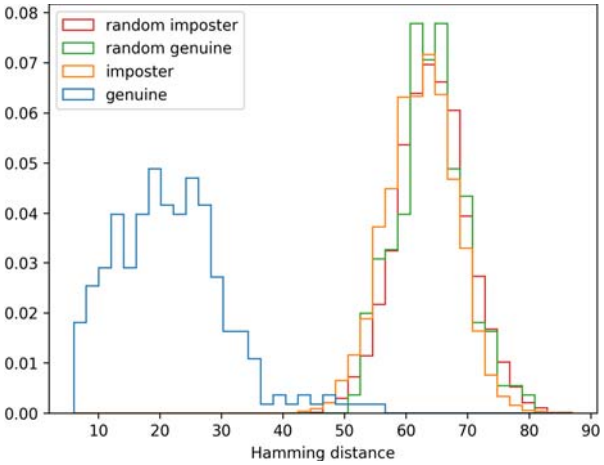


Figure 2 – Histograms of distances between protected templates, where “random” corresponds to choosing a new token each time

In order to see the speed improvement for different layer and block sizes, the amount of projections for the first layer is set as $p_1 = n_p$, which multiplies its output size by n_p as well, and other layers output size is multiplied by n_p. Computation durations of a modified MLP-hash for different parametrisations are listed in Table 5.

Table 5 – Computation time of modified MLP-hash in seconds. Each column corresponds to certain value of n_p, each row – to certain value of b

	1	2	4	8	16
1	0.009750	0.017471	0.107659	1.512096	7.540924
2	0.004071	0.009447	0.037941	0.198579	1.952194
4	0.003696	0.007272	0.019800	0.056675	0.395279
8	0.003670	0.006958	0.014189	0.043949	0.118840
16	0.003275	0.006494	0.013236	0.027612	0.066130

6 DISCUSSION

As can be seen from Table 2 and Table 4, applying l_2-normalization or statistical normalization reduces EER in most cases. From the same tables it can be seen that using leaky ReLU, tanh or f_{custom} leads to smaller EER in most cases. While using tanh or f_{custom}, using simplified projections leads to smaller EER. However, this improvement is not present while using ReLU or Leaky ReLU. In the stolen token scenario, the most optimal parametrisation without simplifying matrices (statistical normalization, f_{custom}) has 1.33 times smaller EER than with original parametrisation (no normalization, ReLU) and 1.14 times smaller EER than with base biohashing. In this same scenario, the most optimal parametrisation with simplified matrices (statistical normalization, f_{custom}) has 1.05 times smaller EER than with same parametrisation without matrix simplification, or 1.19 times smaller than with base biohashing.

From the comparisons above it can be concluded that applying normalization improves EER. Simplification of projection matrices also leads to improvement of EER in some cases, although it is less noticeable and the main advantage of it is increasing computation speed.

DOI 10.15588/1607-3274-2026-1-16

The proposed modification complies with requirements for biometric template protection methods listed in problem statement:

1) non-reversibility: without matrix simplification this method is at least as difficult to partially invert as base biohashing and MLP-hash;

2) accuracy: the modified method does not lead to increased EER in comparison to base biohashing, furthermore, it leads to improvement and it is closer to EER of classifier based on unprotected feature vectors (1.11 times bigger) than biohashing is (1.33 times bigger);

3) diversity: as it can be seen from Fig. 2, the distributions of distances between templates of different users is similar to that of same user but with different tokens, which means that if user creates several templates with different tokens, it will be difficult to decide if these templates are from one user or from several;

4) revocability: a protected biometric template can be replaced by replacing users' secret token *t*, in the same way as it can be replaced while using biohashing or MLP-hash.

As it can be seen from Table 5, the modified method is faster than non-modified (for $n_p = 1$ it is 2.39 times faster with just $b = 2$), and the difference in speed is increasing with increasing block number (for $n_p = 1$ and $b = 16$ modified method is 2.98 times faster) and it is more significant for bigger layer sizes (for $n_p = 16$, which corresponds to layers of size 4096, and $b = 16$ modified method is 114.03 times faster). It should be kept in mind that this modification may make inversion easier and that an analysis of inversion of MLP-hash (and of the modification presented in this work) was not performed, therefore setting a block number to high values may make the biometric system less safe.

CONCLUSIONS

This work is focused on modifying a nonlinear biohashing-based biometric template protection method, further generalizing and improving it.

The scientific novelty of this work is a modification of the MLP-hash method including normalization between layers of MLP-hash and simplification of pseudorandom projection matrices. Results of experiments conducted with feature vectors extracted from fingerprints show that applying l_2 or statistical normalization improves EER of the classifier using protected templates (for example, with statistical normalization and f_{custom} EER is 1.14 times smaller than without normalization and with f_{custom} and 1.33 times smaller than without normalization and with ReLU). Because the classifier based on templates protected by the modified method is more accurate than a classifier based on templates protected by base MLP-hash and biohashing, it can be said that the modified method preserves more distinction between templates. The method is shown to have all properties required from biometric template protection methods. The main advantage of simplifying projection matrices is the decrease of time complexity of the method from $O(m^2n)$ to $O\left(\frac{m^2n}{b^2}+\frac{mn}{b}+n\right)$ and space complexity from $O(mn)$ to $O\left(\frac{mn}{b}+m+n\ln n\right)$. Time complexity difference is also shown by experimental results, with the modified method being 2.39 times faster for $n_p = 1$, $b = 2$ and 114.03 times faster for $n_p = 16$, $b = 16$. Projection simplification also causes a small decrease in EER of a corresponding classifier: 1.05 times smaller while both simplified and non-simplified methods use statistical normalization and f_{custom} or 1.39 times smaller in comparison to base MLP-hash.

The practical significance of obtained results is that the modified method is faster and causes less classifier performance loss than the original. It can be used in authentication systems with high-dimensional feature vectors when users expect an absence of delays, or it can be used with hardware having low computational power and memory. The developed method produces protected templates that are similar to unprotected in overall structure, especially if unprotected templates are binarized before usage, so they can be used in the same applications in similar ways. These applications include authentication, identification, cryptographic key generation or binding (this key may be further transformed to match the system it is being used, for example by being extended or shortened to a bit string of a needed length, integer in some specific range for a cryptographic key exchange based on elliptic curves over finite fields, basis for lattice based cryptography etc.).

Prospects for further research include investigating even more nonlinear transformations and normalization methods, usage of the modified method with other biometric modalities and feature extraction techniques. Invertibility of the original MLP-hash and an effect nonlinear function choice has on it should also be studied, along with the decreasing of inversion difficulty coming from the simplification of projection matrices, in order to determine which parametrisations allow saving computation time without noticeable drop in safety.

ACKNOWLEDGEMENTS

The study was performed without financial support.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: Mykola Onai: conceptualization, methodology, formulation of tasks, software, analysis of research results, formulation of conclusions, review, and editing; Oleksandr Kosenko: conceptualization, methodology, formulation of tasks, software, analysis of research results, formulation of conclusions, review, and editing.

DOI 10.15588/1607-3274-2026-1-16

Data availability: The manuscript has no associated data.

Software availability: The manuscript has associated software in a public repository: https://github.com/KosenkoAlexander/modified-MLP-hash.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

REFERENCES

1. Maltoni D., Maio D., Jain A. K., Prabhakar S. Handbook of Fingerprint Recognition (2nd. ed.). London, Springer, 2009, 494 p. DOI: 10.1007/978-1-84882-254-2
2. Teoh A., Ngo D., Goh A. Biohashing: two factor authentication featuring fingerprint data and tokenised random number, *Pattern Recognition*, 2004, Vol. 37, Issue 11, pp. 2245–2255. DOI: 10.1016/j.patcog.2004.04.011
3. Baier H., Breitinger F., Busch C., Rathgeb C. On application of bloom filters to iris biometrics, *IET Biometrics*, 2014, Vol. 3, Issue 4, pp. 207–218. DOI: 10.1049/iet-bmt.2013.0049
4. Jin Z., Hwang J. Y., Lai Y., Kim S., Teoh A. Ranking-Based Locality Sensitive Hashing-Enabled Cancelable Biometrics: Index-of-Max Hashing, *IEEE Transactions on Information Forensics and Security*, 2018, Vol. 13, Issue 2, pp. 393–407. DOI: 10.1109/TIFS.2017.2753172
5. Lumini A. and Nanni L. An improved BioHashing for human authentication, *Pattern Recognition*, 2007, Vol. 40, Issue 3, pp. 1057–1065. DOI: 10.1016/j.patcog.2006.05.030
6. Durbet A., Grollemund P., Lafourcade P., Migdal D., Thiry-Atighehchi K. Authentication Attacks on Projection-based Cancelable Biometric Schemes, *International Conference on Security and Cryptography (SECRYPT) : 19th International Conference, Lisbon, 11–13 July, 2022 : proceedings*. SCITEPRESS Digital Library, 2022, pp. 568–573. DOI: 10.5220/0011277100003283
7. Otroshi H. S. and Krivokuća V. H. and Marcel S. MLP-Hash: Protecting Face Templates via Hashing of Randomized Multi-Layer Perceptron, *European Signal Processing Conference (EUSIPCO) : 31st European Conference, Helsinki, 4–8 September, 2023 : proceedings.* – Helsinki, IEEE, 2023, pp. 605–609. DOI: 10.23919/EUSIPCO58844.2023.10289780
8. Maio D., Maltoni D., Cappelli R., Wayman J., Jain A. K. FVC2000: Fingerprint verification competition, *Pattern Analysis and Machine Intelligence*, 2002, Vol. 24, Issue 3, pp. 402–412. DOI: 10.1109/34.990140
9. Jain A. K., Prabhakar S., Hong L., Pankanti S. Filterbank-based fingerprint matching, *IEEE transactions on image processing*, 2000, Vol. 9, Issue 5, pp. 846–859. DOI: 10.1109/83.841531
10. Kawagoe M., Tojo A.Fingerprint pattern classification, *Pattern Recognition*, 1984, Vol. 17, Issue 3, pp. 295–303. DOI: 10.1016/0031-3203(84)90079-7
11. Duda R. O., Hart P. E. Pattern classification and scene analysis. New York, Wiley, 1978, 512 p.

Received 31.07.2025.
Accepted 08.01.2026.
Published 27.03.2026.

УДК 004.056.55

МОДИФІКОВАНИЙ МЕТОД ЗАХИСТУ БІОМЕТРИЧНИХ ШАБЛОНІВ З НЕЛІНІЙНИМИ ПЕРЕТВОРЕННЯМИ

Онай М. В. – канд. техн. наук, доцент кафедри програмного забезпечення комп'ютерних систем Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна. ROR: https://ror.org/00syn5v21. ORCID: https://orcid.org/0000-0002-4938-8355.

Косенко О. В. – бакалавр кафедри програмного забезпечення комп'ютерних систем Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна. ROR: https://ror.org/00syn5v21. ORCID: https://orcid.org/0009-0001-2115-7918.

АНОТАЦІЯ

Актуальність. Біометричні дані нерідко використовуються для аутентифікації або ідентифікації. Однак такі дані є вразливими і не є замінюваними у разі викрадення. У літературних джерелах запропоновано кілька методів створення захищених біометричних шаблонів, одним з яких є біогешинг. Однак лінійність біогешинга може бути його вразливістю. MLP-hash є схожим методом, що додає нелінійні перетворення. Цей метод модифікується у даній роботі.

Мета роботи. Метою даної роботи є розроблення модифікації MLP-hash, що є швидшою за оригінал та дозволяє більш чітке розділення користувачів за їх шаблонами.

Метод. Дана робота зосереджена на модифікуванні MLP-hash, варіації біогешингу з нелінійними перетвореннями. однією з запропонованих змін є застосування нормалізації перед застосуванням нелінійного перетворення у кожному шарі MLP-hash. Іншою запропонованою зміною є спрощення псевдовипадкових матриць, що використовуються у кожному шарі MLP-hash. Кожна така матриця замінюється на блочну матрицю, у якій блоки, що лежать на діагоналі, є ортонормальними матрицями, а решта блоків заповнюються нулями. Кожен ненульовий блок генерується з використанням користувацького секретного токена. Для того щоб зробити вплив кожного ненульового блоку менш локалізованим, перед кожним множенням на матрицю та після всіх шарів додаються псевдовипадкові перестановки. Псевдовипадкові перестановки також генеруються з використанням користувацького секретного токена у якості сіда. Застосування запропонованого методу близьке до застосування оригінального MLP-hash та біогешинга: метод приймає користувацький секретний токерн та біометричний вектор фіксованої довжини та повертає бінарний вектор

DOI 10.15588/1607-3274-2026-1-16

фіксованої довжини з такою ж або меншою розмірністю. MLP-hash з блоковими матрицями порівняно з оригіналом при застосуванні різноманітних способів нормалізації та різноманітних нелінійних перетворень.

Результати. Програмно реалізовано запропоновану модифікацію, оригінальний MLP-hash та біогешинг. Порівняно швидкодію та точність розділення користувачів з використанням цих методів на векторах характеристик, виділених з відбитків пальців з використанням фільтрів Габора.

Висновки. Проведені експерименти показали збільшення швидкодії та здатності розділення користувацьких шаблонів у результаті підстановки запропонованих блочних матриць та підвищення здатності розділення користувацьких шаблонів у результаті застосування нормалізації. Крім того, проведено порівняння застосування різних методів нормалізації та різних нелінійних перетворень. Практична цінність розробленого методу полягає в тому, що він є швидшим та може бути використаний у складі програмного забезпечення, користувачі якого очкують на відсутність затримок, зберігаючи при цьому складність інвертування. Перспективи подальших досліджень включають тестування розробленого методу з іншими біометричними модальностями, іншими нелінійними перетвореннями та техніками нормалізації та аналіз складності інвертування розробленого методу у порівнянні з MLP-hash та біогешингом.

КЛЮЧОВІ СЛОВА: біометрія, біометричний шаблон, захист біометричних шаблонів, одностороннє перетворення, біогешинг, еліптична криптографія, скінченні поля.

ЛІТЕРАТУРА

1. Handbook of Fingerprint Recognition (2nd. ed.) / [D. Maltoni, D. Maio, A. K. Jain, S. Prabhakar]. – London : Springer, 2009. – 494 p. DOI: 10.1007/978-1-84882-254-2
2. Teoh A. Biohashing: two factor authentication featuring fingerprint data and tokenised random number / A. Teoh, D. Ngo, A. Goh // Pattern Recognition. – 2004. – Vol. 37, Issue 11. – P. 2245–2255. DOI: 10.1016/j.patcog.2004.04.011
3. On application of bloom filters to iris biometrics / [H. Baier, F. Breitinger, C. Busch, C. Rathgeb] // IET Biometrics. – 2014. – Vol. 3, Issue 4. – P. 207–218. DOI: 10.1049/iet-bmt.2013.0049
4. Ranking-Based Locality Sensitive Hashing-Enabled Cancelable Biometrics: Index-of-Max Hashing / [Z. Jin, J. Y. Hwang, Y. Lai et al.] // IEEE Transactions on Information Forensics and Security. – 2018. – Vol. 13, Issue 2. – P. 393–407. DOI: 10.1109/TIFS.2017.2753172
5. Lumini A. An improved BioHashing for human authentication / A. Lumini and L. Nanni // Pattern Recognition. – 2007. – Vol. 40, Issue 3. – P. 1057–1065. DOI: 10.1016/j.patcog.2006.05.030
6. Authentication Attacks on Projection-based Cancelable Biometric Schemes / [A. Durbet, P. Grollemund, P. Lafourcade et al.] // International Conference on Security and Cryptography (SECRYPT) : 19th International Conference, Lisbon, 11–13 July, 2022 : proceedings. – SCITEPRESS Digital Library, 2022. – P. 568–573. DOI: 10.5220/0011277100003283
7. Otroshi H. S. MLP-Hash: Protecting Face Templates via Hashing of Randomized Multi-Layer Perceptron / H. S. Otroshi and V. H. Krivokuća and S. Marcel // European Signal Processing Conference (EUSIPCO) : 31st European Conference, Helsinki, 4–8 September, 2023 : proceedings. – Helsinki, IEEE, 2023. – P. 605–609. DOI: 10.23919/EUSIPCO58844.2023.10289780
8. FVC2000: Fingerprint verification competition / [D. Maio, D. Maltoni, R. Cappelli et al.] // Pattern Analysis and Machine Intelligence. – 2002. – Vol. 24, Issue 3. – P. 402–412. DOI: 10.1109/34.990140
9. Filterbank-based fingerprint matching / [A. K. Jain, S. Prabhakar, L. Hong, S. Pankanti] // IEEE transactions on image processing. – 2000. – Vol. 9, Issue 5. – P. 846–859. DOI: 10.1109/83.841531
10. Kawagoe M. Fingerprint pattern classification / M. Kawagoe, A. Tojo // Pattern Recognition. – 1984. – Vol. 17, Issue 3. – P. 295–303. DOI: 10.1016/0031-3203(84)90079-7
11. Duda R. O. Pattern classification and scene analysis / R. O. Duda, P. E. Hart. – New York : Wiley, 1978. – 512 p.

