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

Радіоелектроніка Інформатика Управління

Radio Electronics Computer Science Control

2025/3

Міністерство освіти і науки України
Національний університет «Запорізька політехніка»

Радіоелектроніка, інформатика, управління

Науковий журнал
Виходить чотири рази на рік
№ 3(74) 2025
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»
2025

Ministry of Education and Science of Ukraine
National University Zaporizhzhia Polytechnic

Radio Electronics, Computer Science, Control

The scientific journal
Published four times per year
№ 3(74) 2025
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
2025

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 № 1 від 28.08.2025.

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 № 1 dated 28.08.2025.

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 621.396.962

METHOD FOR STUDYING THE TIME-SHIFTED MATHEMATICAL MODEL OF A TWO-FRAGMENT SIGNAL WITH NONLINEAR FREQUENCY MODULATION

Kostyria O. O. – Dr. Sc., Senior Research, Leading Research Scientist, Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine.

Hryzo A. A. – PhD, Associate Professor, Head of the Research Laboratory, Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine.

Trofymov I. M. – PhD, Senior Researcher, Professor of Department, Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine.

Liashenko O. I. – Researcher, Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine.

Biernik Ye. V. – Graduate student, Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine.

ABSTRACT

Context. The further development of the theory and techniques for forming and processing complex radar signals encompasses both the study of existing mathematical models of probing radio signals and the creation of new ones. One of the directions of such research focuses on reducing the maximum side lobe level in the autocorrelation functions of signals with intra-pulse modulation of frequency or phase. In this context, the instantaneous frequency may vary according to either a linear or nonlinear law. Nonlinear frequency modulation laws can reduce the maximum level of side lobes without introducing amplitude modulation in the output signal of the radio transmitting device and, consequently, without causing power loss in the sensing signals. The widespread implementation of nonlinear-frequency-modulated signals in radar technology is constrained by the insufficient development of their mathematical models. Therefore, the development of methods for analyzing existing mathematical models of signals with nonlinear frequency modulation remains an urgent scientific task.

Objective. The purpose of this work is to develop a method for conducting research to evaluate the advantages and disadvantages of a mathematical model of a nonlinear-frequency-modulated signal consisting of two fragments with linear frequency modulation.

Method. This study proposes a method for analyzing mathematical models of signals based on the transition from a shifted time scale to the current time scale. The methodology consists of the following main stages: a formalized description of mathematical models, transition to an alternative time scale, identification of components and determination of their physical essence, and a comparative analysis. The proposed method was validated through simulation modeling.

Results. Using the proposed method, it has been determined that the mathematical operation of time scale shifting is equivalent to the introduction of additional components in the mathematical model. These components simultaneously and automatically compensate for the frequency jump at the junction of fragments, as well as introduce an additional linear phase increment in the second linearly frequency-modulated fragment. This approach provides a clear illustration of the frequency jump compensation mechanism in the studied mathematical model. The applied method enabled the identification of a drawback in the examined mathematical model, namely, the absence of a compensatory component for the instantaneous phase jump during the transition from the first LFM fragment to the second.

Conclusions. A method has been developed to determine the essence and corresponding influence of the components of a mathematical model in a time-shifted, nonlinear, frequency-modulated signal, which consists of two fragments with linear frequency modulation. The model under study is not entirely accurate, as it lacks a component to compensate for the phase jump at the transition from the first signal fragment to the second. The introduction of such a component ensures a further reduction in the maximum level of the side lobes of the signal autocorrelation function.

KEYWORDS: nonlinear-frequency-modulated signals, mathematical model, instantaneous phase jump, autocorrelation function, maximum level of side lobes.

ABBREVIATIONS

ACF is an autocorrelation function;
AFS is an amplitude-frequency spectrum;
ESS is an effective scattering surface;
FM is a frequency modulation;
FMR isa frequency modulation rate;
IPM is an intra-pulse modulation;
LFM is a linear frequency modulation;
MF is a matched filtering;
ML is a main lobe;
NLFM is a non-linear frequency modulation;
MM is a mathematical model;

DOI 10.15588/1607-3274-2025-3-1

MPSLL is a maximum peak side lobe level;
PSLL is a peak side lobe level;
REQ is a radio electronic equipment;
SAR is a synthetic aperture radar;
SL is a side lobe;
WF is a window function.

NOMENCLATURE

f_0 is an initial signal frequency, Hz;

$f(t)$ is an instantaneous frequency of the NLFM signal, Hz;

$f_2(t)$ is an instantaneous frequency of the second fragment of the NLFM signal, Hz;

Δf_1 is a frequency deviation of the first fragments of the NLFM signal, Hz;

Δf_2 is a frequency deviation of the second fragments of the NLFM signal, Hz;

C_2 is a constant of integration;

t is a current time, s;

T_1 is a duration of the first fragments of the NLFM signal, s;

T_2 is a duration of the second fragments of the NLFM signal, s;

$\varphi(t)$ is a instantaneous phase of the NLFM signal, rad;

$\delta\varphi_{12}$ is a compensation component for the instantaneous phase jump at the junction of the first and second LFM fragments, rad;

$\varphi_2(t)$ is an instantaneous phase of the second LFM fragment of the signal, rad;

β_1 is a FMR of the first fragments of the NLFM signal, Hz/s;

β_2 is a FMR of the second fragments of the NLFM signal, Hz/s.

INTRODUCTION

The continuous development of electronic technologies, along with the near-total abandonment of electrovacuum devices in favor of solid-state components, ensures multifunctionality and reduces the weight and dimensions of REQ. Another important direction of expanding the capabilities of REQ is the use of signals with IPM of phase and frequency, commonly referred to as complex signals. Historically, the first systems to incorporate linear frequency-modulated and code-phase manipulated signals emerged [1–10] and are still widely used today. These signals and their MF systems continue to evolve, undergo modifications, and improve [11–22, 24–44].

The implementation of signals with IPM into radio engineering and telecommunications has provided developers of radar systems, radionavigation, and communication technologies with additional opportunities to significantly enhance system performance. This includes extending the operational range of REQ, provided that the peak power of their radio transmitting devices is limited, improving electromagnetic compatibility, enhancing noise immunity and operational security, and increasing the bandwidth of information transmission channels [1–10].

However, despite the numerous advantages of using signals with IPM, a significant drawback occurs – the high MPSLL of the ACF compared to simple signals, which imposes limitations on the achievable dynamic range of the radio receiving device. In a multi-target aerial environment, the reflected signal from a target with a lower ESS may be masked by the side lobes of the signal from a target or a passive obstacle with a higher ESS. Research aimed at reducing MPSLL focuses on both improving MF radio receiving devices and implementing new types of signals with IPM [1–4, 11–44].

Thus, in studies [11–16], methods and devices have been proposed to improve the processing of received echo signals.

In [17–44], several types of NLFM signals have been proposed for use in radio transmitting devices to reduce MPSLL.

A large number of studies have been dedicated to the problem of reducing MPSLL in SAR systems for various purposes [15, 18–22].

Research on the issue of reducing MPSLL by implementing NLFM signals whose fragments have IPM different from LFM has been conducted in studies [23–38].

A number of works are aimed at improving the existing MM of NLFM signals consisting of two or three LFM fragments [39–44].

The authors [45–47] conducted a study analyzing NLFM signal types and estimating their time-frequency parameters in the context of solving problems related to electronic intelligence and REQ suppression.

It should be noted that in the reviewed works, different time scales are used for the formal description of the mathematical model (MM) of pulsed NLFM signals [1–4, 11–44]. In particular, a symmetric time scale relative to zero results in a radio pulse consisting of two opposite fragments. Additionally, a continuous time scale allows for the sequential determination of amplitude values of signal samples in real-time, referred to as the current-time MM. Another approach is to use a mathematical technique in which the mathematical description of both the first and each subsequent fragment of the NLFM signal starts from the zero-time value, thus implementing the time-shifted MM.

In study [40], it is noted that the time-shifted MM of two- and three-fragment NLFM signals has a useful feature, namely, it provides automatic compensation for the frequency jump at the junction of LFM fragments, which occurs at the moment of a change in the FMR value.

A detailed analysis of the compensation mechanism and the study of the peculiarities of MM operation with a time shift, due to the complexity of the mathematical description, have not been conducted in known studies, so it is advisable to conduct such an analysis.

To conduct the study, this paper proposes a method based on the mathematical transformation of a time-shifted MM into an equivalent current-time MM, i.e., the representation of models on a unified time scale, followed

DOI 10.15588/1607-3274-2025-3-1

by a detailed analysis of their fine structure and properties.

Using the example of the MM of a two-fragment NLFM signal, the validity of this approach is demonstrated, and the equivalence of the current-time and time-shifted MM is substantiated. Based on the results of the conducted analysis, previously unaccounted distortion components of the NLFM signal are identified, and a method for their compensation is proposed.

The object of study is the process of synthesizing NLFM signals using a time-shifted MM based on two LFM fragments with different FMR values.

The subject of study is the MM of a time-shifted NLFM signal consisting of two LFM fragments.

The purpose of the work is to identify and analyze the components of the time-shifted MM of a two-fragment NLFM signal by converting it into a current-time model.

1 PROBLEM STATEMENT

For further analysis, we will write the expressions for the instantaneous frequency and phase of the two-fragment NLFM signal in the current time.

$$f(t)=\begin{cases} f_0 \pm \beta_1 t,\ 0 \le t \le T_1; \\ f_0 \pm \beta_1 T_1 \pm \beta_2 t,\ T_1 < t \le T_1+T_2; \end{cases} \quad (1)$$

$$\varphi(t)=2\pi\begin{cases} f_0 t \pm \beta_1 \dfrac{t^2}{2},\ 0 \le t \le T_1; \\ (f_0 \pm \Delta f_1)t \pm \beta_2 \dfrac{t^2}{2},\ T_1 < t \le T_1+T_2, \end{cases} \quad (2)$$

where $\beta_1=\dfrac{\Delta f_1}{T_1};\ \beta_2=\dfrac{\Delta f_2}{T_2}$.

The '+' or '–' sign in (1), (2), and the subsequent expressions for the uncertainty ' ' is chosen depending on whether the frequency of the LFM fragments increases or decreases with time.

The descriptions (1) and (2) are obtained by removing the third fragment of the MM of the NLFM signal introduced in [2].

For describing time-shifted NLFM signals, we will use the MM introduced by the authors in [40, 42–44]. The expression for the instantaneous frequency is as follows:

$$f(t)=\begin{cases} f_0 \pm \beta_1 t, 0 \le t \le T_1; \\ f_0 \pm \beta_1 T_1 \pm \beta_2 (t-T_1), T_1 < t \le T_2, \end{cases} \quad (3)$$

for the description of the instantaneous phase:

$$\varphi(t)=2\pi\begin{cases} f_0 t \pm \dfrac{\beta_1}{2} t^2, 0 \le t \le T_1; \\ (f_0 \pm \Delta f_1)(t-T_1) \pm \beta_2 \left(\dfrac{t^2}{2} - T_1 t \right), \\ T_1 < t \le T_2. \end{cases} \quad (4)$$

The defining difference between (3), (4) and (1), (2) is the use of a different time scale. The initial count of the frequency and phase of the second LFM fragment in (3), (4) is shifted to the zero mark. During the studies conducted earlier [40], it was determined that the MMs (3), (4) provide automatic compensation for the instantaneous frequency jump at the junction of the fragments, but the mechanism of such compensation has not been studied in the known works.

2 REVIEW OF THE LITERATURE

The fundamental principles of the construction and functioning of REQ for various purposes, including the use of signals with IPM, are discussed in papers [1–10]. It is noted that in devices using MF based on the compression of complex signals, various methods for reducing the MPSLL of the ACF are employed.

The time WF method is most widely used in the radio receiving device, which rounds the radio pulse envelope, resulting in a reduction of the MPSLL at the MF output [12, 16, 24].

The authors [1, 2] proposed emitting radio signals with a rounded AFS, which is equivalent to its WF in the time domain.

It is possible to further improve the results of the window function (WF) and achieve lower values of the MPSLL. To this end, signals with polynomial FM [1, 14, 15] and NLFM signals [2, 17–22] are proposed. In papers [1, 2], the achievable MPSLL for the proposed signals was determined by calculation to be –42.8 dB. However, in study [2], it is noted that for low-base signals – those for which the product of signal duration and spectrum width is less than 100 – a MPSLL value of –30.0 dB is considered a significant achievement. The authors [37], by improving the MM of the form [1] and using a genetic algorithm to optimize the time-frequency parameters, achieved an MPSLL for the low-base signal at –34.9 dB.

The distinctive feature of the NLFM signal proposed in [2] is that it consists of three LFM fragments, with the FMR changing as it transitions to a new fragment. Further improvement of the MM of the form [2] was conducted by the authors [34–44]. The main difference in the MMs they propose is the use of a time-shifted scale and fragments with both linear and nonlinear frequency modulation laws.

The authors [34–37, 40–42] have developed MM for two- and three-fragment NLFM signals, in which compensation for frequency-phase distortions at the junctions of the fragments is implemented at the moments when the FMR changes. The fragments of such signals follow both linear and nonlinear frequency modulation laws. These NLFM signals, which simultaneously include fragments with both linear and nonlinear FM, have been proposed to be classified as combined signals [34–36].

It is shown in [43, 44] that the feature of the time-shifted MM is the automatic compensation for the frequency jump at the junction of fragments, which allows the synthesis of signals while adhering to the specified time-frequency parameters without the introduction of

DOI 10.15588/1607-3274-2025-3-1

any additional compensation components. However, the compensation mechanism itself has not been studied.

3 MATERIALS AND METHODS

For further research, we will use the proposed method, which involves replacing the time scale for describing the second fragment (3), i.e., transitioning from shifted time to current time. By expanding the brackets and combining like terms, we obtain:

$$f_2(t) = f_0 \pm \beta_2 t \mp (\beta_2 - \beta_1)T_1, T_1 < t \le T_2. \quad (5)$$

The analysis of (5) indicates that, compared to (1), the expression for the second fragment has undergone a transformation of the constant component of the frequency change from $\beta_1 T_1$ to the component $(\beta_2 - \beta_1)T_1$, which is accounted with the opposite sign relative to the main one and is interpreted as compensatory.

Let us compare (5) with the current-time MM of the two-fragment NLFM signal with compensation for instantaneous phase and frequency jumps at the junction of fragments [41]:

$$f(t) = \begin{cases} f_0 \pm \beta_1 t, \ 0 \le t \le T_1; \\ f_0 \pm \beta_2 t \mp (\beta_2 - \beta_1)T_1, \ T_1 < t \le T_1 + T_2; \end{cases} \quad (6)$$

$$\varphi(t) = 2\pi \begin{cases} f_0 t + \dfrac{\beta_1 t^2}{2}, \ 0 \le t \le T_1; \\ \left[f_0 + (\beta_2 - \beta_1)T_1\right]t + \dfrac{\beta_2 t^2}{2} + \dfrac{(\beta_2 - \beta_1)T_1^2}{2}, \\ T_1 \le t \le T_1 + T_2. \end{cases} \quad (7)$$

The change in instantaneous frequency is defined by (6), from the second expression of which we can observe that the compensation for the instantaneous frequency jump at the junction of the signal fragments occurs due to the component $(\beta_2 - \beta_1)T_1$.

From this, an important conclusion can be drawn: the time-shifting operation in the second expression of the system of equations (3) is equivalent to the introduction of a compensatory component for the frequency jump, which is what needed to be proven.

We now turn to the current time scale in the description of the instantaneous phase of the second LFM fragment of the MM (4). Through simple transformations, we obtain:

$$\varphi_2(t) = 2\pi \begin{cases} \left[f_0 \mp (\beta_2 + \beta_1)T_1\right]t \pm \dfrac{\beta_2 t^2}{2} \mp \Delta f_1 T_1, \\ T_1 < t \le T_1 + T_2. \end{cases} \quad (8)$$

The analysis of (8) shows that the component $(\beta_2 + \beta_1)T_1$ in square brackets compensates for the linear incursion of the instantaneous phase caused by the compensatory component of the frequency jump, while the component $\Delta f_1 T_1$ outside the square brackets compensates for the total phase incursion of the first LFM fragment. That is, as intended by the authors of MM (3), (4), the use of a shifted time scale should ensure a zero initial count for both the frequency and phase of the signal for the second LFM signal fragment. However, in this case, a jump in the instantaneous phase occurs due to the instantaneous frequency jump $T_1(\beta_2 + \beta_1)$ at the junction of the fragments, requiring additional compensation. The development of such an MM was carried out in study [40], resulting in the following:

$$\varphi(t) = 2\pi \begin{cases} f_0 t \pm \dfrac{\beta_1 t^2}{2}, 0 \le t \le T_1; \\ (f_0 \pm \Delta f_1)(t - T_1) \pm \beta_2 \left(\dfrac{t^2}{2} - T_1 t\right) \mp \delta\varphi_{12}, \\ T_1 < t \le T_1 + T_2 \ , \end{cases} \quad (9)$$

where $\delta\varphi_{12} = \dfrac{1}{2}T_1^2(\beta_2 + \beta_1)$. (10)

It should be noted that the time scale shift operation in (10) led to a sign change from '–' to '+' in the parentheses compared to (7).

To derive (9), the graph-analytical method was applied in [40]; let us attempt to obtain this MM purely analytically. As a result of integrating the second equation (3), we obtain:

$$\varphi_2(t) = \begin{cases} 2\pi \int\limits_t f_2(t - T_1)dt = (f_0 \pm \Delta f_1)(t - T_1) \pm \beta_2 \left(\dfrac{t^2}{2} - T_1 t\right) + C_2, \\ T_1 < t \le T_2. \end{cases} \quad (11)$$

The integration constant C_2 is determined considering the initial conditions:

$$C_2 = \varphi_2(t)\Big|_{t=T_1} = \mp\beta_2 \frac{T_1^2}{2}, \quad (12)$$

has the physical meaning of a compensatory component concerning the phase jump at the junction of the fragments caused by the instantaneous frequency jump.

Comparison of (10) and (12) shows that they do not correspond, and therefore we conclude that expression (11) is not entirely correct. In order to compensate for the phase jump at the junction of fragments caused by a frequency jump for a time-shifted MM, it is necessary to integrate the frequency jump component $(\beta_2 + \beta_1)T_1$, which is used to calculate the linear phase incursion in (8). This is quite logical since the time-shift operation ensures a zero initial phase value for the second LFM fragment, while the final phase of the first LFM fragment is not zero.

Thus, it has been established that the application of the time scale shift to a zero point for the second LFM fragment in MM (3) is equivalent to the introduction of a

DOI 10.15588/1607-3274-2025-3-1

compensatory component for the instantaneous frequency jump at the junction of the fragments. The obtained result can be similarly extended to NLFM signals with a greater number of fragments.

As a result of the conducted study, it has been established that MM (4), used by the authors in [43, 44], does not provide complete compensation for the phase distortions of the resulting NLFM signal that occur when transitioning to a new fragment due to the instantaneous change in the value of the FMR. Only the additional linear phase incursion caused by the frequency jump at the moment of this transition is compensated. Therefore, obtaining the resulting signal with a reduced MPSLL is not a general rule and is only achieved for specific sets of frequency-time parameters of the LFM fragments, which are typically determined through selection.

We will verify the obtained theoretical results through simulation modeling. It should be noted that MM (7) and (9) are equivalent and provide identical results, so it is sufficient to use one of them for the simulation.

4 EXPERIMENTS

Mathematical modeling was carried out using the Matlab software package. The studies were conducted through a comparative analysis of the results obtained using MM (4) and (9) for NLFM signals with identical frequency-time parameters, namely: $\Delta f_1 = \Delta f_2 = 200$ kHz, $T_1 = 40$ μs, $T_2 = 100$ μs. A classical LFM signal with a duration of 140 μs and a frequency deviation of 400 kHz was also modeled.

The ACF parameters obtained as a result of modeling the specified signals are summarized in Table 1.

Table 1 – Values of ACF Parameters for the Signals

The name of the ACF parameter	LFM	NLFM (4)	NLFM (9)
Width of the ML of the ACF, μs	2.21	2.37	2.43
MPSLL, dB	–13.5	–14.59	–17.14
Rate of SL decay, dB/decade	19.35	19.8	21.25

The result of using the time-shifted MM without compensation for instantaneous phase jumps (4) is shown in Fig. 1. Figure 1a demonstrates the change in instantaneous frequency without a jump at the junction of the LFM fragments (at the current time of 40 μs). To simplify the analysis of the oscillogram of the resulting signal (Fig. 1b), we take $f_0 = 0$. In the oscillogram, at the moment of transition from the first fragment of the NLFM signal to the second at 40 μs, an instantaneous phase jump is observed, with a magnitude exceeding 180^0. The presence of a significant phase jump causes a dip in the signal spectrum between the LFM components at a frequency of 200 kHz (Fig. 1c), which also results in distortion of its peak and the presence of ripples on the side slopes and "wings" of the spectrum.

The analysis of the ACF of the signal (Fig. 1d) indicates the merging of the ML with the nearby SLs, which led to the expansion of the ML and, consequently, a deterioration in range resolution. Due to instantaneous phase distortion at the junction of the fragments, the resulting MPSLL is higher than expected, and the SL ripples exhibit level fluctuations and an irregular pattern of changes. The MPSLL of the ACF is –14.59 dB, and the ML width at the 0.707 level is 2.37 μs. Compared to the classical LFM signal, the signal of type (4) demonstrated an 8% reduction in the MPSLL of the ACF, a 7% increase in the ML width at the 0.707 level of the maximum, and a slight increase (approximately 2%) in the rate of decline in the SL of the ACF.

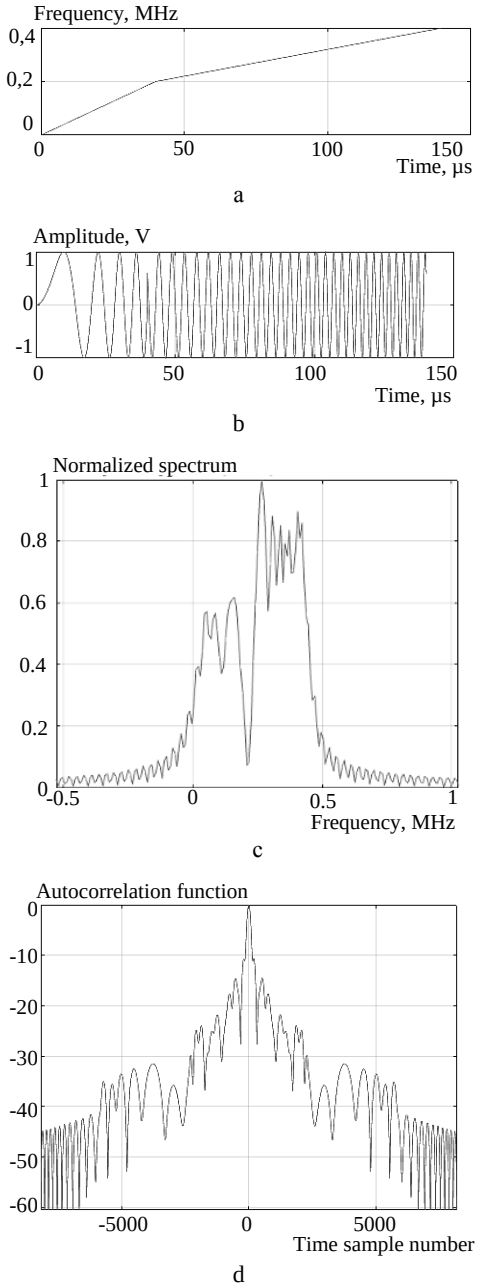


Figure 1 – Oscillogram (a), instantaneous frequency change graph (b), spectrum (c), ACF (d) of the NLFM signal according to the model (4)

DOI 10.15588/1607-3274-2025-3-1

The simulation results corresponding to (9) are presented in Fig. 2. Figure 2a shows that the frequency of the signals (4) and (9) changes within the same range; therefore, their frequency variation graphs are in full compliance. The frequency jump at the junction of the fragments is compensated. The oscillogram in Fig. 2b, in contrast to Fig. 1b, is smooth, i.e., the compensatory phase component was calculated correctly, ensuring this result. Accordingly, the spectrum of the resulting NLFM signal acquired the expected shape. Due to the higher FMR of the first LFM fragment, its power spectral density is lower, as seen in the spectrograms of Figs. 1c and 2c. However, Fig. 2c shows no dips, peak distortion, or ripples on the slopes and 'wings' of the spectrum.

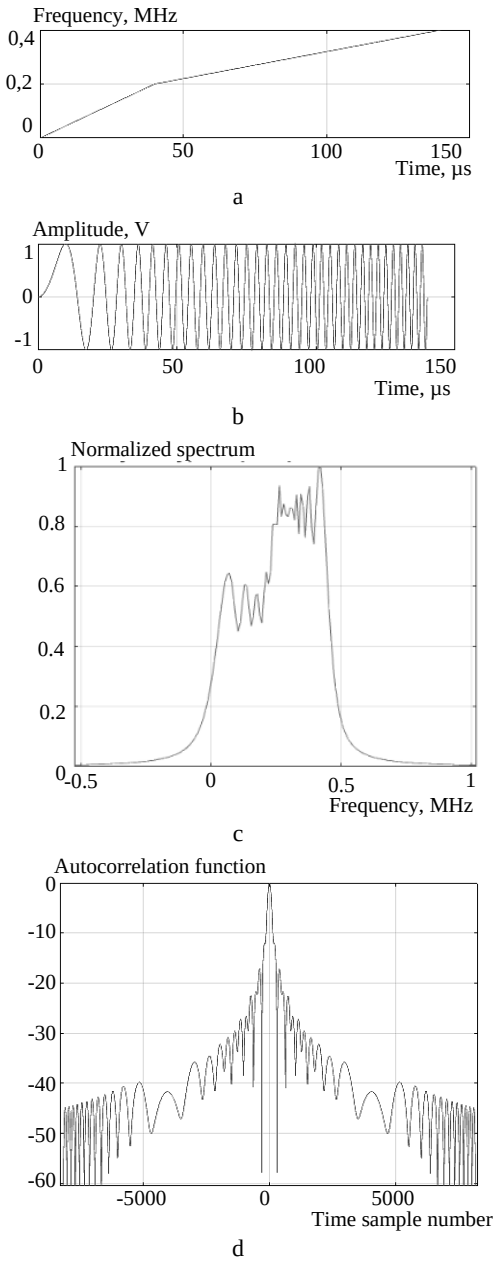


Figure 2 – Oscillogram (a), instantaneous frequency change graph (b), spectrum (c), ACF (d) of the NLFM signal according to the model (9)

The absence of a phase jump at the moment of transition to the second fragment of the signal ensured an improvement in the shape and a decrease in the MPSLL of the ACF, which is –17.14 dB. However, the ML width at the 0.707 level from the maximum increased to 2.43 μs. Compared to the LFM signal, the resulting NLFM signal of type (9) provided a 27% reduction in the MPSLL of the ACF, the ML width at the 0.707 level of the maximum increased by 10%, and the rate of decline in the SL of the ACF increased by 10%.

5 RESULTS

As a result of mathematical modeling, it was established that for the given identical frequency and time parameters of the two-fragment NLFM and classical LFM signals, the use of the studied MM for the NLFM signal (4) compared to the LFM signal resulted in a decrease in the MPSLL by 1.09 dB, a slight increase in the decay rate of the PSLL by 0.45 dB/decade, and a slight expansion of the ML, which corresponds to an increase in the resolving power for a range of 24 m. After transitioning to a continuous time scale, in accordance with the proposed method, a phase jump at the junction of fragments was identified, which had not been taken into account previously. The further introduction of a compensatory component for the phase jump improved the well-known MM (9) and reduced the MPSLL compared to the LFM signal by 3.64 dB. In relation to MM (4), the reduction was more than three times greater. The corresponding PSLL decay rate increased by 1.9 dB/decade, i.e., it increased fourfold compared to MM (4), while the resolving power for a range, compared to the LFM signal, increased by 33 m, which is approximately 1.4 times higher than MM (4).