DOI 10.15588/1607-3274-2026-1-16

UDC 004.8

METHOD FOR CORRECTION OF MULTISUBJECTIVE MULTI-FACTORIAL ENVIRONMENTS OF SOFTWARE COMPLEXES' SUPPORT

Pukach A. I. – PhD, Assistant of Automated Control Systems Department, Institute of Computer Sciences and Informational Technologies, Lviv Polytechnic National University, Lviv, Ukraine. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0009-0001-8563-3311.

Teslyuk V. M. – Doctor of Sciences, Professor, Head of Automated Control Systems Department, Institute of Computer Sciences and Informational Technologies, Lviv Polytechnic National University, Lviv, Ukraine. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0000-0002-5974-9310.

ABSTRACT

Context. The problem of correction of multisubjective multifactorial environments of software complexes' support is considered in this research, necessary to provide the possibility(-ies) of adjusting the perception's subjectivization of the support object (the supported software, as well as related processes of its complex support), caused by the influence of relevant impact factors. **The object of research** is a process of correction of multisubjective multifactorial environments of software complexes' support. **The subject of research** are methods and means of correction of a multisubjective multifactorial environments of software complexes' support, as well as methods of an artificial neural networks (in particular: a multilayer perceptron).

Objective – is the development of method for correction of multisubjective multifactorial environments of software complexes' support.

Method. The development of a method for correction of multisubjective multifactorial environments of software' support is proposed. which provides possibilities for the necessary adjustments of the perception subjectivization of the researched support objects (which could be either the supported software itself, as well as the related processes for its comprehensive support) relevant (directly or directly) interacting subjects, who provide and implement this comprehensive support of the researched supported software product, in order to provide the possibility(-ies) of further automation and intellectualization of its comprehensive support.

Results. The results of functioning of the developed method – are relevant models of adjusted multisubjective multifactorial environments of software complexes' support, obtained in result of solving a relevant scientific and applied problem of adjusting such class of environments. The developed method provides the opportunity(-ies) for studying the processes of collective perception's subjectivization (caused by the influence of existing impact factors) of the objects of comprehensive support by the appropriate related subjects, which directly provide and implement this support, and also facilitates and ensures for further automation and intellectualization of such complex support of various software products and complexes in this separate and exact functional and procedural segment. As a practical approbation of the developed method, – the results of solved applied practical task of determination and further correction the impact factors of maximum imbalance of the researched multisubjective multifactorial environment (representing the technician team of the supported software product) are given.

Conclusions. The developed method solves the declared problem of correction of multisubjective multifactorial environments of software complexes' support. At the same time, the obtained results of performed practical approbation of the developed method confirm its functionality in solving a range of scientific and applied tasks based on the processes of collective perception's subjectivization of support objects (the supported software complexes, as well as the processes of their comprehensive support), which (those tasks), in turn, are included into the cluster of a more valuable scientific and applied problem of software products' comprehensive support automation and intellectualization.

KEYWORDS: software product, comprehensive support, impact factors, automation, correction, multisubjective multifactorial environment, neural networks, multilayer perceptron.

ABBREVIATIONS

ABC is an Artificial Bee Colony Algorithm;
ACO is an Ant Colony Optimization;
AI is an Artificial Intelligence;
ANN is an Artificial Neural Network;
CCS is a Cartesian Coordinate System;
CI/CD is a Continuous Integration / Continuous Delivery;
DevOps is a Development and Operations;
DM is a Data Mining;
EAI is a Explainable Artificial Intelligence;
FA is a Firefly Algorithm;
GA is a Genetic Algorithm;
HA is a Hybrid Algorithms;
HCA is a Hill Climbing Algorithm;
MP is a Multilayer Perceptron;
ML is a Machine Learning;
NLP is a Natural Language Processing;
PCS is a Polar Coordinate System;
PSO is a Particle Swarm Optimization;
SVM is a Support Vector Machine.

NOMENCLATURE

Δ is an adjustment parameter;

$\Delta_{[i][j]}$ is an adjustment parameter component for *j*-th impact factor of personal multifactor portrait of *i*-th subject;

$(\rho^{Dest}_{[i,j]}; \varphi^{Dest}_{[i,j]})$ is a component target coordinates (in the PCS) of the *j*-th impact factor of the *i*-th subject's personal multifactor portrait, which can be achieved using the adjustment parameter $\Delta_{[i,j]}$;

DOI 10.15588/1607-3274-2026-1-17

$(\rho^{Currr}_{[i,j]}; \varphi^{Currr}_{[i,j]})$ is a component current coordinates (in the PCS) of the *j*-th impact factor of the *i*-th subject's personal multifactor portrait;

$FCSP_{[j]}(Obj)$ is a component of the *j*-th impact factor inside the subjectivization function of the personalized perception (of the researched support object) by the current interaction subject;

$Fsubj_{[i]}$ is a nonlinear subjectivization function of personalized perception of the support object by the *i*-th interaction subject;

Obj is a variable-identifier of the researched support object;

$PCoPsMfSE_{[j]}$ is a parametric characteristics of a multisubjective multifactorial environment of the supported software complex;

$(x^{C}_{[i,j]}; y^{C}_{[i,j]})$ is a component current coordinates (in the CCS) of the *j*-th impact factor of the *i*-th subject's personal multifactor portrait;

$(x^{D}_{[i,j]}; y^{D}_{[i,j]})$ is a component target coordinates (in the CCS) of the *j*-th impact factor of the *i*-th subject's personal multifactor portrait, which can be achieved using the adjustment parameter $\Delta_{[i,j]}$;

$xc^{2}_{[i,j]}$ is a *x*-coordinate (in the CCS) of the adjustment parameter's current position (of the *j*-th impact factor of the *i*-th subject's personal multifactor portrait);

$xd^{2}_{[i,j]}$ is a *x*-coordinate (in the CCS) of the adjustment parameter's target position (of the *j*-th impact factor of the *i*-th subject's personal multifactor portrait);

$yc^{2}_{[i,j]}$ is a *y*-coordinate (in the CCS) of the adjustment parameter's current position (of the *j*-th impact factor of the *i*-th subject's personal multifactor portrait);

$yd^{2}_{[i,j]}$ is a *y*-coordinate (in the CCS) of the adjustment parameter's target position (of the *j*-th impact factor of the *i*-th subject's personal multifactor portrait).

INTRODUCTION

One of the major and key components of the life cycle of any software product – is its comprehensive support, which includes, in particular, such elements as: development, testing, implementation, environment configuration, and processing of requests (both external from outside customers' companies, and internal from the inside members of the development company itself).

At the same time, the automation of this complex support of various software products – is a complex scientific and applied problem, which includes a whole range of relevant scientific and applied tasks, including, among others, tasks based on the processes of collective subjectivization of the perception of support objects (the supported software complexes, as well as the processes of their comprehensive support), which arise as a result of presence of various impact factors, that lead to a distortion of the objective perception (of the object of support) by the relevant subjects (e.g. personnel), which, in fact, directly provide and implement this comprehensive support.

Thus, all available subjects of the comprehensive support (of any supported software complex) form a corresponding multisubjective multifactorial support environment, which is synthesized on the basis of a set of their individual multifactor representations of a personalized perception of the same support object.

The object of research is a process of correction of a multisubjective multifactorial environments of software complexes' support.

The subject of research are methods and means of correction of a multisubjective multifactorial environments of software complexes' support, as well as methods of an artificial neural networks (in particular: a multilayer perceptron).

The objective of the research consists in the development of method for correction of a multisubjective multifactorial environments (of software complexes' comprehensive support), which provides/ensures possibilities for the necessary adjustments of the perception's subjectivization of the researched support object (which can act as directly supported software product/complex itself, as well as processes related to its comprehensive support) by the relevant interaction subjects who directly provide and implement this comprehensive support (for the supported researched software product/complex), in order to provide the possibility(-ies) of further automation and intellectualization of suck kind comprehensive support.

1 PROBLEM STATEMENT

Let's consider the formalization of given problem of analysis a multisubjective multifactorial support environment – in the relevant form of a nonlinear polycriterial dependence task.

Thus, in considered case, the input variables of the problem – are nonlinear functions of subjectivization of the personalized perception (of the support object) by each of the interaction subjects: $Fsubj_{[i]}=[FCSP_{[j]}(Obj)]$ ($i \in [1..n]$, $j \in [1..m]$), where: *Obj* – variable-identifier of the researched support object; *n* – number of subjects interacting with the support object; *m* – number of declared impact factors.

The output variables of given problem – are the parametric characteristics of the multisubjective multifactorial support environment: $PCoPsMfSE_{[j]}$ (j=[1..m]), де *m* – number of declared impact factors.

Let us have a set of functions of subjectivization of the personalized perception of the support object by each of the subjects interacting with this object, that form relevant parametric characteristics of the multisubjective multifactorial support environment of the software complex:

DOI 10.15588/1607-3274-2026-1-17

$$PCoPsMfSE_{[j]} = \frac{\sum_{i=1}^{n} Fsubj_{[i]}[FCSP_{[j]}(Obj)]}{n}. \quad (1)$$

The main mandatory and necessary criterion of the given problem – is the finiteness of the set of support subjects, as well as the finiteness of the declared impact factors' set, which is due to the possibility of operating (in scope of given problem) with only a certain constant number of these pre-determined support subjects and impact factors.

Limitation of the problem:

1. The value of a personalized perception (of the support object) by each of the interacting subjects $Fsubj_{[i]}$ must be given as real numbers in a normalized representation form (that is, in the range of values between 0.0 and 1.0): $Fsubj[1..n] \in [0..1]$.

Expression (1) provides the possibility of interpreting the given problem of analyzing a multisubjective multifactorial support environments.

However, such an interpretation requires an additional mechanism to provide the possibility of correction (including balancing, or others) of the obtained multisubjective multifactorial support environments, taking into account the individual specifics and peculiarities of each such environment at the stage of forming/construction its interpretation.

Thus rise corresponding scientific and applied problem of correction of a multisubjective multifactorial environments of software complexes' support, for the purpose of solving which, in fact, a corresponding specialized method has been developed and presented in this research.

The main purpose of this article is to highlight the developed method, as well as the corresponding models for adjusting the investigated multisubjective multifactorial support environments, which together provide the possibility(-ies) of solving given scientific and applied problem of correction a multisubjective multifactorial support environments of software complexes.

2 REVIEW OF THE LITERATURE

The analysis of existing researches and publications was implemented both in the direction of automation of the components of software products' complex support, as well as in the research direction of the process(-es) of perception subjectivization of supported software complexes. Based on the analysis, the following interpreted results and conclusions have been obtained, presented below. In particular, the most common and basic areas of automation of comprehensive support for any software product(s) are: testing automation, DevOps automation, and automation of request processing (both external and internal).

The authors of the work [1] carried out a comprehensive comparison (specifically in the context of software testing automation) of such known Machine Learning and Data Mining algorithms as:

– HCA;
– ABC;
– FA;
– PSO;
– GA;
– ACO;
– ANN;
– SVM;
– HA.

In scope of research [2], author analyzed a batch of works related to application of AI in software testing and debugging, as well as the prospects for the application of artificial intelligence in these areas, and provided a brief summary of the methodologies, methods and approaches currently used in this field, in particular:

– using deep learning for creating test cases;
– usage of EAI for debugging;
– using reinforcement learning for test set optimization;
– usage of NLP for requirements assessment;
– use of artificial intelligence for test automation and continuous testing;
– and finally – usage of CI/CD, in order to simplify the whole process.

Authors of research [3] investigate the application of machine learning methods to increase the efficiency of software testing automation systems, based on a methodology that includes a comparative analysis of conventional testing methods with an integrated approach to machine learning, measuring performance through accuracy, execution speed, and resource utilization indicators, and, as a result, – authors confirm a significant increase in testing efficiency with using machine learning approach, methods and means.

The study [4] evaluates various AI techniques (including ML and NLP) and their application for test cases creation, software testing process optimization, and software defect(s) prediction, and the obtained results – highlight the efficiency and quality improvements achieved through software testing using AI.

Continuing an overview, authors of research [5] explore key aspects of AI and ML usage in DevOps, especially such kind of usage as: automated source code quality analysis, as well as predictive analytics for deployment and self-healing systems, and also study of tools and technologies that facilitate DevOps based on artificial intelligence, including, in particular, relevant machine learning frameworks (such as TensorFlow), and observation platforms (such as Datadog).

The work [6] explores the contribution of artificial intelligence to various aspects of DevOps, including source code management, CI/CD pipelines, deployment infrastructure, software testing infrastructure, logging mechanisms, data analysis tools, and comprehensive reporting systems, and also studies the impact of artificial intelligence on team communication, collaboration, and workflow orchestration in DevOps environments.

Authors of research [7] have conducted a systematic literature review using the PSALSAR Framework as a

DOI 10.15588/1607-3274-2026-1-17

tool for researching these relevant sources of information based on the SCOPUS database (including the Elsevier, Research Gate, and Semantic Scholar databases), starting from 2012 to 2022, which represents a comprehensive picture of usage of an artificial intelligence and machine learning technologies precisely in the context of automation of users' request(s) processing.

In the context of work [8], the integration of Apache Kafka (which is an existing platform for instant data streaming) with complex machine learning methods is considered – in order to ensure adaptive change(s) and improve customer support responses, which allows to significantly improve the efficiency and customization of contacts with customers (end users).

However, in the context of existing researches, there is no appropriate analysis (as well as, in fact, synthesis) of relevant methods and means of correction a multisubjective multifactorial environments for supporting software complexes, for which appropriate automation technologies (of their comprehensive support) are being implemented, including: testing, DevOps, or processing requests (and/or appeals) from clients/customers and end users.

In turn, this leads to the emergence of an appropriate relevant actual scientific and applied problem of correction of multisubjective multifactorial environments of software complexes' support, necessary to provide the possibility(-ies) for correction the perception's subjectivization of the support object (the supported software, as well as related processes of its complex support), caused by the influence of numerous relevant existing impact factors.

3 MATERIALS AND METHODS

The formation of a multisubjective multifactorial support environments (for software complexes' comprehensive support) is carried out on the basis of relevant personal multifactor portraits of separate subjects [9], who directly "shape" (i.e. form) the researched support environment. Thus, a multifactor portraits of support subjects (e.g. support personnel), in fact, – constitute an elementary structural and functional component of the relevant researched multisubjective multifactorial support environments of software complexes.

Accordingly, all corrections of the support environment are carried out directly through these structural and functional elementary components, which are – the relevant personal multifactor portraits of the existing subjects of each particular researched support environment. In the context of the conducted research, two variants of adjustment models (of the researched multisubjective multifactorial environments of software complexes' comprehensive support) have been developed and proposed, presented below.

In particular, one of the model options involves adjustment of the researched multisubjective multifactorial support environment – by direct adjustment of personal multifactor portraits of those subjects (e.g. support personnel) who form this support environment.

The main advantage of this option is the absence of need to encapsulate any additional (e.g.: "new", "non-native", "external" or "foreign") subjects, maintaining, in such way, the conservatism and integrity of the researched environment(s).

This particular version of the model (of adjusting multisubjective multifactorial environments of software complexes comprehensive support) is represented by the following expression (2) given below:

$$PCoPsMfSE_{[j]} = \frac{\sum_{i=1}^{n} Fsubj_{[i]}[FCSP_{[j]}(Obj) + \Delta_{[i][j]}]}{n}, \quad (2)$$

At the same time, various/different systems of distribution of the value of the correction parameter Δ – are possible.

In particular, within the framework of this research, the following variants have been developed and proposed, based on:

1) the PCS;
2) the CCS;
3) a vector system;
4) the Archimedes spiral.

Let's consider in more detail each of the proposed variants of systems for distributing the value of the correction parameter Δ.

The polar coordinate system variant declares the adjustment parameter Δ as follows (3):

$$\Delta_{[i,j]} = (\rho_{[i,j]}^{Dest}; \varphi_{[i,j]}^{Dest}) - (\rho_{[i,j]}^{Currr}; \varphi_{[i,j]}^{Currr}). \quad (3)$$

The main feature of the variant which is based on the polar coordinate system – is its advantages in solving situations and tasks when the adjustment of the component of required impact factor (of the investigated subject's personal multifactor portrait) is easier (and/or more convenient) to interpret as a vector of the deviation angle between the current and target values of the researched impact factor, since in this case there is no need to interpret such a relationship by using more complex trigonometric equations.