6 DISCUSSION

The study of MM (3), (4) conducted using the proposed method enabled a comparison with MM (1), (2) and analogues [2, 43, 44].

The analysis of MM (3), (4), carried out using the current time scale, allowed for the formalization of the analytical expressions of the compensatory components:

– the frequency jump at the junction of fragments in the second expression of (3);

– the linear increment of the instantaneous phase caused by the frequency jump at the junction of fragments in the second expression of (4).

The applied approach provides a clear illustration of the frequency jump compensation mechanism in MM (3), (4).

The applied method enabled the identification of a drawback in the examined MM, namely, the absence of a compensatory component for the instantaneous phase jump at the moment of transition from the first LFM fragment to the second (second expression in (4)).

This can be explained by the fact that when the time scale is shifted to zero, the initial phase of the second LFM fragment becomes zero, while the final phase of the first LFM fragment can take any value within the interval from 0 to 2π, which explains the mechanism of the phase

DOI 10.15588/1607-3274-2025-3-1

jump occurrence. This is confirmed by the fact that direct integration of the second expression in model (3) does not allow for obtaining the complete set of compensatory components needed to calculate the instantaneous phase values of the second LFM fragment of the signal.

In addition to directly introducing a compensatory component for the phase jump, it is possible to avoid such a jump by using the method of providing an integer value for the number of periods of the LFM oscillation in the first fragment of the NLFM signal, as proposed in [42].

CONCLUSIONS

The study is based on the use of the proposed method for transitioning to a continuous current time scale, followed by a comparative analysis of the MM of NLFM signals, which consist of two LFM fragments. The study was made possible through the use of results obtained in [40, 41].

This approach is advisable for comparing different MMs of NLFM signals with equivalent time-frequency parameters. The feasibility of the method is confirmed through mathematical calculation and verified by simulation modeling.

The scientific novelty. A method for studying the time-shifted MMs of NLFM signals is proposed, which involves transitioning to the current time scale followed by a detailed analysis of its structure and properties.

For the first time, the mechanism of automatic compensation for the instantaneous frequency jump at the junction of fragments is explicitly highlighted. It is shown that performing the time scale shift operation for the second LFM fragment in (3) and (4) is equivalent to the emergence of compensatory components for the frequency jump and the linear increment of the instantaneous phase. This ensures the absence of a gap in the spectrum of the resulting signal and the specified frequency deviation.

The use of the proposed method allowed for determining the advantages and disadvantages of the time-shifted MM (3), (4) relative to the MM (1), (2). It is shown that the studied time-shifted MM (4) does not function correctly in the process of determining the instantaneous phase increments of the second LFM fragment, and the reason for this is the failure to account for the instantaneous phase jump at the junction of the fragments.

The practical significance of the obtained results lies in providing the scientific community with a new mechanism for studying time-shifted MM. This mechanism involves applying a mathematical technique to transform such a model into a current-time MM, which facilitates the simplification and detailed analysis of the model's components.

Prospects for further research involve applying the proposed approach to the study of time-shifted MM for NLFM signals with a greater number of LFM fragments and with FM laws different from linear in one or both fragments.

ACKNOWLEDGEMENTS

We thank the management of Ivan Kozhedub Kharkiv National Air Force University for the opportunity to conduct scientific research.

REFERENCES

1. Skolnik M. Radar Handbook. Editor in Chief. Boston, McGraw-Hill Professional, second edition, 1990, 846 p.
2. Cook C., Bernfeld M. Radar Signals: An Introduction to Theory and Application. Boston, Artech House, 1993, 552 p.
3. Van Trees H. L. Detection, Estimation, and Modulation Theory, Part III: Radar-Sonar Processing and Gaussian Signals in Noise. John Wiley & Sons, Inc., 2001, 643 p. DOI: 10.1002/0471221090
4. Levanon N., Mozeson E. Radar Signals. New York, John Wiley & Sons, 2004, 403 p.
5. Barton D. K. Radar System Analysis and Modeling. Boston, London, Artech House Publishers, 2004, 566 p.
6. He You, Jianjuan Xiu, Xin Guan Radar Data Processing with Applications. Publishing House of Electronics Industry, 2016, 536 p. DOI: 10.1002/9781118956878
7. Melvin W. L., Scheer J. A. Principles of modern radar. Vol. II: Advanced techniques, Sci Tech Publishing, 2013, 846 p.
8. Richards M. A., Scheer J., Holm W. A. Principles of Modern Radar: Basic Principles, SciTech Pub., 2010, 924 p.
9. Blackman S. S., Popoli R. F. Design and Analysis of Modern Tracking Systems. Boston, London, Artech House, 1999, 1230 p.
10. McDonough R. N., Whalen A. D. Detection of Signals in Noise (second ed.). San Diego, Academic Press, Inc., USA, 1995, 495 p.
11. Argenti F., Facheris L. Radar pulse compression methods based on nonlinear and quadratic optimization, *IEEE transactions on geoscience and remote sensing*, 2020, Vol. 59, Issue 5, pp. 3904–3916. DOI:10.1109/TGRS.2020.3010414
12. Muralidhara N., Velayudhan V., Kumar M. Performance Analysis of Weighing Functions for Radar Target Detection, *International Journal of Engineering Research & Technology (IJERT)*, 2022, Vol. 11, Issue 03, pp. 161–165.
13. Azouz A. , Abosekeen A., Nassar S. et al. Design and Implementation of an Enhanced Matched Filter for Sidelobe Reduction of Pulsed Linear Frequency Modulation Radar, *Sensors*, 2021, Vol. 21(11), Article № 3835. DOI:10.3390/s21113835
14. Xie Q., Zeng H., Mo Z. et al. Two-Step Optimization Framework for Low Sidelobe NLFM Waveform Using Fourier Series, *IEEE Geoscience and Remote Sensing Letters.* – 2022, Vol. 19, Article № 4020905. DOI:10.1109/LGRS.2022.3141081
15. Ghavamirad J. R., Sadeghzadeh R. A., Sebt M. A. Sidelobe Level Reduction in the ACF of NLFM Signals Using the Smoothing Spline Method, *Electrical Engineering and Systems Science, Signal Processing: arXiv:2501.06657 [eess. SP]*, 2025, 5 p. DOI:10.48550/arXiv.2501.06657
16. Wei W., Wang Y., Wang Z. et al. Novel Nonlinear Frequency Modulation Waveform With Low Sidelobes Applied to Synthetic Aperture Radar, *IEEE Geoscience and Remote Sensing Letters*, 2022, Vol. 19, Article № 4515405, pp. 1–5. DOI: 10.1109/LGRS.2022.3216340
17. Zhang Y., Deng Y., Zhang Z. et al. Parametric NLFM Waveform for Spaceborne Synthetic Aperture Radar, *IEEE Transactions on Geoscience and Remote Sensing*, 2022,

DOI 10.15588/1607-3274-2025-3-1

Vol. 60, Art no. 5238909, pp. 1–9. DOI: 10.1109/TGRS.2022.3221433
18. Xu W., Zhang L., Fang C. et al. Staring Spotlight SAR with Nonlinear Frequency Modulation Signal and Azimuth Non-Uniform Sampling for Low Sidelobe Imaging, *Sensors*, 2021, Vol. 21, Article № 6487. DOI:10.3390/s21196487
19. Jiang T., Li B., Li H. et al. Design and implementation of spaceborne NLFM radar signal generator, *Second IYSF Academic Symposium on Artificial Intelligence and Computer Engineering*, 2021, Vol. 12079, Article № 120792S, 6 p. DOI: 10.1117/12.2623222
20. Zhang Y., Deng Y., Zhang Z. et al. Analytic NLFM Waveform Design with Harmonic Decomposition for Synthetic Aperture Radar, *IEEE Geoscience and Remote Sensing Letters*, 2022, Vol. 4, Art no. 4513405. DOI:10.1109/lgrs.2022.3204351
21. Zhaoa Yu., Ritchieb M., Lua X. et al. Non-Continuous Piecewise Nonlinear Frequency Modulation Pulse with Variable Sub-Pulse Duration in a MIMO SAR Radar System, *Remote Sensing Letters*, 2020, Vol. 11, Issue 3, pp. 283–292. DOI: 10.1080/2150704X.2019.1711237
22. Anoosha C., Krishna B. T. Comparison on Radar Echo Cancellation Techniques for SAR Jamming, *Lecture Notes in Electrical Engineering Microelectronics, Electromagnetics and Telecommunications*. Springer, Singapore, 2020, pp. 651–658. DOI:10.1007/978-981-15-3828-5_67
23. Zhang, Y. Wang W., Wang R. et al.] A novel NLFM waveform with low sidelobes based on modified Chebyshev window, *IEEE Geosci. Remote Sens. Lett*., 2020, Vol. 17, Issue 5, pp. 814–818. DOI: 10.1109/LGRS.2019.2930817
24. Roy A., Nemade H. B., Bhattacharjee R. Radar waveform diversity using nonlinear chirp with improved sidelobe level performance, *AEU – International Journal of Electronics and Communications*, 2021, Vol. 136(11), Article № 153768. DOI:10.1016/J.AEUE.2021.153768
25. Saleh M., Omar S.-M., Grivel E. et al. A Variable Chirp Rate Stepped Frequency Linear Frequency Modulation Waveform Designed to Approximate Wideband Non-Linear Radar Waveforms, *Digital Signal Processing*, 2021, Vol. 109, №102884, 19 p. DOI:10.1016/j.dsp.2020.102884
26. Van-Zyl A. C., Wiehahn E. A., Cillers J. E. et al. Optimized Multi-Parameter NLFM Pulse Compression Waveform for low Time-Bandwidth Radar, *International Conference on Radar Systems (RADAR 2022)*, 2022, pp. 289–294. DOI: 10.1049/icp.2022.2332
27. Li J., Wang P., Zhang H. et al. A Novel Chaotic-NLFM Signal under Low Oversampling Factors for Deception Jamming Suppression, *Remote Sens*, 2024, № 1, P. 35. DOI: 10.3390/rs16010035
28. Fan Z., Meng H. Coded Excitation with Nonlinear Frequency Modulation Carrier in Ultrasound Imaging System, *2020 IEEE Far East NDT New Technology & Application Forum (FENDT)*. Kunming, Yunnan province, China: conference paper, IEEE, 2020. pp. 31–35. DOI: 10.1109/FENDT50467.2020.9337517
29. Xu Z. Wang X., Wang Y. Nonlinear Frequency-Modulated Waveforms Modeling and Optimization for Radar Applications, *Mathematics*, 2022, Vol. 10, P. 3939. DOI:10.3390/math10213939
30. Singh A.K. Bae K.-B. , Park S.-O. NLFM Pulse Radar for Drone Detection using Predistortion Technique, *Journal of Electromagnetic Waves and Applications*, 2021, Vol. 35, pp. 416–429. DOI:10.1080/09205071.2020.1844598
31. Ping T., Song C., Qi Z. et al. PHS: A Pulse Sequence Method Based on Hyperbolic Frequency Modulation for Speed Measurement, *International Journal of Distributed Sensor Networks*, 2024, Vol. 2024, Article № 6670576, 11 p. DOI: 10.1155/2024/6670576
32. Shuyi L., Jia Y., Liu Y. et al. Research on Ultra-Wideband NLFM Waveform Synthesis and Grating Lobe Suppression, *Sensors*, 2022, № 24, Article № 9829. DOI:10.3390/s22249829
33. Zhuang R., Fan H., Sun Y. et al. Pulse-Agile Waveform Design for Nonlinear FM Pulses Based on Spectrum Modulation, *IET International Radar Conference*, 2021, pp. 964–969. DOI: 10.1049/icp.2021.0700.
34. Kostyria O. O., Hryzo A. A., Dodukh O. M. Combined two-fragment radar signals with linear and exponential frequency modulation laws, *Systems of Arms and Military Equipment*, 2024, № 4 (76), pp. 58–64. DOI: 10.30748/soivt.2023.76.06
35. Kostyria O. O., Hryzo A. A., Solomonenko Yu. S. et al. Mathematical Model of Shifted Time of Combined Signal as Part of Fragments with Linear and Quadratic Frequency Modulation, *Visnyk NTUU KPI Seriia – Radiotekhnika Radioaparatobuduvannia*, 2024, Vol. 97, pp. 5–11. DOI: 10.20535/RADAP.2024.97.5-11
36. Kostyria O. O., Hryzo A. A., Dodukh O. M. Synthesis time-shifted mathematical model of a combined signal with linear and cubic frequency modulation, *Information Processing Systems*, 2024, № 1(176), pp. 73–81. DOI: 10.30748/soi.2024.176.09
37. Hryzo A. A., Kostyria O. O., Khudov H. V. et al. Implementation of Structural-Parametric Synthesis of a Nonlinear Frequency Modulated Signal on the Basis of a Genetic Algorithm, *Science and Technology of the Air Force of Ukraine*, 2024, № 1(54), pp. 77–82. DOI:10.30748/nitps.2024.54.10
38. Cheng Z., Sun Z., Wang J. et al. Magneto-Acousto-Electrical Tomography using Nonlinearly Frequency-Modulated Ultrasound, *Phys Med Biol*,2024, Vol. 69(8), PMID: 38422542. DOI: 10.1088/1361-6560/ad2ee5
39. Septanto H., Sudjana O., Suprijanto D. A Novel Rule for Designing Tri-Stages Piecewise Linear NLFM Chirp, *2022 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET) 6–7 December 2022: proceedings*. Bandung, Indonesia, IEEE, 2022, pp. 62–67. DOI: 10.1109/ICRAMET56917.2022.9991201
40. Kostyria O. O., Hryzo A. A., Dodukh O. M. et al. Improvement of Mathematical Models with Time-Shift of Two- and Tri-Fragment Signals with Non-Linear Frequency Modulation, *Visnyk NTUU KPI Seriia – Radiotekhnika Radioaparatobuduvannia*, 2023, Vol. 93, pp. 22–30. DOI: 10.20535/RADAP.2023.93.22-30
41. Kostyria O. O., Hryzo A. A., Dodukh O. M. et al. Mathematical model of a two-fragment signal with a non-linear frequency modulation in the current period of time, *Visnyk NTUU KPI Seriia – Radiotekhnika Radioaparatobuduvannia*, 2023, Vol. 92, pp. 60–67. DOI:10.20535/RADAP.2023.92.60-67
42. Kostyria O. O., Hryzo A. A., Khizhnyak I. A. et al.] /Implementation of the Method of Minimizing the Side Lobe Level of Autocorrelation Functions of Signals with Nonlinear Frequency Modulation, *Visnyk NTUU KPI Seriia – Radiotekhnika Radioaparatobuduvannia*, 2024, Vol. 95, pp. 22–30. DOI:10.20535/RADAP.2023.93.22-30
43. Adithya valli N., Elizabath rani D., Kavitha C. Performance Analysis of NLFM Signals with Doppler Effect and Background Noise, *International Journal of Engineering and*

DOI 10.15588/1607-3274-2025-3-1

Advanced Technology (IJEAT), 2020, Vol. 9, Issue 3, pp. 737–742. DOI:10.35940/ijeat.B3835.029320
44. Anoosha Ch., Krishna B. T. Peak Side Lobe Reduction Analysis of NLFM and Improved NLFM Radar Signal with Non-Uniform PRI, *Aiub Journal of Science and Engineering (AJSE)*, 2022, Vol. 21, Issue 2, pp. 125–131. DOI: 10.53799/ajse.v21i2.440
45. Swiercz E., Janczak D., Konopko K. Estimation and Classification of NLFM Signals Based on the Time-Chirp, *Sensors*, 2022, Vol. 22, Issue 21, Article № 8104. DOI:10.3390/s22218104
46. Milczarek H., Le´snik C., Djurovi´c I. et al. Estimating the Instantaneous Frequency of Linear and Nonlinear Frequency Modulated Radar Signals – A Comparative Study, *Sensors*. 2021, Vol. 21(8), Article № 2840. DOI: 10.3390/s21082840
47. Anoosha Ch., Krishna B. T. Non-Linear Frequency Modulated Radar Echo Signal Cancellation using Interrupted Sampling Repeater Jamming, *International Journal of Performability Engineering*, 2021, Vol. 17(4), pp. 404–410. DOI: 10.23940/ijpe.21.04.p8.404410

Received 04.04.2025.
Accepted 26.06.2025.

УДК 621.396.962

МЕТОД ДОСЛІДЖЕННЯ МАТЕМАТИЧНОЇ МОДЕЛІ ЗСУНУТОГО ЧАСУ ДВОФРАГМЕНТНОГО СИГНАЛУ З НЕЛІНІЙНОЮ ЧАСТОТНОЮ МОДУЛЯЦІЄЮ

Костиря О. О. – д-р техн. наук, старший науковий співробітник, провідний науковий співробітник Харківського національного університету Повітряних Сил імені Івана Кожедуба, Харків, Україна.

Гризо А. А. – канд. техн. наук, доцент, начальник НДЛ Харківського національного університету Повітряних Сил імені Івана Кожедуба, Харків, Україна.

Трофимов І. М. – канд. техн. наук, старший дослідник, професор кафедри Харківського національного університету Повітряних Сил імені Івана Кожедуба, Харків, Україна.

Ляшенко О. І. – науковий співробітник Харківського національного університету Повітряних Сил імені Івана Кожедуба, Харків, Україна.

Бєрнік Є. В. – ад'юнкт Харківського національного університету Повітряних Сил імені Івана Кожедуба, Харків, Україна.

АНОТАЦІЯ

Актуальність. Подальший розвиток теорії та техніки формування і оброки складних радіолокаційних сигналів передбачає дослідження існуючих та створення нових математичних моделей зондувальних радіосигналів. Один із напрямків таких досліджень спрямовується на зниження максимального рівня бічних пелюсток автокореляційних функцій сигналів з внутрішньо імпульсною модуляцією частоти або фази. При цьому миттєва частота може змінюватися за лінійним або ж нелінійним законом. Нелінійні закони частотної модуляції можуть забезпечити зниження максимального рівня бічних пелюсток без амплітудної модуляції вихідного сигналу радіопередавального пристрою, а значить без втрат потужності зондувальних сигналів. Широке запровадження нелінійно-частотно модульованих сигналів в радіолокаційну техніку стримується недостатньою проробкою їх математичних моделей. Тому розроблення методів для дослідження існуючих математичних моделей сигналів з нелінійною частотною модуляцією є актуальною науковою задачею.

Метою роботи є розробка методу для виконання досліджень стосовно визначення переваг та недоліків математичної моделі нелінійно-частотно модульованого сигналу у складі двох фрагментів з лінійною модуляцією частоти.

Метод. У цьому дослідженні запропоновано метод аналізу математичних моделей сигналів, який базується на переході від шкали зсунутого часу до шкали поточного часу. Методологія включає такі основні етапи: формалізований опис математичних моделей, перехід до іншої шкали часу, виділення складових та визначення їх фізичної сутності, проведення порівняльного аналізу. Перевірку працездатності методу виконано шляхом імітаційного моделювання.

Результати. З використанням запропонованого методу визначено, що математична операція зсуву шкали часу є еквівалентною появі в математичній моделі додаткових складових, що здійснюють одночасну автоматичну компенсацію стрибка частоти на стику фрагментів, а також додаткового лінійного приросту фази у другому лінійно-частотно модульованому фрагменті. Застосований підхід забезпечує наочну ілюстрацію механізму компенсації стрибка частоти у математичної моделі, що досліджувалася. Використаний метод дозволив виявити недолік розглянутої математичної моделі, який полягає у відсутності компенсаційної складової стрибка миттєвої фази у момент переходу від першого ЛЧМ фрагменту до другого.

Висновки. Розроблено метод для визначення сутності та відповідного впливу складових математичної моделі у зсунутому часі нелінійно-частотно модульованого сигналу, до складу якого входять два фрагменти з лінійною модуляцією частоти. Досліджувана модель є не зовсім коректною, оскільки не має у собі складової для компенсації стрибка фази у момент переходу від першого фрагменту сигналу до другого. Введення такої складової забезпечує подальше зниження максимального рівня бічних пелюсток автокореляційної функції сигналу.

КЛЮЧОВІ СЛОВА: сигнали з нелінійною частотною модуляцією; математична модель; стрибок миттєвої фази; автокореляційна функція; максимальний рівень бічних пелюсток.

ЛІТЕРАТУРА

1. Skolnik M. Radar Handbook. Editor in Chief / M. Skolnik. – Boston : McGraw-Hill Professional, second edition, 1990. – 846 p.
2. Cook C. Radar Signals: An Introduction to Theory and Application / C. E. Cook, M. Bernfeld. – Boston : Artech House, 1993. – 552 p.
3. Van Trees H. L. Detection, Estimation, and Modulation Theory, Part III: Radar-Sonar Processing and Gaussian Sig-

DOI 10.15588/1607-3274-2025-3-1

nals in Noise / H. L. Van Trees. – John Wiley & Sons, Inc., 2001. – 643 p. DOI: 10.1002/0471221090
4. Levanon N. Radar Signals / N. Levanon, E. Mozeson. – New York : John Wiley & Sons, 2004. – 403 p.
5. Barton D. K. Radar System Analysis and Modeling / D. K. Barton. – Boston, London : Artech House Publishers, 2004. – 566 p.
6. He You. Radar Data Processing with Applications / He You, Xiu Jianjuan, Guan Xin. – Publishing House of Electronics Industry, 2016. – 536 p. DOI: 10.1002/9781118956878
7. Melvin W. L. Principles of modern radar. Vol. II: Advanced techniques / W. L. Melvin, J. A. Scheer. – Sci Tech Publishing, 2013. – 846 p.
8. Richards M. A. Principles of Modern Radar: Basic Principles / M. A. Richards, J. Scheer, W. A. Holm. – SciTech Pub., 2010. – 924 p.
9. Blackman S. S. Design and Analysis of Modern Tracking Systems / S. S. Blackman, R. F. Popoli. – Boston, London : Artech House, 1999. – 1230 p.
10. McDonough R. N. Detection of Signals in Noise (second ed.) / R. N. McDonough, A. D. Whalen. – San Diego : Academic Press, Inc., USA, 1995. – 495 p.
11. Argenti F. Radar pulse compression methods based on nonlinear and quadratic optimization / F. Argenti, L. Facheris // IEEE transactions on geoscience and remote sensing. – 2020. – Vol. 59, Issue 5. – P. 3904–3916. DOI:10.1109/TGRS.2020.3010414
12. Muralidhara N. Performance Analysis of Weighing Functions for Radar Target Detection / N. Muralidhara, V. Velayudhan, M. Kumar // International Journal of Engineering Research & Technology (IJERT). – 2022. – Vol. 11, Issue 03. – P. 161–165.
13. Design and Implementation of an Enhanced Matched Filter for Sidelobe Reduction of Pulsed Linear Frequency Modulation Radar / [A. Azouz, A. Abosekeen, S. Nassar et al.] // Sensors. – 2021. – Vol. 21(11). – Article № 3835. DOI:10.3390/s21113835
14. Two-Step Optimization Framework for Low Sidelobe NLFM Waveform Using Fourier Series / [Q. Xie, H. Zeng, Z. Mo et al.] // IEEE Geoscience and Remote Sensing Letters. – 2022. – Vol. 19. – Article № 4020905. DOI:10.1109/LGRS.2022.3141081
15. Ghavamirad J. R. Sidelobe Level Reduction in the ACF of NLFM Signals Using the Smoothing Spline Method / J. R. Ghavamirad, R. A. Sadeghzadeh, M. A. Sebt // Electrical Engineering and Systems Science, Signal Processing: arXiv:2501.06657 [eess. SP]. – 2025. – 5 p. DOI:10.48550/arXiv.2501.06657
16. Novel Nonlinear Frequency Modulation Waveform With Low Sidelobes Applied to Synthetic Aperture Radar / [W. Wei, Y. Wang, Z Wang et al.] // IEEE Geoscience and Remote Sensing Letters. – 2022. – Vol. 19. – Article № 4515405. – P. 1–5. DOI: 10.1109/LGRS.2022.3216340
17. Parametric NLFM Waveform for Spaceborne Synthetic Aperture Radar / [Y. Zhang, Y. Deng, Z. Zhang et al.] // IEEE Transactions on Geoscience and Remote Sensing. – 2022. – Vol. 60. – Art no. 5238909. – P. 1–9. DOI: 10.1109/TGRS.2022.3221433
18. Staring Spotlight SAR with Nonlinear Frequency Modulation Signal and Azimuth Non-Uniform Sampling for Low Sidelobe Imaging / [W. Xu, L. Zhang, C. Fang et al.] // Sensors. –2021. –Vol. 21. – Article № 6487. DOI:10.3390/s21196487
19. Design and implementation of spaceborne NLFM radar signal generator / [T. Jiang, B. Li, H. Li et al.] // Second IYSF Academic Symposium on Artificial Intelligence and Computer Engineering. – 2021. – Vol. 12079. – Article № 120792S. – 6 p. DOI: 10.1117/12.2623222
20. Analytic NLFM Waveform Design with Harmonic Decomposition for Synthetic Aperture Radar / [Y. Zhang, Y. Deng, Z. Zhang et al.] // IEEE Geoscience and Remote Sensing Letters. – 2022. – Vol. 4. – Art no. 4513405. DOI:10.1109/lgrs.2022.3204351
21. Non-Continuous Piecewise Nonlinear Frequency Modulation Pulse with Variable Sub-Pulse Duration in a MIMO SAR Radar System / [Yu. Zhaoa, M. Ritchieb, X. Lua et al.] // Remote Sensing Letters. – 2020. – Vol. 11, Issue 3. – P. 283–292. DOI: 10.1080/2150704X.2019.1711237
22. Anoosha, C. Comparison on Radar Echo Cancellation Techniques for SAR Jamming / C. Anoosha, B.T. Krishna // Lecture Notes in Electrical Engineering Microelectronics, Electromagnetics and Telecommunications. – Springer, Singapore. – 2020. – P. 651–658. DOI:10.1007/978-981-15-3828-5_67
23. A novel NLFM waveform with low sidelobes based on modified Chebyshev window / [Y. Zhang, W. Wang, R. Wang et al.] // IEEE Geosci. Remote Sens. Lett. – 2020. – Vol. 17, Issue 5. – P. 814–818. DOI: 10.1109/LGRS.2019.2930817
24. Roy A. Radar waveform diversity using nonlinear chirp with improved sidelobe level performance / A. Roy, H. B. Nemade, R. Bhattacharjee // AEU – International Journal of Electronics and Communications. – 2021. – Vol. 136(11). – Article № 153768. DOI:10.1016/J.AEUE.2021.153768
25. A Variable Chirp Rate Stepped Frequency Linear Frequency Modulation Waveform Designed to Approximate Wideband Non-Linear Radar Waveforms / [M. Saleh, S.-M. Omar, E. Grivel et al.] // Digital Signal Processing. – 2021. – Vol. 109. – №102884. – 19 p. DOI:10.1016/j.dsp.2020.102884
26. Optimized Multi-Parameter NLFM Pulse Compression Waveform for low Time-Bandwidth Radar / [A. C. Van-Zyl, E. A. Wiehahn, J. E. Cillers et al.] // International Conference on Radar Systems (RADAR 2022). – 2022. – P. 289–294. DOI: 10.1049/icp.2022.2332
27. A Novel Chaotic-NLFM Signal under Low Oversampling Factors for Deception Jamming Suppression / [J. Li, P. Wang, H. Zhang et al.] // Remote Sens. – 2024. – № 1. – P. 35. DOI:10.3390/rs16010035
28. Fan Z. Coded Excitation with Nonlinear Frequency Modulation Carrier in Ultrasound Imaging System / Z. Fan, H. Meng // 2020 IEEE Far East NDT New Technology & Application Forum (FENDT). – Kunming, Yunnan province, China: conference paper. – IEEE. – 2020. P. 31–35. DOI: 10.1109/FENDT50467.2020.9337517
29. Xu Z. Nonlinear Frequency-Modulated Waveforms Modeling and Optimization for Radar Applications / Xu Z., X. Wang, Wang Y. // Mathematics. – 2022. – Vol. 10. – P. 3939. DOI:10.3390/math10213939
30. Singh A.K. NLFM Pulse Radar for Drone Detection using Predistortion Technique / A.K. Singh, K.-B. Bae, S.-O. Park // Journal of Electromagnetic Waves and Applications. – 2021. – Vol. 35. – P. 416–429. DOI:10.1080/09205071.2020.1844598
31. PHS: A Pulse Sequence Method Based on Hyperbolic Frequency Modulation for Speed Measurement / [T. Ping, C. Song, Z. Qi et al.] // International Journal of Distributed Sensor Networks. – 2024. – Vol. 2024. – Article № 6670576. – 11 p. DOI: 10.1155/2024/6670576