Another variant which is based on the Cartesian coordinate system is a little bit similar to the previous one, with the only significant difference consisting in a fact that the target and the current coordinates (of the component of *j*-th impact factor of the *i*-th subject's personal multifactor portrait) are actually given in the Cartesian coordinate system, and, in particular, can be represented by expression (4) below:

$$\Delta_{[i,j]} = (x_{[i,j]}^{D}; y_{[i,j]}^{D}) - (x_{[i,j]}^{C}; y_{[i,j]}^{C}). \quad (4)$$

The main feature of the option which is based on the CCS – is its advantages in solving tasks and situations in which the adjustment of the required impact factor's component can be carried out with a minimum set of iterations of simple movement between balancing points with previously known (given) exact coordinates.

DOI 10.15588/1607-3274-2026-1-17

The variant which is based on the vector system – represents the distribution of the value of correction parameter Δ as the interference of decomposition vectors of the components representing impact factors of personal multifactor portrait of the researched subject, and is described by the following expression (5):

$$\Delta_{[i,j]} = \sum_{k=1}^{m-1} Fr\left(\sqrt{((v_k)^2 + (v_{k+1})^2 - 2 \cdot v_k \cdot v_{k+1} \cdot \cos(\alpha_k^{k+1}))}\right), \quad (5)$$

where m – number of interference vectors; k – current interference vector's pointer counter; Fr – the recursive function of the interference of the current and the next vectors of the set; v_k – k-th vector (from the declared set of vectors) of influence onto the researched subject's personal multifactor portrait; α_k^{k+1} – the angle between the vectors v_k та v_{k+1}.

The main feature of the considered variant of the adjustment parameter Δ value distribution – is its advantages in solving those adjustment problems where the adjustment parameter is most expedient and/or convenient to be displayed precisely as an interference (superposition, or summation) of vectors representing appropriate decomposition components of the impact factors of the researched subject's personal multifactor portrait.

The variant which is based on the Archimedes spiral – actually represents a radial coordinate system, where the distribution of correction parameter Δ value is possible only within the framework of its existence on the plane (or in space) of the concentric circles of the Archimedes spiral, the transition between which is possible only along a radial trajectory, which, in fact, describes the dynamics of the: correction parameter's value distribution (relative to the investigated object and the subject); as well as the existing impact factors.

Accordingly, this variant of the adjustment parameter's distribution can be described by the following expression (6):

$$\Delta_{[i,j]} = \left(\sqrt{xd_{[i,j]}^2 + yd_{[i,j]}^2}\right) - \left(\sqrt{xc_{[i,j]}^2 + yc_{[i,j]}^2}\right). \quad (6)$$

The main feature of the considered variant (based on the Archimedes spiral) of the correction parameter's Δ value distribution – is its advantages in solving a cluster of problems related to the correction of those multisubjective multifactorial environments in which the trajectory of change (e.g. dynamics) of impact factors has a spiral (or radial, or concentric) form, due to the peculiarity of their interaction both: with each other, as well as with the subject and the object of the researched environment.

Thus, the proposed variants of the adjustment parameter's Δ value distribution (for the above-presented model option, which provides the adjustment of the researched multisubjective multifactorial support environment by direct adjusting the particular personal multifactor portraits of those subjects who form this environment, which is implemented through corresponding adjustment parameter Δ), are considered.

While another model option involves adjusting existing researched multisubjective multifactorial support environment by encapsulating (into this environment) additional new subjects with such personal multifactor portraits, which would ensure a shift in the common/general portrait of the entire environment (into which they are encapsulated) in the required vector/direction.

In turn, main advantage of this option is the preservation of the primary (original) portraits of already existing subjects that form the researched environment, that is ensuring a conservatism in relation to these subjects.

Accordingly, this version of the model for adjusting multisubjective multifactorial environments (of software complexes' comprehensive support) is represented by expression (7) below:

$$PCoPsMfSE_{[j]} = \frac{\sum_{i=1}^{n+n'} Fsubj_{[i]}[FCSP_{[j]}(Obj)]}{n+n'}, \quad (7)$$

where n' – number of additional new subjects, encapsulated into existing currently researched multisubjective multifactorial environment of complex support, necessary for appropriate adjustment(s) of this environment.

4 EXPERIMENTS

The experiment consists of the constructing a multisubjective multifactorial environment (for the researched support object), which is performed on the basis of the preliminary constructed personal multifactor portraits [9] of each of the subjects who form this entire environment. After that, the obtained multisubjective multifactorial environment is analyzed for necessity of its correction(s), and the required (appropriate) option and variant of correction are selected (in accordance to the characteristics of both the environment itself and the correction tasks) in order to select the most optimal option and variant. At the final stage of the experiment, the obtained identification results are presented in an arbitrary (convenient) form.

5 RESULTS

The main task/purpose of the developed method, presented in this research, is obtaining an adjusted multisubjective multifactorial environment according to set/predefined requirements (assessments, criteria, tasks, etc.). Let's consider obtained results on the example of solving a practical applied task of determination and further correction the impact factors of maximum imbalance of the researched multisubjective multifactorial environment (representing the technician team of the supported software product).

Figure 1 below presents a visualization of personal multifactor portraits of all subjects – team members of the investigated support environment. It should be also noted that such visualization actually represents nothing more than an expanded form of representation of the researched multisubjective multifactorial environment itself, with a detailed palette of all its constituent components of decomposition (both impact factors and subjects which "shape"/form it).

DOI 10.15588/1607-3274-2026-1-17

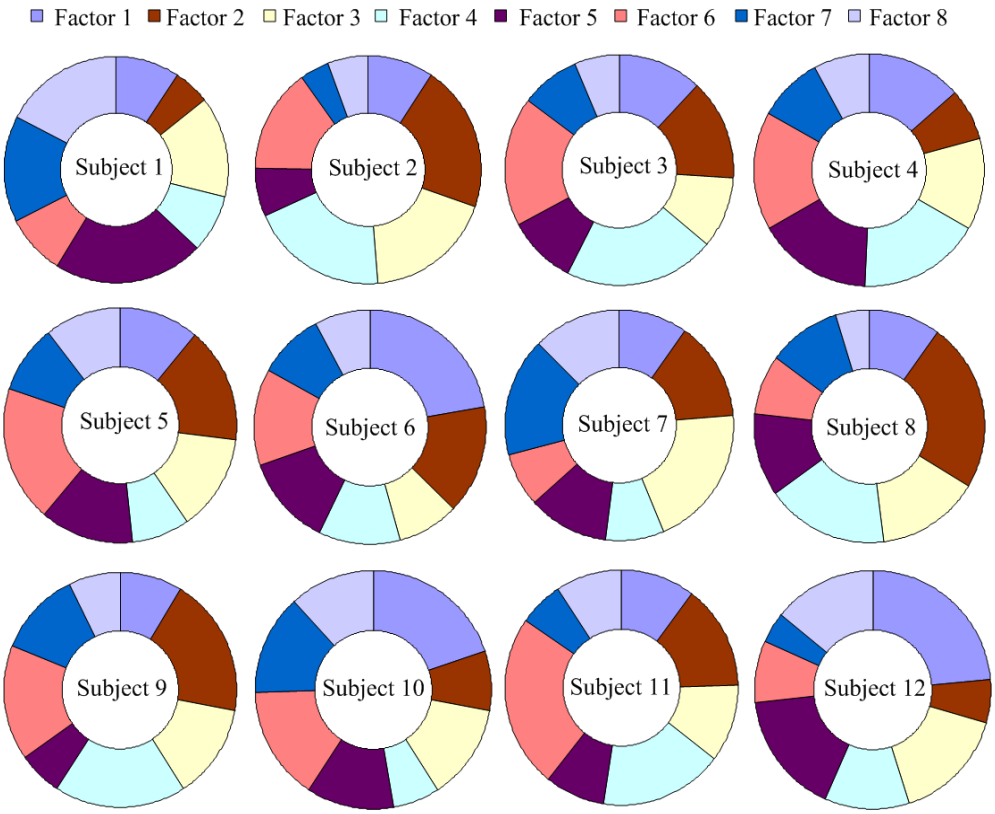


Figure 1 – Visualization of personal multifactor portraits of all subjects – team members of the investigated support environment.

In turn, the following Figure 2 represents the dynamics of the polysubject distribution of impact factors (within the framework of currently researched multisubjective multifactorial environment).

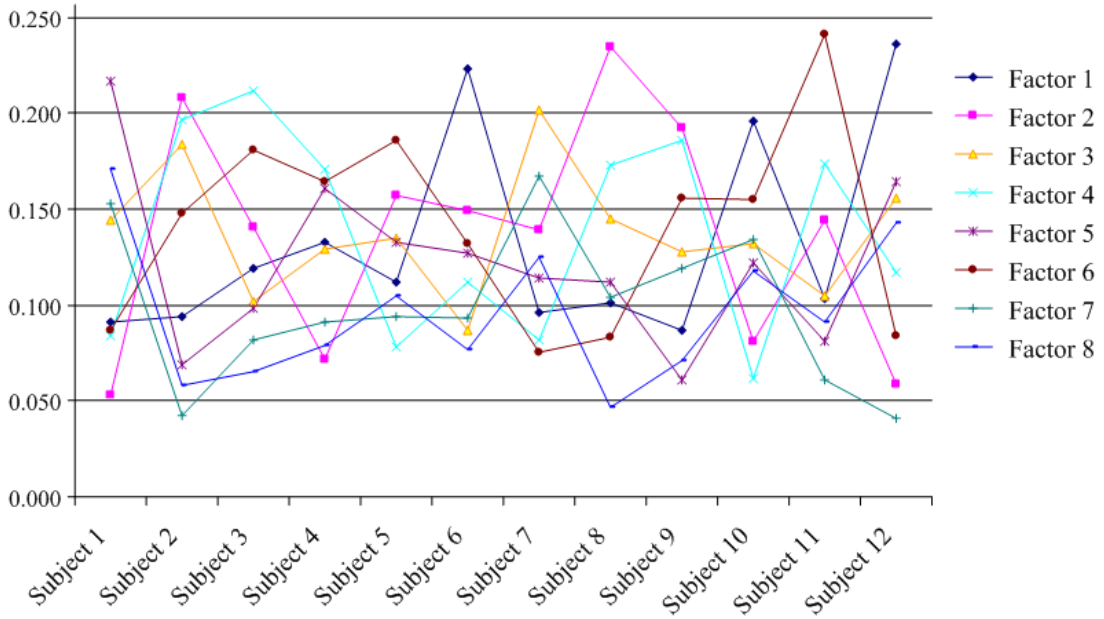


Figure 2. – Visualization of the dynamics of the polysubject distribution of impact factors (within the framework of currently researched multisubjective multifactorial environment)

Table 1 below displays the numerical data obtained for the personal multifactor portraits of all subjects – team members of the investigated support environment. In turn, the last row (“Average”) of Table 1 contains the average values for all considered impact factors.

DOI 10.15588/1607-3274-2026-1-17

Table 1 – Numerical data obtained for the personal multifactor portraits of all subjects – team members of the investigated support environment

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
Subject 1	0.091	0.053	0.144	0.084	0.217	0.087	0.153	0.171
Subject 2	0.094	0.208	0.184	0.197	0.069	0.148	0.042	0.058
Subject 3	0.119	0.141	0.102	0.212	0.098	0.181	0.082	0.065
Subject 4	0.133	0.072	0.129	0.171	0.161	0.164	0.091	0.079
Subject 5	0.112	0.157	0.135	0.078	0.133	0.186	0.094	0.105
Subject 6	0.223	0.149	0.087	0.112	0.127	0.132	0.093	0.077
Subject 7	0.096	0.139	0.202	0.082	0.114	0.075	0.167	0.125
Subject 8	0.101	0.235	0.145	0.173	0.112	0.083	0.104	0.047
Subject 9	0.087	0.192	0.128	0.186	0.061	0.156	0.119	0.071
Subject 10	0.196	0.081	0.132	0.062	0.122	0.155	0.134	0.118
Subject 11	0.103	0.144	0.105	0.174	0.081	0.241	0.061	0.091
Subject 12	0.236	0.059	0.156	0.117	0.164	0.084	0.041	0.143
Average	0.133	0.136	0.137	0.137	0.122	0.141	0.098	0.096

At the same time, Table 2 below presents corresponding calculation results of the differences in values of each of the impact factors (for each separate subject's individual multifactor portrait) from the average value (of the same impact factor among all these subjects).

Table 2 – Calculation results of the differences in values of each impact factors (for each separate subject's individual multifactor portrait) from the average value (of the same impact factor among all these subjects)

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8		
Subject 1	–0.042	***–0.083***	0.007	–0.053	***0.095***	–0.054	0.055	0.075		
Subject 2	–0.039	0.072	0.047	0.060	–0.053	0.007	–0.056	–0.038		
Subject 3	–0.014	0.005	–0.035	0.075	–0.024	0.040	–0.016	–0.031		
Subject 4	0.000	–0.064	–0.008	0.034	0.039	0.023	–0.007	–0.017		
Subject 5	–0.021	0.021	–0.002	–0.059	0.011	0.045	–0.004	0.009		
Subject 6	***0.090***	0.013	–0.050	–0.025	0.005	–0.009	–0.005	–0.019		
Subject 7	–0.037	0.003	0.065	–0.055	–0.008	–0.066	0.069	0.029		
Subject 8	–0.032	***0.099***	0.008	0.036	–0.010	–0.058	0.006	–0.049		
Subject 9	–0.046	0.056	–0.009	0.049	–0.061	0.015	0.021	–0.025		
Subject 10	0.063	–0.055	–0.005	***–0.075***	0.000	0.014	0.036	0.022		
Subject 11	–0.030	0.008	–0.032	0.037	–0.041	***0.100***	–0.037	–0.005		
Subject 12	***0.103***	***–0.077***	0.019	–0.020	0.042	–0.057	–0.057	0.047		
							*recommended values of adjustment parameter: Δmin=–0.083–(–0.061);	Δmax=0.103–0.085	min/ max	aver– age
Min	*–0.046*	*–0.083*	*–0.050*	*–0.075*	*–0.061*	*–0.066*	*–0.057*	*–0.049*	–0.083	–0.061
Max	*0.103*	*0.099*	*0.065*	*0.075*	*0.095*	*0.100*	*0.069*	*0.075*	0.103	0.085

At the same time, the last two rows ("Min" and "Max") of Table 2 contain (appropriately): the minimal and the maximal values present in each column of the previous rows of the same table (i.e., among the values of the obtained differences of each of the impact factors).

Among all the results obtained (and presented in Table 2) of the differences between the values of each of the impact factors and the average value of the same impact factor, in the context of given practical applied task (determination and further correction the impact factors of maximum imbalance of the researched multisubjective multifactorial environment representing the technician team of the supported software product) – the maximum level of interest is represented by those values that represent two key indicators, namely: the minimum difference value, as well as the maximum difference value. Because these values, actually, indicate those impact factors which bring the maximal imbalance within the framework of the expanded representation form of the researched multisubjective multifactorial environment (that represents the technician team of the supported software product).

Accordingly, correction of these indicators makes it possible to ensure an increase in the balance level of the entire investigated multisubjective multifactorial environment of the researched software complex's support team. At the same time, the recommended value of the adjustment parameter will be the difference between the current value (of one of the two key indicators) and the corresponding arithmetic mean value (for the same key indicator).

Figure 3 below provides a corresponding graphical interpretation, visualized in the form of a relevant histogram, of the data from Table 2, which represents the results of calculating the differences in values between each of the impact factors and the average value of the same impact factor.