DOI 10.15588/1607-3274-2025-3-1

32. Research on Ultra-Wideband NLFM Waveform Synthesis and Grating Lobe Suppression / [L. Shuyi, Y. Jia, Y. Liu et al.] // Sensors. – 2022. – № 24. – Article № 9829. DOI:10.3390/s22249829
33. Pulse-Agile Waveform Design for Nonlinear FM Pulses Based on Spectrum Modulation / [R. Zhuang, H. Fan, Y. Sun et al.] // IET International Radar Conference. – 2021. – P. 964–969. DOI: 10.1049/icp.2021.0700.
34. Kostyria O. O. Combined two-fragment radar signals with linear and exponential frequency modulation laws / O. O. Kostyria, A. A. Hryzo, O. M. Dodukh // Systems of Arms and Military Equipment. – 2024. – № 4 (76). – P. 58–64. DOI: 10.30748/soivt.2023.76.06
35. Mathematical Model of Shifted Time of Combined Signal as Part of Fragments with Linear and Quadratic Frequency Modulation / [O. O. Kostyria, A. A. Hryzo, Yu. S. Solomonenko et al.] // Visnyk NTUU KPI Seriia – Radiotekhnika Radioaparatobuduvannia. – 2024. – Vol. 97. – P. 5–11. DOI: 10.20535/RADAP.2024.97.5-11
36. Kostyria O. O. Synthesis time-shifted mathematical model of a combined signal with linear and cubic frequency modulation / O. O. Kostyria, A. A. Hryzo, O. M. Dodukh // Information Processing Systems. – 2024. – № 1(176). – P. 73–81. DOI: 10.30748/soi.2024.176.09
37. Implementation of Structural-Parametric Synthesis of a Nonlinear Frequency Modulated Signal on the Basis of a Genetic Algorithm / [A. A. Hryzo, O. O. Kostyria, H. V. Khudov et al.] // Science and Technology of the Air Force of Ukraine. – 2024. – № 1(54). – P. 77–82. DOI:10.30748/nitps.2024.54.10
38. Magneto-Acousto-Electrical Tomography using Nonlinearly Frequency-Modulated Ultrasound / [Z. Cheng, Z. Sun, J. Wang et al.] // Phys Med Biol. – 2024. – Vol. 69(8). – PMID: 38422542. DOI:10.1088/1361-6560/ad2ee5
39. Septanto H. A Novel Rule for Designing Tri-Stages Piecewise Linear NLFM Chirp / H. Septanto, O. Sudjana, D. Suprijanto // 2022 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET) 6–7 December 2022: proceedings. – Bandung, Indonesia. – IEEE, 2022. – P. 62–67. DOI: 10.1109/ICRAMET56917.2022.9991201
40. Improvement of Mathematical Models with Time-Shift of Two- and Tri-Fragment Signals with Non-Linear Frequency Modulation / [O. O. Kostyria, A. A. Hryzo, O. M. Dodukh et al.] // Visnyk NTUU KPI Seriia – Radiotekhnika Radioaparatobuduvannia. – 2023. – Vol. 93. – P. 22–30. DOI: 10.20535/RADAP.2023.93.22-30
41. Mathematical model of a two-fragment signal with a nonlinear frequency modulation in the current period of time / [O. O. Kostyria, A. A. Hryzo, O. M. Dodukh et al.] // Visnyk NTUU KPI Seriia – Radiotekhnika Radioaparatobuduvannia. – 2023. – Vol. 92. – P. 60–67. DOI:10.20535/RADAP.2023.92.60-67
42. Implementation of the Method of Minimizing the Side Lobe Level of Autocorrelation Functions of Signals with Nonlinear Frequency Modulation / [O. O. Kostyria, A. A. Hryzo, I. A. Khizhnyak et al.] // Visnyk NTUU KPI Seriia – Radiotekhnika Radioaparatobuduvannia. – 2024. – Vol. 95. – P. 22–30. DOI: 10.20535/RADAP.2023.93.22-30
43. Adithya valli N. Performance Analysis of NLFM Signals with Doppler Effect and Background Noise / N. Adithya valli, D. Elizabath rani, C. Kavitha // International Journal of Engineering and Advanced Technology (IJEAT). – 2020. – Vol. 9, Issue 3. – P. 737–742. DOI:10.35940/ijeat.B3835.029320
44. Anoosha Ch. Peak Side Lobe Reduction Analysis of NLFM and Improved NLFM Radar Signal with Non-Uniform PRI / Ch. Anoosha, B. T. Krishna // Aiub Journal of Science and Engineering (AJSE). – 2022. – Vol. 21, Issue 2. – P. 125–131. DOI: 10.53799/ajse.v21i2.440
45. Swiercz E. Estimation and Classification of NLFM Signals Based on the Time-Chirp / E. Swiercz, D. Janczak, K. Konopko // Sensors. – 2022. – Vol. 22. – Issue 21. – Article № 8104. DOI:10.3390/s22218104
46. Estimating the Instantaneous Frequency of Linear and Nonlinear Frequency Modulated Radar Signals – A Comparative Study / [H. Milczarek, C. Le´snik, I. Djurovi´c et al.] // Sensors. – 2021. – Vol. 21(8). – Article № 2840. DOI: 10.3390/s21082840
47. Anoosha Ch. Non-Linear Frequency Modulated Radar Echo Signal Cancellation using Interrupted Sampling Repeater Jamming / Ch. Anoosha, B.T. Krishna // International Journal of Performability Engineering. – 2021. – Vol. 17(4). – P. 404–410. DOI: 10.23940/ijpe.21.04.p8.404410

DOI 10.15588/1607-3274-2025-3-1

МАТЕМАТИЧНЕ ТА КОМП'ЮТЕРНЕ МОДЕЛЮВАННЯ

MATHEMATICAL AND COMPUTER MODELING

UDC 004.42

EVALUATING FAULT RECOVERY IN DISTRIBUTED APPLICATIONS FOR STREAM PROCESSING APPLICATIONS: BUSINESS INSIGHTS BASED ON METRICS

Bashtovyi A. V. – Post-graduate student of the Department of Software, Lviv Polytechnic National University, Lviv, Ukraine.

Fechan A. V. – Dr. Sc., Professor of the Software Department, Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. Stream processing frameworks are widely used across industries like finance, e-commerce, and IoT to process real-time data streams efficiently. However, most benchmarking methodologies fail to replicate production-like environments, resulting in an incomplete evaluation of fault recovery performance. The object of this study is to evaluate stream processing frameworks under realistic conditions, considering preloaded state stores and business-oriented metrics.

Objective. The aim of this study is to propose a novel benchmarking methodology that simulates production environments with varying disk load states and introduces SLO-based metrics to assess the fault recovery performance of stream processing frameworks.

Method. The methodology involves conducting a series of experiments. The experiments were conducted on synthetic data generated by application using Kafka Streams in a Docker-based virtualized environment. The experiments evaluate system performance under three disk load scenarios: 0%, 50%, and 80% disk utilization. Synthetic failures are introduced during runtime, and key metrics such as throughput, latency, and consumer lag are tracked using JMX, Prometheus, and Grafana. The Business Fault Tolerance Impact (BFTI) metric is introduced to aggregate technical indicators into a simplified value, reflecting the business impact of fault recovery.

Results. The developed indicators have been implemented in software and investigated for solving the problems of Fisher's Iris classification. The approach for evaluating fault tolerance in distributed stream processing systems has been implemented, additionally, the investigated effect on system performance under different disk utilization.

Conclusions. The findings underscore the importance of simulating real-world production environments in stream processing benchmarks. The experiments demonstrate that disk load significantly affects fault recovery performance. Systems with disk utilization exceeding 80% show increased recovery times by 2.7 times and latency degradation up to fivefold compared to 0% disk load. The introduction of SLO-based metrics highlights the connection between system performance and business outcomes, providing stakeholders with more intuitive insights into application resilience. The findings underscore the importance of simulating real-world production environments in stream processing benchmarks. The BFTI metric provides a novel approach to translating technical performance into business-relevant indicators. Future work should explore adaptive SLO-based metrics, framework comparisons, and long-term performance studies to further bridge the gap between technical benchmarks and business needs.

KEYWORDS: fault-tolerance, Kafka Streams, benchmarking, distributed systems, performance measurement, stream processing, SLO(Service level objectives).

ABBREVIATIONS

BFTI – Business Fault Tolerance Impact;
SLA – Service Level Agreement;
SLO – Service Level Objective;
IoT – Internet of Things;
JMX – Java Management Extensions;
CPU – Central Processing Unit;
RAM – Random Access Memory;
e2e – End-to-End;
API – Application Programming Interface;
VM – Virtual Machine;
I/O – Input/Output.

NOMENCLATURE

BFTI is a Business Fault Tolerance Impact, aggregated metric for evaluating the business impact of fault tolerance based on SLO indicators;

D is a disk usage for the state store in stream processing application.

$L(t)$ is a lag at a specific point of time t;

$L_{normalised}(t)$ is a normalised consumer lag at a specific t point;

$L_{\max}$ is a max allowed lag defined by stakeholders;

Latency (t) is a event processing delay at a specific t point;

DOI 10.15588/1607-3274-2025-3-2

$\Delta Latency$ is a latency difference;

$Latency_{actual}$ is an actual latency limit during faults;

$Latency_{SLO}$ is a maximum allowed latency as per the SLO;

n is a number of measurements of lag;

t_i is a time point during the total measured time;

T_{total} is the total time of an experiment;

$Throughput(t)$ is a number of events processed by the system at a specific t point;

$\Delta Throughput$ is a throughput difference;

$Throughput_{normal}$ is an average number of events/records processed per second when the system is operating normally;

$Throughput_{fault}$ is a number of events processed per second during the fault period;

$\Delta(t)$ is a vector of normalized deviations for metrics;

Δt is an interval between measurements;

S is a distributed processing system;

V_{lag} is a SLO based lag;

w_1 is a weight coefficient for V_{lag} in BFTI formula;

w_2 is a weight coefficient for calculated throughput in BFTI formula;

w_3 is a weight coefficient for latency in BFTI formula;

X_{target} is a vector target values for metrics;

$X(t)$ is a single value that represents how systems performs across different(latency, throughput, consumer lag) specific t point.

INTRODUCTION

In today's data-driven world, stream processing frameworks have become essential for handling real-time data streams across industries such as finance, e-commerce, and IoT. Fragkoulis et al.in their work [1] talk about how increasing volume and velocity of data have driven significant advancements in stream processing technologies, including the adoption of serverless computing, edge streaming, enhanced query capabilities, and hardware acceleration. While serverless architectures offer flexibility and cost efficiency, they also introduce challenges in state management and low-latency processing. Similarly, edge computing reduces latency by bringing processing closer to data sources but requires lightweight techniques, particularly for IoT environments [2]. Hardware acceleration using GPUs, FPGAs, and near-memory computing further enhances performance, making stream processing frameworks increasingly powerful and widely adopted. As stream processing applications gain traction, ensuring their performance, reliability, scalability, and fault tolerance becomes critical. They are even considered as a supportive tool in migration to distributed systems[3]. Metrics such as throughput, latency, and resource utilization are key indicators of system efficiency and stability, aiding in detecting and mitigating failures that could impact business operations. Benchmarking is a common approach used to evaluate these metrics, providing insights into system behavior under various conditions. Fault tolerance, in particular, is a crucial aspect of stream processing, enabling systems to recover from failures and maintain uninterrupted data processing [4]. Techniques such as checkpointing, state recovery, and replication are widely used to ensure data continuity and resilience [5][6]. This resilience is increasingly essential as applications expand in scale and complexity, underscoring the need for robust benchmarking to evaluate recovery times and fault-tolerance effectiveness [7]. However, despite advancements in benchmarking methodologies, many existing studies focus on clean-state conditions, often using synthetic data that does not accurately reflect real-world production environments. Current benchmarking approaches often fail to consider the impact of preloaded state stores on system performance, leading to an incomplete understanding of how state accumulation affects latency and recovery times. The approach[8] presented by Van Dongen G, et al., lacks representation of real-world scenarios where state stores are progressively populated, impacting latency and recovery time. Another research on stream-processing cost tracking [9] in the same manner does not consider the effect of compound time on the system performance.

Introducing a benchmark methodology that simulates production-like environments, including preloaded state stores, would allow for a more accurate assessment of system performance under realistic load conditions, which is crucial for decision-making regarding resource allocation and infrastructure needs [10]. Our study addresses this gap by introducing a benchmarking methodology that evaluates stream processing performance under varying state loads, simulating real-world conditions more accurately. Additionally, while most research primarily tracks technical metrics such as latency and throughput, there is a lack of business-oriented insights that connect system performance with user experience and operational efficiency. To resolve the existing gap, we integrate SLO-based metrics, providing a framework that translates technical performance indicators into business-relevant insights.

The object of study is the process of evaluation of fault recovery in distributed stream processing applications under realistic conditions, considering preloaded state stores and business-oriented metrics.

The subject of study is benchmarking methodologies for assessing fault tolerance in stream processing frameworks, focusing on the impact of state store accumulation, synthetic failures, and business-driven SLO metrics.

The purpose of the work is to develop and validate a benchmarking methodology that simulates production environments with varying state loads, integrates SLO-based metrics, and provides insights into the business impact of fault recovery in stream processing applications.

1 PROBLEM STATEMENT

Let's assume that S distributed stream processing system is provided, which constantly ingests and processes events in real time. The system maintains local state in an embedded state store that resides on disk. The system's performance is defined by the following primary metrics:

DOI 10.15588/1607-3274-2025-3-2

– $Latency\,(t) \in \mathrm{R}_{\geq 0}$;
– $Throughput\,(t) \in \mathrm{R}_{\geq 0}$;
– $L\,(t) \in \mathrm{R}_{\geq 0}$.

Each of these metrics can degrade under fault conditions and contributes to the system's overall performance loss. However, monitoring them separately requires detailed technical analysis, making it difficult for non-technical stakeholders to evaluate the system's health quickly and effectively.

Let's assume that target values for these metrics must be:

$$X_{target} = [\ Latency_{SLO}\ ,\ Throughput_{normal},\ L_{\max}\ ]. \qquad (1)$$

Let's define the vector of normalized deviations from expected behavior:

$$\Delta(t) = [\Delta Latency\,(t),\ \Delta Throughput(t),\ L_{normalised}\,(t)]. \qquad (2)$$

This vector represents the normalized deviations from the target values. By aggregating these deviations into a single value, we aim to simplify the monitoring process. It is required to build a mathematical model of a single normalised value $X(t) \in \mathrm{R}_{\geq 0}$ that encapsulates the system's overall performance at time t and provides a single value which represents value of three metrics in the system at point t. Which must be adjustable from the priorities and requirements perspective.

The second objective is to analyze the behavior of S under varying levels of disk usage D in the embedded state store. The research's goal is to evaluate how changes in disk state store utilization D influence the system's real-time performance metrics and the resulting value of $X(t)$, providing insights for optimizing resilient stream processing in production-like environments.

2 REVIEW OF THE LITERATURE

SLO metrics are commonly referenced in fault tolerance research, they are rarely explicitly defined or structured to provide meaningful insights for stakeholders. Most studies focus on system stability from an engineering perspective, overlooking how technical failures translate into business impacts such as service availability and user satisfaction [11]. Existing benchmarking methodologies do not simulate production environments where state stores are preloaded and continuously evolving.

There are studies that have explored benchmarking methodologies for stream processing frameworks, with a strong emphasis on scalability and fault tolerance. The paper [12] offers provides insights into the scalability of frameworks like Apache Flink, Kafka Streams, and Hazelcast Jet within cloud-native microservice architectures. The study focuses on scaling challenges, particularly in dynamic resource allocation and fault recovery. However, it primarily addresses short-term scalability and does not extensively explore fault tolerance under long-term operational conditions, where large state accumulation and multi-fault scenarios may arise. This limitation highlights the need for research that considers fault recovery in systems with extensive state persistence. Another study [13] provides a comprehensive classification of fault tolerance techniques in stream processing systems, emphasizing their importance in preventing erroneous results and system unavailability. The authors highlight that failures in processing nodes or communication networks can lead to severe disruptions, impacting user experience and causing financial losses. The study introduces an evaluation framework for fault tolerance mechanisms in Apache Flink assessing efficiency in failure recovery. Key future research directions include adaptive checkpointing, integration with modern hardware, and parallel recovery mechanisms. Despite these advancements, the study does not account for real-world scenarios where applications run continuously, accumulating state over time.

3 MATERIALS AND METHODS

In this section, we propose our own method for benchmarking stream processing applications based on the basic metrics. In the end, evaluation and experiment details are presented. As we discussed previously, stakeholders may be interested in knowing how stream processing systems perform in general without technical details in debt. We concentrate on the business-related SLO metric, which is supposed to be straightforward and representative of business and engineering needs. Despite the fact, that the basic metrics like throughput, and latency are less conductive; they represent core system performance. In the following sections, we define core metrics of the system which are used for formulating our SLO metric.

The throughput metric represents the number of events per certain time mark. The metric is considered to be a status quo for most event-based applications. We define throughput as a number of processed events per second on the instance in general. For our experiment, we had to understand the change in throughput under different states of the application. The change is calculated based on the difference between throughput under normal conditions and throughput when some parts of the system are under a fault. In this way, we can define how faults affect the throughput. The change is stated as throughput difference which is declared by the formula(3):

$$\Delta Throughput = Throughput_{normal} - Throughput_{fault}. \qquad (3)$$

The latency metric represents the time delay from when an event is generated or received to when it is fully processed and produces a result. In event-based applications, latency is a critical measure of system responsiveness. Increased latency impacts customer experience and can lead to missed business opportunities in real-time applications. We describe latency as the average time taken to process each event from the moment it enters the system to when it completes processing on the instance. As with throughput, we calculate $\Delta Latency$ as the difference in latency observed under normal conditions versus

DOI 10.15588/1607-3274-2025-3-2

when fault conditions are introduced into the system by the formula(4):

$$\Delta Latency = Latency_{actual} - Latency_{SLO}. \quad (4)$$

This comparison will allow us to quantify the influence of faults on latency. In addition to $\Delta Latency$, we defined $Latency_{SLO}$, which describes a certain latency threshold acceptable by business requirements. Some applications, like the critical financial sector strictly require minimum latency for the operations in a system. This value is quite subjective and depends on the business needs. The best way to establish Latency SLO latency is by defining limitations for a certain business process by stakeholders based on the monitoring and recording of an average Latency for a specific time frame. Total latency is calculated by the formula(5):

$$Latency = \frac{\Delta Latency}{Latency_{SLO}}. \quad (5)$$

Consumer lag refers to the difference between the last message produced to a Kafka topic and the last message consumed by a downstream consumer. It indicates how far behind the consumer is in processing the data, which can occur due to factors such as high data production rates, network bottlenecks, slow processing by the application, failures, network delays, or other reasons. The actual consumer lag $L(t)$ at time t is normalized and defined by formula(6):

$$L_{\text{normalaised}}(t) = \min\left(\frac{L(t)}{L_{\max}}, 1\right). \quad (6)$$

Basically, $L_{normalised}(t) = 0$ when there is no lag and $L_{normalised}(t) = 1$ when the lag reaches or exceeds $L_{\max}$.

The formula is V_{lag} designed to quantify the proportion of time during which a Kafka Streams application experiences consumer lag violations relative to a defined SLO threshold, specifically , which is defined by the stakeholders based on both business expectations and tracking average or common lag in the production system. The actual formula(7):

$$V_{lag} = \frac{1}{T_{total}}\left(\sum_{i=1}^{n} L_{normalized}(t_i) \times \Delta t\right). \quad (7)$$

This metric allows for precise monitoring of application performance by assessing how often and for how long the system fails to meet the lag criteria, which is critical for maintaining real-time processing guarantees. The use of time-weighted integration ensures that the metric reflects the severity and duration of SLO violations, enabling more accurate diagnosis and optimization of stream processing topologies. This makes it an essential tool for evaluating fault tolerance and ensuring that the application meets business-critical requirements. Since the formula uses time-weighted experiments measurements must be conducted at least for two time points with the specified time differences. By normalizing lag values to the range [0, 1] and aggregating them over time, the formula provides a clear, comparable, and actionable measure of performance degradation under varying workloads or fault scenarios. The resulting value 0 means there is no lag on the consumer while 1 means lag is severe respectively to the defined lag threshold L_{max}. Based on the previously mentioned formulas, we defined Business Fault Tolerance Impact (BFTI) metric which is designed to quantify the overall business impact of a fault in Kafka Streams by considering the direct effects on SLOs, recovery time, throughput reduction, and their subsequent impact on operational costs, which is described by the formula(8):

$$BFTI = w_1 \times V_{lag} + w_2 \times \left(\frac{Throughput}{Throughput_{normal}}\right) + \\ + w_3 \times (Latency). \quad (8)$$

This metric provides a view of how system performance during failures translates to operational impact on the business. The results of the formula are represented in a value that is in the range of [0,1]. Lower BFTI values indicate high fault tolerance, meaning the application recovers quickly from failures with minimal impact on throughput, latency, or SLO violations. Higher BFTI values suggest poor fault tolerance, indicating significant performance degradation during failures, such as prolonged recovery times, high lag, or unacceptably high latency. In the Table 1 we have defined a reference table to simplify the interpretation of results.

Table 1 – BFTI formula results interpretation

BFTI output	Explanation	Description
(0 – 0.3]	Excellent fault tolerance	The system performs reliably under failure conditions. No immediate action is required.
(0.3–0.6]	Good fault tolerance	Minor impact on performance during failures. Monitor specific bottlenecks (e.g., lag or latency).
(0.6–0.8]	Moderate fault tolerance	Noticeable performance degradation. Review system capacity and failure recovery mechanisms.
(0.8–1.0]	Poor fault tolerance	Critical issues with fault handling. Immediate optimization or adding new instance is required.

DOI 10.15588/1607-3274-2025-3-2

Additionally, the formula introduces weighted prioritization based on business-critical metrics. Weights related to 3 main measurements: throughput, latency, and V_{lag}. A larger weight coefficient means a larger impact and priority for the respective measurement. If throughput is more critical than latency and V_{lag} than $w_2 > w_3$ and $w_2 > w_1$.

4 EXPERIMENTS

In this section we describe our technical setup used for experiments, actual experiment methodology and architecure. The technologies were selected based on our experience, usage in production, popularity, and expertise. Moreover, the selected tools mimic real-world production deployments. Our experiment replicated real-world stream processing environments using a Docker-based virtual machine setup managed via Docker Compose (v2.32.4). Containers were configured with 8 GB RAM, 40 GB disk space, and an 8-core CPU. Apache Kafka (vcp-kafka:7.1.0-1-ubi8) served as the message broker, while Kafka Streams (v3.8.1) with Spring Boot (v3.4.1) handled stream processing. Metrics were tracked using JMX, with Prometheus(v2.54.1) for data collection and Grafana for real-time visualization. The experiments ran on two Kafka client instances to evaluate distributed processing performance, ensuring a high-throughput and reproducible benchmarking environment.

For the experiment, we created a sample infrastructure based on the state-of-the-art technologies discussed above. Fig 1. shows the architectural solution. Kafka Streams was chosen as the core client library for our stream processing experiments due to our experience with it, its seamless integration with Kafka-based ecosystems, and its suitability for projects requiring rapid deployment [14]. As a lightweight, client-side library, it simplifies real-time data processing without the need for additional infrastructure, unlike other solutions that may require dedicated servers. Kafka Streams also supports essential features such as fault tolerance, stateful processing, and windowing, making it a practical and efficient choice for high-velocity data streams in agile business environments. We defined producer application and consumer application, message broker, metrics aggregator, and visualization of the metrics. To collect data and state we set JMX exporters that publish metrics from consumer applications and message brokers to monitor the metrics defined above.

For the producer, we created a data-generator app, that operated continuously throughout the experiment, providing a consistent workload that enabled detailed tracking of individual metrics across iterations. The data generator generates synthetic data at a rate of 700 events per second for two specific topics. This rate was chosen to ensure optimal resource utilization and maintain clarity in experimental conditions. In our experiment, these are tutor-in and lesson-in topics, ids for the models are generated iteratively, beginning from 0. We took a synthetic example from the educational domain, tutors can have multiple lessons attached to them.

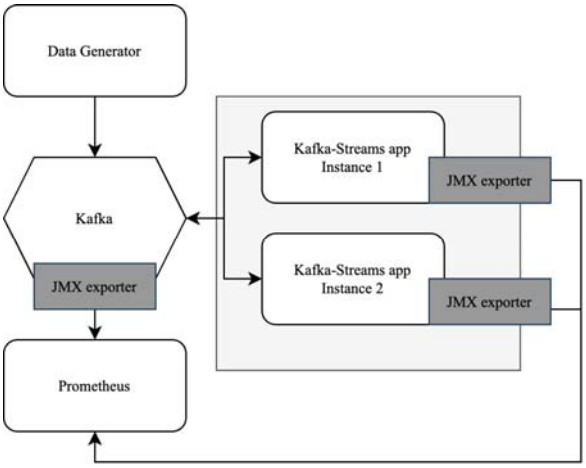


Figure 1 – The architecture of the experimental application

Tutor topic events are not necessarily related to unique tutors, so multiple events can be connected to a single business entity. The app allows a flexible configuration change for the rate of the events and other parameters. Fig. 2 shows the Kafka Stream topology used by consumers in our application. The topology consists of two source processors, multiple intermediary processors, and one sink processor where eventual results are recorded. The application consumes input events from two different topic which are populated by the producer app. There are two live instances of consumer applications. The metrics we measure are calculated based on the values from two instances, hence the moments one instance is not working metrics should reflect that accordingly. Consumer applications expose all existing build-in metrics via JMX to Prometheus.

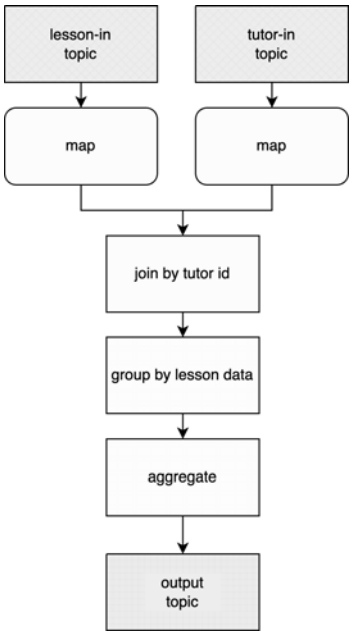


Figure 2 – Synthetic topology architecture

DOI 10.15588/1607-3274-2025-3-2

In order to measure the effect of storage capacity on the processing throughput our topology contains stateful operation – aggregation. Stateful processing helps us understand the relationships between events and leverage these relationships for more advanced stream processing use cases. To support stateful operations, we need a way of storing and retrieving the remembered data, or state, required by each stateful operator in our application (e.g., count, aggregate, join, etc.). The storage abstraction that addresses these needs in Kafka Streams is called a state store, and since a single Kafka Streams application can leverage many stateful operators, a single application may contain several state stores. There are two options for how Kafka Streams can handle stateful operations, in memory and disk storage state. The memory state is efficient in terms of processing and fast operations. On the other hand disk storage has significant benefits in comparison to the in-memory: A state can exceed the size of available memory. In the event of failure, persistent stores can be restored quicker than in-memory stores. Since the application state is persisted to disk, Kafka Streams does not need to replay the entire topic to rebuild the state store whenever the state is lost (e.g., due to system failure, instance migration, etc.). It just needs to replay whatever data is missing between the time the application went down and when it came back up.