DOI 10.15588/1607-3274-2026-1-17

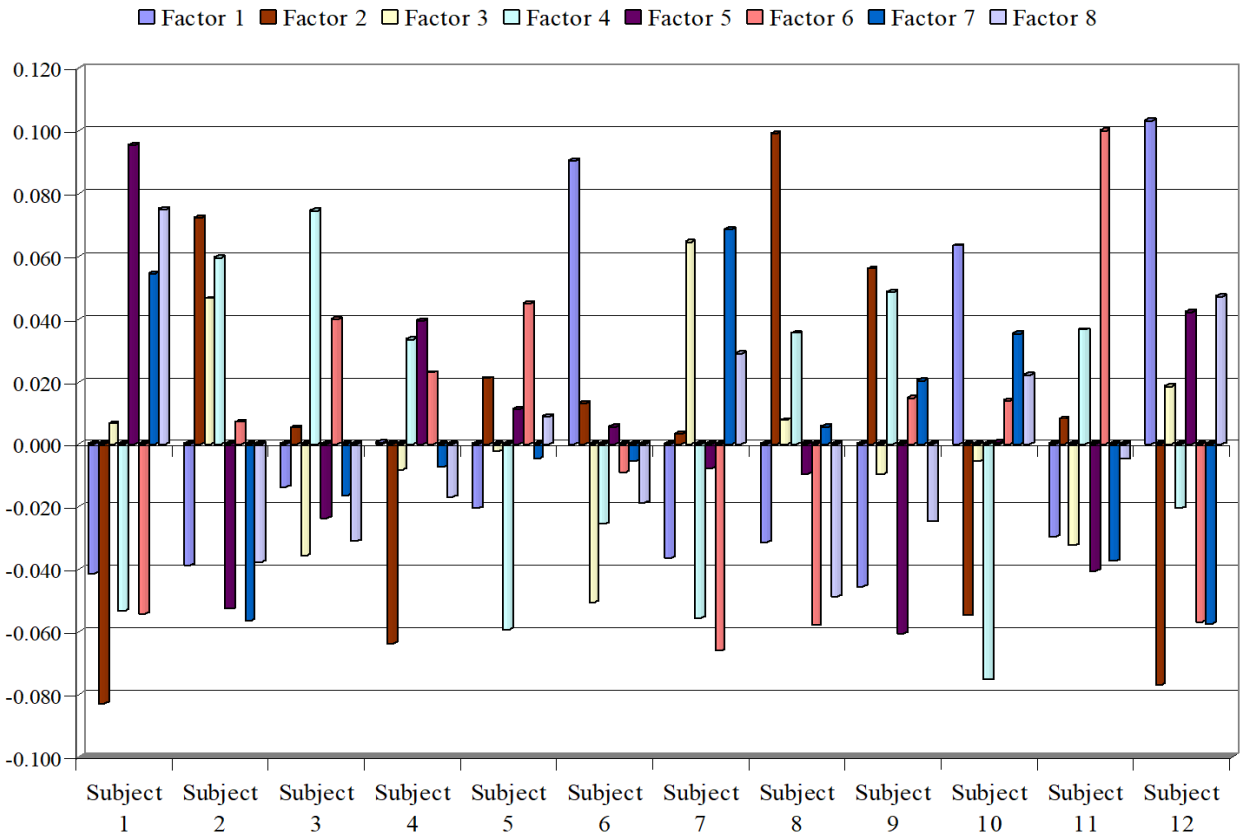


Figure 3 – Visualization of calculation results of the differences in values between each of the impact factors and the average value of the same impact factor.

Besides that, it is also possible to additionally establish a certain (necessary or expedient) threshold deviation value (both in the absolute and relative scales), above which the value of the corresponding impact factor of the relevant subject will be considered as critically affecting the imbalance of the investigated multisubjective multifactorial support environment – and, therefore, requires correction(s).

Thus, for example, based on the data represented above in Table 2, as well as in Figure 3, the identified impact factors which have been considered as such that should be corrected (in the context of given / "being solved" practical applied task of determination and further correction the impact factors of maximum imbalance of the researched multisubjective multifactorial environment representing the technician team of the supported software product) are the following:

– factor 1 within subject 12;
– factor 2 within subject 1;
– factor 2 within subject 12;
– factor 2 within subject 8;
– factor 5 within subject 1;
– factor 6 within subject 11;
– factor 1 within subject 6;
– factor 4 within subject 10.

Thus, determining the relevant impact factors influencing the maximum imbalance of the existing multisubjective multifactorial environment (of the researched software complex' support team), as well as determining the recommended value of the correction parameter of these impact factors – both together provide the possibility of their further correction using the developed method for correction of multisubjective multifactorial environments of software complexes' support.

Figure 4 below presents a visualization of calculation results of the differences in values for all considered impact factors (from the average value of the same impact factor) after the relevant correction(s) have been done for the previously determined impact factors (of maximal unbalancing of the investigated multisubjective multifactorial environment representing the research software complex' support team).

Therefore, as follows from the obtained results of the histogram visualized in Figure 4, as a result of performed correction and application of the developed method, – a reduction of the differences (between the values of each of the impact factors and the average value of the same impact factors) was ensured.

For a more clear understanding of the significance of obtained results, – Figure 5 below presents: the absolute values of the differences for each of the identified impact factors (for which correction has been applied), as well as the differences between the minimal / maximal values and the corresponding arithmetic mean value.

DOI 10.15588/1607-3274-2026-1-17

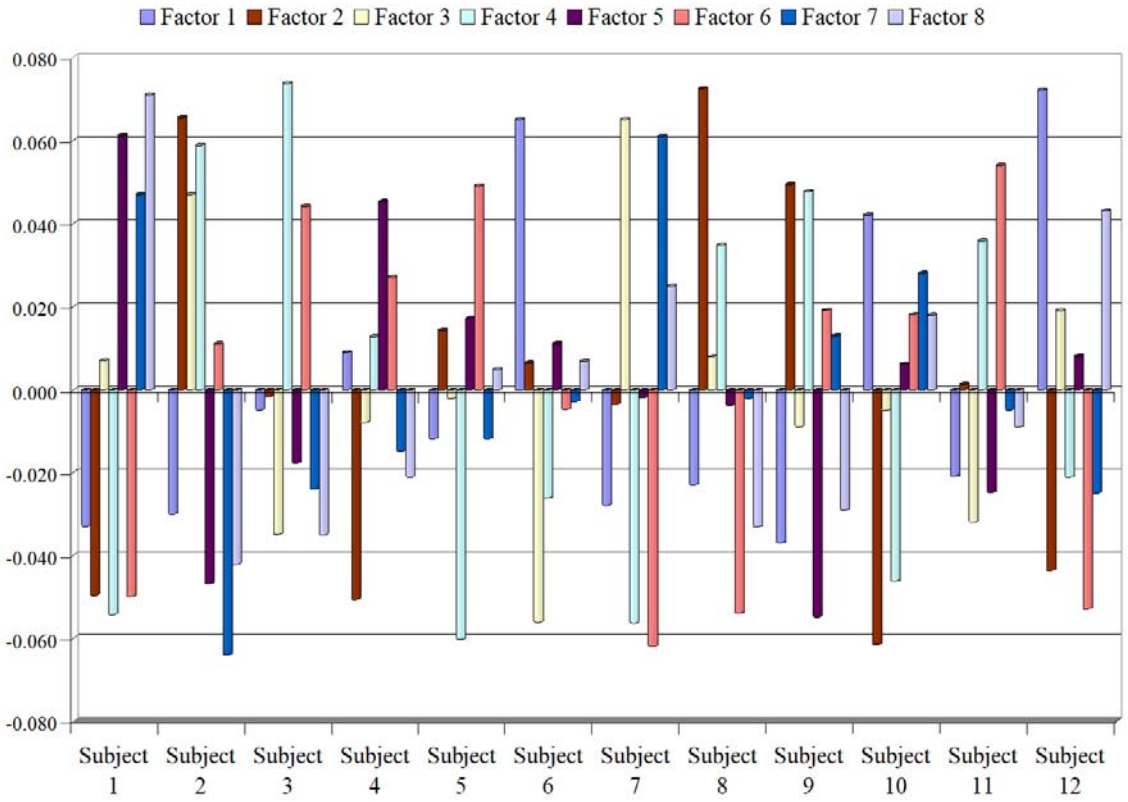


Figure 4 – Visualization of calculation results of the differences in values for all considered impact factors (from the average value of the same impact factor) after the relevant correction(s) have been done for the previously determined impact factors (of maximal unbalancing of the investigated multisubjective multifactorial environment representing the research software complex' support team)

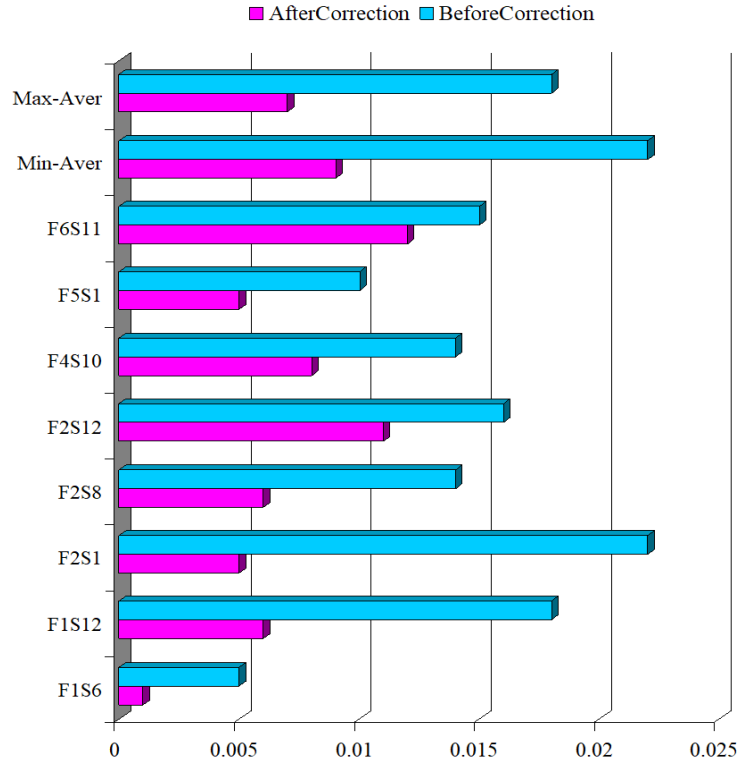


Figure 5 – Comparison visualization between differences' values before and after correction of the identified maximal imbalance's impact factors of the researched multisubjective multifactorial environment

DOI 10.15588/1607-3274-2026-1-17

Thus, as a result of applying the developed method (for correction of multisubjective multifactorial environments of software complexes' support) for solving the given relevant example practical applied task (of determination and further correction the impact factors of maximum imbalance of the researched multisubjective multifactorial environment representing the technician team of the supported software product) a significant improvement(s) in the indicators of the difference between the values of each of the identified "maximum imbalance" impact factors and the corresponding arithmetic mean values of these same factors, have been obtained.

In particular, the following specific improvement achievements have been obtained:

– for factor 1 within subject 6 – 80% improvement;
– for factor 1 within subject 12 – 67% improvement;
– for factor 2 within subject 1 – 77% improvement;
– for factor 2 within subject 8 – 57% improvement;
– for factor 2 within subject 12 – 31% improvement;
– for factor 4 within subject 10 – 43% improvement;
– for factor 5 within subject 1 – 50% improvement;
– for factor 6 within subject 11 – 20% improvement;
– in the context of the minimal deviation value – 59% improvement;
– in the context of the maximal deviation value – 61% improvement.

In addition, the developed method can be used (both in the context of solving the given practical applied tasks, as well as in general) using a recursive (cyclic) approach, where (at each separate cycle of recursion) the determination and correction of the corresponding necessary impact factors are implemented, thereby providing the possibility of "branching", parallelization and research of alternative variants of a multicriterial optimization (including balancing, for example) of any studied multisubjective multifactorial environments.

6 DISCUSSION

Research [10] confirms an importance of taking into account the impact factors (in particular, such global groups of factors as: human and social, as well as technical) on such processes of comprehensive software support as Requirements Engineering. It provides some idea(s) about the importance's distribution structure (of the influence of each of the categories of declared impact factors onto these processes), forming a more or less holistic picture of the environment for interaction of various subjects involved in these processes.

Another research [11] provides a clear list of key factors that affect the productivity of teams who provide and implement such a component of comprehensive software product support as Agile. At the same time, the peculiarity of this research is that it studies the factors that influence the productivity of entire teams (i.e., groups of subjects), but not a separate individuals.

At the same time, research [12] proposes a model (which is based on three dimensions and twelve key factors) that significantly affect Agile software development, and the results of the conducted research provide structuring an Agile environment(s) by taking into account all critical factors of success. At the same time, the presence of dimensions allows to adapt the proposed model in accordance with such important parameters of the developed and supported software products as their: size, nature and financial constraints.

So, absolutely all of the above studies, unfortunately, do not reveal the issues of forming, or, even more so: correcting environments based on these present and declared impact factors, while, to a greater extent, they just focus their main attention on these factors as a separate components, regardless of their integration with each other, and with the researched object(s).

However, at the same time, absolutely all of these studies confirm the criticality of the need of considering and taking into account factors influencing various processes, which (those processes) are the components of comprehensive support for software products.

While the developed and presented in current research method for correction of multisubjective multifactorial environments of software complexes' support allows not only to take into account these impact factors (by the way: without any restrictions on their set, equally processing each necessary and previously declared set of factors), but also to combine them into more significant components – a multisubjective multifactorial environments, but also: to provide the possibility(-ies) of correcting such environments, thereby influencing the quantitative and qualitative indicators of comprehensive support of the investigated supported software products and complexes.

As a further application of the developed method, we see its perspectives and promises both in solving a number of relevant practical applied tasks and problems, including, in particular, the problems of identifying and further correcting problematic factors of influence and/or subjects of the studied multisubjective multifactorial support environments, as well as scientific problems of intellectualizing the processes of software complexes' automated support of various both existing and developed software products.

Thus, given the wide range of relevant and related practical and/or applied problems, the feasibility of further investigation in the considered direction is pretty justified. Furthermore, the proposed method can be adapted for use in other areas of science which study and research any polysubject (or heterogeneous) multifactor environments, such as, for example: socionics, conflictology, social psychology, etc.

CONCLUSIONS

The method for correction of multisubjective multifactorial environments of software complexes' support has been proposed, developed and represented in this research. A key scientific & applied problem, which has been successfully resolved by the proposed method, – is the problem of correction of a multisubjective multifactorial environments of software complexes' support is considered in this research, necessary to provide the possibility(-ies) of adjusting the perception's subjectivization of

DOI 10.15588/1607-3274-2026-1-17

the support object (the supported software, as well as related processes of its complex support), caused by the influence of relevant impact factors.

Input data for the proposed method – are individual multifactor portraits of those researched subjects, who form the corresponding investigated multisubjective multifactorial support environment, as well as the corresponding tasks, defined for correcting this environment.

Two variants of models for adjusting the researched multisubjective multifactorial environments (of software complexes' comprehensive support) have been introduced, namely:

– variant by direct adjusting the personal multifactor portraits of those existing subjects who already form this environment;

– as well as a variant of adjustment the entire environment by encapsulating (into this environment) additional new subjects with such personal multifactor portraits that would ensure a "shift" in the portrait of the entire environment (into which they will be encapsulated) in the required correction vector/direction.

Besides that, additional variants of the correction parameter's Δ value distribution's system have been presented, in particular: based on the polar coordinate system; based on the Cartesian coordinate system; based on the vector system; and based on the Archimedean spiral, and their main features and practical advantages have been given as well.

The developed method provides the possibility(-ies) for studying the processes of collective perception's subjectivization (caused by the influence of various existing impact factors) of objects (of software complexes' comprehensive support) by the relevant subjects (e.g.: support personnel) who directly provide and implement this comprehensive support, and also makes it possible to further automate this complex support of software products in scope of the considered functional and procedural activity segment.

The scientific novelty consists in development of method for correction of multisubjective multifactorial environments of software complexes' support, which ensures possibility(-ies) for researching the factors influencing the processes of complex support of various software products, and also provides opportunities for further intellectualization of the automation processes of such comprehensive support. **The practical significance** consists in development of basic models for adjusting the researched multisubjective multifactorial environments (of software complexes' comprehensive support), as well as related variants of the necessary and suitable distribution system(s) for the value(s) of the adjustment parameter Δ for its better adaptation to the given tasks.

Prospects for further research consist in the designing of an extra specialized dedicated algorithmic and software supply for modeling the correction processes (for example, balancing, and others) of the studied multisubjective multifactorial environments (of software products' and complexes' comprehensive support), as well as in further application of the developed method in solving a range of relevant practical applied problems and tasks, including, in particular, the problems of identifying and further correcting problematic factors of influence and/or subjects of the studied multisubjective multifactorial support environments, as well as scientific tasks and problems of intellectualizing the processes of automation of the software products' and complexes' comprehensive support.

DOI 10.15588/1607-3274-2026-1-17

ACKNOWLEDGEMENTS

This research is proactive. It was carried out as a part of the scientific activity of the authors outside of the working hours at their main positions.

DECLARATIONS

Conflict of interest: The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors' contributions: Andrii Pukach: the method for correction of multisubjective multifactorial environments of software complexes' support, methodology, resources, data curation, project administration, modeling, visualization, correspondence; Vasyl Teslyuk: conceptualization, supervision.

Data availability: The original contributions presented in this study are included in the article.

Software availability: The manuscript has no associated software.