For our application, every individual Kafka Streams instance preserves its own disk-based state store. When an individual application fails on the restart it checks out local state stores and restores all of the relevant values. If an application cannot be started on the same instance then a new instance has to be started. In this scenario, the state store will not be replicated on the new instance automatically and the application has to restore a state from scratch. Since every state store is backed by Kafka Topics it is quite a simple task to do. The application has to replay all of the messages available in backed topics and restore the state. In order to speed up instance state store recovery there are multiple techniques that can be used. One of them is to have disk memory saved somewhere in the durable storage so after instance replacement there is a way to simply use the existing disk instead of replaying Kafka messages. Since the focus of the experiment is to provide insights on the efficient and rapid state recovery we based our experiment on the persistent disk state store for the applications. We conducted a detailed series of experiments over a total execution time of more than 24 hours to evaluate system performance and fault tolerance under varying conditions. The experiment consisted of six iterations, with each iteration running for 2 hours to gather metrics systematically. Within each iteration, we assessed fault tolerance by introducing synthetic failures, randomly terminating one instance multiple times during the execution period to simulate real-world disruption scenarios. We determined the instance exactly after 30 minutes of beginning an interaction. To ensure robust and unbiased results, we repeated the experiments under three different disk load conditions, with two iterations per load state. Eventually, we took average results per two interactions. For the first set of iterations, the disk state was empty, representing 0% capacity utilization. The subsequent set of iterations simulated a slightly(0%), partially(about 50%), and heavily(more than 80%) loaded state. This approach allowed us to observe the system's behavior across varying levels of resource utilization, ensuring the reliability and accuracy of the evaluation outcomes. In order to load a disk for our experiment we decided to generate data synthetically, thereby populating individual instance stores with relevant events which are aggregated on the instances. This mimics the real-world scenarios when an application works for days and stores a lot of data in state storage. To monitor disk load during the experiments, we utilized the Prometheus metrics disk_free_bytes and disk_total_bytes. These metrics allowed us to calculate the percentage of disk capacity utilized at any given time, providing a clear representation of the disk load for each test scenario. This approach ensured precise tracking of disk usage, which was critical for evaluating the impact of varying disk load states on system performance and fault tolerance.

Throughput generally refers to the rate at which a system processes data. In Kafka consumers, we measured it as the number of records consumed per second from our source topics. Meaning that if records were consumed successfully then the records will be processed by the next nodes in the topology. For the throughput, we decided to measure throughput for individual instances and sum this value up. We selected kafka_consumer_records_consumed_total, which is a counter metric that increments whenever a Kafka consumer successfully processes a record. It directly reflects the number of records consumed over time, making it a reliable proxy for measuring throughput. Since the metric is cumulative we decided to take a rate of this metric and measure the number of records consumed per 1 minute. End-to-end (e2e) latency measures the total time taken for a record to traverse from the source topic to the sink topic within a Kafka Streams topology. This metric captures the processing delays introduced at each node in the topology, including computation, state store interactions, and any intermediate transformations. Measuring e2e latency helps us understand the overall performance of the data pipeline and identify potential bottlenecks. For this measurement, we chose the kafka_stream_processor_node_record_e2e_latency_avg metric, which represents the average latency for records processed by each processor node in the topol-

DOI 10.15588/1607-3274-2025-3-2

ogy. This metric provides granularity at the processor node level, allowing us to analyze and aggregate the latency for the entire topology. To calculate the e2e latency for individual instances and across the application, we averaged the metric across all processor nodes and instances. Since latency is not cumulative, we used an average calculation instead of a rate: avg(kafka_stream_processor_node_record_e2e_latency _avg). This setup allows us to evaluate the latency at each processor node. To calculate the overall e2e latency for the entire topology, we aggregated the metric across all nodes and instances: This approach ensures that we capture the average e2e latency for processing records, giving us insights into the system's overall responsiveness and helping to pinpoint areas for optimization.

The V_{lag} metric uses the kafka_consumer_fetch_manager_records_lag_avg metric, which represents the average consumer lag in Kafka consumers. This metric is aggregated and normalized over time to reflect the proportion of lag violations across the system. The base metric kafka_consumer_fetch_manager_records_lag_avg is collected for each consumer group and topic. We applied the metric for our two input topics. Since the metric is tracked across two running instances it is aggregated by topic only and normalized for value to be in [0,1] range.

Implementing the BFTI formula within Prometheus involved monitoring and calculating essential metrics directly from the Kafka Streams application. It is designed to evaluate the system's fault tolerance based on three critical components: SLO-based lag, throughput degradation, and latency increase. These components are derived from the metrics above and provide a quantitative measure of the system's performance under fault conditions. Each component incorporates arguments that must be defined collaboratively by stakeholders and the engineering team. These arguments are based on system behaviour during normal operation and the tolerable thresholds established for each metric. For instance, tolerable limits for latency or throughput degradation may reflect SLO agreements or operational baselines. The table 2, shows values we have set for these arguments, ensuring alignment with real-world operational expectations and providing a framework for accurately assessing the system's resilience. This structured approach allows for reproducible evaluation and fosters a deeper understanding of the system's fault tolerance characteristics.

Table 2 – Input arguments for the experiment

Argument	Value	Description
w_1, w_2, w_3	1	Determine the relative importance of each factor lag violation, throughput degradation, and latency increase.
$L_{\max}$	1000 messages	Lag SLO Threshold. The maximum acceptable consumer lag beyond which SLO is considered violated.
$Latency_{SLO}$	35sec	The end-to-end latency threshold that defines acceptable performance.
$Throughput_{normal}$	500 messages/sec	The normal or expected throughput of the Kafka Streams application under fault-free conditions

5 RESULTS

This section presents the findings from our experiments. Our analysis showed a positive correlation between system failures and changes in key performance metrics, including those measured by the SLO metric. In the first iteration, no synthetic data was preloaded onto the hard drive, resulting in a 0% load from experimental data. However, due to Kafka's internal metadata storage, the actual disk utilization was approximately 24%.

Metrics gathered under these baseline conditions served as benchmarks for defining SLO constraints during subsequent experiments. As illustrated on Fig. 3a, e2e latency significantly increased during an instance shutdown, doubling its normal value during the restoration period of the second instance. This spike is consistent with expectations, as live instances require additional time to rebalance and synchronize with the data generation rate during failures. Once the disrupted instance was restored, latency returned to baseline levels, reflecting the system's recovery capabilities under fault conditions. Initially, the system maintained a stable load, with throughput exhibiting relatively consistent, non-volatile values. Additionally, we can see(Fig. 3b), throughput was not significantly impacted by instance failures. In fact, brief increases in throughput were observed, indicating that the active instances temporarily accelerated processing to catch up on event backlogs. Figure 4 illustrates the effect of instance failures on SLO-based lag. The experiment reveals a dramatic spike in consumer lag, with values increasing from a regular average of 112 messages to over 15,000 messages in a short period. This behaviour aligns with expectations, as the number of incoming events remained constant while the number of active instances available to process these events was temporarily reduced. Consequently, the accumulation of unprocessed messages caused the lag to escalate rapidly during the failure period.

DOI 10.15588/1607-3274-2025-3-2

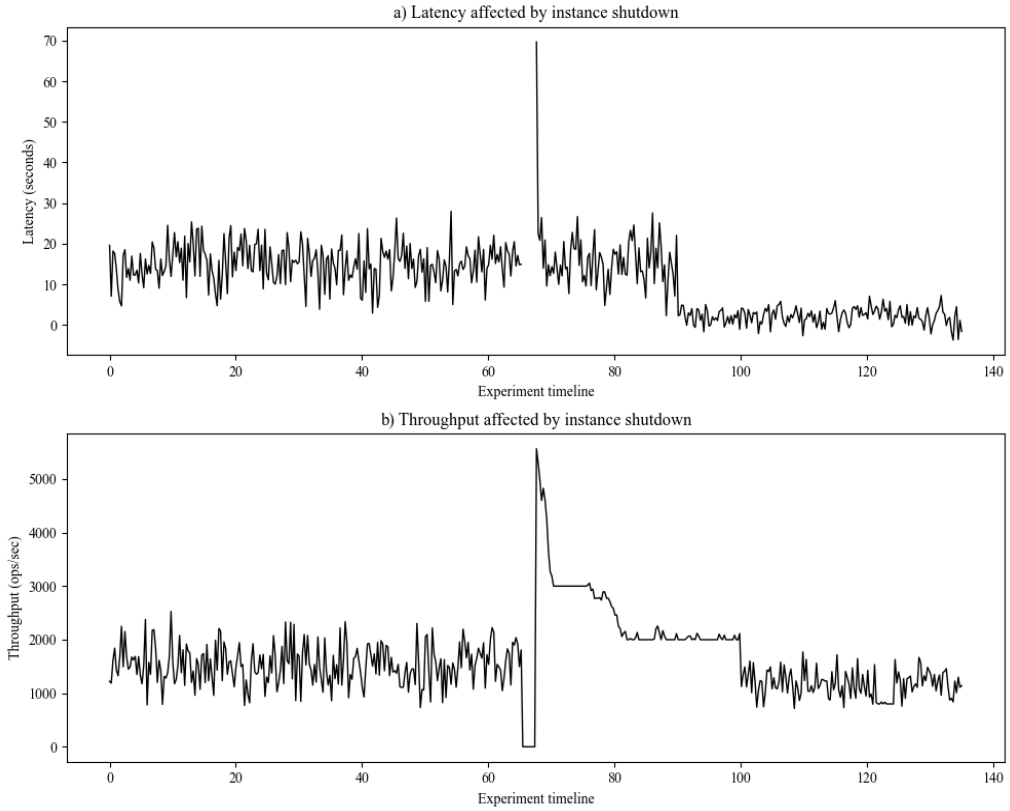


Figure 3 – Latency and throughput affected by instance shutdown at 66 minute of experiment

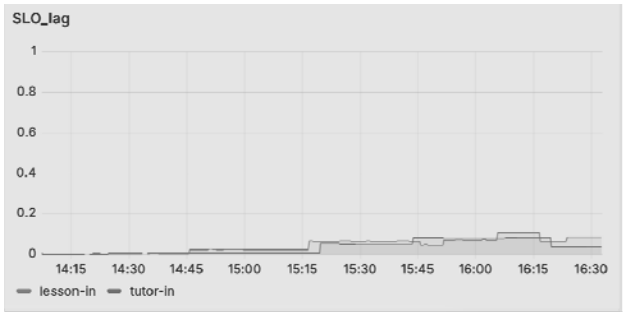


Figure 4 – Failures effect on V_{lag} latency

In the next phase of the experiment, we introduced varying levels of disk load and conducted additional iterations using the same methodology described earlier. The results, summarized in Table 3, reveal the impact of disk load on throughput, latency, and overall application performance. Notably, the system maintained stable performance when disk utilization remained below 80%, with state restoration occurring relatively quickly and without significant degradation in the basic operational characteristics of the instances. However, when disk utilization exceeded 80%, the system exhibited marked performance degradation. Latency following a fault increased nearly fivefold compared to regular latency under minimal disk load conditions. Additionally, recovery time was prolonged, taking approximately 2.7 times longer than under a 0% disk load. Throughput also declined significantly, dropping to less than half of the normal rate observed under lighter disk load conditions.

Table 3 – Experiment results for the defined metrics

Metric	Disk load		
	0%	<50%	>80%
Latency, sec	87	121	445
Throughput, ops/sec	912	819	331
V_{lag}	0.21	0.34	0.54
BFTI	0.31	0.41	1

6 DISCUSSION

The findings in the section above highlight the critical role of disk utilization in maintaining the performance and fault recovery capabilities of the system. While author[15] extensively evaluated stream processing frameworks under failure scenarios, their benchmark primarily focused on stateless or short-lived workloads with empty state stores. In contrast, our approach simulates production-grade conditions with preloaded persistent states and long-running streams, uncovering fault recovery delays that were not visible in prior benchmarks. This highlights both a methodological extension and deeper insight into the behavior of stateful applications under disk-intensive loads. Additionally, no authors provide a standardized methods for stream processing app tracking. Based on the experiment results represented in Table 3 BFTI formula provides stakeholders with a comprehensive measure of the overall fault tolerance and performance stability based

DOI 10.15588/1607-3274-2025-3-2

on the single output. We can see the BFTI definition correlates with increased latency, lower throughput, and increased V_{lag}. The formula simply represents an aggregation of all of the results in a simple manner. It simplifies decisions and saves time for interested people mainly because they can take a look at individual value instead of monitoring the whole system for different metrics. If the following and deeper analysis is required then engineering teams can elaborate on discovery and investigate various metrics of the system. However, it is the next step after monitoring BFTI results. The impact of a failure on one instance is straightforward: the application processes fewer events at a slower pace. But the long-term effects of such failures are more critical. Our analysis shows that latency and throughput remain degraded for a considerable period even after the affected instance has restarted. This occurs because the data generator continues to produce events during the failure, leading to a backlog of unprocessed events. As a result, some events remain delayed while the failed instance recovers and the system rebalances. For stakeholders, this means the system handles business tasks at a reduced efficiency, potentially affecting operational timelines and customer satisfaction. The metrics reveal that it took over an hour for the application to recover to its initial performance levels, indicating that fault recovery is not instantaneous and can disrupt normal operations for an extended period. This recovery lag underscores the importance of considering fault-tolerance strategies to minimize downtime and ensure smoother operations. One potential mitigation strategy is scaling the application dynamically in response to failures. For example, when an instance fails, a new instance could be launched alongside restarting the failed one. However, this approach has limitations it requires the number of Kafka partitions to be equal to or greater than the total number of instances, and scaling down later can introduce additional overhead due to rebalancing. Moreover, this solution may be resource-intensive and could temporarily impact throughput and overall performance. These findings highlight the need for careful planning of fault-tolerance mechanisms that balance recovery speed, resource allocation, and system performance, ensuring minimal disruption to both operations and stakeholder expectations.

CONCLUSIONS

The study addressed critical gaps in benchmarking methodologies for stream processing frameworks by simulating production-like environments and introducing SLO-based metrics to evaluate fault-tolerance performance. Key findings demonstrate that while systems maintain stable performance under moderate disk loads, performance degrades significantly when disk utilization exceeds 80%. The increased latency, throughput reduction, and prolonged recovery times observed under heavy disk loads underline the importance of robust fault-tolerance mechanisms. Furthermore, the incorporation of SLO-based metrics provided meaningful insights into how technical disruptions affect business outcomes, emphasizing the value of bridging the gap between engineering metrics and stakeholder objectives.

The scientific novelty of the obtained results lies in the proposed methodology for evaluating fault recovery in stream processing applications using preloaded state stores and business-driven performance indicators for the first time. Unlike existing approaches, this method integrates SLO-based metrics to quantify the business impact of failures, providing a novel perspective on fault tolerance assessment. Additionally, our work introduces a new benchmarking framework that considers varying state loads, which enables a more realistic evaluation of stream processing resilience in production environments.

The practical significance of the results is that the developed benchmarking methodology and BFTI metric allow practitioners to assess the reliability of stream processing applications with greater precision. The methodology was validated through experiments, demonstrating its applicability for real-world deployments. The proposed approach can be directly integrated into performance monitoring systems, aiding decision-makers in optimizing resource allocation, failure recovery strategies, and system resilience.

Prospects for further research include refining the proposed benchmarking methodology to accommodate different types of state store implementations, such as distributed file systems or cloud-based storage solutions. Future work should also explore the development of adaptive SLO-based metrics that dynamically adjust based on workload variations and user-defined business priorities. Additionally, extending the study to compare fault recovery across multiple stream processing frameworks, such as Apache Flink and Spark Streaming, would provide deeper insights into optimizing real-time data processing for various industry applications.

REFERENCES

1. Fragkoulis M., Carbone P., Kalavri V. et al. A survey on the evolution of stream processing systems, *The VLDB Journal*, 2024, Vol. 33, № 2, pp. 507–541. DOI: 10.1007/s00778-023-00819-8
2. Sasaki Y. A survey on IoT big data analytic systems: Current and future, *IEEE Internet of Things Journal*, 2022, Vol. 9, № 2, pp. 1024–1036. DOI: 10.1109/JIOT.2021.3131724
3. Bashtovyi A., Fechan A. Change data capture for migration to event-driven microservices: Case study, *Proc. of the IEEE Int. Conf. on Computer Science and Information Technologies (CSIT)*, 2023, pp. 1–4. DOI: 10.1109/CSIT61576.2023.10324262
4. Vogel A., Henning S., Perez-Wohlfeil E. et al. A comprehensive benchmarking analysis of fault recovery in stream processing frameworks, *Proc. of the 18th ACM Int. Conf. on Distributed and Event-Based Systems*, 2024, pp. 171–182. DOI: 10.48550/arXiv.2404.06203
5. Marcotte P., Grégoire F., Petrillo F. Multiple fault-tolerance mechanisms in cloud systems: A systematic review, *2019 IEEE Int. Conf. on Software Quality, Reliability and Security Companion (QRS-C)*, 2019, pp. 337–344. DOI: 10.1109/ISSREW.2019.00104

DOI 10.15588/1607-3274-2025-3-2

6. Friedman E., Tzoumas K. Introduction to Apache Flink: Stream Processing for Real Time and Beyond. Sebastopol, O'Reilly Media, 2016, 322 p.
7. Wu H., Shang Z., Peng G., Wolter K. A reactive batching strategy of Apache Kafka for reliable stream processing in real-time, *2020 IEEE 31st Int. Symp. on Software Reliability Engineering (ISSRE)*, 2020, pp. 252–261. DOI: 10.1109/ISSRE5003.2020.00028
8. Van Dongen G., Van den Poel D. Evaluation of stream processing frameworks for fault tolerance and performance metrics, *IEEE Access*, 2021, Vol. 9, pp. 102349–102365. DOI: 10.1109/TPDS.2020.2978480
9. Venkataraman S., Yang Z., Parashar M. et al. Cost of fault-tolerance on data stream processing, *Proc. of the VLDB Endowment*, 2017, Vol. 10, № 11, pp. 1478–1491. DOI: 10.1007/978-3-030-10549-5_2
10. Grambow M. Benchmarking Microservice Platforms and Applications in the Cloud. Berlin, TU Berlin, 2024. [in press].
11. Henning S., Hasselbring W. Benchmarking scalability of stream processing frameworks deployed as microservices in the cloud, *Journal of Systems and Software*, 2024, Vol. 208, pp. 111879. – DOI: 10.1016/j.jss.2023.111879
12. Wang X., Zhang C., Fang J. et al. A comprehensive study on fault tolerance in stream processing systems, *Frontiers of Computer Science*, 2022, Vol. 16, P. 162603. DOI: 10.1007/s11704-020-0248-x
13. Hoseiny Farahabady M. R., Taheri J., Zomaya A. Y. et al. A dynamic resource controller for resolving quality of service issues in modern streaming processing engines, *2020 IEEE 19th Int. Symp. on Network Computing and Applications (NCA)*, 2020, pp. 1–8. DOI: 10.1109/NCA51143.2020.9306697
14. Van Dongen G., Van den Poel D. A performance analysis of fault recovery in stream processing frameworks, *IEEE Access*, 2021, Vol. 9, pp. 93745–93763. DOI: 10.1109/ACCESS.2021.3093208
15. Van Dongen G. Open stream processing benchmark: an extensive analysis of distributed stream processing frameworks : Master's thesis. Ghent, Ghent University, Faculty of Economics and Business Administration, 2021, 112 p.

Accepted 18.03.2025.
Received 11.06.2025.

УДК 004. 42

ОЦІНКА ВІДНОВЛЕННЯ РОЗПОДІЛЕНИХ СИСТЕМ ПІСЛЯ ЗБОЇВ У ДОДАТКАХ ПОТОКОВОЇ ОБРОБКИ ДАНИХ: РОЗУМІННЯ МЕТРИК З ТОЧКИ ЗОРУ БІЗНЕСУ

Баштовий А. В. – аспірант кафедри програмного забезпечення національного університету "Львівська політехніка", Львів, Україна.

Фечан А. В. – д-р техн. наук, професор кафедри програмного забезпечення, Національний університет "Львівська політехніка", Львів, Україна.

АНОТАЦІЯ

Актуальність. Фреймворки потокової обробки даних широко використовуються в галузях фінансів, електронної комерції та IoT для ефективної обробки потоків даних у реальному часі. Проте більшість методологій тестування не відтворюють умови реальної роботи після впровадження, що призводить до неповної оцінки продуктивності відновлення після збоїв. Об'єктом дослідження є оцінка фреймворків потокової обробки у реалістичних умовах з урахуванням попередньо завантажених сховищ даних та бізнес-орієнтованих метрик.

Мета роботи. Розробка нової методології оцінювання продуктивності відновлення після збоїв у фреймворках потокової обробки, яка імітує виробничі умови з різними рівнями завантаження диска та вводить SLO-орієнтовані метрики для оцінки.

Метод. Методологія передбачає серію експериментів із використанням Kafka Streams у віртуалізованому середовищі на базі Docker. Експерименти оцінюють продуктивність системи при трьох рівнях завантаження диска: 0%, 50% та 80%. Під час роботи вводяться синтетичні збої, а ключові метрики, такі як пропускна здатність, затримка та відставання споживачів, відстежуються за допомогою JMX, Prometheus та Grafana. Запропонована метрика Впливу Бізнесу на Толерантність до Збоїв (BFTI) агрегує технічні показники у спрощене значення, що відображає бізнес-ефекти відновлення після збоїв.

Результати. Експерименти показують, що рівень завантаження диска суттєво впливає на продуктивність відновлення. При завантаженні диска понад 80% час відновлення збільшується у 2,7 рази, а затримка зростає до п'яти разів у порівнянні з 0% завантаження. Введення SLO-орієнтованих метрик підкреслює зв'язок між продуктивністю системи та бізнес-результатами, надаючи зацікавленим сторонам більш інтуїтивну оцінку стійкості програми.

Висновки. Отримані результати підкреслюють важливість моделювання реальних виробничих умов у тестуванні фреймворків потокової обробки. Метрика BFTI пропонує новий підхід до перетворення технічних показників у бізнес-орієнтовані індикатори. Подальші дослідження повинні включати адаптивні SLO-метрики, порівняння фреймворків та дослідження продуктивності на довготривалих інтервалах для подальшого усунення розриву між технічними показниками та бізнес-потребами.

КЛЮЧОВІ СЛОВА: потокова обробка даних, відмовостійкість, Kafka Streams, зняття метрик, розподілені системи, цілі рівня обслуговування (SLO), вимірювання продуктивності.

DOI 10.15588/1607-3274-2025-3-2

ЛІТЕРАТУРА

1. A survey on the evolution of stream processing systems / [M. Fragkoulis, P. Carbone, V. Kalavri et al.] // The VLDB Journal. – 2024. – Vol. 33, № 2. – P. 507–541. DOI: 10.1007/s00778-023-00819-8
2. Sasaki Y. A survey on IoT big data analytic systems: Current and future / Y. Sasaki // IEEE Internet of Things Journal. – 2022. – Vol. 9, № 2. – P. 1024–1036. DOI: 10.1109/JIOT.2021.3131724
3. Bashtovyi A. Change data capture for migration to event-driven microservices: Case study / A. Bashtovyi, A. Fechan // Proc. of the IEEE Int. Conf. on Computer Science and Information Technologies (CSIT). – 2023. – P. 1–4. DOI: 10.1109/CSIT61576.2023.10324262
4. A comprehensive benchmarking analysis of fault recovery in stream processing frameworks / [A. Vogel, S. Henning, E. Perez-Wohlfeil et al.] // Proc. of the 18th ACM Int. Conf. on Distributed and Event-Based Systems. – 2024. – P. 171–182. DOI: 10.48550/arXiv.2404.06203
5. Marcotte P. Multiple fault-tolerance mechanisms in cloud systems: A systematic review / P. Marcotte, F. Grégoire, F. Petrillo // 2019 IEEE Int. Conf. on Software Quality, Reliability and Security Companion (QRS-C). – 2019. – P. 337–344. DOI: 10.1109/ISSREW.2019.00104
6. Friedman E. Introduction to Apache Flink: Stream Processing for Real Time and Beyond / E. Friedman, K. Tzoumas. – Sebastopol : O'Reilly Media, 2016. – 322 p.
7. Wu H. A reactive batching strategy of Apache Kafka for reliable stream processing in real-time / H. Wu, Z. Shang, G. Peng, K. Wolter // 2020 IEEE 31st Int. Symp. on Software Reliability Engineering (ISSRE). – 2020. – P. 252–261. – DOI: 10.1109/ISSRE5003.2020.00028
8. Van Dongen G. Evaluation of stream processing frameworks for fault tolerance and performance metrics / G. Van Dongen, D. Van den Poel // IEEE Access. – 2021. – Vol. 9. – P. 102349–102365. DOI: 10.1109/TPDS.2020.2978480
9. Venkataraman S. Cost of fault-tolerance on data stream processing / [S. Venkataraman, Z. Yang, M. Parashar et al.] // Proc. of the VLDB Endowment. – 2017. – Vol. 10, № 11. – P. 1478–1491. DOI: 10.1007/978-3-030-10549-5_2
10. Grambow M. Benchmarking Microservice Platforms and Applications in the Cloud / M. Grambow. – Berlin : TU Berlin, 2024. – [in press].
11. Henning S. Benchmarking scalability of stream processing frameworks deployed as microservices in the cloud / S. Henning, W. Hasselbring // Journal of Systems and Software. – 2024. – Vol. 208. – P. 111879. DOI: 10.1016/j.jss.2023.111879
12. A comprehensive study on fault tolerance in stream processing systems / [X. Wang, C. Zhang, J. Fang et al.] // Frontiers of Computer Science. – 2022. – Vol. 16. – P. 162603. DOI: 10.1007/s11704-020-0248-x
13. A dynamic resource controller for resolving quality of service issues in modern streaming processing engines / [M. R. Hoseiny Farahabady, J. Taheri, A. Y. Zomaya et al.] // 2020 IEEE 19th Int. Symp. on Network Computing and Applications (NCA). – 2020. – P. 1–8. DOI: 10.1109/NCA51143.2020.9306697
14. Van Dongen G. A performance analysis of fault recovery in stream processing frameworks / G. Van Dongen, D. Van den Poel // IEEE Access. – 2021. – Vol. 9. – P. 93745–93763. DOI: 10.1109/ACCESS.2021.3093208
15. Van Dongen G. Open stream processing benchmark: an extensive analysis of distributed stream processing frameworks : Master's thesis / G. Van Dongen. – Ghent : Ghent University, Faculty of Economics and Business Administration, 2021. – 112 p.

DOI 10.15588/1607-3274-2025-3-2

UDC 004.93

METHODS AND ALGORITHMS OF BUILDING A 3D MATHEMATICAL MODEL OF THE SURROUNDING SPACE FOR AUTOMATIC LOCALIZATION OF A MOBILE OBJECT

Korpan Ya. W. – PhD, Associate Professor, Associate Professor of the Department of Robotics and Specialized Computer Systems, Cherkassy State Technological University, Cherkassy, Ukraine.

Nechyporenko O. V. – PhD, Associate Professor, Associate Professor of the Department of Informational Security and Computer Engineering, Cherkassy State Technological University, Cherkassy, Ukraine.

Fedorov E. E. – Dr. Sc., Associate Professor, Professor of the Department of Statistics and Applied Mathematics, Cherkassy State Technological University, Cherkassy, Ukraine.

Utkina T. Yu. – PhD, Associate Professor, Associate Professor of the Department of Robotics and Specialized Computer Systems, Cherkassy State Technological University, Cherkassy, Ukraine.

ABSTRACT

Context. The task of automating the positioning of a mobile object in a closed space under the condition of its partial or complete autonomy is considered. The object of study is the process of automatic construction of a 3D model of the surrounding space.

Objective. The goal of the work is the develop an algorithm for creating a 3D model of the surrounding space for further localization of a mobile object in conditions of its partial or complete autonomy.

Method. The results of the study of the problem of localization of a mobile object in space in real time are presented. The results of the analysis of existing methods and algorithms for creating mathematical models of the surrounding space are presented. Algorithms that are widely used to solve the problem of localization of a mobile object in space are described. A wide range of methods for constructing a mathematical model of the surrounding space has been researched – from methods that use the comparison of successive point clouds of the object of the surrounding space to methods that use a series of snapshots of characteristic points and comparison of information about them in different snapshots at points that are as similar as possible according to the parameter vector.

Results. The method for three-stage construction of a 3D model of the surrounding space is proposed for solving the problem of localization of a mobile object in a closed space.

Conclusions. The conducted experiments have confirmed the possibility of the proposed algorithm for three-stage construction of a mathematical model of the environment to determine the position of a mobile object in space. The methods used in the algorithm allow obtaining information about the surrounding space, which allows localizing a mobile object in a closed space. Prospects for further research may lie in the integration of information flows about the position of the object from different devices, depending on the type of data acquisition, into a centralized information base for solving a wide range of tasks performed by automatic mobile objects (robots).

KEYWORDS: mathematical 3D model, localization method, SLAM method algorithms, position determination, mobile object.

ABBREVIATIONS

DATMO is a detection and tracking of movingobjects;
EKF is a extended Kalman filter;
ICP is a iterative closest point;
NDT is a normal-distributions transform;
NNS is a nearest neighbor search;
LS3D is a least squares 3D surface matching;
ML-NDT is a multi-layered NDT;
PCL is a point cloud library;
PMD is a photonic mixer device;
RANSAC is a random sample consensus;
SLAM is a simultaneous localization and mapping;
SURF is a speeded up robust features;
ToF is a time-of-flight.

NOMENCLATURE

T_n is a environmental space;
x_n is a scan area size along the x-axis;
y_n is a scan area size along the y-axis;
z_n is a scan area size along the z-axis;
$z_{t,n}$ is a beam data obtained using a laser range scanner;
$c_{t,n}$ characterizes the erroneous measurements in the data;
k is a field covered by the beam;
m is a model that maximizes the probability of the data;
x_t is a robot position at a certain point in time;
f is a function that returns for each position x_t of the robot, each beam with number n and each field $k \le z_{t,n}$.;
ζ is a indicator variables that are equal to 1 if and only if $z_{t,n}$ is the maximum value of the range, otherwise – 0;
s_i is a given point in space R^3, for which it is necessary to find the shortest distance to the set T;
τ is a the point that gives the smallest value to the functional S;
τ_j is a nearest pair of points for point s_i;
J is a functional for a given displacement vector t and rotation matrix R;
R is a real orthogonal matrix with a determinant equal to one, that is, it belongs to a special orthogonal group of rotation $SO(3)$;
t is a vector representing the displacement of a set of points along vector $[x,y,z]^T$;

DOI 10.15588/1607-3274-2025-3-3

$\|\cdot\|_2^2$ is a square of the norm or the square of the Euclidean distance between two points;

S_i is a point cloud block;

ε is a point cloud threshold;

$T_7(p,x)$ is a spatial transformation function;

p is a vector;

t is a displacement;

r is a axis of rotation;

ϕ is a angle of rotation.

INTRODUCTION

Today, there is a rapid development of mobile autonomous systems. Such systems, for the most part, are increasingly based on the use of intelligent systems. One of the main functions of modern intelligent mobile systems is their autonomous navigation. Implementation of such a function is possible with a clear understanding of the surrounding environment, that is, it is necessary to transform information about the surrounding world into a mathematical model convenient for processing by these systems.

The object of study is the process of automatic construction of a 3D model of the surrounding space.

Construction of a mathematical model of the surrounding environment using modern circuit solutions is possible when using SLAM and DATMO algorithms [1].

The subject of study is methods of building a mathematical model of the environment at the expense of a set of received spatial points that give an idea of their relative location in space.

The purpose of the work is to develop a method for creating a 3D model of the surrounding space for further localization of a mobile object in conditions of its partial or complete autonomy.

1 PROBLEM STATEMENT

Let's assume that there is a mobile object (robot) that is in a closed space under conditions of partial or complete autonomy. The task facing the system of such a mobile object is that, based on the received data, it should be able to form a 3D model of the surrounding space, localize the object in space, create (or update or supplement) a "map" of movement and to decide on further actions. It should be borne in mind that many methods of creating a mathematical model of the environment are designed with powerful hardware and computing resources in mind, and this often does not depend on the complexity of the task assigned to the mobile device.

Let's assume that the environmental space T_n (x_n, y_n, z_n). For a given space, the task of determining the point clouds that characterise the objects located in the given area can be solved in two ways: the first is by directly calculating the point clouds in the entire space T_n and the second is by dividing the space into parts $T_{n/m}$, where m characterizes the share into which the space is divided (in this case, the calculation of the point clouds is carried out separately in each part). The division into parts and division options are set by the user and can look like this, for example:

– into two parts – $T_{n/2}(x_{n/2}, y_n, z_n)$ or $T_{n/2}(x_n, y_n, z_{n/2})$;
– into three parts – $T_{n/3}(x_{n/3}, y_n, z_n)$ or $T_{n/3}(x_n, y_n, z_{n/3})$;
– into six parts – $T_{n/6}(x_{n/3}, y_n, z_{n/2})$ or $T_{n/6}(x_{n/2}, y_n, z_{n/3})$;
– into nine parts – $T_{n/9}(x_{n/3}, y_n, z_{n/3})$.

This division is set by the user taking into account the possible complexity of the "contour" of the surrounding space and the available computing power. That is, the task of determining the need for division (or its expediency) should be determined by the nature of the objects in the scanning area T_n (x_n, y_n, z_n) and the degree of complexity of calculating the parts m.

So, in the final result, using the methods and algorithms of SLAM and DATMO, the mobile object system, when using limited hardware and software capabilities, should produce a convenient model of the environment for further processing.

2 REVIEW OF THE LITERATURE

The work [1] presents a description of the system operation method intended for simultaneous localization and mapping, as well as detection and tracking of objects moving in dynamic environments. It is known that for more accurate localization and mapping, it is necessary to carry out a detailed reconstruction of the surrounding environment.

All approaches to creating three-dimensional modeling ("reconstruction") are of two types: passive and active. Passives do not affect the object to be reconstructed, unlike actives. In work [2], two approaches to the reconstruction of a three-dimensional model are distinguished.

The first approach. Three-dimensional scanning, which refers to active types of reconstruction and is carried out using special scanners. This method is characterized by high accuracy and does not depend on weather conditions, but it also has disadvantages, such as expensive and hard-to-find equipment, as well as a large amount of time needed to develop the model. These problems can be solved by reducing the quality of the original 3D model for simple objects that do not have clear requirements for detail and the difference in quality will not be very noticeable.

The second approach. Photogrammetric approach, which belongs to the passive type. It consists of determining characteristic points on a series of images and comparing information about them with the points that are as similar as possible according to the vector of parameters. The approach is characterized by the ability to reconstruct complex objects of any level of complexity without the use of special equipment, but it requires a lot of time and depends on weather conditions. Reducing the influence of the number of provided reference images and weather conditions on the quality of the original 3D model can be achieved by using the stereoscopic parallax algorithm and stereo images.

Active methods [3] of obtaining a mathematical description of an object include any methods that emit

DOI 10.15588/1607-3274-2025-3-3

any waves. Such methods include obtaining object characteristics using PMD-cameras, lasers, echo sounders, etc.

The principle of operation of PMD-cameras is based on ToF measurements, i.e. the measurement of the time it takes for light to move from the camera to the object and back after reflection from the object to a special light-sensitive matrix. The distance can be calculated from the equation for an ideal camera.

The article [4] compares distance determination methods using a PMD-camera and stereo vision. Possible deviations of the distance depending on the angle of inclination of the cameras are indicated. In conclusion, the distance is determined more accurately by the PMD-camera, the disadvantages are the low resolution of the PMD-camera, which leads to a lower quality of information compared to stereo vision, therefore, for the purposes of surface reconstruction, the use of both methods would be desirable.

The work [5] is devoted to the combination of active and passive methods of determining object coordinates, that is, the use of both PMD-cameras and stereo vision.

In [6], it is proposed to use a combination of a PMD-camera with a high-resolution RGB camera to improve the quality of object visualization. The accuracy of using PMD-cameras is considered, but the accuracy of the resulting combination is not specified.

The authors of [7] conduct a comparison of different ToF cameras, based on the distance determination error, depending on the installation angle, as well as the quality of the averaged frames at each distance.

The application of the laser is described in detail in the works [8, 9] for the composition of spectral portraits of objects, use for navigation of a mobile robot and for 3D modeling of an object when using a system of four cameras, respectively. The use of lasers is associated with the high accuracy of determining the points of the object's surface, but this leads to a significant increase in the price of the system for finding coordinates, modeling and visualization of objects.

3 MATERIALS AND METHODS

At the moment, there are many different SLAM algorithms, which differ both in the type of input information, the representation of the surrounding space in the form of a map, and in the methods of processing this information. The work [10] presents the classification of localization algorithms according to the dimension of the fixed space:

– two-dimensional localization on the plane (2D-SLAM);

– three-dimensional localization in space (3D-SLAM);

– color localization by R, G, B image components (RGB-D SLAM);

– color three-dimensional localization in space (6D-SLAM).

These characteristics depend directly on the type of sensor used. For example, when using simple laser rangefinders, the input information about the surrounding space is a set of grid maps, accordingly, 2D-SLAM is used for processing. In the presence of an additional scanning axis, a set of spatial points can be obtained, which gives a representation of the objects of the room taking into account their relative location in space, so 3D-SLAM can be applied here. Color localization algorithms evaluate the state of the robot based on the image from the color video camera installed on it. 6D-SLAM algorithms are used when using sensors that allow obtaining a three-dimensional color image of objects for the purpose of localization and map construction. It should be noted that the vast majority of localization algorithms on the plane can be extended to three-dimensional space.

An important feature of SLAM is that most of the algorithms can be implemented only in a static environment, that is, the room or area where the robot is located should not change.

The 2D-SLAM algorithm is used, as a rule, in application of laser rangefinders. But when processing the received data, especially in the presence of dynamic objects, it is also necessary to take into account the probability of their position changing [11]:

$$p(z_{t,n} \mid c_{t,n}, x_t, m) = \left[\prod_{k=0}^{z_{t,n}-1} (1 - m_{f(x_t,n,k)}) \right]^{\zeta_{t,n}} \times$$

$$\times \left[[m_{f(x_t,n,z_{t,n})}]^{c_{t,n}} \times [1 - m_{f(x_t,n,z_{t,n})}]^{(1-c_{t,n})} \times \right.$$

$$\left. \times \prod_{k=0}^{z_{t,n}-1} (1 - m_{f(x_t,n,k)}) \right]^{(1-\zeta_{t,n})}.$$

The first term in this equation determines the probability that the distance specified by the beam is the maximum scan range. In such a situation, the probability is calculated as the product of the probabilities that the beam covered the region from 0 to $z_{t,n-1}$.The second term of the equation indicates what to do in the case when the maximum range of the beam is not displayed. If $z_{t,n}$ is not reflected by a dynamic object, i.e. $c_{t,n}$=1, then the probability is equal to $m_{f(x,n,k)}$. If, on the contrary, $z_{t,n}$ is reflected by a dynamic object, then probability takes value is $1 - m_{f(x,n,k)}$.

The built model, when using this approach, should take into account the probability of the appearance of false measurements when building the map.

Using the 3D-SLAM algorithm has a number of advantages:

– the complete vector of the position and orientation of the mobile robot in space is known;

– measurement data obtained from the sensors do not depend on the shape of the surface on which the object is moving;

– 3D reconstruction of the room in which the moving object is located is possible.

The disadvantages of this type of algorithms include the limited speed of model building, which is associated with a large flow of information from sensors and the

DOI 10.15588/1607-3274-2025-3-3

need to process it. This problem can be partially solved using such algorithms as ICP, 3D-NDT, ML-NDT.

As in the two-dimensional version, the ICP algorithm is based on finding pairs of matching points between the current and reference scans.

The ICP algorithm [13] can be conditionally divided into four stages.

The first stage is finding the matching closest pair τ_j for the point s_i, such that

$$S(\tau)=\left\|\tau_j - s_i\right\|_2,$$

$$(\tau) = \underset{s_i\in S,\ \tau_j\in T}{\arg\min}\ S(\tau).$$

The second stage is calculating the displacement vector t and the rotation matrix R, which deliver the minimum functionality

$$J(R,t) = \sum_{i=1}^{N}\left\|(Rs_i + t) - \tau_i\right\|_2^2,$$

$$(R,t) = \underset{R\in SO(3),\ t\in R^3}{\arg\min}\ J(R,t).$$

The third stage is converting the block of transforming point cloud using the found rotation matrix of the displacement vector into a new point cloud

$$S_i = RS_i + t.$$

The fourth stage is repeating the entire iterative process of the algorithm until $J(R,t) \geq \varepsilon$, where the transforming point cloud is the point cloud obtained at the previous stage.

One of the main problems of this algorithm is the limited area of convergence: the algorithm works only under the condition that the point clouds are not significantly shifted from each other.

3D-NDT is an algorithm for three-dimensional transformation of normal distributions. The main difference between the 3D-NDT algorithm and the two-dimensional algorithm is the type of coordinate transformation functions $T(p, x)$ and its partial derivatives [14]. A general rotation in 3D is more complicated. A robust 3D rotation representation requires both an axis and an angle. A simple way to represent a general 3D transformation is to use seven parameters – three parameters for displacement, three for the rotation axis, and one for the rotation angle. Using a right-handed coordinate system and counterclockwise rotation, the 3D transformation of a point x by a parameter vector p can be formulated as

$$T_7(p,x) = \begin{bmatrix} er_x^2 + c & er_xr_y - sr_z & er_xr_z + sr_y \\ er_xr_y + sr_z & er_y^2 + c & er_yr_z - sr_x \\ er_xr_z - sr_z & er_yr_z + sr_x & er_z^2 + c \end{bmatrix} x + \begin{bmatrix} t_x \\ t_y \\ t_z \end{bmatrix}.$$

$$p = [\ t \mid r \mid \phi\ ],$$

$$t = [\ t_x, t_y, t_z\ ],$$

$$r = [\ r_x, r_y, r_z\ ],$$

$$s = \sin\phi,\ c = \cos\phi,\ e = 1 - \cos\phi.$$

In the 3D-NDT algorithm, the correct choice of cell size is very important. If the cell is too large, many other details will not be taken into account and the localization accuracy will decrease; if a very small cell is selected, it will be described quite clearly, but for the convergence of the algorithm, it is necessary to choose an initial approximation close to the real position, which cannot always be implemented. The optimal size of the cell depends on the shape and size of the room in which the moving object is located. Therefore, a structure for storing cells with normal distributions with adaptive spaces discretization, depending on the detail of the scanned areas, is needed. The work [14] presents several options for solving this problem, namely: the use of an octal tree to divide spaces into octants, additive distribution, iterative distribution, as well as the use of connected cells and cells with infinite boundaries.

ML-NDT algorithm is an extension of 3D-NDT that improves convergence speed and long-distance measurement [15]. This effect is achieved due to the automatic assignment of the cell size for each reference scan. This approach was presented as an eight-cylinder tree model, but the mathematical expectation vector and covariance matrix of each cell is stored in all layers if it contains 5 or more points. The essence of this method consists in the sequential comparison of first general forms, then more detailed and, finally, small features of the object.

In addition, a different description of the matching functions of the scan and NDT maps is presented than in the original algorithm, using the Newton and Levenberg-Marquardt iterative optimization method. The identified main drawback of the algorithm is the expansion of the necessary memory for storing layers. During experimental verification, it was found that the convergence speed increased by an order of magnitude compared to the original 3D-NDT algorithm.

Works [16, 17] present one of the methods of implementing the color localization system based on the R, G, B components of the image – RGB-D SLAM. The work algorithm is presented in the following steps:

– extracting the SURF function from current input color images [12];

– comparing the obtained functions with the functions of previous images – obtaining characteristic points;

– evaluating image depth at the locations of characteristic points (obtaining 3D correspondence points between two frames);

– estimating the relative transformation between frames using RANSAC;

DOI 10.15588/1607-3274-2025-3-3

– improving the initial estimate using the ICP algorithm [13];

– optimizing the resulting position graph.

As a result, a global consistent model of the environment is obtained, which is presented in the form of a colored point cloud.

Input information for 6D-SLAM is typically 3D laser range finder data and locations obtained using an EKF that measure odometry and metrics such as: yaw, pitch, roll, acceleration, roll rates [18].

When measuring, it is advisable to use equipment that allows you to obtain data while the robot is moving. But it is more convenient for the calculation to use a laser with a rotating profile, which requires a static 3D measurement (the robot does not move during the measurement). The initial trajectories and the 3D data associated with various positions from that trajectory are the input data for the iterative data logging component. So six registration algorithms are available:

– ICP with parallel implementation of NNS nearest neighbor procedure.

– ICP – implementation of PCL;

– ICP indicates a projection with parallel NNS;

– Semantic ICP with parallel division of points into four classes (plane, edge, floor/ground, ceiling);

– LS3D is the smallest surface area that coincides with the parallel computation of the surface representation;

– NDT – implementation of PCL.

4 EXPERIMENTS

The goal set in the work is complicated by the fact that mobile devices used indoors to capture the surrounding space use optical systems that are sensitive to the level of illumination, which increases the ambiguity of detailing the construction of a 3D model of the surrounding environment.

The research was carried out in a laboratory room with optimal filling of the space with objects of varying perceptual complexity by an optical device. The laboratory room made it possible to change the intensity of illumination of the surfaces.

A mobile object with an optical system attached to it was used during the research. The experiment was conducted at different mounting heights (from 40 to 100 cm), as well as at different tilt angles (from –30° to 40°) of the optical system.

Since it was necessary to use low-powered systems for the task at hand, i.e. the use of algorithms and methods that require minimal hardware and software resources, two systems were used:

1) a platform based on an Intel Core 2 Duo E6400 processor with a video core based on NVIDIA GeForce GT220M (1Gb), 4 Gb RAM;

2) a platform based on an Intel(R) Core(TM) i5-9300H processor with a video core based on NVIDIA GeForce GTX 1650 (4Gb), 8 Gb RAM.

5 RESULTS

The purpose of the research was to expand the capabilities of existing 2D SLAM methods to perform 3D probabilistic SLAM. The main specificity consists in supplementing the existing stages with the stages of data segmentation and scanning compliance analysis. The scanned data of the three-dimensional range is presented in the form of individual three-dimensional point clouds, which are correlated with the location of the scanning device.

The results of one of the calculations are presented in Fig. 1 and Fig. 2.

When creating the environmental model, the results of object fixation were analyzed, which were pre-ordered according to the distance from the object of movement to the object of fixation. Fig. 1 shows the results of creating a 3D model of individual elements. Fig. 1a, 1c, 1e show the results of data processing by the first computing platform. Fig. 1b, 1d, 1f show the results of data processing by the second computing platform.

Fig. 2 presents the results of data processing for all fixation components with their combination into a single 3D model. Fig. 2a shows the model built using the first computing platform, Fig. 2b – using the second computing platform.

A further study of the results showed that as the distance at which an object is fixed increases, the detail of its 3D reconstruction decreases. However, more emphasis is placed on the contours of the object. Therefore, when building a generalized model, it is advisable to use the results of calculations of individual objects with their subsequent combination. This approach will allow not only to automatically localize the mobile device in space, but also, if necessary, to analyze in more detail the objects near which it is located. This is necessary for a correct analysis of the current situation, especially for confined spaces.

The analysis of the results of the work of the first and second computing platforms (described in section 4) showed that there are no significant differences that would later influence the model creation process. The main difference lies in the time required to process the database and create point cloud. According to this indicator, the first platform spends almost three times more time processing these data sets. The total data processing time on both platforms depends on two factors:

– how many objects are located on the analyzed territory and their overall dimensions;

– how detailed a 3D model of the surrounding space needs to be built.

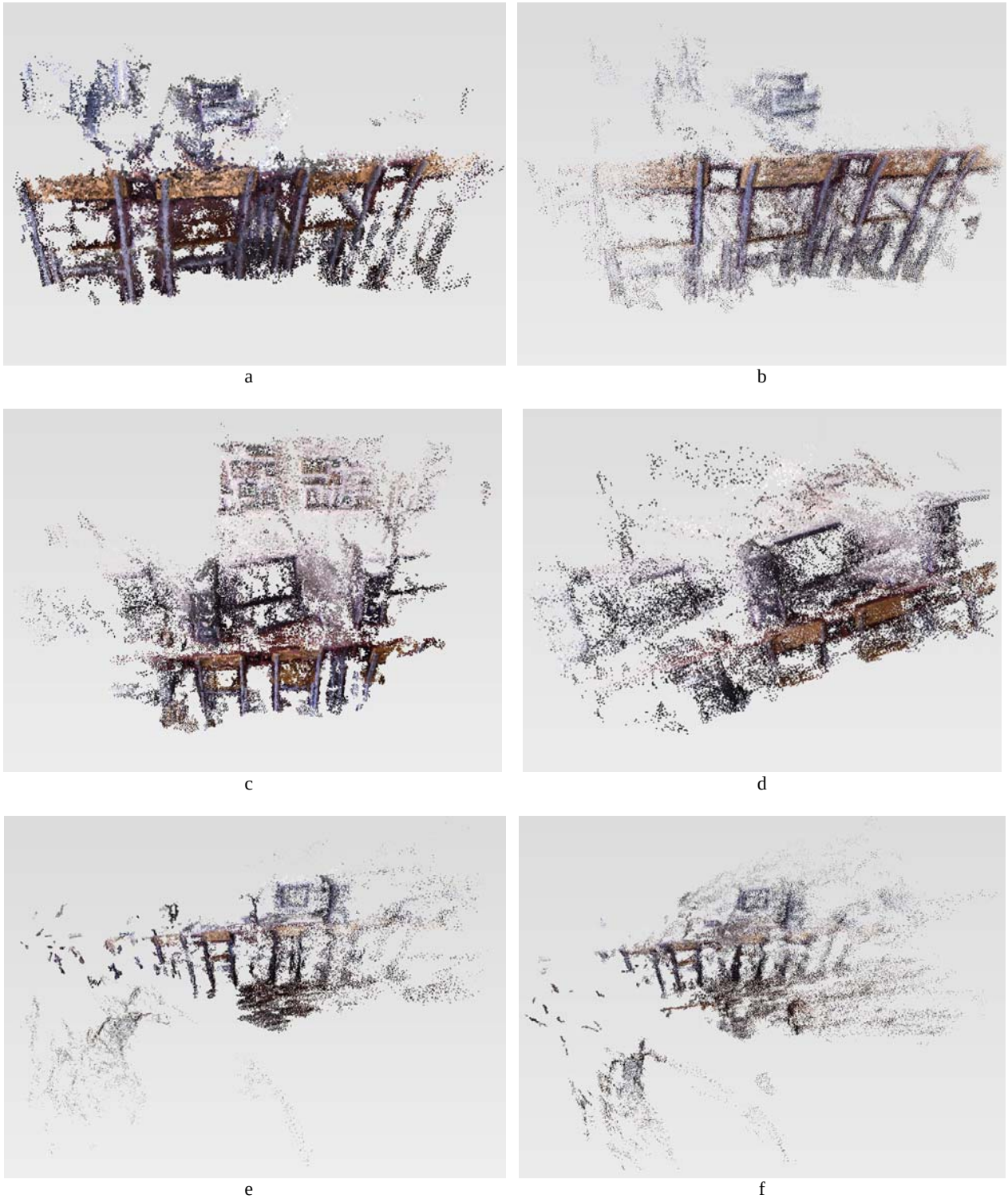
The algorithm for recreating the surrounding space is proposed to be implemented in three stages:

– initial – small elements that are near the mobile object are recreated;

– intermediate – combining small elements into a 3D model of the space near the mobile object;

– generalized – combining all intermediate objects into a generalized scheme/system.

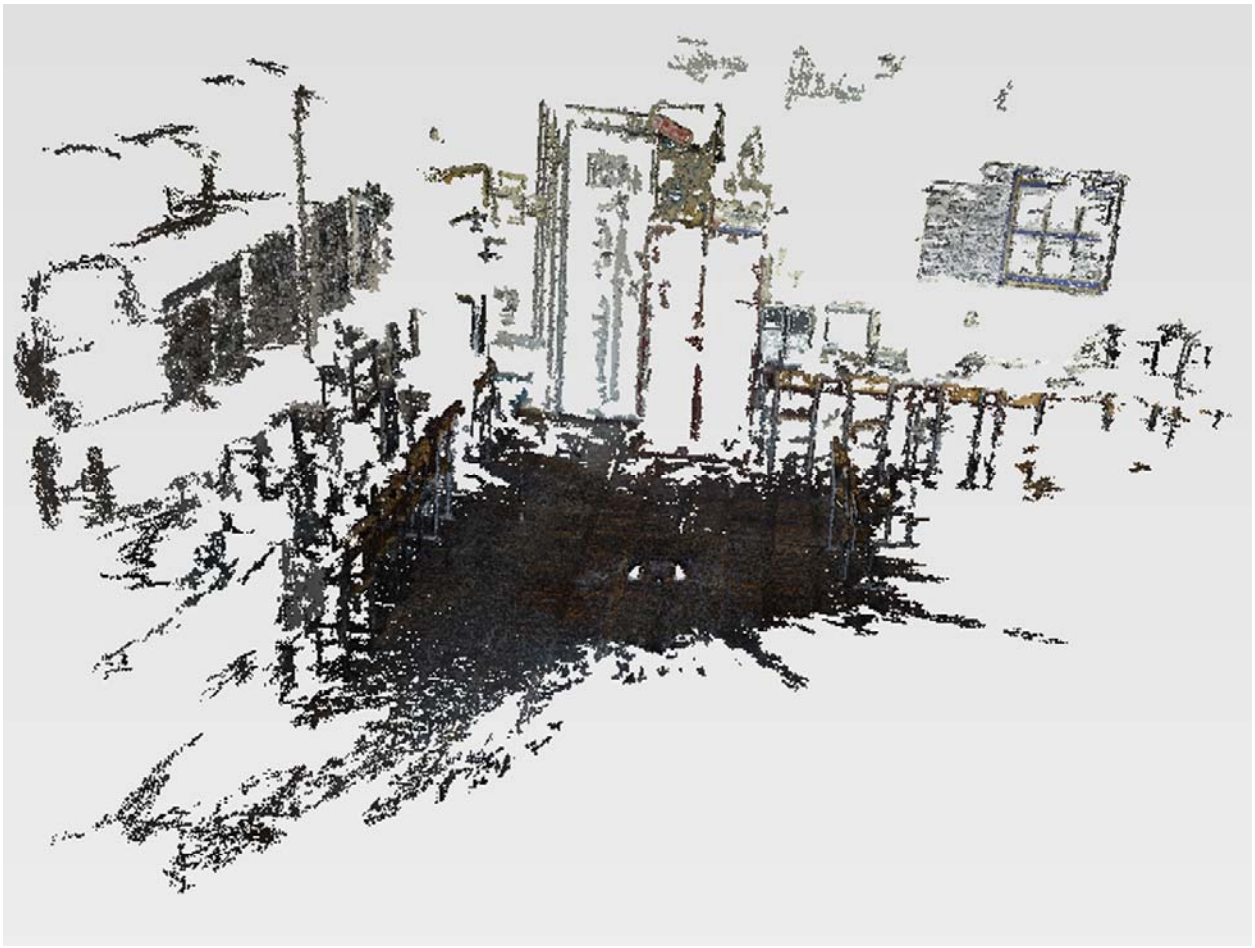
DOI 10.15588/1607-3274-2025-3-3



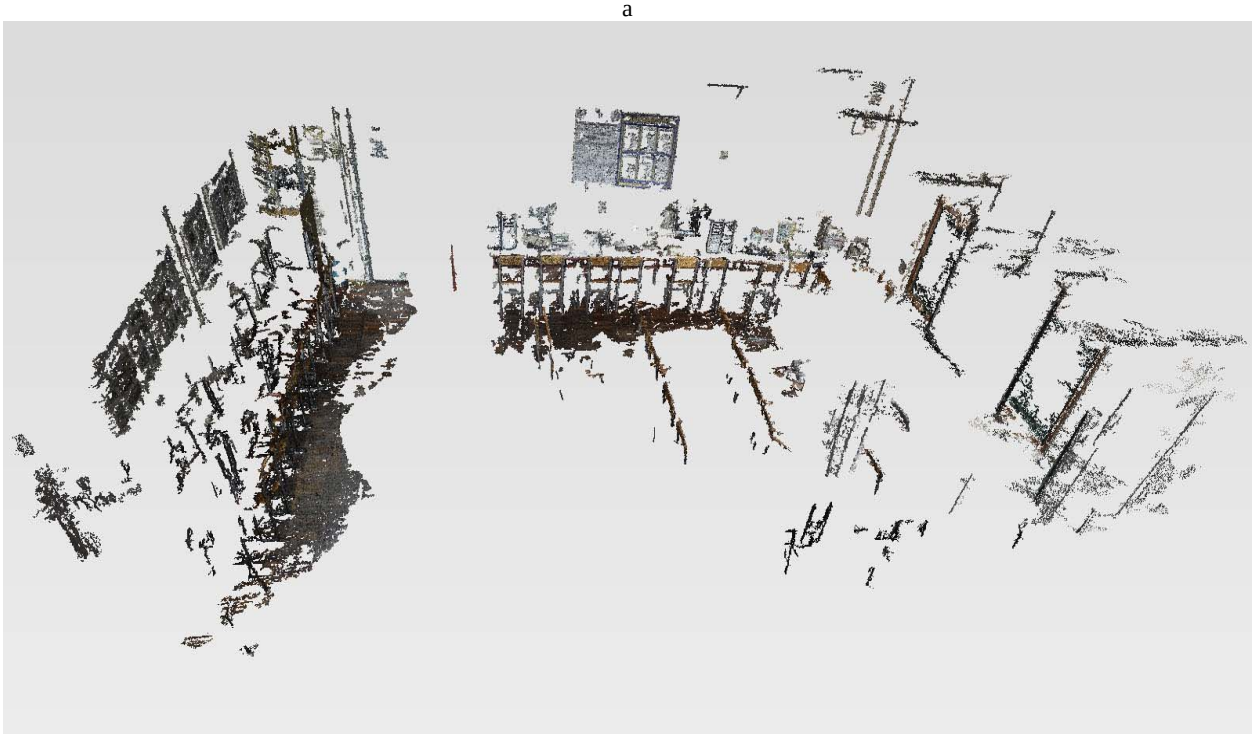
a b c d e f

Figure 1 – The result of the initial stage of creating a 3D model of a separate object:
a, c, e – the result of data processing by the first computing platform;
b, d, f – the result of data processing by the second computing platform

DOI 10.15588/1607-3274-2025-3-3



a



b

Figure 2 – The result of the intermediate stage of creating a generalized 3D model of the surrounding space:
a – the result of data processing by the first computing platform;
b – the result of data processing by the second computing platform

DOI 10.15588/1607-3274-2025-3-3

This approach will make it possible to obtain not only a generalized result of object localization, but also, if necessary, more detailed information about a separate small element of space. The difference in the results of data processing at these stages lies in the details. That is, at the initial stage, we receive more detailed information about small objects, but there is no general information about the position of the mobile object. At the generalized stage, generalized information about individual elements of the surrounding space is obtained, but there is information for accurate localization of the mobile object.