Use of artificial intelligence tools: The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

REFERENCES

1. Salam M., Abdel-Fattah M., Moemen M. A survey on software testing automation using machine learning techniques, *International Journal of Computer Applications*, 2022, Vol. 183, Iss. 51, pp. 12–19. DOI: 10.5120/ijca2022921919
2. Garg U. Exploring the Use of Artificial Intelligence for Software Testing and Debugging, *International Journal of Electrical Engineering and Technology*, 2020, Vol. 11, Iss. 1, P. 94–102. – DOI: 10.17605/OSF.IO/43AE2
3. Kathiriya S., Karangara R., Challla N. Optimizing automated software testing with Machine Learning Techniques, *International Journal of Science and Research*, 2018, Vol. 7, Iss. 3, pp. 1960–1964. DOI: 10.21275/sr24304113021
4. Prathyusha N. Integrating AI in testing automation: Enhancing test coverage and predictive analysis for improved software quality, *World Journal of Advanced Engineering Technology and Sciences*, 2024, Vol. 13, Iss. 1, pp. 769–782. DOI: 10.30574/wjaets.2024.13.1.0486
5. Oyeniran O. C., Adewusi A. O., Adeleke A. G. et al. AI-driven DevOps: Leveraging machine learning for automated software deployment and maintenance [Electronic resource], *Engineering Science & Technology Journal*, 2023, Vol. 4, Iss. 6, pp. 728–740. Mode of access: https://www.researchgate.net/profile/Adams-Adeleke-2/publication/383915954_AI-driven_devops_Leveraging_machine_learning_for_automated_software_deployment_and_maintenance/links/66e0bf2a64f7bf7b19a5c2ed/AI-driven-devops-Leveraging-machine-

learning-for-automated-software-deployment-and-maintenance.pdf (date of access: 14.03.2025). Title from screen.
6. Ghantous G.B. Enhancing DevOps using AI [Electronic resource], *Information Control Systems & Technologies*, 2024, pp. 1–13. Mode of access: https://ceur-ws.org/Vol-3790/paper12.pdf (date of access: 14.03.2025). – Title from screen.
7. Wijaya A. S., Oktavia T. Machine learning approaches for helpdesk ticketing system: a systematic literature review [Electronic resource], *Journal of Theoretical and Applied Information Technology*, 2024, Vol. 102, No. 5, pp. 1831–1842. Mode of access: https://www.jatit.org/volumes/Vol102No5/14Vol102No5.pdf (date of access: 14.03.2025). – Title from screen.
8. Upadhyaya N. Enhancing real-time customer service through Adaptive Machine Learning, *International Journal of Advanced Research in Science, Communication and Technology*, 2024, Vol. 4, Iss. 1, pp. 630–636. – DOI: 10.48175/ijarsct-19381
9. Pukach A. I., Teslyuk V. M. Method of forming multifactor portraits of the subjects supporting software complexes, using a multilayer perceptron, *Radio Electronics, Computer Science, Control*, 2025, Vol. 1, pp. 130–141. DOI: 10.15588/1607-3274-2025-1-12
10. Hidellaarachchi D., Grundy J., Hoda R. et al.The Influence of Human Aspects on Requirements Engineering: Software Practitioners Perspective, *ACM Transactions on Software Engineering and Methodology*, 2021, Vol. 1, No. 1, pp. 1–37. DOI: 10.1145/1122445.1122456
11. Turic M., Celar S., Dragicevic S. Productivity Factors in Agile Software Development Projects, *Proceedings of the 34th DAAAM International Symposium on Intelligent Manufacturing and Automation*, 2023, P. 0004–0008. DOI: 10.2507/34th.daaam.proceedings.001
12. Muhammad A. et al. Investigating crucial factors of Agile software development through composite approach, *Intelligent Automation & Soft Computing*, 2021, Vol. 27, Iss. 1, pp. 15–34. DOI: 10.32604/iasc.2021.014427

Received 11.05.2025.
Accepted 09.01.2026.
Published 27.03.2026.

УДК 004.8

МЕТОД КОРЕКЦІЇ ПОЛІСУБ'ЄКТНИХ МУЛЬТИФАКТОРНИХ СЕРЕДОВИЩ ПІДТРИМКИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ (КОМПЛЕКСІВ, СИСТЕМ, ПРОДУКТІВ, ТОЩО)

Пукач А. І. – канд. техн. наук, асистент кафедри Автоматизованих Систем Управління Інституту Комп'ютерних Наук та Інформаційних Технологій Національного Університету «Львівська Політехніка», Львів, Україна. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0009-0001-8563-3311.

Теслюк В. М. – д-р техн. наук, професор, завідувач кафедри Автоматизованих Систем Управління Інституту Комп'ютерних Наук та Інформаційних Технологій Національного Університету «Львівська Політехніка», Львів, Україна. ROR: https://ror.org/0542q3127. ORCID: https://orcid.org/0000-0002-5974-9310.

АНОТАЦІЯ

Актуальність. Розглянуто задачу корекції полісуб'єктних мультифакторних середовищ підтримки програмного забезпечення (комплексів, систем, продуктів, тощо), необхідної для забезпечення можливості коригування суб'єктивізації сприйняття об'єкта підтримки (підтримуваного програмного комплексу, або процесів його комплексної підтримки), зумовленого дією відповідних факторів впливу. **Об'єктом дослідження** є процес корекції полісуб'єктних мультифакторних середовищ підтримки програмного забезпечення (комплексів, систем, продуктів, тощо). **Предметом дослідження** є методи та засоби корекції полісуб'єктних мультифакторних середовищ підтримки програмного забезпечення (комплексів, систем, продуктів, тощо).

Метою роботи – є розроблення методу корекції полісуб'єктних мультифакторних середовищ підтримки програмного забезпечення (комплексів, систем, продуктів, тощо).

Метод. Запропоновано розроблення методу корекції полісуб'єктних мультифакторних середовищ підтримки програмного забезпечення (комплексів, систем, продуктів, тощо), що дає змогу здійснювати необхідні коригування суб'єктивізації сприйняття досліджуваного об'єкта підтримки (в якості якого може виступати як безпосередньо саме підтримуване програмне забезпечення, а також процеси дотичні до його комплексної підтримки) відповідними суб'єктами взаємодії (як прямої, так і опосередкованої), котрі безпосередньо забезпечують та здійснюють цю комплексну підтримку досліджуваного програмного продукту, з метою забезпечення можливостей подальшої автоматизації й інтелектуалізації цієї підтримки.

Результати. Результатами роботи розробленого методу є моделі відкоригованих полісуб'єктних мультифакторних середовищ підтримки програмного забезпечення (комплексів, систем, продуктів, тощо), отриманих врезультаті розв'язання релевантної науково-прикладної задачі коригування середовищ даного класу. Розроблений метод забезпечує можливість дослідження процесів колективної суб'єктивізації сприйняття (зумовленої дією наявних факторів впливу) об'єктів комплексної підтримки відповідними суб'єктами, що безпосередньо забезпечують та реалізують цю підтримку, а також дає змогу здійснити подальшу автоматизацію та інтелектуалізацію цієї комплексної підтримки програмних продуктів саме в даному функціонально-процесуальному сегменті. В якості практичної апробації розробленого методу, наведено результати розв'язаної прикладної практичної задачі визначення та подальшої корекції факторів впливу максимального розбалансування полісуб'єктного мультифакторного середовища команди підтримки досліджуваного програмного комплексу.

Висновки. Розроблений метод вирішує поставлену задачу корекції полісуб'єктних мультифакторних середовищ підтримки програмного забезпечення (комплексів, систем, продуктів, тощо). Водночас, отримані результати практичної апробації розробленого методу підтверджують його функціональність при вирішенні спектру науково-прикладних задач на основі процесів колективної суб'єктивізації сприйняття об'єктів підтримки (як підтримуваного програмного забезпечення, так і процесів дотичних до його комплексної підтримки), які (ці задачі), в свою чергу, входять до кластеру більш глобальної науково-прикладної проблеми автоматизації й інтелектуалізації комплексної підтримки програмних продуктів.

КЛЮЧОВІ СЛОВА: програмний продукт, комплексна підтримка, фактори впливу, автоматизація, корекція, полісуб'єктне мультифакторне середовище, нейронні мережі, багатошаровий перцептрон.

DOI 10.15588/1607-3274-2026-1-17

ЛІТЕРАТУРА

1. Salam M. A survey on software testing automation using machine learning techniques / M. Salam, M. Abdel-Fattah, M. Moemen // International Journal of Computer Applications. – 2022. – Vol. 183, Iss. 51. – P. 12–19. – DOI: 10.5120/ijca2022921919
2. Garg U. Exploring the Use of Artificial Intelligence for Software Testing and Debugging / U. Garg // International Journal of Electrical Engineering and Technology. – 2020. – Vol. 11, Iss. 1. – P. 94–102. – DOI: 10.17605/OSF.IO/43AE2
3. Kathiriya S. Optimizing automated software testing with Machine Learning Techniques / S. Kathiriya, R. Karangara, N. Challla // International Journal of Science and Research. – 2018. – Vol. 7, Iss. 3. – P. 1960–1964. – DOI: 10.21275/sr24304113021
4. Prathyusha N. Integrating AI in testing automation: Enhancing test coverage and predictive analysis for improved software quality / N. Prathyusha // World Journal of Advanced Engineering Technology and Sciences. – 2024. – Vol. 13, Iss. 1. – P. 769–782. – DOI: 10.30574/wjaets.2024.13.1.0486
5. AI-driven DevOps: Leveraging machine learning for automated software deployment and maintenance [Electronic resource] / [O. C. Oyeniran, A. O. Adewusi, A. G. Adeleke et al.] // Engineering Science & Technology Journal. – 2023. – Vol. 4, Iss. 6. – P. 728–740. – Mode of access: https://www.researchgate.net/profile/Adams-Adeleke-2/publication/383915954_AI-driven_devops_Leveraging_machine_learning_for_automated_software_deployment_and_maintenance/links/66e0bf2a64f7bf7b19a5c2ed/AI-driven-devops-Leveraging-machine-learning-for-automated-software-deployment-and-maintenance.pdf (date of access: 14.03.2025). – Title from screen.
6. Ghantous G. B. Enhancing DevOps using AI [Electronic resource] / G. B. Ghantous // Information Control Systems & Technologies. – 2024. – P. 1–13. – Mode of access: https://ceur-ws.org/Vol-3790/paper12.pdf (date of access: 14.03.2025). – Title from screen.
7. Wijaya A. S. Machine learning approaches for helpdesk ticketing system: a systematic literature review [Electronic resource] / A. S. Wijaya, T. Oktavia // Journal of Theoretical and Applied Information Technology. – 2024. – Vol. 102, No. 5. – P. 1831–1842. – Mode of access: https://www.jatit.org/volumes/Vol102No5/14Vol102No5.pdf (date of access: 14.03.2025). – Title from screen.
8. Upadhyaya N. Enhancing real-time customer service through Adaptive Machine Learning / N. Upadhyaya // International Journal of Advanced Research in Science, Communication and Technology. – 2024. – Vol. 4, Iss. 1. – P. 630–636. – DOI: 10.48175/ijarsct-19381
9. Pukach A. I. Method of forming multifactor portraits of the subjects supporting software complexes, using a multilayer perceptron / A. I. Pukach, V. M. Teslyuk // Radio Electronics, Computer Science, Control. – 2025. – Vol. 1. – P. 130–141. – DOI: 10.15588/1607-3274-2025-1-12
10. Hidellaarachchi D. The Influence of Human Aspects on Requirements Engineering: Software Practitioners Perspective / [D. Hidellaarachchi, J. Grundy, R. Hoda et al.] // ACM Transactions on Software Engineering and Methodology. – 2021. – Vol. 1. No. 1. – P. 1–37. – DOI: 10.1145/1122445.1122456
11. Turic M. Productivity Factors in Agile Software Development Projects / M. Turic, S. Celar, S. Dragicevic // Proceedings of the 34th DAAAM International Symposium on Intelligent Manufacturing and Automation. – 2023. – P. 0004–0008. – DOI: 10.2507/34th.daaam.proceedings.001
12. Investigating crucial factors of Agile software development through composite approach / A. Muhammad et al. // Intelligent Automation & Soft Computing. – 2021. – Vol. 27, Iss. 1. – P. 15–34. – DOI: 10.32604/iasc.2021.014427

DOI 10.15588/1607-3274-2026-1-17

УПРАВЛІННЯ У ТЕХНІЧНИХ СИСТЕМАХ

CONTROL IN TECHNICAL SYSTEMS

UDC 621.51

OPTIMIZATION OF FUEL CONSUMPTION IN THE PROBLEM OF STABILIZING THE ANGULAR POSITION OF AN AXISYMMETRIC SPACECRAFT

Stenin A. A. – Dr. Sc., Professor of Department of Technical Cybernetics National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kiev, Ukraine. ROR: https://ror.org/00syn5v21. ORCID https://orcid.org/0000-0001-5836-9300.

Pasko V. P. – PhD, Associate Professor of Department of Information Systems and Technologies, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kiev, Ukraine. ROR: https://ror.org/00syn5v21. ORCID https://orcid.org/0000-0001-9011-7218.

Soldatova M. O. – PhD, Assistant professor of department of information systems and technologies, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kiev, Ukraine. ROR: https://ror.org/00syn5v21. ORCID https://orcid.org/0000-0003-1233-1272.

Drozdovych I. G. – PhD, Senior Researcher of the Department of Natural Resources, Institute for Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine, Kiev, Ukraine. ROR: https://ror.org/04qbxr381. ORCID https://orcid.org/0000-0002-4216-2417.

ABSTRACT

Context. The problem of maintaining the angular orientation of a spacecraft is critical, especially when subjected to impulsive external disturbances that cause sharp deviations in angular velocities. The relevance of solving this problem is determined by the limited fuel supply on board, particularly for the class of spacecraft designed to provide artificial gravity, where precise and efficient control is paramount.

Objective. The main objective of this work is to minimize the consumption of energy resources (fuel) for the stabilization of the angular position of a specific class of spacecraft. This goal is achieved through the sequential execution of two interrelated tasks: 1) damping sharp deviations in the spacecraft's angular velocities; 2) stabilizing the final angular position.

Method. A two-stage approach is proposed. To solve the first task (damping), optimal control is synthesized using a combination of Pontryagin's maximum principle and the phase plane method. This allows for the creation of optimal switching curves that divide the phase plane into regions with corresponding optimal control values. To solve the second task (stabilization), a modal approach based on a proposed method of indeterminate coefficients is used, which ensures the specified dynamic characteristics of the transient stabilization processes.

Results. Modeling of the dynamics of the spacecraft's angular motion was carried out. The simulation results confirm the high effectiveness of using the proposed combined approach for solving the problem of stabilizing the angular position of the spacecraft after significant external disturbances.

Conclusion. The joint application of Pontryagin's maximum principle with the phase plane method for fuel-efficient damping of angular velocities, followed by the implementation of an optimal stabilization law based on the proposed method of indeterminate coefficients, represents an effective procedure for controlling the orientation and stabilization of a spacecraft with minimal fuel consumption.

KEYWORDS: axisymmetric spacecraft, maximum principle, phase plane, optimal switching and disconnection lines, modal synthesis, method of undetermined coefficients.

NOMENCLATURE

$x(t)$ – n-dimensional state vector;
$u(t)$ – m-dimensional control vector;
$A(t)$ – matrices of size parameters $n \times n$;
$B(t)$ – matrices of size parameters $n \times m$ respectively;
$F(x, u)$ – quality functional;
$H(x, u)$ – Hamilton function;
K – gravitational parameter of the Earth;
$r(\boldsymbol{n})$ – radius-vector of the center of mass of the spacecraft;
$F[N]$ – sum of external forces;
m [kg] – mass of the spacecraft;
I_X, I_Y, I_Z (kg m^2 – the main central moments of inertia of the spacecraft;

DOI 10.15588/1607-3274-2026-1-18

ω_i – projections of the angular velocities and the moments of the spacecraft;

ψ – Euler angles (roll) rotation of the apparatus around its transverse axis *x*;

φ – Euler angles (pitch) rotation of the apparatus around its longitudinal axis *x*;

θ – Euler angles (yaw), rotation of the apparatus around its transverse axis y which runs from one wing to the other;

$M_i[N \cdot m]$ – control moments;

$M_{i3}[N \cdot m]$ – perturbing moments;

T [s] – time interval of the stabilization process;

T_{min} [s] – stabilization time in a system that is optimal in terms of minimizing the time of the transient process;

$\psi_i(t)$ – variables determined from the solution of the conjugate system;

α – coefficient in the equations of motion;

β – coefficient under control actions;

ε[rad/s] – specified area for the final state of angular velocities;

k – weighting coefficient in the quality functional;

c[rad/s] – constant angular velocity of the spacecraft;

λ – roots (poles) of the characteristic polynomial of the closed system

G – diagonal matrices with constant coefficients of the corresponding dimension , penalizes deviations from the target state (zero angles and velocities);

Q – diagonal matrices with constant coefficients of the corresponding dimension, penalizes the expenditure of the control resource (fuel);

a^i_j and b^i_j – coefficients, corresponding the coordinates of the center of that cicle arc, which is the optimal trajectory of switching;

R^i_j – coefficients that determines the radius of the circle arc;

x^T – spacecraft angular motion variables;

f_j – coefficient at λj in *i*-th considered determinant.