6 DISCUSSION

The results show that when the position of the scanning device changes, that is, a new 3D point cloud is formed and fixed at that position, odometry provides mutually exclusive transformations between them. The intermediate transformations can then be used to augment the delayed-state EKF with additional degrees of freedom. After each state change, successive point clouds can be registered together, using the odometry as an initial estimate, using a 6 degrees of freedom registration algorithm. High-precision transformations can later be used to update the EKF state.

It is recommended to use an integrity check to determine the coincidence of scans approaching false local minima. This factor is especially important when the initial estimates of the transformation are erroneous.

To extend the methods described above, integrity check quality metrics have been added to the registration process itself. This step can deepen the convergence volume reading to the desired convergence properties.

CONCLUSIONS

The urgent task of determining the optimal method and algorithm for building a mathematical 3D model of the surrounding space for automatic localization of a mobile object in space is solved.

The scientific novelty of the obtained results is that a method of three-stage construction of a 3D model of the surrounding space is proposed to solve the problem of localization of a mobile object in a closed space. This method will make it possible to direct information flows about the object's position from different devices, depending on the type of data acquisition, into a centralized information base for solving a wide range of tasks performed by automatic mobile objects (robots). Combining information flows will allow creating a centralized information base, which will not only position the mobile object in space, but also allow mapping and localization on the terrain with great accuracy.

The practical significance of the obtained results is that the proposed method of building a mathematical model of the environment for determining the position of a mobile object in space allows, regardless of the complexity of the task set before the mobile device and the use of limited hardware and software capabilities, to ultimately produce an easy-to-process model of the environment.

Prospects for further research lie in studying the proposed method to extend the capabilities of existing 2D SLAM methods to perform 3D probabilistic SLAM.

ACKNOWLEDGEMENTS

The work is the result of research conducted at the Faculty of information technologies and systems of Cherkasy state technological university, Department of robotics and specialized computer systems.

REFERENCES

1. Azim A., Aycard O. Detection, classification and tracking of moving objects in a 3d environment, *In Intelligent Vehicles Symposium (IV), IEEE*, 2012, pp. 802–807. DOI: 10.1109/IVS.2012.6232303.
2. DosuzhIy O. O., Savchuk T. O. PIdhId do rekonstruktsIYi 3d-modelI zI stereo-zobrazhennya, *MaterIali XLVI naukovo-tehnIchnoYi konferentsIYi pIdrozdIlIv VNTU*, VInnitsya (2017).
3. Lade S. Pawale S., Patil A. GPU Accelerated Simulation of Scene Generation of 3D Photonic Mixer Device Camera, *International Journal on Recent and Innovation Trends in Computing and Communication*, 2023, No. 11, pp. 254–258. DOI**:** 10.17762/ijritcc.v11i9.8341.
4. Beder C., Bartczak B., Koch R. A comparison of pmd-cameras and stereo-vision for the task of surface reconstruction using patchlets, *In Computer Vision and Pattern Recognition. CVPR'07. IEEE Conference on IEEE*, 2007, pp. 1–8. DOI: 10.1109/CVPR.2007.383348.
5. Kuhnert K.-D., Langer M., Stommel M., Kolb A. Dynamic 3D-Vision, *Vision Systems: Applications*, 2007, pp. 311–334. DOI: 10.5772/4995.
6. Reulke R. Combination Of Distance Data With High Resolution Images, *Image Engineering and Vision Metrology (IEVM)*, 2006, pp. 86–92.
7. Langmann B., Hartmann K., Loffeld O. Depth Camera Technology Comparison and Performance Evaluation, *In Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, 2012, pp. 438–444. DOI: 10.5220/0003778304380444.
8. Karimpour O. Implementation of SLAM, Navigation, Obstacle Avoidance, and Path Planning of a Robust Mobile Robot Using 2D Laser Scanner (version 1). Toronto Metropolitan University, 2023. DOI: 10.32920/ryerson.14643939.v1.
9. Sasaki M., Tsuda Y., Matsushita K. Development of Autonomous Mobile Robot with 3DLidar Self-Localization Function Using Layout Map, *Electronics*, 2024, No. 13, P. 1082. DOI: 10.3390/electronics13061082.
10. Marek J., Chmelař P. Survey of Point Cloud Registration Methods and New Statistical Approach, *Mathematics*, 2023, No. 11, P. 3564. DOI: 10.3390/math11163564.
11. Hahnel D., Triebel R., Burgard W., Thrun S. Map building with mobile robots in dynamic environments, *In 2003 IEEE Int. Conf. Robot. Autom. (Cat. No.03CH37422)*, 2003, Volume 2, pp. 1557–1563. DOI**:** 10.1109/ROBOT.2003.1241816.
12. Bay H., Ess A., Tuytelaars T., Gool L. V. Surf: Speeded Up Robust Features, *Computer Vision and Image Understanding*, 2008, No.10, pp. 346–359.
13. Cao L., Zhuang S., Tian S., Zhao Z., Fu C., Guo, Y. Wang D. A Global Structure and Adaptive Weight Aware ICP Algorithm for Image Registration, *Remote Sensing*, 2023, No. 15. 3185. – DOI**:** 10.3390/rs15123185.
14. Magnusson M., Lilienthal A., Duckett T. Scan Registration for Autonomous Mining Vehicles Using 3D-NDT, *Journal of Field Robotics*, 2007, Vol. 24(10), pp. 803–827. DOI: 10.1002/rob.20204.
15. Ulaş C., Temeltaş H. 3D Multi-Layered Normal Distribution Transform for Fast and Long Range Scan Matching, *Journal of Intelligent & Robotic Systems*, 2013, Vol. 71, pp. 85–108. DOI**:** 10.1007/s10846-012-9780-8.
16. Nechyporenko O., Korpan Y. Research of methods and technologies for determining the position of the mobile object in space, *Technology Audit and Production Reserves*, 2018, No. 6 (2(44)), pp. 4–10. DOI: 10.15587/2312-8372.2018.147861.
17. Engelhard N., Endres F., Hess J., Sturm J., Burgard W. Real-time 3D visual SLAM with a hand-held RGB-D camera, *In Proceedings of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum*. Vasteras, Sweden, 6–8 April 2011, pp. 1–15.

DOI 10.15588/1607-3274-2025-3-3

18. Bedkowski J., Röhling T., Hoeller F., Shulz D., Schneider F. E. Benchmark of 6D slam (6D simultaneous localisation and mapping) algorithms with robotic mobile mapping systems, *Foundations of Computing and Decision Sciences*, 2017, Vol. 42(3), pp. 275–295. DOI: 10.1515/fcds-2017-0014.

Received 13.03.2025.
Accepted 26.06.2025.

УДК 004.93

МЕТОДИ ТА АЛГОРИТМИ ПОБУДОВИ МАТЕМАТИЧНОЇ 3D-МОДЕЛІ НАВКОЛИШНЬОГО ПРОСТОРУ ДЛЯ АВТОМАТИЧНОЇ ЛОКАЛІЗАЦІЇ МОБІЛЬНОГО ОБ'ЄКТА

Корпань Я. В. – канд. техн. наук, доцент, доцент кафедри робототехніки та спеціалізованих комп'ютерних систем Черкаського державного технологічного університету, Черкаси, Україна.

Нечипоренко О. В. – канд. техн. наук, доцент, доцент кафедри інформаційної безпеки та комп'ютерної інженерії Черкаського державного технологічного університету, Черкаси, Україна

Федоров Є. Є. – д-р техн. наук, доцент, професор кафедри статистики та прикладної математики Черкаського державного технологічного університету, Черкаси, Україна.

Уткіна Т. Ю. – канд. техн. наук, доцент, доцент кафедри робототехніки та спеціалізованих комп'ютерних систем Черкаського державного технологічного університету, Черкаси, Україна.

АНОТАЦІЯ

Актуальність. Розглянуто задачу автоматизації позиціонування мобільного об'єкта в замкненому просторі при умові його часткової або повної автономності. Об'єктом дослідження є процес автоматичної побудови 3D-моделі навколишнього простору.

Мета роботи – розробка методу створення 3D моделі навколишнього простору для подальшої локалізації мобільного об'єкта в умовах його часткової або повної автономності.

Метод. Приведено результати дослідження проблеми локалізації мобільного об'єкта в просторі в реальному часі. Приведено результати аналізу існуючих методів та алгоритмів створення математичних моделей навколишнього простору. Описані алгоритми, які широко використовуються для вирішення проблеми локалізації мобільного об'єкта в просторі. Проведено дослідження широкого спектру методів побудови математичної моделі навколишнього простору – від методів, які використовують співставлення послідовних хмарин точок об'єкта навколишнього простору до методів, які використовують серії знімків характеристичних точок та порівнянні інформації про них на різних знімках в точках, максимально схожих за вектором параметрів.

Результати. Запропоновано метод трьохетапної побудови 3D моделі навколишнього простору для вирішення задачі локалізації мобільного об'єкта в замкненому просторі.

Висновки. Проведені експерименти підтвердили можливість запропонованого алгоритму трьохетапної побудови математичної моделі навколишнього середовища для визначення положення мобільного об'єкта у просторі. Методи, які використовуються в алгоритмі дозволяють отримати інформацію про навколишній простір, що дозволяє провести локалізацію мобільного об'єкту в замкненому просторі. Перспективи подальших досліджень можуть полягати в інтеграції інформаційних потоків про положення об'єкта з різних, за типом отримання даних, приладів в централізовану інформаційну базу для вирішення широкого спектру задач, які виконують автоматичні мобільні об'єкти (роботи).

КЛЮЧОВІ СЛОВА: математична 3D-модель, метод локалізації, алгоритми методу SLAM, визначення положення, мобільний об'єкт.

ЛІТЕРАТУРА

1. Azim A. Detection, classification and tracking of moving objects in a 3d environment / A. Azim, O. Aycard // In Intelligent Vehicles Symposium (IV), IEEE. – 2012. – P. 802–807. DOI: 10.1109/IVS.2012.6232303.
2. DosuzhIy O. O. PIdhId do rekonstruktsIYi 3d-modelI zI stereo-zobrazhennya / O. O. DosuzhIy, T. O. Savchuk // MaterIali XLVI naukovo-tehnIchnoYi konferentsIYi pIdrozdIlIv VNTU, VInnitsya (2017).
3. Lade S. GPU Accelerated Simulation of Scene Generation of 3D Photonic Mixer Device Camera. / S. Lade, S. Pawale, A. Patil // International Journal on Recent and Innovation Trends in Computing and Communication. – 2023. – No. 11. –P. 254–258. DOI: 10.17762/ijritcc.v11i9.8341.
4. Beder C. A comparison of pmd-cameras and stereo-vision for the task of surface reconstruction using patchlets / C. Beder, B. Bartczak, R. Koch // In Computer Vision and Pattern Recognition. CVPR'07. IEEE Conference on. IEEE. – 2007. – P. 1–8. DOI: 10.1109/CVPR.2007.383348.
5. Dynamic 3D-Vision / [K.-D. Kuhnert, M. Langer, M. Stommel, A. Kolb] // Vision Systems: Applications. – 2007. P. 311–334. DOI: 10.5772/4995.
6. Reulke R. Combination Of Distance Data With High Resolution Images / R. Reulke // Image Engineering and Vision Metrology (IEVM). – 2006. – P. 86–92.
7. Langmann B. Depth Camera Technology Comparison and Performance Evaluation. / B. Langmann, K. Hartmann, O. Loffeld // In Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods. – 2012. – P. 438–444. DOI: 10.5220/0003778304380444.
8. Karimpour O. Implementation of SLAM, Navigation, Obstacle Avoidance, and Path Planning of a Robust Mobile Robot Using 2D Laser Scanner (version 1) / O. Karimpour. – Toronto Metropolitan University, 2023. DOI: 10.32920/ryerson.14643939.v1.
9. Sasaki M. Development of Autonomous Mobile Robot with 3DLidar Self-Localization Function Using Layout Map / M. Sasaki, Y. Tsuda, K. Matsushita // Electronics. – 2024. – No. 13. – P. 1082. DOI: 10.3390/electronics13061082.
10. Marek J. Survey of Point Cloud Registration Methods and New Statistical Approach. / J. Marek, P. Chmelař // Mathematics. – 2023. – No. 11. – P. 3564. DOI: 10.3390/math11163564.
11. Map building with mobile robots in dynamic environments. / [D. Hahnel, R. Triebel, W. Burgard, S. Thrun] // In 2003 IEEE Int. Conf. Robot. Autom. (Cat. No.03CH37422). – 2003. – Vol. 2. – P. 1557–1563. DOI: 10.1109/ROBOT.2003.1241816.
12. Surf: Speeded Up Robust Features / [H. Bay, A. Ess, T. Tuytelaars, L. V. Gool] // Computer Vision and Image Understanding. – 2008. – No. 10. –P. 346–359.
13. Cao L. A Global Structure and Adaptive Weight Aware ICP Algorithm for Image Registration. / [L. Cao, S. Zhuang, S. Tian et al.] // Remote Sensing. – 2023. – No. 15. – P. 3185. – DOI: 10.3390/rs15123185.
14. Magnusson M. Scan Registration for Autonomous Mining Vehicles Using 3D-NDT / M. Magnusson, A. Lilienthal, T. Duckett // Journal of Field Robotics. – 2007. – Vol. 24(10). –P. 803–827. DOI: 10.1002/rob.20204.
15. Ulaş C. 3D Multi-Layered Normal Distribution Transform for Fast and Long Range Scan Matching / C. Ulaş, H. Temeltaş // Journal of Intelligent & Robotic Systems. – 2013. – Vol. 71. – P. 85–108. DOI: 10.1007/s10846-012-9780-8.
16. Nechyporenko O. Research of methods and technologies for determining the position of the mobile object in space / O. Nechyporenko, Y. Korpan // Technology Audit and Production Reserves. – 2018. – No. 6(2(44)). – P. 4–10. DOI: 10.15587/2312-8372.2018.147861.
17. Real-time 3D visual SLAM with a hand-held RGB-D camera / N. Engelhard, F. Endres, J. Hess et al.] // In Proceedings of the RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum. – Vasteras, Sweden, 6–8 April 2011. – P. 1–15.
18. Benchmark of 6D slam (6D simultaneous localisation and mapping) algorithms with robotic mobile mapping systems. / [J. Bedkowski, T. Röhling, F. Hoeller et al.] // Foundations of Computing and Decision Sciences. – 2017. – Vol. 42(3). – P. 275–295. DOI: 10.1515/fcds-2017-0014.

DOI 10.15588/1607-3274-2025-3-3

UDC 519.2+004.8

THE METHOD OF ADAPTATION OF THE PARAMETERS OF ALGORITHMS FOR THE DETECTION AND CLEANING OF A STATISTICAL SAMPLE FROM ANOMALIES FOR DATA SCIENCE PROBLEMS

Pysarchuk O. O. – Dr. Sc., Professor, Professor of the Department of Computer Engineering, Faculty of Informatics and Computing, National Technical University of Ukraine "Ihor Sikorskyi Kyiv Polytechnic Institute".

Pavlova S. O. – Student of the Faculty of Informatics and Computing, National Technical University of Ukraine "Ihor Sikorskyi Kyiv Polytechnic Institute".

Baran D. R. – Assistant of the Department of Computer Engineering, Faculty of Informatics and Computing, National Technical University of Ukraine "Ihor Sikorskyi Kyiv Polytechnic Institute".

ABSTRACT

Context. Popularization of the Data Science for the tasks of e-commerce, the banking sector of the economy, for the tasks of managing dynamic objects – all this actualizes the requirements for indicators of the efficiency of data processing in the Time Series format. This also applies to the preparatory stage of data analysis at the level of detection and cleaning of statistical samples from anomalies such as rough measurements and omissions.

Objective. The development of the method for adapting the parameters of the algorithms for detecting and cleaning the statistical sample of the Time Series format from anomalies for Data Science problems.

Method. The article proposes a method for adapting the parameters of algorithms for detecting and cleaning a statistical sample from anomalies for data science problems. The proposed approach is based on and differs from similar practices by the introduction of an optimization approach in minimizing the dynamic and statistical error of the model, which determines the parameters of settings of popular algorithms for cleaning the statistical sample from anomalies using the Moving Window Method.

Result. The introduction of the proposed approach into the practice of Data Science allows the development of software components for cleaning data from anomalies, which are trained by parameters purely according to the structure and dynamics of the Time Series.

Conclusions. The key advantage of the proposed method is its simple implementation into existing algorithms for clearing the sample from anomalies and the absence of the need for the developer to select parameters for the settings of the cleaning algorithms manually, which saves time during development. The effectiveness of the proposed method is confirmed by the results of calculations.

KEYWORDS: anomaly detection, dynamic error, statistical error, model optimization, Moving Window, Data Science, Big Data, time series.

ABBREVIATIONS

AM is an Abnormal Measurements;

ARIMA is an Autoregressive Integrated Moving Average;

MAE is a mean absolute error.

NOMENCLATURE

θ is a dynamic error;

σ is a standard deviation (square root of variance);

σ^2 is a variance of the error sample ($D[y \mid x]$);

σ_y^2 is a variance of the dependent variable sample ($D[y]$);

$D[y]$ is a variance of the sample;

$D[y \mid x]$ is a conditional variance of the dependent variable given factors x (variance of the model error);

e_i is a model error;

n is a number of observations in the sample;

R^2 is a coefficient of determination;

$threshold$ is a threshold for anomaly detection;

$window\_size$ is a size of the sliding window;

x_i is an predicted values of the variable;

y_i is an actual values of the variable.

INTRODUCTION

At present, Data Science tasks have gained immense popularity, as they allow the use of large amounts of data (Big Data) to obtain valuable information and make informed decisions [1–14]. It should be noted that this trend has been maintained for many years, which is due to the development of information technologies and their implementation in many areas.

One of the areas of Data Science is the processing of data in the format of a time series, which characterizes the studied processes with a discrete series of values that change in time or depending on another argument (variable). Examples of time series processing tasks are the analysis of changes in time and forecasting: indicators of economic efficiency of trading companies; weather indicators; changes in exchange rate fluctuations; global statistical indicators of the state's economy – production of agricultural products, population growth, subsistence minimum, morbidity of the population, etc.; navigation parameters of the movement of dynamic objects – airplanes, cars, robotic/unmanned aerial vehicles and many other industries.

DOI 10.15588/1607-3274-2025-3-4

These examples are focused on high accuracy of time series processing. This is achieved by considering the heterogeneity of the input data due to the presence of abnormal measurements (AM). The problem of clearing the time series from AM is quite common [1–3, 6–10]. Since, depending on the ratio of the number and magnitude of AM to the number of measurements in the time series, anomalies can significantly distort the processing results. However, this is an additional stage, which on Big Data is critical to the conflict of attracting resource space and the efficiency of obtaining the result. Therefore, more often they prefer simple but effective algorithms built on the principles of a sliding window [3, 4, 8–14]. Here, simplicity is a positive and negative property at the same time. The negative is manifested in fixing the parameters of such algorithms. But this does not allow you to adapt to dynamic data properties. This phenomenon is significantly manifested by data with significant nonlinearities, seasonal variations, etc. That is, the algorithms for clearing the time series from AM with fixed parameters are fast, but "blind" to the dynamics of data changes. This leads to the need to support program implementations of such approaches, which is not always justified and possible.

One of the basic approaches to time series processing is statistical training methods. But they apply pre-prepared data through AM cleanup.

In connection with the above, the task of developing effective (in terms of speed and accuracy) approaches to adapting the parameters of algorithms for detecting and cleaning the statistical sample from anomalies for data science problems **is relevant**.

The object of study is the process of purifying the statistical sample from anomalous measurements

The subject of study is methods for cleaning the statistical sample from anomalous measurements.

The purpose of the work is to develop a method for adapting the parameters of algorithms for detecting and cleaning the statistical sample of the Time Series format from anomalies for Data Science problems.

1 PROBLEM STATEMENT

Time series processing methods are quite common and are represented by algorithms such as ARIMA, regression analysis and statistical training (such as the method of least squares (LSM) and others), deep learning using artificial neural networks [2–4]. The quality of application of all these approaches is largely determined by the quality of data preparation for processing. One of the stages of data preparation is to clean them from anomalous measurements – those that differ significantly in their values from other measurements and disrupt the dynamics of the time series, as well as data omissions. Depending on the absolute values of AM and the ratio of the number of AMs to the sample size of the time series, anomalies can distort the processing results quite strongly [6–9]. Therefore, in the process of data preparation, it is necessary to provide for the stage of clearing the sample of measurements from AM. In turn, the process of clearing time series from AM is and remains one of the most difficult and time-consuming tasks in the field of Data Science. This is due to the complex nature of the reasons for the appearance of AM and their negative impact on the results of processing. At the same time, quite high requirements for performance are put forward to the algorithms for clearing time series from AM (especially on Big Data arrays) and to autonomous adaptation (adaptation of parameters by "self-learning" depending on the properties of the Time Series – the nature of the trend, statistical characteristics, etc.).

Let us assume that a set of measurements y_i ,that form the Time Series. It is known that the measurements are distributed with normal law and a contain certain percentage of anomalous measurements. Detection of anomalous measurements is carried out using a sliding window algorithm, the efficiency of which is determined by the parameters $threshold$ – the anomaly detection threshold and $window\_size$ – the size of the sliding window.

It is necessary to develop a method for adapting the parameters of the algorithms for detecting and cleaning the statistical sample from anomalies – $threshold$, $window\_size$ to the properties of a specific sample of measurements.

Criteria and limitations. The method under development should ensure the minimization of dynamic θ and stochastic σ estimation of errors on a limited set of time series measurements n .

2 REVIEW OF THE LITERATURE

In the problems of clearing the sample from anomalies, there are quite a lot of varieties of methods and algorithms based on different approaches and principles [6–14]. All known approaches are based on unitary and/or combinatorial analysis of AM features. In general, there are AM of the rough dimensions and AM of the omission type. In both cases, the signs of AM are a change in the dynamics of the time series (dynamic properties); a difference in the value of a single dimension compared to other dimensions (properties of absolute values measurements); changes in the statistical properties of the sample in the presence of AM (statistical properties).

Depending on the signs used to detect AM and the principles of their detection, the following classes should be distinguished:

– methods of clustering according to the principles of machine learning [5, 6];

– methods for analyzing the dynamic properties of the time series [8];

– methods for analyzing the statistical properties of the time series [8, 9].

Despite the versatility and wide representation of these approaches, their key drawback is the empirical (research) adjustment of their parameters, depending on the nature of the properties of the time series. This may not be acceptable,

DOI 10.15588/1607-3274-2025-3-4

as finding the best solutions can take a significant amount of time during the development phase. It also complicates their practical implementation and scalability for time series with a wide range of properties that sometimes change during the operation of the software system. The disadvantages of known approaches to training according to the parameters of algorithms for clearing the sample from AM are also in the complexity of their implementation on Big Data arrays with significant nonlinearities and seasonalities.

3 MATERIALS AND METHODS

The method under development is aimed at supplementing the known time series cleaning algorithms based on the principle of a sliding window, for example: Moving Window Method, Median Filtering algorithm or Least Squares Method [9].

The main idea of the proposed method is as follows.

The parameters to be determined are the size of the sliding window and the threshold value (sensitivity) of the algorithms for detecting and cleaning the time series from the anomalies. These parameters are determined from the list of discrete values that ensure a minimum of dynamic and statistical error in the model of the results of statistical selection after cleaning the time series from the anomalies. The method of statistical learning is used as the Least Squares Method [9].

It is advisable to put forward the following requirements for the method of adaptation of the parameters of the algorithms for detecting and cleaning the statistical sample from anomalies:

1) The use of the method of parameter adaptation should lead to an improvement in the results of cleaning the sample from anomalies in accordance with the quality metrics of the statistical learning model given below.

2) The method of parameter adaptation should be based on the choice of a statistical learning model with the minimum combination of dynamic and statistical error.

3) Sample cleaning by the developed method should not remove structurally important properties of the sample.

We will introduce model quality indicators to understand how successful data cleaning from anomalies was. We will take the mean absolute error (MAE) and the coefficient of determination (R^2) as such metrics.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} = \frac{\sum_{i=1}^{n}|e_i|}{n}, \quad (1)$$

$$R^2 = 1 - \frac{D[y \mid x]}{D[y]} = 1 - \frac{\sigma^2}{\sigma_y^2}, \quad (2)$$

where $D[y] = e_y^2$ is the variance of the random error of the measured sample y, and $D[y \mid x] = \sigma^2$ is the conditional (by factors x) variance of the dependent variable (variance of the model error).

Considering the above requirements and model quality indicators, a method for adapting the parameters of algorithms for detecting and cleaning a statistical sample from anomalies has been developed, based on minimizing the dynamic and statistical error of the statistical learning model.

A method for adapting the parameters of the algorithms for detecting and cleaning up the sample anomalies.