INTRODUCTION

At present, space technology is represented by a wide range of vehicles that differ in their intended use, overall weight characteristics, and the composition of onboard equipment. New tasks solved by spacecraft put forward high demands on onboard systems, in particular, the attitude control and attitude stabilization system. Since the accuracy of stabilization of the angular position of spacecraft is significantly affected by external disturbances, in particular, vibrations of attached elastic elements, such as solar panels, antennas, etc., as well as the damping of the angular oscillations of the spacecraft after separation from the launch vehicle or upper stage. it is necessary to improve the methods and algorithms for controlling the orientation and stabilization of the spacecraft, taking into account the specifics of the tasks it solves [1–3].

Energy consumption is one of the most important characteristics of the spacecraft control systems. Accordingly, the consumption of the working body of the engines to maintain the required angular position of the spacecraft in the control mode of its orientation should be minimal. This problem is quite relevant for modern cosmonautics, and its solution based mainly on the methods of the theory of automatic control, and, in particular, the methods of optimal control. It should be noted that, in contrast to the task of minimizing the time of transients, three-level control with zero control value zone is implemented in fuel consumption optimization tasks [4].

To implement these tasks, the most effective control system, which most often used in practice, is the active control system for jet nozzles. The control moment in this system occurs when the mass of the working fluid ejected from the nozzle of a small jet engine, the axis of which does not pass through the center of mass of the spacecraft. The control torque depends on the flow rate of the working fluid, as well as on the size of the lever acting on the engine's tractive effort.

The specified position of the spacecraft is determined in a certain coordinate system, the direction of the axes of which in space known in advance. This coordinate system called the basic reference system. Axes of this system must be set on board the spacecraft using special devices and devices. The second Newton's law used to create the equations of motion of the centers of mass of the spacecraft [1]:

$$m \cdot \frac{d^2 r}{dt^2} = -\frac{K}{r^2} \cdot \frac{\vec{r}}{r} + \vec{F}. \qquad (1)$$

In this paper the equations of rotational motion of the spacecraft are described by the Euler dynamic equations look like [5, 6]:

$$\begin{aligned} I_X \cdot \frac{d\omega_X}{dt} + \omega_Y \cdot \omega_Z \cdot (I_Z - I_Y) &= \sum M_X; \\ I_Y \cdot \frac{d\omega_Y}{dt} + \omega_Y \cdot \omega_Z \cdot (I_Z - I_Y) &= \sum M_Y; \\ I_Z \cdot \frac{d\omega_Z}{dt} + \omega_X \cdot \omega_Z \cdot (I_Y - I_X) &= \sum M_Z. \end{aligned} \qquad (2)$$

$$\begin{aligned} \omega_x &= coa\varphi \ \cos\psi \cdot \frac{d\theta}{dt} - \sin\varphi \frac{d\psi}{dt}; \\ \omega_y &= coa\theta \frac{d\psi}{dt} - \sin\varphi \ \cos\psi \frac{d\theta}{dt}; \\ \omega_z &= \frac{d\varphi}{dt} - \sin\psi \frac{d\theta}{dt}. \end{aligned} \qquad (3)$$

The object of study is an axisymmetric cylindrical spacecraft, a simplified diagram of which with the location of jet engines is shown in Fig. 1. Angular control moments $\sum M_i (I = x, y, z)$ creates by jet engines.

DOI 10.15588/1607-3274-2026-1-18

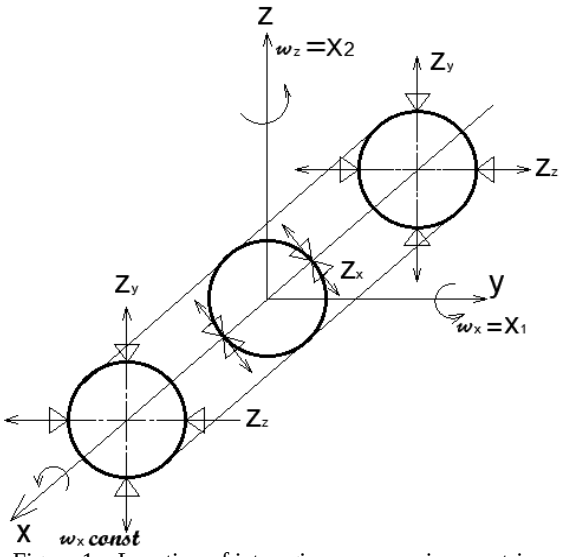


Figure 1 – Location of jet engines on an axisymmetric spacecraft

Introduce the following assumptions:

– spacecraft is axisymmetric with respect to the axis $0x(Iz = IY = I; M_y = M_z = M)$;

– the perturbing forces of the M_{ip} in comparison with the control moments can be ignored $M_{ip} = 0(i = x, y, z)$;

– the angular velocity of the spacecraft around the symmetry axis 0x is constant $(\omega_x(t) = \omega_{x0} = \text{const} = c)$.

Such assumptions may be due to the creation of artificial gravity on the spacecraft. The idea of artificial gravity due to the rotation of an axisymmetric cylindrical spacecraft based on the principle of equivalence of the force of gravity and the force of inertia. In other words, if the inert mass and gravitational mass are equal, it is impossible to distinguish which force acts on the body-the gravitational or inertia force, i.e. the centrifugal force will "push" the astronaut away from the center of rotation, and he will be able to stand on the "floor".

This article considers the problem of maintaining the angular orientation of the spacecraft during sudden deviations of the angular velocities from impulsive external disturbances. Its solution is proposed to be carried out with the sequential execution of two interrelated tasks that guarantee the minimization of fuel consumption for control:

– damping sudden deviations of the angular velocities of the spacecraft (task 1);

– stabilization of the angular position of the spacecraft (task 2).

The relevance of this problem statement is obvious, since the spacecraft have a limited supply of fuel.

It should be noted that this approach is explained by the fact that in practice, due to the presence of measurement errors and delays in the processing of control signals, it is impossible to reduce strictly to zero the impulse perturbations of the angular velocities. In real conditions, with relay control, this leads to the occurrence of undamped self-oscillations at the end point of control, and, moreover, does not guarantee the preservation of the original angular orientation of the spacecraft.

1 PROBLEM STATEMENT

Task 1. To solve task 1, only the system of equations (2) is considered. The system of equations (2) is sufficient to describe the angular movements of the spacecraft, if the influence of internal moments acting on it can be ignored [3, 5, 6].

Introduce the following notation:

$$x_1 = \omega_y; x_2 = \omega_z; a = \omega_{x0}\frac{(I - I_x)}{I}; \quad \beta = \frac{M}{I}; u_1 = \frac{M_y}{M}; u_2 = M_z / M. \tag{4}$$

Taking into account (4), the system (2) will take the form:

$$\frac{dx_1(t)}{dt} = ax_1(t) + \beta u_1(t); \tag{5}$$

$$\frac{dx_2(t)}{dt} = ax_2(t) + \beta u_2(t). \tag{6}$$

The boundary conditions of the optimization problem are:

$$x_1(0) = x_{10}; x_2(0) = x_{20}; x_1(T) = x_2(T) = 0. \tag{7}$$

Finally, as is customary in many works on optimal control, we will consider normalized controls for the convenience of analyzing the results obtained:

$$|u_1(t)| \le 1; |u_2(t)| \le 1. \tag{8}$$

Task 1 of the optimal fuel consumption damping of sharp deviations of the angular velocities of the spacecraft is formulated as follows: to determine from the permissible range (8) the control actions $u_1(t)$ and $u_2(t)$ and the boundary conditions (7) that ensure the transfer of the system (5), (6) at the stabilization time interval $0 \le t \le T$ from an arbitrary initial state x_{10}, x_{20} to a given final state, defined by some area $| x_1(T) | \le \varepsilon$, $| x_2(T)| \le \varepsilon$, and minimize the quality functional:

$$I_K = \int_0^T \left[k + \sum_{i=1}^{2} |u_i(t)|\right] dt, \quad T - not \quad fixed, 0 \le k \le \infty. \tag{9}$$

In other words, the task 1 is to extinguish sharp deviations of the angular velocities ω_y and ω_z from their zero values.

Task 2. To solve task 2 the system of equations (2) and (3) for the spacecraft (Fig. 1) is considered under assumptions (4) and the following notation:

$$x_1 = \omega_y; x_2 = \omega_z; x_3 = \psi; x_4 = \varphi; \quad \alpha = \omega_{x0}\frac{(I - I_x)}{I}; \beta = \frac{M}{I}. \tag{10}$$

DOI 10.15588/1607-3274-2026-1-18

Taking into account (10) and the values of trigonometric functions for small values of their parameters, system (2), (3) takes the form:

$$\begin{aligned}\frac{dx_1}{dt} &= ax_2 + \beta u_1;\\ \frac{dx_2}{dt} &= -ax_1 + \beta u_2;\\ \frac{dx_3}{dt} &= x_1 + cx_4;\\ \frac{dx_4}{dt} &= x_2 - cx_3.\end{aligned} \tag{11}$$

We accept as boundary conditions:

$$x_i(0) = x_{i0}; x_i(T) = 0 (i = 1,2,3,4). \tag{12}$$

We also assume that system (11) is completely controllable and completely observable. To assess the quality of transient processes of spacecraft orientation stabilization, we will use a quadratic functional of the form

$$\begin{aligned}x_1(0) = x_{10}; x_2(0) = x_{20}; x_1(T) = x_2(T) = 0,\\ F(x,u) = \int_0^\infty (x^T Gx + u^T Qu)dt.\end{aligned} \tag{13}$$

Task 2 is formulated as follows: to find the optimal values of the control u_1, u_2, which transfer the system (11) from the given initial values of the variables to the final ones according to the boundary conditions (12) and minimize the functional (13).

2 REVIEW OF THE LITERATURE

Quite a lot of works are devoted to the problems of optimal control of space vehicles. In particular, in [7], an optimal control law was obtained that stabilizes the program motion of the spacecraft. The stabilizing properties of the proposed regulators are proved by the Lyapunov method. In [8], an approximate optimal method for stabilizing the relative motion of a spacecraft is proposed, based on the dynamic programming method and the averaging method. In [9] the problem of control efficiency was solved with priorities in managing based on Pontryagin maximum and the mathematical apparatus of quaternions. In [10], a block diagram was developed for monitoring the angle of rotation of the steering mechanism and its angular velocity for the operation of disturbing forces. In [11], an anti-perturbing inverse control scheme for the movement and rotation of a rigid spacecraft with external disturbances and drive limitation was implemented. In [12], the maneuver of satellites with a minimum orientation time is studied using Control Moment Gyros (CMG) gyroscopes. In [13], an optimal energy-saving problem for a rendezvous mission based on linear-quadratic optimization is considered. In [14], a control law with output feedback was proposed for the problem of spacecraft reorientation based on the Lyapunov method. In [15], a significant control of the engine switch was implemented for an accurate study of the spacecraft. The article [16] proposes a PID controller architecture for controlling the attitude of a spacecraft, and the concepts of the nonlinear control theory H_∞ are applied to obtain stability properties.

It should be noted that despite the significant number of articles related to the control of the angular motion of spacecraft, in contrast to the problem of linear quadratic optimization and minimum transient time, there are practically no articles on the problems of optimizing fuel consumption during orientation and stabilization of the angular position of spacecraft. In this sense, we can note the work [17], which considers the problem of the optimal turn of the spacecraft. Turnaround time is kept to a minimum, as is the functionality that matters in terms of fuel consumption.

This article proposes algorithm for stabilizing the initial orientation of a spacecraft in the event of sudden disturbances based on the sequential solution of the above two interrelated tasks.

3 MATERIALS AND METHODS

The choice of functional (9) is scientifically attractive in the sense that it provides the necessary compromise between the fuel consumption and the stabilization time of the angular position of the spacecraft by the given value k. In this case, for k=0, the problem of pure fuel consumption is solved, and for $k\to\infty$, the problem of minimizing the transition process time is solved. It should be noted that in the case of pure fuel consumption, when k=0 in the functional (9), the condition for existence of optimal control is the condition $T >$, $T_{\min}$, when $k\to\infty$ in functional (9).

To solve the task 1 use the mathematical apparatus of the maximum principle (in our case, the minimum principle) [4, 9] and the phase space method (in our case, the phase plane) [18, 19, 24]. According to the minimum principle [4], the Hamiltonian of this task has the form:

$$H = k + \beta|u_1| + \beta|u_2| + \psi_1(ax_2 + \beta u_1) + \psi_2(-ax_2 + \beta u_2). \tag{14}$$

It follows from the analysis of the Hamiltonian (14) that the controls that minimize it satisfy the following conditions:

$$u_i(t) = 0, if\ |\psi_i(t)| \le 1; \tag{15}$$

$$u_i(t) = -sig\psi_i(t), if\ |\psi_i(t)| < 1; \tag{16}$$

$$0 \le u_i(t) \le 1, if\ |\psi_i(t)| = -1; \tag{17}$$

$$-1 \le u_i(t) \le 0, if\ |\psi_i(t)| = 1. \tag{18}$$

Due to the non-degeneracy of this problem, which is quite easy to justify in accordance with [4], conditions (17) and (18) are excluded from consideration. The optimal values of the control actions $u_1(t), u_2(t)$ from (15), (16) are clearly illustrated in Fig. 2, from which it follows that the following control sequences are optimal

$$\begin{aligned}&\ldots -1.\ -1\to -1,0\to -1,1\to 0,1\to -1,1\to 0,1\\ &\to 1,1\to 1,0\to 1,\ -1\to 0,\ -1\to -1,1\ldots\end{aligned} \tag{19}$$

DOI 10.15588/1607-3274-2026-1-18

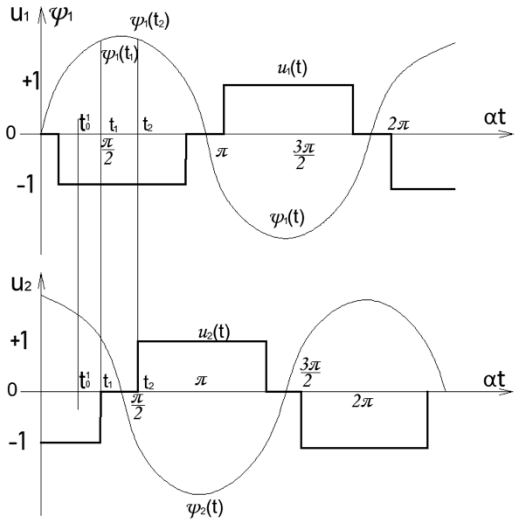


Figure 2 – Optimal sequence of values of control actions

It should be noted that [4] provides a strict justification for the structure of optimal fuel consumption processes with three-level control. Moreover, as in our case, in the presence of complex-conjugate roots, the number of switches depends on the initial conditions. By entering the inverse time $z=T-t$, integrate the system for $u_1=+/-1, u_2=+/-1$ and, excluding z in the obtained solutions, obtain circles on the phase plane (αx_1, αx_2), which are described by equations:

$$[ax_1-(2i-1)\beta u_2]^2+(ax_2+\beta u_1)^2=2\beta^2; u_1=-u; \quad (20)$$

$$(ax_1-\beta u_2)^2[ax_2-(2i-1)\beta u_1]^2= \quad (21)$$
$$=2\beta^2; u_1=-u_2=+/-1.$$

The area of turning off one of the control actions in sequence (19) can be found as the locus of displaying the points of the switching lines found above for the time-optimal control system (Fig. 3). Omitting intermediate mathematical calculations, the optimal disabling curves of one of the controls in the sequence (19) are determine in the coordinate system $(\tilde{x}_1, \tilde{x}_2)$

$$\begin{aligned}\tilde{x}_1&=ax_1\cos\varphi+ax_2\sin\varphi;\\ \tilde{x}_2&=ax_2\cos\varphi+ax_1\sin\varphi;\\ \varphi&=\frac{\operatorname{arctg}\beta}{k+\beta},\end{aligned} \quad (22)$$

i.e. they are described as arcs of circles

$$\left(\tilde{x}_2+\alpha_j^i u_l^i\right)^2+\left(\tilde{x}_1+b_j^i u_2^i\right)^2=\left(R_j^i\right)^2. \quad (23)$$

On Fig. 3 the disabling curves of one of the control actions in the sequence (19) are denoted as $0Q_j^iK_j^iL_j^i...(i=1,2,3,4)$. Here are the values of optimal controls for each of the regions of the phase plane (ax_1, ax_2). The value of the coefficient k in the functional (9) is selected based on the practical requirements for fuel consumption and stabilization time. In addition, Fig. 3 also shows the optimal phase trajectory, which is a spiral from point $A(ax_{20}, ax_{10})$ to the origin. It consists of arcs of a circle with a center and radius determined by the values of optimal controls from the area of the phase plane, where the point of the current values of the angular velocities of the spacecraft is located.