The size of the sliding window and the threshold for detecting the anomalies are subject to adaptation based on the characteristics of the input sample. This is done by finding a balance between the dynamic and statistical errors of the statistical learning model [3–5]. The dynamic error is the previously defined metric of mean absolute error (MAE), and the statistical error is the coefficient of determination (R^2).

For a representative sample, the mean absolute error is minimal, and the coefficient of determination is close to one. A small MAE value guarantees minimal discrepancy between data without anomalies and the results of anomaly removal algorithms. An R^2 value close to one means that the model reproduces the data well and considers all its variability.

If the parameters of the algorithms for detecting and cleaning the sample from the outliers are incorrectly defined, this will result in the outliers remaining in a posteriori sample. The presence of anomalies in the sample will lead to an increase in the mean absolute error (MAE) and a decrease in the coefficient of determination (R^2), which can be used as feedback for evaluating the next combination of parameters of the algorithms for detecting and cleaning the sample from the outliers.

Thus, the problem of parameter adaptation is reduced to minimizing the result of the expression calculation: $MAE + (1 - R^2)$ over the course of the values of the parameters of the algorithms for detecting and cleaning the sample from the anomalies. Reducing the result of calculating this expression means that the model has smaller errors (low MAE) and at the same time explains the data well (high R^2). This is a consequence of the quality of the algorithm for cleaning the sample from outliers.

The stages of the method of adapting the parameters of the algorithms for detecting and cleaning the sample from the anomalies include the following.

1. Determine the range for optimizing the $window\_size$ and $threshold$ parameters within the specified limits. The boundaries, i.e. the minimum and maximum values of $window\_size$ and $threshold$, are determined using the statistical parameters of the input sample – sample size, standard deviation, etc.

a) The $window\_size$ parameter affects how many neighboring values will be considered during data cleaning. Determining the optimal window size allows you to balance data smoothing and detail preservation.

DOI 10.15588/1607-3274-2025-3-4

b) The *threshold* parameter defines the acceptable level of deviation for anomaly detection. Values above this threshold are considered anomalies. Determining the optimal threshold allows you to effectively detect and eliminate anomalies without unnecessarily deleting correct data.

2. A nested loop is executed for *window_size* and *threshold*. At each iteration of the loop, one of all possible combinations of *window_size* and *threshold* within the previously defined limits is considered.

3. For each combination of *window_size* and *threshold*, one of the following data cleaning algorithms is used: Moving Window Method, Median Filtering, or Least Squares Method.

4. For each combination of *window_size* and *threshold*, the $MAE + (1 - R^2)$ values are calculated for the original and cleaned data.

5. For each combination of *window_size* and *threshold*, the $MAE + (1 - R^2)$ values of the current combination are compared with the best values of the previous combinations. If the current values are better (less *MAE* and more R^2), they become the best values.

6. The result is a combination of *window_size* and *threshold* parameters with the lowest $MAE + (1 - R^2)$ value.

To implement these stages of the method of adapting the parameters of the algorithms for detecting and cleaning the sample from anomalies, a software script was developed in the Python programming language with the numpy [11], pandas [10], and matplotlib libraries.

To evaluate the effectiveness of the proposed solutions, several computational experiments were conducted. The essence of the experiments is to process a stochastic sample with anomalies by a known algorithm and an algorithm using the developed method. The analysis of the results was carried out by comparing the initial and final characteristics of the sample obtained using the traditional and the proposed approaches.

4 EXPERIMENTS

We will conduct a series of experiments to evaluate the effectiveness of the method of adapting the parameters of the Moving Window Method [9], Least Squares Method, and Median Filtering algorithms for the task of cleaning the sample from anomalies.

A statistical sample of $n = 21$ measurements was subject to modeling. The basis was real data: statistics on Russian army losses for 1–21 September 2023. The data is presented by category: personnel, armored combat vehicles, tanks, artillery, aircraft, helicopters, ships.

For the Mowing Window Method, the standard deviation in the input sample is $\sigma = 0.2800$, and the dynamic error is $\theta = 0.2287$. After modeling the addition of 10% of anomalies, which are uniformly distributed over the sample, we have the following characteristics of the statistical sample: $\sigma = 0.3763, \theta = 0.2863$.

For the Least Squares Method, the input sample contained: $\sigma = 0.2669$, $\theta = 0.2033$. The sample with anomalies: $\sigma = 0.4115$, $\theta = 0.3209$.

For the Median Filtering algorithm, the input sample contained: $\sigma = 0.2710$, $\theta = 0.2035$. Sample with anomalies: $\sigma = 0.3463$, $\theta = 0.2705$.

5 RESULTS

The results of the study of the method of parameter adaptation based on the Mowing Window Method are shown in Fig. 1.

Fig. 1a shows the sample plot (dependence of the value of the controlled parameter “Values” on time “Time”) after using the well-known Moving Window Method: $\sigma = 0.1840$, $\theta = 0.1398$. The model quality indicators mean absolute error $MAE = 0.2876$, coefficient of determination $R^2 = 0.7649$.

Instead, Fig. 1b shows the sample plot after using the developed method of parameter adaptation, which has error values: $\sigma = 0.2535$, $\theta = 0.1933$. The following model quality indicators were obtained: $MAE = 0.2569$, $R^2 = 0.7761$.

The statistical characteristics show that the algorithm without a method of parameter adaptation also removes structurally important data. While the proposed approach allows preserving the structure of the input sample and provides better model accuracy.

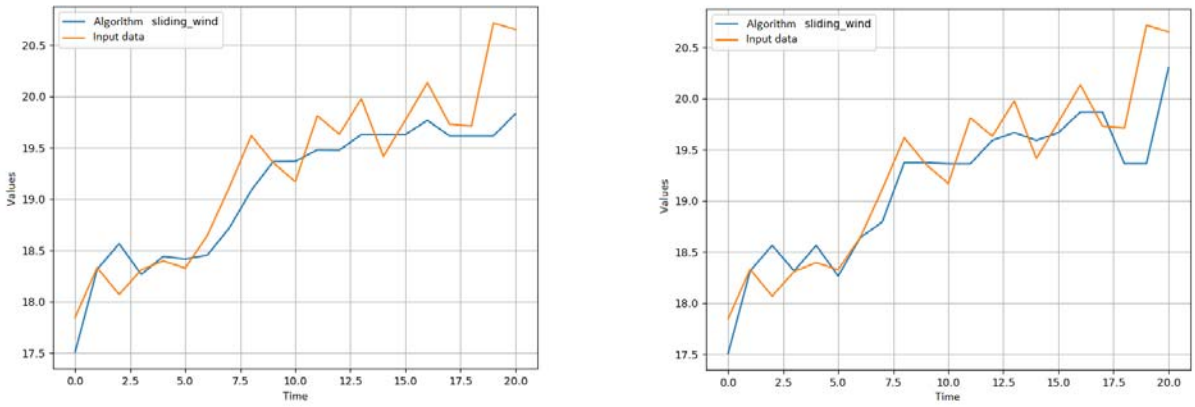
The results of the study of the method of parameter adaptation based on Least Squares Method are shown in Fig. 2, where the notation is like that of Fig. 1.

The use of the well-known Least Squares Method algorithm gave the following results: $\sigma = 0.1220$, $\theta = 0.0694$. When applying the developed method, we have: $\sigma = 0.0845, \theta = 0.0304$.

Comparison of the graphs of Fig. 2, and Fig. 2 b and the model quality criteria indicate that the sample is still representative when using the developed adaptation method. While the well-known Least Squares Method algorithm focuses more on the initial values of the sample.

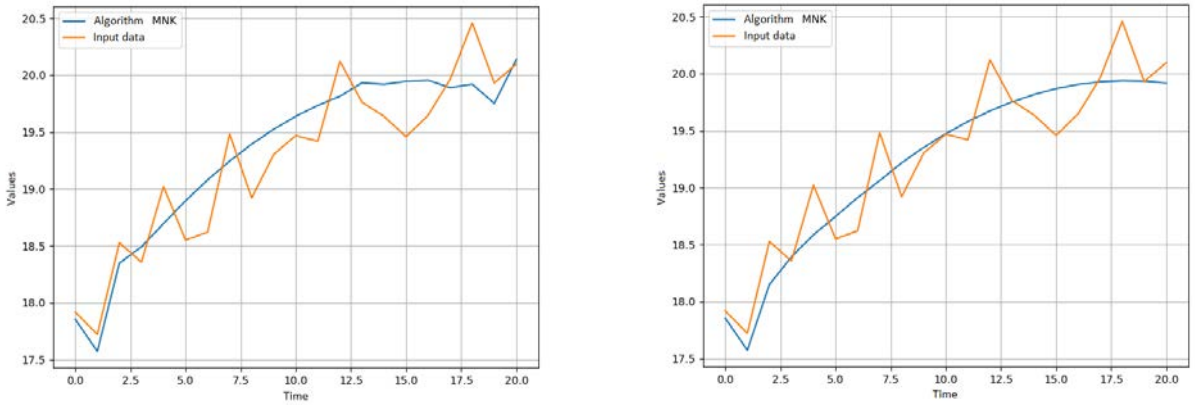
The results of the study of the method of parameter adaptation based on the Median Filtering are presented in Fig. 3, where the notation is like that of Fig. 1. The well-known Median Filtering showed the following results: $\sigma = 0.1653, \theta = 0.1229$. Whereas the optimized algorithm is: $\sigma = 0.1910, \theta = 0.1532$. Comparison of the graphs of Fig. 3a and Fig. 3b also demonstrates a decrease in the average absolute error and increase in the coefficient of determination when using the proposed approach, which indicates its effectiveness.

DOI 10.15588/1607-3274-2025-3-4



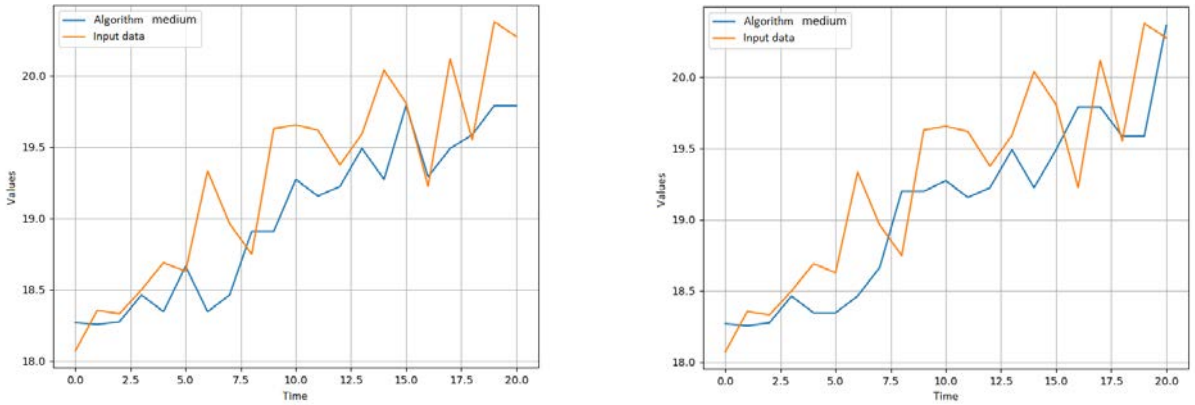
a b

Figure 1 – Results of the study of the method of parameter adaptation based on the Moving Window Method through comparison of the input and cleaned samples: a –using a well-known Moving Window Method ($MAE = 0.2876$, $R^2 = 0.7649$), b – using adaptation of Moving Window Method ($MAE = 0.2569$, $R^2 = 0.7761$)



a b

Figure 2 – The results of the study of the method of parameter adaptation based on the Least Squares Method: a – using well-known Least Squares Method ($MAE = 0.2598$, $R^2 = 0.8353$), b – using adaptation of Least Squares Method parameters ($MAE = 0.2154$, $R^2 = 0.8609$)



a b

Figure 3 – Results of the study of the method of parameter adaptation based on the Median Filtering: a – using well-known Median Filtering ($MAE = 0.3243$, $R^2 = 0.5809$), b – using adapting the parameters of the Median Filtering ($MAE = 0.2630$, $R^2 = 0.7685$)

DOI 10.15588/1607-3274-2025-3-4

Table 1 – Generalized statistical characteristics of the approaches under study

Algorithm	Input sample	A sample with anomalies	Standard algorithm	Minimization method
Moving Window Method	$\sigma = 0.2800$, $\theta = 0.2287$	$\sigma = 0.3763$, $\theta = 0.2863$	$\sigma = 0.1840$, $\theta = 0.1398$	$\sigma = 0.2535$, $\theta = 0.1933$
Least Squares Method	$\sigma = 0.2669$, $\theta = 0.2033$	$\sigma = 0.4115$, $\theta = 0.3209$	$\sigma = 0.1220$, $\theta = 0.0694$	$\sigma = 0.0845$, $\theta = 0.0304$
Median Filtering	$\sigma = 0.2710$, $\theta = 0.2035$	$\sigma = 0.3463$, $\theta = 0.2705$	$\sigma = 0.1653$, $\theta = 0.1229$	$\sigma = 0.1910$, $\theta = 0.1532$

Table 2 – Summary of model quality indicators

Algorithm	Standard algorithm	Minimization method
Moving Window Method	$MAE = 0.2876$, $R^2 = 0.7649$	$MAE = 0.2569$, $R^2 = 0.7761$
Least Squares Method	$MAE = 0.2598$, $R^2 = 0.8353$	$MAE = 0.2154$, $R^2 = 0.8609$
Median Filtering	$MAE = 0.3243$, $R^2 = 0.5809$	$MAE = 0.2630$, $R^2 = 0.7685$

6 DISCUSSION

A summary of the statistical characteristics of the approaches studied is presented in Table 1. The generalized quality indicators of the models are presented in Table 2. The analysis of the data in Tables 1 and 2 allows us to conclude that the application of the developed method of parameter adaptation for the Least Squares Method algorithm is not the best choice to preserve the statistical properties of the sample. Figures 2.a and 2.b show excessive smoothing of the data and, accordingly, the loss of their features, which is also demonstrated by the results presented in Table 1. However, even with such a loss of features, the use of the developed approach demonstrates the best quality of the model among the three algorithms considered the properties of time series [9]. The calculation results have proved the effectiveness of the proposed approach.

It is worth noting that the values of the dynamic and random component errors in the estimates of the controlled parameters after the applied solutions are sufficient to be no worse than the known analogues. This statement is true because the main advantage of the proposed approach is the adaptation of the parameters of the anomaly detection methods to the properties of the input sample. Therefore, the fact of preserving accuracy along with the adaptive properties of the proposed approach is evidence of achieving the goal and conditions and limitations of the pre-face part of the work.

CONCLUSIONS

The work solves the problem of developing a method for adapting the parameters of algorithms for detecting and cleaning statistical samples from anomalies for data science tasks.

The scientific novelty of the obtained results lies in the implementation of an optimization approach in minimizing the dynamic and statistical error of the model, which determines the parameters (sliding window size and sensitivity coefficient) of known algorithms for cleaning statistical samples from anomalies according to the principles of the sliding window.

The practical value of the proposed solution for Data Science tasks lies in the possibility of developing software components for cleaning data from anomalies, which are trained according to the parameters taking into account the structure and dynamics of changes in the time series. At the same time, high accuracy rates of estimation for the dynamic and stochastic components of errors are maintained. The advantage of the proposed method is also its simplicity and implementation into existing algorithms.

Prospects for further research lie in expanding the list of anomaly indicators (for example, to dynamic and influential) for multifactorial optimization of the parameters of detection algorithms.

ACKNOWLEDGEMENTS

These studies were conducted for educational purposes at the Department of Computer Engineering at the Faculty of Informatics and Computer Engineering of the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”.

The results of these studies were used in scientific research project “Nonlinear and multicriterial mathematical models for Data Science and Embedded Systems technologie” (state registration number 0124U003323).

REFERENCES

1. Kumar J., Kumar A., Kumar R. Big Data and Analytics: The key concepts and practical applications of big data analytics. BPB Publications, 2024, 232 p.
2. Dietrich D., Heller B., Yang B. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Present-

DOI 10.15588/1607-3274-2025-3-4

ing Data. Indianapolis, Indiana, John Wiley & Sons, 2015, 420 p.
3. Provost F., Fawcett T. Data Science for Business. New York: O'Reilly Media, Inc, 2013, 409 p.
4. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition. New York, Springer, 2017, 763 p.
5. Brockwell P. J., Davis R. A. Introduction to Time Series and Forecasting. New York, Springer, 2016, 439 p.
6. Tuhanskykh O., Baran D., Pysarchuk O. Method for Statistical Evaluation of Nonlinear Model Parameters in Statistical Learning Algorithms, *Proceedings of Ninth International Congress on Information and Communication Technology*. – Springer, Singapore, 2024, No. 1013, pp. 265–274. (Series "Lecture Notes in Networks and Systems"). DOI: 10.1007/978-981-97-3559-4_21
7. Nassif A. B., Talib M. A., Nasir Q., Dakalbab F. M. Machine Learning for Anomaly Detection: A Systematic Review, *IEEE Access*, 2021, No. 9, pp. 78658–78700. DOI: 10.1145/3439950.
8. Pang G., Shen C., Cao L., Hengel A. Deep Learning for Anomaly Detection, *ACM Computing Surveys*, 2021, Vol. 54(2), pp. 1–38. DOI: 10.1145/3439950.
9. Song X., Wu M., Jermaine C., Ranka S. Conditional Anomaly Detection, *IEEE Transactions on Knowledge and Data Engineering*, 2007, No. 19, pp. 631–645. DOI: 10.1109/TKDE.2007.1009.
10. Pysarchuk O., Baran D., Mironov Y., Pysarchuk I. Algorithms of statistical anomalies clearing for data science applications, *System research and information technologies.* – 2023, No. 1, pp. 78–84. DOI: 10.20535/SRIT.2308-8893.2023.1.06.
11. Mehrotra K. G., Mohan C. K., Huang H. Anomaly Detection Principles and Algorithms. Switzerland, Springer, 2017, 229 p.
12. McKinney W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media, 2017, 550 p.
13. Nelli F. Python Data Analytics: With Pandas, NumPy, and Matplotlib, 2nd ed. Edition. Apress, 2018, 588 p.
14. Raschka S., Mirjalili V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, Second Edition. Packt Publishing, 2017, 622 p.
15. Joshi P. Artificial Intelligence with Python: A Comprehensive Guide to Building Intelligent Apps for Python Beginners and Developers. Packt Publishing, 2017, 466 p.

Received 04.04.2025.
Accepted 29.06.2025.

УДК 004.5

СПОСІБ АДАПТАЦІЇ ПАРАМЕТРІВ АЛГОРИТМІВ ВИЯВЛЕННЯ ТА ОЧИЩЕННЯ СТАТИСТИЧНОЇ ВИБІРКИ ВІД АНОМАЛІЙ ДЛЯ ЗАДАЧ DATA SCIENCE

Писарчук О. О. – д-р техн. наук, професор, професор кафедри обчислювальної техніки, факультету інформатики та обчислювальної техніки, Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського».

Павлова С. О. – студентка факультету інформатики та обчислювальної техніки, Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського».

Баран Д. Р. – асистент кафедри обчислювальної техніки факультету інформатики та обчислювальної техніки, Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського».

АНОТАЦІЯ

Актуальність. Популяризація задачі Data Science для завдань електронної комерції, банківського сектору економіки, для задач управління динамічними об'єктами – актуалізує вимоги до показників ефективності обробки даних формату Time Series.Зазначене стосується і підготовчого етапу аналізу даних на рівні виявлення та очищення статистичних вибірок від аномалій типу грубі виміри та пропуски.

Метою роботи є розробка способу адаптації параметрів алгоритмів виявлення та очищення статистичної вибірки формату Time Series від аномалій для задач Data Science.

Метод. У статті запропоновано спосіб адаптації параметрів алгоритмів виявлення та очищення статистичної вибірки від аномалій для задач data science. Запропонований підхід базується та відрізняється від аналогічних практик запровадженням оптимізаційного підходу в мінімізації динамічної та статистичної похибки моделі, що визначає параметри налаштувань популярних алгоритмів очищення статистичної вибірки від аномалій з використанням ковзного вікна (Moving Window Method).

Результат. Запровадження запропонованого підходу в практику Data Science дозволяє розробляти програмні компонентидля очищення даних від аномалій, що навчаються за параметрами суто за структурою та динамікою Time Series. Це забезпечує підтримку широкого кола задач з нелінійними властивостями та сезонними закономірностями у даних. Отже спрощується процес супроводження подібних продуктів після впровадження їх в практику застосування.

Висновки. Ключовою перевагою запропонованого методу є його проста імплементації в існуючі алгоритми очищення вибірки від аномалій та відсутність необхідності розробнику підбирати параметри налаштувань алгоритмів очищення вручну, що економить час при розробці. Ефективність запропонованого способу підтверджується результатами розрахунків.

КЛЮЧОВІ СЛОВА: аномальні виміри, динамічна похибка, статистична похибка, оптимізація моделі, Moving Window, Data Science, Big Data, Time Series.

DOI 10.15588/1607-3274-2025-3-4

ЛІТЕРАТУРА

1. Kumar J. Big Data and Analytics: The key concepts and practical applications of big data analytics / J. Kumar, A. Kumar, R. Kumar. – BPB Publications, 2024. – 232 p.
2. Dietrich D. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data / D. Dietrich, B. Heller, B. Yang. – Indianapolis, Indiana : John Wiley & Sons, 2015. – 420 p.
3. Provost F. Data Science for Business / F. Provost, T. Fawcett. – New York : O'Reilly Media, Inc, 2013. – 409 p.
4. Hastie T. The Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition / T. Hastie, R. Tibshirani, J. Friedman. – New York : Springer, 2017. – 763 p.
5. Brockwell P. J. Introduction to Time Series and Forecasting / P. J. Brockwell, R. A. Davis. – New York : Springer, 2016. – 439 p.
6. Tuhanskykh O. Method for Statistical Evaluation of Nonlinear Model Parameters in Statistical Learning Algorithms / O. Tuhanskykh, D. Baran, O. Pysarchuk // Proceedings of Ninth International Congress on Information and Communication Technology. – Springer, Singapore. – 2024. – 1013. – P. 265–274. (Series "Lecture Notes in Networks and Systems"). DOI: 10.1007/978-981-97-3559-4_21
7. Machine Learning for Anomaly Detection: A Systematic Review / [A. B. Nassif, M. A. Talib, Q. Nasir, F. M. Dakalbab] // IEEE Access. – 2021. – No. 9. – P: 78658–78700. DOI: 10.1145/3439950.
8. Deep Learning for Anomaly Detection: / [G. Pang, C. Shen, L. Cao, A. Hengel] // ACM Computing Surveys. – 2021. – No. 54(2). – P. 1–38. DOI: 10.1145/3439950.
9. Conditional Anomaly Detection [X. Song, M. Wu, C. Jermaine, S Ranka] // IEEE Transactions on Knowledge and Data Engineering. – 2007. – No. 19. – P. 631–645. DOI: 10.1109/TKDE.2007.1009.
10. Algorithms of statistical anomalies clearing for data science applications / [O. Pysarchuk, D. Baran, Y. Mironov, I. Pysarchuk] // System research and information technologies. – 2023. – No. 1. – P. 78–84. DOI: 10.20535/SRIT.2308-8893.2023.1.06.
11. Mehrotra K. G. Anomaly Detection Principles and Algorithms / K. G. Mehrotra, C. K. Mohan, H. Huang. – Switzerland : Springer, 2017. – 229 p.
12. McKinney W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython / W. McKinney. – O'Reilly Media, 2017. – 550 p.
13. Nelli F. Python Data Analytics: With Pandas, NumPy, and Matplotlib, 2nd ed. Edition / F. Nelli. – Apress, 2018. – 588 p.
14. Raschka S. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, Second Edition / S. Raschka, V. Mirjalili. – Packt Publishing, 2017. – 622 p.
15. Joshi P. Artificial Intelligence with Python: A Comprehensive Guide to Building Intelligent Apps for Python Beginners and Developers / P. Joshi. – Packt Publishing, 2017. – 466 p.

DOI 10.15588/1607-3274-2025-3-4

НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

УДК 004.8:004.032.26

ШВИДКА НЕЙРОННА МЕРЕЖА ТА ЇЇ АДАПТИВНЕ НАВЧАННЯ В ЗАДАЧАХ КЛАСИФІКАЦІЇ

Бодянський Є. В. – д-р техн. наук, професор, професор кафедри штучного інтелекту, Харківський національний університет радіоелектроніки, Харків, Україна.

Шафроненко Є. О. – асистент кафедри медіаінженерії та інформаційних радіоелектронних систем, Харківський національний університет радіоелектроніки, Харків, Україна.

Бродецький Ф. А. – старший викладач кафедри інформатики, Харківський національний університет радіоелектроніки, Харків, Україна.

Танянський О. С. – аспірант кафедри інформатики, Харківський національний університет радіоелектроніки, Харків, Україна.

АНОТАЦІЯ

Актуальність. Для вирішення широкого класу задач обробки інформації і, перш за все, розпізнавання образів за умов суттєвої нелінійності широке розповсюдження одержали штучні нейронні мережі, завдяки своїм універсальним апроксимуючим властивостям та здатності до навчання на основі тренувальних навчальних вибірок. Найбільшого розповсюдження отримали глибокі нейронні мережі, які дійсно демонструють дуже високу якість розпізнавання, але потребують надвеликих обсягів навчальних даних, які не завжди є доступними. За цих умов ефективними можуть бути, так звані, машини опорних векторів найменших квадратів, які не потребують великих обсягів навчальних вибірок, однак можуть навчатися лише у пакетному режимі і є достатньо громіздкими у чисельній реалізації. Тому достатньо актуальною є задача навчання LS-SVM у послідовному режимі за умов суттєвої нестаціонарності даних, що послідовно у онлайн режимі надходять на опрацювання у нейронну мережу.

Мета. Мета роботи полягає у запровадженні підходу до адаптивного навчання LS-SVM, що дозволяє відмовитися від перетворення зображень у векторні сигнали.

Метод. Запропоновано підхід для розпізнавання образів-зображень за допомогою машини опорних векторів найменших квадратів (LS-SVM) за умов, коли дані на обробку надходять у послідовному онлайн режимі. Перевагою запропонованого підходу є скорочення часу вирішення задачі розпізнавання образів-зображень, а також дозволяє реалізувати процес навчання на нестаціонарних тренувальних вибірках. Особливістю запропонованого методу є обчислювальна простота і висока швидкодія, пов'язана з тим, що кількість нейронів у мережі не змінюється з часом, тобто архітектура залишається фіксованою у процесі налаштування.

Результати. Запропонований підхід до адаптивного навчання LS-SVM спрощує чисельну реалізацію нейронної мережі та дозволяє підвищити швидкість обробки інформації і, перш за все, налаштування її синаптичних ваг.

Висновки. Розглянута задача розпізнавання образів-зображень за допомогою машини опорних векторів найменших квадратів (LS-SVM) за умов, коли дані на обробку надходять у послідовному онлайн режимі. Процес навчання реалізується на ковзному вікні, що веде до того, що кількість нейронів у мережі не змінюється з часом, тобто архітектура залишається фіксованою у процесі налаштування. Такий підхід спрощує чисельну реалізацію системи та дозволяє реалізувати процес навчання на нестаціонарних тренувальних вибірках. Розглянута можливість навчання у ситуаціях, коли навчальні образи задані не лише у векторній формі, а й матричній, що дозволяє відмовитися від перетворення зображень у векторні сигнали.

КЛЮЧОВІ СЛОВА: адаптивне навчання, класифікація, швидка нейронна мережа, машина опорних векторів.

АБРЕВІАТУРИ

SVM – машина опорних векторів;

LS-SVM – машина опорних векторів найменших квадратів.