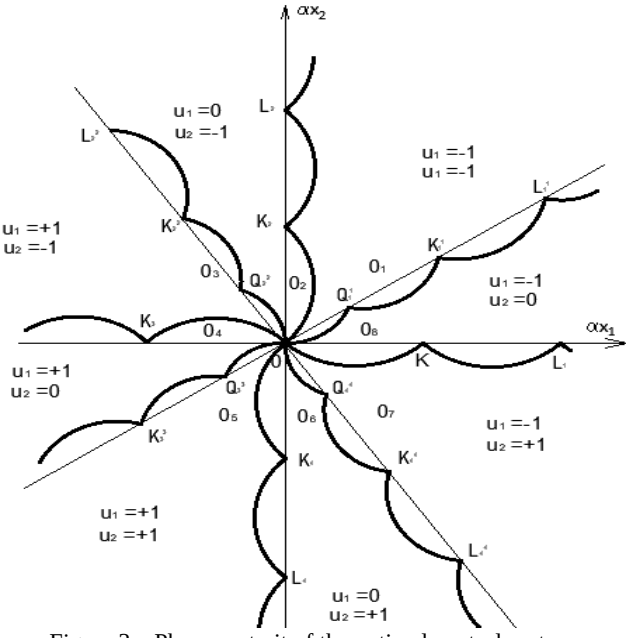


Figure 3 – Phase portrait of the optimal control system

In task 2 the controls u_1, u_2 are determined by a linear combination of deviations *xi* (i=1,…,4), i.e:

$$u_k=\sum p_{ki}x_i (i=1,...,4; k=1,2); \quad (24)$$

It is known that the main problem of the analytical design of optimal controllers is the absence of a direct relationship between the coefficients of the functional and the dynamic indicators of transient stabilization processes. Therefore, in this paper, it is proposed to use the modal synthesis of optimal control based on the method of indefinite coefficients [20], which in this article is generalized to the vector case of control using the superposition principle. Write system (11) in the following vector form:

$$\frac{dx}{dt}=Ax+Bu. \quad (25)$$

Without losing the generality of the results obtained below, we use the principle of superposition to determine the required coefficients $p_{ki}(i=1,...,4; k=1,2)$.

Step 1. Accept $(u=u_{1,0}), u_1=u$. It is known that for completely observable systems of the form (25) in the case of a quadratic performance criterion (13), the extre-

DOI 10.15588/1607-3274-2026-1-18

mal control u is a linear state function. We write control of the system (25) in vector form:

$$u = \bar{p}^T \bar{x}; \tag{26}$$

To determine the vector of feedback coefficients $\bar{p}$, we use the modal approach proposed by the authors in [20] based on the method of indeterminate coefficients.

Modal synthesis is based on the fact that the vector of feedback coefficients $\bar{p}$ can be chosen in such a way that the poles of the closed system (25) will be located at given arbitrary points that provide the required dynamic properties of transient processes of stabilizing the angular orientation of the spacecraft [21, 22].

In the general case, it was proved in [20] that the unknown coefficients in the optimal control law (26) enter linearly into the expression for the coefficients of the characteristic polynomial of the closed system (25), i.e.:

$$H(\lambda) = \lambda^n + \left(\sum_{i=1}^{n} c_{n-1,i}\, p_i + d_{n-1}\right)\lambda^{n-1} + \ldots + \left(\sum_{i=1}^{n} c_{0,i}\, p_i + d_0\right), \tag{27a}$$

or

$$H(\lambda) = \lambda^n + \left(\bar{c}_{n-1}^T \bar{p} + d_{n-1}\right)\lambda^{n-1} + \ldots + \left(\bar{c}_0^T \bar{p} + d_0\right). \tag{27b}$$

Characteristic determinant of the closed system (25) has next form:

$$\det(\lambda) = \left|A + B_p^{-T} - I\right|\lambda = \begin{vmatrix} a_{11} + b_1 p_1 - \lambda \ldots a_{1j} p_j \ldots a_{1n} + b_1 n \\ a_{ji} + b_j p_1 \ldots a_{jj} + b_j p_j - \lambda \ldots a_{jn} + b_j p_n \\ a_{n1} + b_n p_1 \ldots a_{nj} + b_n p_j \ldots a_{nn} - \lambda \end{vmatrix}. \tag{28}$$

Define the unknown parameters $c_{ji}, d_j\,(j = \overline{0,n-1}; i = \overline{1,n})$ using the undetermined coefficients method [23]. To do this, we put $p_i = 0 (i = \overline{1,n})$ in the characteristic determinant (28) at the first step and explanded characteristic determinant (28) by one of the known numerical methods [23]. The coefficients found for different powers of λ determine the unknown coefficients $d_i = (j = \overline{01-n})$ in the expressions for the characteristic polynomial of the closed system for the corresponding powers of λ. In the next n steps, setting sequentially one of the coefficients $p_i = (i = \overline{1,n})$ equal to one while others remain zero and explanded the characteristic determinant, we obtain expressions for the unknown parameter c_{ji} for the corresponding power $\lambda^j = (j = \overline{0,n-1})$ in the characteristic polynomial of the closed system:

$$c_{ji} = (f_j - d_i), \tag{29}$$

On the other hand, the characteristic polynomial of the closed system (25) with desired roots $\lambda_1, \lambda_2, \ldots \lambda_n$ has the form:

$$F(\lambda) = \prod_{i=1}^{n} (\lambda - \lambda_i) = \sum_{j=0}^{n-1} 1_j \lambda^j + \lambda^n. \tag{30}$$

As a result, taking into account (27b), (29), (30) to determine the feedback coefficients, we obtain a system of linear algebraic equations:

$$col\left(\bar{c}^T, \bar{c}^T, \ldots, \bar{c}^T\right)\bar{p} = \bar{1} - \bar{d}, \tag{31}$$

where $i = (1_{n-1}, 1_{n-1}, \ldots, 1_0) d = (d_{n-1}, d_{n-2}, \ldots, d_0)$.

According to this approach, for the chosen poles $\lambda_1, \ldots \lambda_4$ of a closed optimal stabilization system, the control u1 is defined as:

$$u_2, p_{11} x + \ldots + p_{14} x_4. \tag{32}$$

The resulting controls (32) and (33) are components of the optimal vector control (26).

Step 2. We close system (25) to the found control u1 from (32). We accept u=(0,u2, 0,…, 0), u2=u. Similarly to step 1, using the modal approach based on the method of uncertain coefficients, we obtain for the same poles the optimal control u2 in the form:

$$u_2 = p_{11} x_1 + \ldots + p_{24} 4 x_4. \tag{33}$$

4 EXPERIMENTS

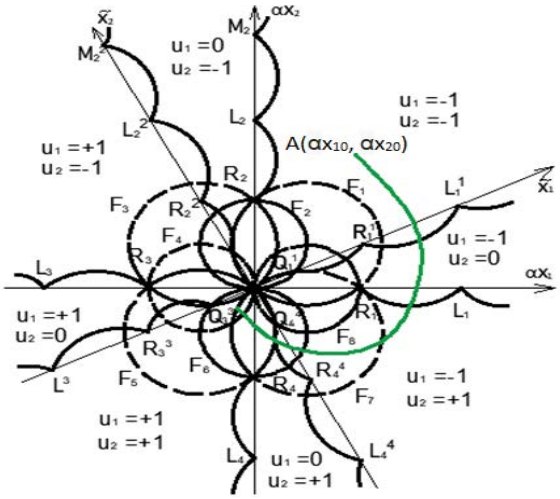
Without losing the generality of the results obtained, to simulate the optimal stabilization problems considered above, the following values of the parameters of equations (5), (6), (11) α = β =c= 1 were accepted.

The simulation was aimed at research in order to confirm the effectiveness of the proposed combined approach to the angular stabilization of the spacecraft, as well as the impossibility of maintaining the original angular orientation only by solving task 1. In addition, the processes of angular stabilization were studied in the presence of errors in assessing the state of the spacecraft and delays in the control loops u_1, u_2. The simulation was carried out in the Mathcad environment. Evaluate the process of angular stabilization of the spacecraft when the values of angular velocities enter the region of specified values of task 2.

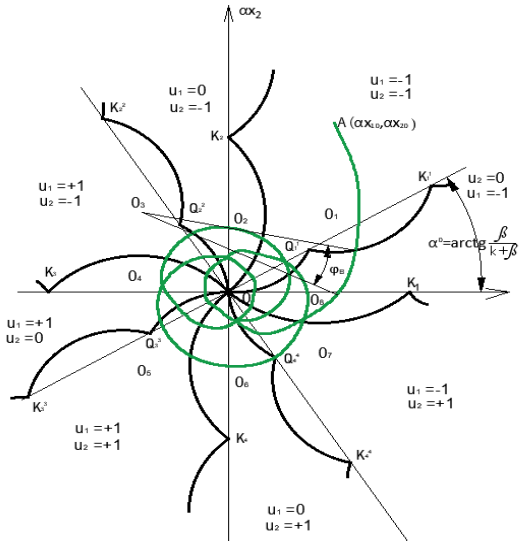
DOI 10.15588/1607-3274-2026-1-18

5 RESULTS

For task 1, based on equations (5), (6) at α = β = 1, the dynamics of the transition process was simulated from the given initial conditions $x_1(0) = 3[0/s]; x_2(0) = 4[0/s]; x_3(0) = 0; x_4(0) = 0$ to zero final ones $x_1(T) = x_2(T) = 0[0/s]$ (Fig. 4).



a



b

Figure 4 – Phase portraits and phase trajectories of the transient process in the absence (Fig. 4a) and presence (Fig. 4b) of delay in the control loop

Also, for the initial ones $x_1(0) = 3[0/s]; x_2(0) = 4[0/s]; x_3(0) = 0; x_4(0) = 0,$ and final $x_1(T) = x_2(T) = 0[0/s]$ conditions, based on equation (11), modeling of the dynamics of the transient process in the time domain was carried out (Fig. 5 and Fig. 6), if there is an error in assessing the state spacecraft. Here $z_{n,j} = x_i$, and the real time t_{real} is determined through the model t_{mod} as $t_{real} = 0.02t_{\text{mod}} = 0.02t$.

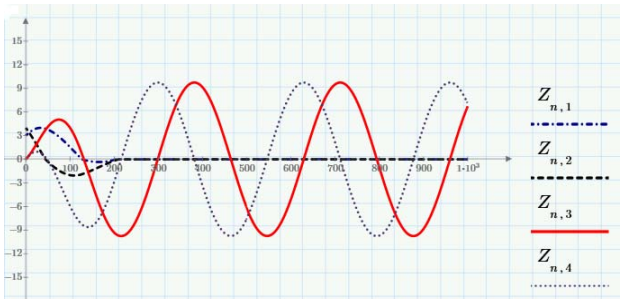


Figure 5 – Continuous self-oscillations x_3, x_4

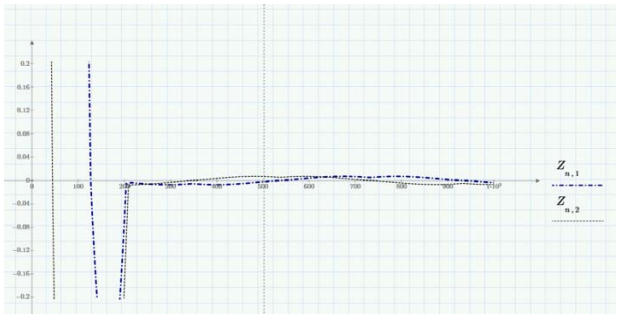
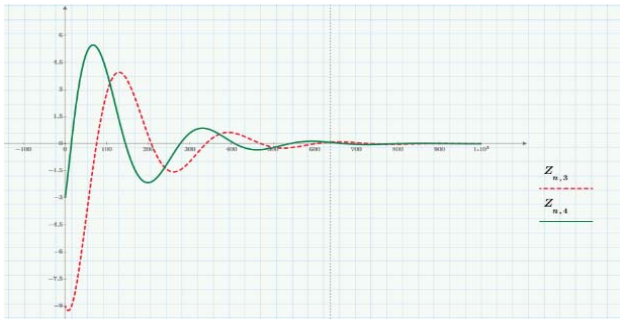
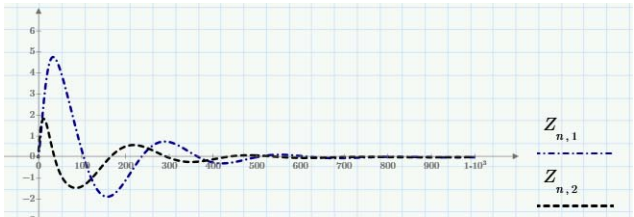


Figure 6 – Continuous self-oscillations x_1, x_2

For task 2, the system of equations (11) was simulated with the values $\alpha - \beta = c = 1; \varepsilon = 0.2[0/s]$ with initial conditions that were final for task 1, i.e. $x_1(T) \le 0.2[0/s]; x_2(T) \le 0.2[0/s]$; and finite zero coordinates $x_1(\infty) = x_2(\infty) = x_3(\infty) = x_4(\infty) = 0$ At the same time, a version of the optimal control actions u_1 and u_2 was synthesized, obtained for the desired spectrum of roots $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\} = \{-0.1, -0.5, -1, -2\}$ based on the above modal approach. Graphs of transient processes are shown in Fig. 7a, b.



a



b

Figure 7 – Transient graphs: a) x_3, x_4; b) x_1, x_2

6 DISCUSSION

With all the obvious advantages of the synthesized optimal fuel consumption law for stabilizing the angular position of the spacecraft (task 1), in which, by choosing

DOI 10.15588/1607-3274-2026-1-18

the value of k in (9), it is possible to vary between the time of the transition process and fuel consumption, in real conditions with relay control in the region of zero coordinates undamped self-oscillations may occur in the presence of unaccounted errors in assessing the state of the spacecraft (Fig. 5) and/or delay (Fig. 4b) in the formation of optimal control actions u_1 and u_2. The error in assessing the state of the spacecraft is simulated by approximating the optimal switching curves with straight lines. In Fig. 6 for greater clarity the presence of self-oscillations x_1 and x_2 , the scale of Fig. 5 along the vertical axis. In addition, the use of only the algorithm of task 1 does not guarantee the preservation of the original angular orientation of the spacecraft in the region of zero values x_1 and x_2 at $t \approx 4$c. (Fig. 5). When implementing task 2, it is possible to avoid the occurrence of undamped self-oscillations, ensure the specified dynamic properties of transient processes in the region of zero coordinates and restore the stable initial angular orientation of the spacecraft, i.e. $x_1(0) = 0 (i = 1,2,2,4)$. Thus, the combined approach proposed in the article to solving the problem of optimal stabilization of the angular position of the spacecraft allows to optimally damp sudden deviations in angular velocities in terms of fuel consumption and to maintain the initial orientation of the spacecraft.

CONCLUSION

This article discusses the problem of maintaining the angular orientation of a spacecraft during sharp deviations of angular velocities from pulsed external disturbances. Its solution is proposed to be carried out by sequentially performing two interrelated tasks: damping sharp deviations in the angular velocities of the spacecraft (task 1); stabilization of the angular position of the spacecraft in the region of zero coordinates (task 2). As part of the solution to task 1, based on the methods of the Pontryagin maximum principle and the phase space (in this case, the phase plane), optimal switching curves were synthesized that unambiguously divide the phase plane into regions with the corresponding values of optimal controls. To solve task 2, a modal approach based on the method of undetermined coefficients is proposed. This approach makes it possible to provide specified dynamic indicators of transient processes for stabilizing the angular position of the spacecraft.

ACKNOWLEDGMENT

The work was carried out within the framework of the state budget research project “Development of a monitoring and control system for robotic mobile devices for integrated monitoring of the state of the environment and ground objects” (state registration number: 0113U003351). The authors express their deep gratitude to Corresponding Member of the National Academy of Sciences of Ukraine Alexander Trofimchuk for valuable and constructive advice and comments during the preparation of this article.

DECLARATIONS

Conflict of interest. The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship, or otherwise, that could affect the research and its results presented in this paper.

Authors’ contributions. **Alexandr Stenin** developed the conceptual framework and methodology based on Pontryagin's maximum principle and provided overall supervision. **M. Soldatova** performed the formal mathematical analysis and investigated the integration of the two-stage stabilization approach. **V. P. Pasko** was responsible for the writing of the original draft and describing the spacecraft orientation processes. **Drozdovich** developed to performed the visualization of results, and served as the project administrator for the final manuscript preparation.

Data availability. The manuscript has no associated data

Software availability. The manuscript has no associated software

Use of artificial intelligence tools. The authors confirm that they did not use artificial intelligence technologies in creating the submitted work.

REFERENCES

1. Chernousko F. L., Akulenko L. D., Leshchenko D. D. Evolution of motions of a rigid body about its center of mass. Cham, Springer International Publishing AG, 2017, 242 p. DOI: 10.1007/978-3-319-53928-7
2. Markley F. L., Crassidis J. L. Fundamentals of spacecraft attitude determination and control. New York, Springer Science+Business Media, 2014, 485 p. DOI: 10.1007/978-1-4939-0802-8
3. Sands T. Advances in spacecraft attitude control. London, IntechOpen Limited, 2020, 274 p. DOI: 10.5772/intechopen.77574
4. Athans M., Falb P. L.Optimal control: an introduction to the theory and its applications. Mineola, Courier Corporation, 2006, 879 p.
5. Yang Y. Spacecraft attitude and reaction wheel desaturation combined control method, *IEEE Transactions on Aerospace and Electronic Systems*, 2017, Vol. 53, № 1, pp. 286–295. DOI: 10.1109/TAES.2017.2650158
6. Kienitz K. H. Attitude stabilization with actuators subject to switching restrictions: an approach via exact relay control methods, *IEEE Transactions on Aerospace and Electronic Systems*, 2006, Vol. 42, № 4, pp. 1485–1492. DOI: 10.1109/TAES.2006.314588
7. El-Gohary A. Optimal control of a rigid spacecraft programmed motion without angular velocity measurements, *European Journal of Mechanics – A/Solids*, 2006, Vol. 25, № 5, pp. 854–866. DOI: 10.1016/j.euromechsol.2005.11.009
8. Zabolotnov Yu. Approximate optimal method for controlling the angular motion of a spacecraft as part of an orbital tether system, *Conference Series Materials Science and Engineering*, 2020, Vol. 984, Art. no. 012024. DOI: 10.1088/1757-899X/984/1/012024
9. Levskii M. V. Optimization problem of attitude control of a spacecraft with bounded rotary energy using quaternions, *International Robotics & Automation Journal*, 2021, Vol. 7, № 2, pp. 63–73. DOI: 10.15406/iratj.2021.07.00228
10. Avdejev V. Reduced observer in stabilizing system of a rocket motion, *Radio Electronics, Computer Science, Control*, 2020, № 2, pp. 165–172. DOI: 10.15588/1607-3274-2020-2-17

DOI 10.15588/1607-3274-2026-1-18

11. Pukdeboon Ch. Anti-disturbance inverse optimal control for spacecraft position and attitude maneuvers with input saturation, *Advances in Mechanical Engineering*, 2016, Vol. 8, № 5, pp. 1–14. DOI: 10.1177/1687814016649887
12. Tsuchiya M., Higuchi T. Semi-optimal control for minimum-time maneuver of satellite with variable speed pyramid type SGCMGs, *Trans. JSASS Aerospace Tech. Japan*, 2021, Vol. 19, № 1, pp. 24–33. DOI: 10.2322/tastj.19.24
13. Moon G.-H., Lee B.-Y., Tahk M.-J., Lee J.-H. Optimal rendezvous guidance using linear quadratic control, *MATEC Web of Conferences*, 2016, Vol. 54, P. 09002. DOI: 10.1051/matecconf/20165409002
14. Guan T., Li B. Output feedback attitude control for rigid spacecraft under attitude constraints, *Journal of Industrial and Management Optimization*, 2023, Vol. 19, № 7, pp. 5294–5305. DOI: 10.3934/jimo.2022173
15. Leomanni M., Garulli A., Giannitrapani A., Pugi A. Minimum switching thruster control for spacecraft precision pointing, *IEEE Transactions on Aerospace and Electronic Systems*, 2017, Vol. 53, № 2, pp. 683–697. DOI: 10.1109/TAES.2017.2655120
16. Show L. L., Juang J. C., Lin C. T., Jean J. H. Spacecraft robust attitude tracking design: PID control approach, *Proceedings of the American Control Conference, Anchorage, 8–10 May 2002 : proceedings*. Piscataway, IEEE, 2002, pp. 836–841. DOI: 10.1109/ACC.2002.1023210
17. Levskii M. V. Optimal control of a programmed turn of a spacecraft, *Cosmic Research*, 2003, Vol. 41, pp. 178–192. DOI: 10.1023/A:1023391232053
18. Stenin O. A., Pasko V. P., Drozdovich I. G., Stenin O. O. Optimal damping of deviations of angular velocities of an axisymmetric spacecraft, *Space Science and Technology*, 2021, Vol. 27, № 4 (131), pp. 21–31. DOI: 10.15407/knit2021.04.021
19. Datta A. K. On optimization of a second-order non-linear control system for various performance criteria, *International Journal of Control*, 1967, Vol. 5, № 3, pp. 269–287. DOI: 10.1080/00207176708921760
20. Stenin A., Drozdovych I., Soldatova M. Method of uncertain coefficients in problems of optimal stabilization of technological processes, *Radio Electronics, Computer Science, Control*, 2020, № 1(52), pp. 209–217. DOI: 10.15588/1607-3274-2020-1-21
21. Zubov, N. E., Mikrin E. A., Misrikhanov M. Sh., Ryabchenko V. N. Synthesis of spacecraft control laws that ensure optimal placement of poles by a closed-loop control system, *Izvestiya of the Russian Academy of Sciences. Theory and Control Systems*, 2012, № 3, pp. 98–111.
22. Zabolotnov Yu. M., Lobankov A. A. On the problem of optimal stabilization of the angular motion of a small spacecraft during the deployment of an orbital tether system, *Vestnik of the Samara State Aerospace University*, 2016, Vol. 15, № 1, pp. 46–54. DOI: 10.18287/2412-7329-2016-15-1-46-54
23. Epperson J. F. An introduction to numerical methods and analysis, 2nd ed. Hoboken, NJ, John Wiley & Sons, Inc., 2013, 591 p.
24. Choi D.-S., Kim S.-J., Ha I.-J. A phase-plane approach to time-optimal control of single-DOF mechanical systems with friction, *Automatica*, 2003, Vol. 39, № 8, pp. 1407–1415. DOI: 10.1016/S0005-1098(03)00112-2

Received 21.08.2025.
Accepted 02.02.2026.
Published 27.03.2026.

УДК 621.51

ОПТИМІЗАЦІЯ ВИТРАТ ПАЛИВА В ЗАДАЧІ СТАБІЛІЗАЦІЇ КУТОВОГО ПОЛОЖЕННЯ ОСЕСИМЕТРИЧНОГО КОСМІЧНОГО АПАРАТУ

Стенін А. А. – д-р техн. наук, професор кафедри технічної кібернетики Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна. ROR: https://ror.org/00syn5v21. ORCID https://orcid.org/0000-0001-5836-9300.

Пасько В. П. – канд. техн. наук, доцент кафедри інформаційних систем і технологій Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна. ROR: https://ror.org/00syn5v21. ORCID https://orcid.org/0000-0001-9011-7218.

Солдатова М. О. – канд. техн. наук, доцент кафедри інформаційних систем і технологій Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна. ROR: https://ror.org/00syn5v21. ORCID https://orcid.org/0000-0003-1233-1272.

Дроздович І. Г. – канд. техн. наук, старший науковий співробітник відділу природних ресурсів Інституту телекомунікацій та глобального інформаційного простору НАН України, Київ, Україна. ROR: https://ror.org/04qbxr381. ORCID https://orcid.org/0000-0002-4216-2417.

АНОТАЦІЯ

Актуальність. Проблема підтримки кутової орієнтації космічного апарата є критичною, особливо в умовах імпульсних зовнішніх збурень, що спричиняють різкі відхилення кутових швидкостей. Актуальність розв'язання цієї задачі визначається обмеженим запасом палива на борту, зокрема для класу космічних апаратів, призначених для забезпечення штучної гравітації, де точне та ефективне керування має першочергове значення.

Мета роботи. Основною метою цієї роботи є мінімізація споживання енергетичних ресурсів (палива) для стабілізації кутового положення певного класу космічних апаратів. Ця мета досягається шляхом послідовного виконання двох взаємопов'язаних завдань: 1) демпфування різких відхилень кутових швидкостей космічного апарата; 2) стабілізації кінцевого кутового положення.

Метод. Запропоновано двоетапний підхід. Для розв'язання першого завдання (демпфування) синтезується оптимальне керування з використанням комбінації принципу максимуму Понтрягіна та методу фазової площини. Це дозволяє побудувати оптимальні криві перемикання, які однозначно поділяють фазову площину на області з відповідними значеннями оптимальних керувань. Для розв'язання другого завдання (стабілізації) використовується модальний підхід на основі запропонованого методу невизначених коефіцієнтів, що забезпечує задані динамічні показники перехідних процесів стабілізації.

Результати. Було проведено моделювання динаміки кутового руху космічного апарата. Результати моделювання підтверджують високу ефективність використання запропонованого комбінованого підходу для розв'язання задачі стабілізації кутового положення космічного апарата після значних зовнішніх збурень.

Висновки. Спільне застосування принципу максимуму Понтрягіна та методу фазової площини для паливно-ефективного демпфування кутових швидкостей, з подальшою реалізацією оптимального закону стабілізації на основі запропонованого методу невизначених коефіцієнтів, є ефективною процедурою керування орієнтацією та стабілізацією космічного апарата з мінімальними витратами палива.

КЛЮЧОВІ СЛОВА: осесиметричний КА, принцип максимуму, фазова площина, оптимальні лінії перемикання та відключення, модальний синтез, метод невизначених коефіцієнтів.

DOI 10.15588/1607-3274-2026-1-18

ЛІТЕРАТУРА

1. Chernousko F. L. Evolution of motions of a rigid body about its center of mass / F. L. Chernousko, L. D. Akulenko, D. D. Leshchenko. – Cham : Springer International Publishing AG, 2017. – 242 p. DOI: 10.1007/978-3-319-53928-7
2. Markley F. L. Fundamentals of spacecraft attitude determination and control / F. L. Markley, J. L. Crassidis. – New York : Springer Science+Business Media, 2014. – 485 p. DOI: 10.1007/978-1-4939-0802-8
3. Sands T. Advances in spacecraft attitude control / T. Sands. – London : IntechOpen Limited, 2020. – 274 p. DOI: 10.5772/intechopen.77574
4. Athans M. Optimal control: an introduction to the theory and its applications / M. Athans, P. L. Falb. – Mineola : Courier Corporation, 2006. – 879 p.
5. Yang Y. Spacecraft attitude and reaction wheel desaturation combined control method / Y.Yang // IEEE Transactions on Aerospace and Electronic Systems. – 2017. – Vol. 53, № 1. – P. 286–295. DOI: 10.1109/TAES.2017.2650158
6. Kienitz K. H. Attitude stabilization with actuators subject to switching restrictions: an approach via exact relay control methods / K. H. Kienitz // IEEE Transactions on Aerospace and Electronic Systems. – 2006. – Vol. 42, № 4. – P. 1485–1492. DOI: 10.1109/TAES.2006.314588
7. El-Gohary A. Optimal control of a rigid spacecraft programmed motion without angular velocity measurements / A. El-Gohary // European Journal of Mechanics – A/Solids. – 2006. – Vol. 25, № 5. – P. 854–866. DOI: 10.1016/j.euromechsol.2005.11.009
8. Zabolotnov Yu. Approximate optimal method for controlling the angular motion of a spacecraft as part of an orbital tether system / Yu. Zabolotnov // Conference Series Materials Science and Engineering. – 2020. – Vol. 984. – Art. no. 012024. DOI: 10.1088/1757-899X/984/1/012024
9. Levskii M. V. Optimization problem of attitude control of a spacecraft with bounded rotary energy using quaternions / M. V. Levskii // International Robotics & Automation Journal. – 2021. – Vol. 7, № 2. – P. 63–73. DOI: 10.15406/iratj.2021.07.00228
10. Avdejev V. Reduced observer in stabilizing system of a rocket motion / V. Avdejev // Radio Electronics, Computer Science, Control. – 2020. – № 2. – P. 165–172. DOI: 10.15588/1607-3274-2020-2-17
11. Pukdeboon Ch. Anti-disturbance inverse optimal control for spacecraft position and attitude maneuvers with input saturation / Ch. Pukdeboon // Advances in Mechanical Engineering. – 2016. – Vol. 8, № 5. – P. 1–14. DOI: 10.1177/1687814016649887
12. Tsuchiya M. Semi-optimal control for minimum-time maneuver of satellite with variable speed pyramid type SGCMGs / M. Tsuchiya, T. Higuchi // Trans. JSASS Aerospace Tech. Japan. – 2021. – Vol. 19, № 1. – P. 24–33. DOI: 10.2322/tastj.19.24
13. Optimal rendezvous guidance using linear quadratic control / [G.-H. Moon, B.-Y. Lee, M.-J. Tahk, J.-H. Lee] // MATEC Web of Conferences. – 2016. – Vol. 54. – P. 09002. DOI: 10.1051/matecconf/20165409002
14. Guan T. Output feedback attitude control for rigid spacecraft under attitude constraints / T. Guan, B. Li // Journal of Industrial and Management Optimization. – 2023. – Vol. 19, № 7. – P. 5294–5305. DOI: 10.3934/jimo.2022173
15. Minimum switching thruster control for spacecraft precision pointing / [M. Leomanni, A. Garulli, A. Giannitrapani, A. Pugi] // IEEE Transactions on Aerospace and Electronic Systems. – 2017. – Vol. 53, № 2. – P. 683–697. DOI: 10.1109/TAES.2017.2655120
16. Spacecraft robust attitude tracking design: PID control approach / [L. L. Show, J. C. Juang, C. T. Lin, J. H. Jean] // Proceedings of the American Control Conference, Anchorage, 8–10 May 2002 : proceedings. – Piscataway : IEEE, 2002. – P. 836–841. DOI: 10.1109/ACC.2002.1023210
17. Левский М. В. Оптимальное управление программным разворотом космического аппарата / М. В. Левский // Космические исследования. – 2003. – Т. 41, № 2. – С. 195–210.
18. Оптимальне демпфірування відхилень кутових швидкостей вісесиметричного космічного апарата / [О. А. Стенін, В. П. Пасько, І. Г. Дроздович, О. О. Стенін] // Космічна наука і технологія. – 2021. – Т. 27, № 4. – С. 21–31. DOI: 10.15407/knit2021.04.021
19. Datta A. K. On optimization of a second-order non-linear control system for various performance criteria / A. K. Datta // International Journal of Control. – 1967. – Vol. 5, № 3. – P. 269–287. DOI: 10.1080/00207176708921760
20. Стенін О. А. Метод невизначених коефіцієнтів у задачах оптимальної стабілізації технологічних процесів / О. А. Стенін, І. Г. Дроздович, М. В. Солдатова // Радіоелектроніка, інформатика, управління. – 2020. – № 1(52). – С. 209–217. DOI: 10.15588/1607-3274-2020-1-21
21. Синтез законів керування космічним апаратом, що забезпечують оптимальне за швидкодією розміщення полюсів замкнутої системи керування / [Н. Є. Зубов, Є. А. Мікрін, М. Ш. Місріханов, В. Н. Рябченко] // Вісті РАН. Теорія та системи керування. – 2012. – № 3. – С. 98–111.
22. Заболотнов Ю. М. До задачі про оптимальну стабілізацію кутового руху малого космічного апарата під час розгортання орбітальної тросової системи / Ю. М. Заболотнов, А. А. Лобанков // Вісник Самарського державного аерокосмічного університету імені академіка С. П. Корольова (Національного дослідного університету). – 2016. – Т. 15, № 1. – С. 46–54. DOI: 10.18287/2412-7329-2016-15-1-46-54
23. Epperson J. F. An introduction to numerical methods and analysis / J. F. Epperson. – 2nd ed. – Hoboken, NJ : John Wiley & Sons, Inc., 2013. – 591 p.
24. Choi D.-S. A phase-plane approach to time-optimal control of single-DOF mechanical systems with friction / D.-S. Choi, S.-J. Kim, I.-J. Ha // Automatica. – 2003. – Vol. 39, № 8. – P. 1407–1415. DOI: 10.1016/S0005-1098(03)00112-2

DOI 10.15588/1607-3274-2026-1-18

Наукове видання

Радіоелектроніка,

інформатика,

управління

№ 1/2026

Науковий журнал

Головний редактор – д-р техн. наук С. О. Субботін

Заст. головного редактора – д-р техн. наук Т. А. Максимюк

Комп’ютерне моделювання та верстання — С. В. Зуб
Редактор англійських текстів — С. О. Субботін

Оригінал-макет підготовлено у редакційно-видавничому відділі НУ «Запорізька політехніка»

Реєстрація суб’єкта у сфері друкованих медіа:
Рішення Національної ради України з питань телебачення і радіомовлення № 3040 від 07.11.2024 року
Ідентифікатор медіа: R30-05582

Підписано до друку 25.02.2026. Формат 60×84/8.
Папір офс. Різогр. друк. Ум. друк. арк. 26,04.
Тираж 300 прим. Зам. № 116.

69063, м. Запоріжжя, НУ «Запорізька політехніка», друкарня, вул. Жуковського, 64

Свідоцтво суб’єкта видавничої справи
ДК № 6952 від 22.10.2019.