НОМЕНКЛАТУРА

w – вектор синаптичних ваг;

k – номер вектору-спостереження;

$x(k)$ – вектор-спостереження;

$\hat{y}(x)$ – вихідний сигнал мережі;

$\varphi(x)$ – вектор, утворений ядерними дзвонуватими активаційними функціями;

DOI 10.15588/1607-3274-2025-3-5

$\varphi_j(x)$ – значення j-ї ядерної дзвонуватої активаційної функції;

w_0 – пороговий параметр вибірки;

w_k – значення k-го вектора-спостереження;

$x(l)$ – вхідні вектори;

$(n\times 1)$ – вектор вхідних сигналів в мережу;

$\lambda(l)$ – множники Лагранжа;

E – цільова функція;

L^{SV} – квадратичний критерій;

$e(l)$ – середня похибка;

$y(l)$ – зовнішній навчальний сигнал;

$(\bullet)^+$ – символ псевдообернення за Муром – Пенроузом;

γ – параметр регуляризації;

I_{kk} – одинична $(k\times k)$ матриця;

$I_k - (k\times 1)$ – вектор, утворений одиницями;

$\Lambda(k)$ – вектор, утворений множниками Лагранжа;

$Y(k)$ – вектор, утворений вихідними сигналами мережі;

$\Omega(k)$ – активаційна ядерна функція;

σ^2 – параметр рецепторного поля активаційної функції;

$Sp(\bullet)$ – символ сліду матриці;

K – традиційний гауссіан;

$P(k)$ – обернення блочних матриць.

ВСТУП

Проблема класифікації – розпізнавання образів, у тому числі зображень, є однією з найважливіших проблем Data Mining, Data Stream Mining, Big Data Mining. В даний час існує багато підходів для її вирішення. На сьогодні найбільш ефективними з них є Deep Neural Networks, які демонструють дійсно вражаючі результати саме в задачах обробки зображень довільної природи. Водночас ці мережі мають ряд істотних недоліків, які обмежують їх використання, особливо в задачах, коли дані надходять на обробку послідовно, спостереження за спостереженням, можливо, в реальному часі. Це пояснюється тим, що глибокі мережі є досить повільними системами, які навчаються за допомогою зворотного поширення помилок протягом багатьох епох у пакетному (мініпакетному) режимі, що займає багато часу та вимагає великих обсягів навчальних даних, що не завжди підходить для реальних програм.

За цих умов ефективними можуть бути, так звані, машини опорних векторів найменших квадратів [1–3], які не потребують великих обсягів навчальних вибірок, однак можуть навчатися лише у пакетному режимі і є достатньо громіздкими у чисельній реалізації. Тому достатньо актуальною є задача навчання LS-SVM у послідовному режимі за умов суттєвої нестаціонарності даних, що послідовно у онлайн режимі надходять на опрацювання у нейронну мережу.

Об’єкт дослідження. Розпізнавання образів-зображень за допомогою машини опорних векторів найменших квадратів (LS-SVM) за умов, коли дані на обробку надходять у послідовному онлайн режимі.

Предмет дослідження. Підхід до адаптивного навчання LS-SVM.

Мета роботи полягає у запровадженні підходу до адаптивного навчання LS-SVM, що дозволяє відмовитися від перетворення зображень у векторні сигнали.

1 ПОСТАНОВКА ЗАВДАННЯ

Як вже відзначалося, навчання традиційних машин опорних векторів є достатньо громіздким з обчислювальної точки зору і пов’язане з вирішенням задачі нелінійного програмування з обмеженнями, а розмірність цієї задачі визначається обсягом навчальної вибірки. Ще раз підкреслимо, що навчання відбувається у пакетному офлайн режимі.

Нелінійне перетворення, що реалізуються нейронною мережею опорних векторів має вигляд

$$\hat{y}(x) = w^T \varphi(x) + w_0,$$

де $w = (w_1,..,w_l,...,w_k)^T$ – вектор синаптичних ваг, розрахований на основі навчальної вибірки $x(1),...,x(l),...,x(k)$; $x(l) = (x_1(l),...,x_i(l),...x_n(l))^T$ – $(n\times 1)$ – вектор вхідних сигналів у мережу; $\varphi(x) = (\varphi_1(x),...,\varphi_l(x),...,\varphi_k(x))^T$ – вектор, утворений ядерними дзвонуватими активаційними функціями, центри яких визначаються вхідними векторами $x(l), l = 1,...,k$.

2 ОГЛЯД ЛІТЕРАТУРИ

Для вирішення широкого класу задач обробки інформації і, перш за все, розпізнавання образів за умов суттєвої нелінійності широке розповсюдження одержали штучні нейронні мережі [4], завдяки своїм універсальним апроксимуючим властивостям та здатності до навчання на основі тренувальних навчальних вибірок. Тут найбільше розповсюдження отримали глибокі нейронні мережі і, перш за все, згорткові мережі [5], які дійсно демонструють дуже високу якість розпізнавання, але потребують надвеликих обсягів навчальних даних, які не завжди є доступними при вирішенні реальних конкретних задач. Для розглянутої задачі ефективним може бути використання опорної векторної машини (SVM) [6–8], яка оптимізує емпіричний критерій навчання ризику та коригує його параметри як на основі традиційного навчання під керівництвом, так і на основі «нейронів у точках даних» поняття [9]. Навчання SVM можна значно прискорити та звести до розв’язування систем лінійних рівнянь за допомогою так званих опорних векторних машин найменших квадратів (LS-SVM) [10].

DOI 10.15588/1607-3274-2025-3-5

У загальному випадку машини опорних векторів є окремим класом нейронних мереж, заснованих на мінімізації, так званого, емпіричного ризику [6] та налаштовуються у режимі контрольованого навчання, при цьому центри їх активаційних функцій розташовуються за принципом «Нейрони в точках даних». Ключовим моментом тут є, так звані, опорні (крайні) вектори, що формують достатньо компактну множину найбільш інформативних спостережень з навчальних даних. Слід також відмітити, що «класичні» SVM навчаються у пакетному режимі, при цьому навчальна вибірка повинна бути сформована заздалегідь.

3 МАТЕРІАЛИ І МЕТОДИ

При навчанні штучних нейронних мереж найчастіше використовується традиційний квадратичний критерій

$$E^{LS}(k)=\sum_{l=1}^{k}e^2(l),$$

де $e(l)=y(l)-\hat{y}(x(l)), l=1,2,...,k$.

Мінімізація цього критерія веде до стандартної оцінки найменших квадратів

$$w^{LS}=\left(\sum_{l=1}^{k}\varphi\left(x(l)\right)\varphi^T\left(x(l)\right)\right)^{+}\varphi\left(x(l)\right)y(l).$$

Значно кращі результати можуть бути отримані на основі критерія емпіричного ризику

$$E^{SV}(k)=\frac{1}{2}w^Tw+\frac{\gamma}{2}\sum_{l=1}^{k}e^2(l)$$

за наявності k лінійних обмежень-рівностей

$$\begin{cases}y(1)=w^T\varphi(x(1))+w_0+e(1),\\ ...\\ y(l)=w^T\varphi(x(l))+w_0+e(l),\\ ...\\ y(k)=w^T\varphi(x(k))+w_0+e(k).\end{cases}$$

При навчанні у пакетному режимі оцінювання синаптичних ваг машини опорних векторів найменших квадратів пов'язане із знаходженням сідлової точки функції Лагранжа

$$L^{SV}=\left(w,w_0,e,\lambda\right)=E^{SV}(k)+\sum_{l=1}^{k}\lambda(l)\left(y(l)-w^T\varphi(x(l))-w_0-e(l)\right)=\frac{1}{2}w^Tw+\frac{\gamma}{2}\sum_{l=1}^{k}\lambda(l)\left(y(l)-w^T\varphi(x(l))-w_0-e(l)\right),$$

при цьому в процесі оптимізації повинні бути знайдені не лише синаптичні ваги w,w_0, але й невизначені множники Лагранжа $\lambda(l), l=1,2,...,k$.

Система рівнянь Куна-Такера для введеної функції Лагранжа має вигляд

$$\begin{cases}\nabla_w L^{SV}\left(w,w_0,e,\lambda\right)=w-\sum_{l=1}^{k}\lambda(l)\varphi\left(x(l)\right)=\vec{0},\\ \dfrac{\partial L^{SV}\left(w,w_0,e,\lambda\right)}{\partial w_0}=-\sum_{l=1}^{k}\lambda(l)=0,\\ \dfrac{\partial L^{SV}\left(w,w_0,e,\lambda\right)}{\partial e(l)}=\gamma e(l)-\lambda(l)=0,\\ \dfrac{\partial L^{SV}\left(w,w_0,e,\lambda\right)}{\partial \lambda(l)}=y(l)-w^T\varphi\left(x(l)\right)-w_0-e(l)=0\end{cases}$$

(тут $\vec{0}$ – $\left(k\times1\right)$ – вектор утворений нулями), або

$$\begin{cases}w=\sum_{l=1}^{k}\lambda(l)\varphi\left(x(l)\right),\\ \sum_{l=1}^{k}\lambda(l)=0,\\ \lambda(l)=\gamma e(l),\\ y(l)=w^T\varphi\left(x(l)\right)-w_0-e(l)=0.\end{cases}$$

Нескладно помітити, що синаптичні ваги повністю визначаються невизначеними множниками Лагранжа (перше рівняння системи), а сама ця система може бути переписана у компактній векторно-матричній формі

$$\begin{pmatrix}0 & I_k^T\\ I_k & \Omega(k)+\gamma^{-1}I_{kk}\end{pmatrix}\begin{pmatrix}w_0\\ \Lambda(k)\end{pmatrix}=\begin{pmatrix}0\\ Y(k)\end{pmatrix},$$

де $\Lambda(k)=\left(\lambda(1),...,\lambda(l),...,\lambda(k)\right)^T$; $Y(k)=\left(y(1),...,y(l),...,y(k)\right)^T$, $\Omega(k)=\left\{\Omega_{ij}=\varphi^T\left(x(i)\right)\varphi(x(j))=K\left(x(i),x(j)\right)\right\}$ – активаційна ядерна функція, найчастіше традиційний гауссіан у формі

$$K\left(x(i),x(j)\right)=e^{-\frac{\|x(i)-x(j)\|^2}{2\sigma^2}}.$$

Тут цікаво відмітити, що вхідним сигналом у мережу традиційно є $\left(n\times1\right)$ – вектор $x(l)$, при цьому, якщо вирішується задача розпізнавання образів-зображень, це зображення що має форму $\left(n_1\times n_2\right)$ – матриці, попередньо повинно бути перетворено у $\left(n\times1\right)$ – вектор за допомогою операцій згортки та субдискретизації.

Використання гауссіанів в якості активаційних функцій у LS-SVM дозволяє використовувати в якості

DOI 10.15588/1607-3274-2025-3-5

вхідного сигналу безпосередньо матрицю-зображення, при цьому

$$K\left(x(i),x(j)\right)=e^{-\frac{Sp(x(i)-x(j))(x(i)-x(j))^T}{2\sigma^2}},$$

при цьому в якості відстані використовується не традиційна евклідова відстань, а матрична метрика Фробеніуса, що є узагальненням евклідової метрики на матричний випадок.

Таким чином, навчання LS-SVM пов'язане з вирішенням систем рівнянь

$$\begin{pmatrix} w_0 \\ \Lambda(k) \end{pmatrix}=\begin{pmatrix} 0 & I_k^T \\ I_k & \Omega(k)+\gamma^{-1}I_{kk} \end{pmatrix}^{-1}\begin{pmatrix} 0 \\ Y(k) \end{pmatrix}=P(k)\begin{pmatrix} 0 \\ Y(k) \end{pmatrix}.$$

Ця система описує процес навчання у пакетному офлайн режимі, коли вся навчальна вибірка сформована заздалегідь і не змінюється з часом.

Із зростанням навчальної вибірки при великих k використання цієї формули, яка хоча і видається дещо громіздкою, спрощує процес обернення матриць великої розмірності.

Із зростанням обсягів навчальної вибірки при великих k процес навчання LS-SVM стає громіздким, тому може потребувати досить багато часу. Крім того, у нестаціонарних ситуаціях, коли характеристики навчальної вибірки змінюються з часом, доцільно організувати «забування» застарілої інформації. При цьому можна організувати процес навчання на «ковзному вікні», коли з надходженням нового спостереження виключається одне застаріле. При цьому при надходженні нового $(k+1)$-го спостереження із нейронної мережі вилучається $(k-s)$-та ядерна активаційна функція (тут s – розмір ковзного вікна, що включає в себе тільки значущі спостереження). Таким чином, у системі фіксується кількість активаційних функцій, яка визначається розміром вікна s.

Введемо далі у розгляд $(s\times s)$ – матрицю ядерних активаційних функцій

Нескладно також організувати цей процес у онлайн режимі, коли у вже навчену на k спостереженнях систему надходить $(k+1)$-ше спостереження $y(k+1)$.

Тоді для $(k+1)$-го відліку можна записати [11]:

$$\begin{pmatrix} w_0 \\ \Lambda(k+1) \end{pmatrix}=\begin{pmatrix} 0 & I_{k+1}^T \\ I_k & \Omega(k+1)+\gamma^{-1}I_{k+1,k+1} \end{pmatrix}^{-1}\begin{pmatrix} 0 \\ Y(k+1) \end{pmatrix},$$

Звідки

$$\begin{pmatrix} 0 \\ Y(k) \\ y(k+1) \end{pmatrix}=\begin{pmatrix} P^{-1}(k) & \vec{K}\left(x(i),x(k+1)\right) \\ \vec{K}^T\left(x(i),x(k+1)\right) & 1+\gamma^{-1} \end{pmatrix}^{-1}\begin{pmatrix} \vec{Y}(k) \\ y(k+1) \end{pmatrix},$$

де

$\vec{K}^T\left(x(i),x(k+1)\right)=\left(1,K(x(1),x(k+1)),...,K(x(k),x(k+1))\right)^T$,

$\vec{Y}(k)=\left(0,Y^T(k)\right)$, після чого, використовуючи формулу обернення блочних матриць, отримуємо

$$P\left(k+1\right)=\begin{pmatrix} P(k)+P(k)\vec{K}(x(i)x(k+1))\bullet & \\ \bullet\vec{K}^T(x(i)x(k+1))P(k)\bullet & -(P(k)\vec{K}(x(i)x(k+1)))\bullet \\ \bullet\left(1+\gamma^{-1}-\vec{K}^T(x(i)x(k+1))\right)\bullet & \bullet\left(1+\gamma^{-1}-\vec{K}^T(x(i)x(k+1))\right)\bullet \\ \bullet P(k)\vec{K}(x(i)x(k+1))^{-1} & \bullet P(k)\vec{K}(x(i)x(k+1))^{-1} \\ -\left(\vec{K}^T(x(i)x(k+1))P(k)\right)\bullet & \\ \bullet\left(1+\gamma^{-1}-\vec{K}^T(x(i)x(k+1))\right)\bullet & \left(1+\gamma^{-1}-\vec{K}^T(x(i)x(k+1))\right)\bullet \\ \bullet P(k)\vec{K}(x(i)x(k+1))^{-1} & \bullet P(k)\vec{K}(x(i)x(k+1))^{-1} \end{pmatrix}.$$

$\Omega(k,s)=\left\{\Omega_{ij}=\varphi^T\left(x(i)\varphi(x(j))\right)=K\left(x(i),x(j),s\right)\right\}$, $i=k-s+1,k-s+1,...,k$; $j=k-s+1,k-s+2,...,k$; на основі яких формується вихідний сигнал мережі

$$\hat{y}(x,k,s)=\sum_{l=k-s+1}^{k}\lambda(l,s)K\left(x,x(l),s\right)+w_0(k,s).$$

Параметри $\lambda(l,s)$, $w_0(k,s)$ можуть бути знайдені шляхом вирішення матричного рівняння

$$\begin{pmatrix} w_0 \\ \Lambda(k,s) \end{pmatrix}=\begin{pmatrix} 0 & I_s^T \\ I_s & \Omega(k,s)+\gamma^{-1}I_{s,s} \end{pmatrix}^{-1}\begin{pmatrix} 0 \\ Y(k,s) \end{pmatrix}=$$
$$=P\left(k,s\right)\begin{pmatrix} 0 \\ Y(k,s) \end{pmatrix},$$

де $\Lambda(k,s)=\left(\lambda(k-s+1,s),\lambda(k-s+2,s),...,\lambda(k,s)\right)^T$,

$Y(k,s)=\left(y(k-s+1),y(k-s+2),...,y(k)\right)^T$.

DOI 10.15588/1607-3274-2025-3-5

З находженням $(k+1)$-го спостереження, його необхідно включити в матрицю $\Omega(k+1,s)$, одночасно з тим, виключаючи спостереження, що відповідає $(k-s)$-му моменту часу.

Нескладно бачити, що оновлені параметри LS-SVM визначаються за допомогою співвідношенням

$$\begin{pmatrix} w_0(k+1,s) \\ \Lambda(k+1,s) \end{pmatrix} = \begin{pmatrix} 0 & I_s^T \\ I_s & \Omega(k+1,s)+\gamma^{-1}I_{s,s} \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ Y(k+1,s) \end{pmatrix} =$$

$$= P(k+1,s)\begin{pmatrix} 0 \\ Y(k+1,s) \end{pmatrix},$$

де

$$\Lambda(k+1,s) = \left(\lambda(k-s+1,s),\lambda(k-s+2,s),...,\lambda(k+1,s)\right)^T,$$

$$Y(k+1,s) = \left(y(k-s+1),y(k-s+2),...,y(k+1)\right)^T.$$

Тут важливо відзначити, що така організація процесу навчання LS-SVM, дозволяє не тільки «придушувати» застарілу інформацію, але й зберігає архітектуру нейронної мережі з часом, що суттєво спрощує її чисельну реалізацію.

4 ЕКСПЕРИМЕНТИ

Для перевірки запропонованого підходу було використано набір даних «Fashion-MNIST» [12] – це набір даних із зображеннями статей Zalando, що складається з навчального набору з 60000 прикладів і тестового набору з 10000 прикладів. Кожен приклад – це зображення у відтінках сірого 28x28, пов'язане з міткою з 10 класів. Приклади спостережень наведено на рисунку 1.

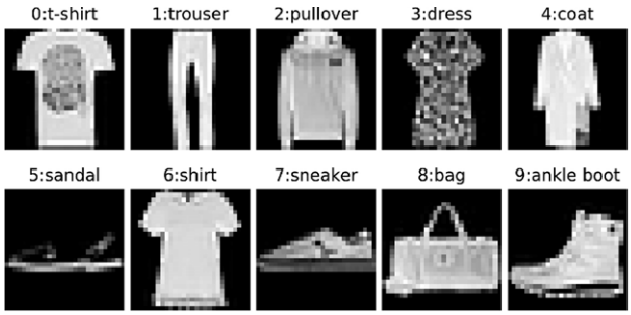


Рисунок 1 – Екземпляри вибірки «Fashion MNIST»

Для експериментальних досліджень та порівняльного аналізу запропонованого підходу були обрані методи K-NN та RCNN з високою швидкістю навчання, оскільки специфікою онлайн-обробки є їх апріорна та поточна невизначеність, тому швидкість навчання системи повинна бути високою.

Для аналізу швидкості та точності підходів, вибірку даних «Fashion-MNIST» було розбито на декілька підмасивів: 5000, 10000, 15000 спостережень відповідно. Результати наведені на рисунку 2.

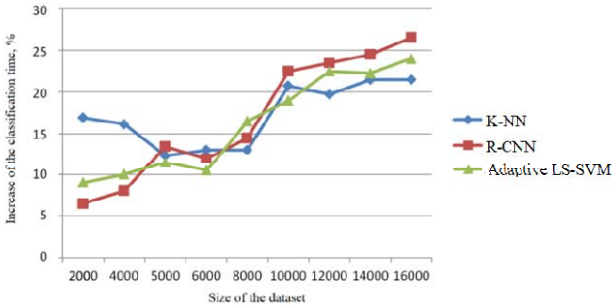


Рисунок 2 – Графік залежності часу класифікації від розміру вхідних даних

Якість класифікації вибірки «Fashion MNIST» був проведений в декілька етапів на 5000, 10000 та 15000 спостережень. Результати наведені в таблиці 1.

Таблиця 1 – Оцінка якості класифікації вибірки «FMNIST»

Методи	SI	CHI	DBI
	5000 спостережень		
K-NN	0,3324	928,01	1,3
R-CNN	0,3335	974,42	1,10
Adaptive LS-SVM	0,3665	1420,28	1,15
	10000 спостережень		
K-NN	0,6045	1460,65	1,5
R-CNN	0,6122	1945,5	1,25
Adaptive LS-SVM	0,7222	2800,05	1,32
	15000 спостережень		
K-NN	0,8324	1560,00	1,8
R-CNN	0,8735	2000,28	1,15
Adaptive LS-SVM	0,9885	2985,2	1,28

Індекс силуету (SI) вимірює, наскільки кожен об'єкт в класі подібний до інших об'єктів у тому ж класі порівняно з об'єктами інших класів. За значенням від 0 до 1, вищі значення SI вказують на кращу якість класифікації.

Індекс Калінські-Харабаса (CHI) вимірює, наскільки добре класи відокремлені один від одного. Високі значення CHI вказують на кращу якість класифікації.

Індекс Девіса-Болдуїна (DBI) вимірює середню відстань між класами і внутрішньокласову відстань. Низькі значення DBI свідчать про кращу якість класифікації.

6 ОБГОВОРЕННЯ

Аналізуючи результати отриманих експериментальних досліджень та порівняльного аналізу роботи запропонованого підходу при вирішенні завдань розпізнавання образів з хорошою точністю та швидкістю в умовах обмеженого навчального набору даних, обробки зображень в онлайн-режимі.

В таблиці 1 метод адаптивної LS-SVM має найвищі значення SI для всіх трьох обсягів даних, що свідчить про його здатність ефективно відокремлювати класи навіть у великих обсягах даних, а також має

DOI 10.15588/1607-3274-2025-3-5

найвищі значення CHI для всіх обсягів даних, що означає, що класи, сформовані Adaptive LS-SVM, добре відокремлені один від одного.

Цей підхід є особливо корисним при роботі з об'ємними даними, забезпечуючи високу точність класифікації та ефективність в умовах зміни типів вхідних даних. Його гнучкість та здатність до адаптації роблять з нього потужний інструмент для різноманітних завдань у сфері автоматичного розпізнавання образів-зображень. Запропонований підхід призначений для вирішення достатньо великого класу проблем у загальних рамках Data Stream Mining і Big Data Mining.

ВИСНОВКИ

Розглянута задача розпізнавання образів-зображень за допомогою машини опорних векторів найменших квадратів (LS-SVM) за умов, коли дані на обробку надходять у послідовному онлайн режимі. Процес навчання реалізується на ковзному вікні, що веде до того, що кількість нейронів у мережі не змінюється з часом, тобто архітектура залишається фіксованою у процесі налаштування. Такий підхід спрощує чисельну реалізацію системи та дозволяє реалізувати процес навчання на нестаціонарних тренувальних вибірках. Розглянута можливість навчання у ситуаціях, коли навчальні образи задані не лише у векторній формі, а й матричній, що дозволяє відмовитися від перетворення зображень у векторні сигнали.

Запропонований підхід до адаптивного навчання LS-SVM спрощує чисельну реалізацію нейронної мережі та дозволяє підвищити швидкість обробки інформації і, перш за все, налаштування її синаптичних ваг.

Наукова новизна: вперше запропоновано підхід до адаптивного навчання LS-SVM за умов, коли дані на обробку надходять у послідовному онлайн режимі.

Практичне значення: результати експерименту дозволяють рекомендувати запропонований підхід для використання на практиці для вирішення проблем автоматичного розпізнавання образів-зображень.

Перспективи подальших досліджень швидкі нейронні мережі розпізнавання образів-зображень для широкого класу практичних задач Data Stream Mining і Big Data Mining.

ПОДЯКА

Робота виконана в рамках науково-дослідного проєкту державного бюджету Харківського національного університету радіоелектроніки «Адаптивний бегінг гібридних систем обчислювального інтелекту на основі оптимального за швидкодією онлайн навчання» (ДР №0124U000363).

ЛІТЕРАТУРА

1. Goodfellow I. Deep learning / I. Goodfellow, J. Begin and A. Courville. –The MIT Press, 2016.
2. Graupe D. Deep learning neural networks: design and case studies / D. Graupe. – World Scientific Publishing Company, 2016. https://doi.org/10.1142/10190
3. Neural networks and deep learning / C. C. Aggarwal et al. – Cham : Springer, 2018. – T. 10. № 978. https://doi.org/10.1007/978-3-319-94463-0
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. – P. 273–297, https://doi.org/10.1007/bf00994018.
8. Steinwart I. Support Vector Machines. / I. Steinwart and A. Christmann. – New York : Springer, 2008.
9. Pattern recognition using radial basis function network / [D. R. Zahirniak, R. Chapman, S. K. Rogers et al.] // Aerospace Application of Artificial Intelligence. – 1990. – P. 249–260.
10. Least Squares Support Vector Machines / [J. Vandewalle, B. D. Moor, T. V. Gestel et al.]. – World Scientific Publishing Company, 2003.
11. Adaptive least-squares support vector machine and its online learning / [Y. Bodyanskiy, A. Deineko, F. Brodetskyi, and D. Kosmin] // CEUR Workshop Proceedings. – Nov. 2020. – Vol. 2762, Art. no. 3. http://ceur-ws.org/Vol-2762/paper3.pdf
12. Xiao H. Fashion-mist: a novel image dataset for benchmarking machine learning algorithms / H. Xiao, K. Rasul, R. Vollgraf. – Available: https://arxiv.org/abs/1708.07747.

Стаття надійшла до редакції 02.04.2025.
Після доробки 12.06.2025.

DOI 10.15588/1607-3274-2025-3-5

UDC 004.8:004.032.26

FAST NEURAL NETWORK AND ITS ADAPTIVE LEARNING IN CLASSIFICATION PROBLEMS

Bodyanskiy Ye. V. – Dr. Sc., Professor at the Department of Artificial Intelligence, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Shafronenko Ye. O. – Assistant at the Department of Media Engineering and Information Radio Electronic Systems, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Brodetskyi F. A. – PhD Student, Assistant at the Department of Informatics, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Tanianskyi O. S. – Postgraduate student at the Department of Informatics, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

ABSTRACT

Context. To solve a wide class of information processing tasks and, above all, pattern recognition under conditions of significant nonlinearity, artificial neural networks have become widely used, due to their universal approximating properties and ability to learn based on training training samples. Deep neural networks have become the most widespread, which indeed demonstrate very high recognition quality, but require extremely large amounts of training data, which are not always available. Under these conditions, the so-called least squares support vector machines can be effective. They do not require large amounts of training samples but can be trained only in batch mode and are quite cumbersome in numerical implementation. Therefore, the problem of training LS-SVM in sequential mode under conditions of significant non-stationarity of data that are sequentially fed online to the neural network for processing is quite relevant.

Objective. The aim of the work is to introduce an approach to adaptive learning of LS-SVM, which allows us to abandon the conversion of images into vector signals.

Method. An approach for image recognition using a least squares support vector machine (LS-SVM) is proposed under conditions when data for processing is received in a sequential online mode. The advantage of the proposed approach is that reduces the time to solve the image recognition problem and allows the implementation of the learning process on non-stationary training samples. A feature of the proposed method is computational simplicity and high speed since the number of neurons in the network does not change over time, i.e., the architecture remains fixed during the tuning process.

Results. The proposed approach to adaptive learning of LS-SVM simplifies the numerical implementation of the neural network and allows for an increase in the speed of information processing and, above all, the tuning of its synaptic weights.

Conclusions. The problem of pattern recognition using the least squares support vector machine (LS-SVM) is considered under conditions when data for processing is received in a sequential online mode. The training process is implemented on a sliding window, which leads to the fact that the number of neurons in the network does not change over time, i.e. the architecture remains fixed during the tuning process. This approach simplifies the numerical implementation of the system and allows the training process to be implemented on non-stationary training samples. The possibility of training in situations where training images are given not only in vector form but also in matrix form allows us to abandon the conversion of images into vector signals.

KEYWORDS: Adaptive learning, classification, fast neural network, support vector machine.

REFERENCES

1. Goodfellow I., Begin J. and Courville A. Deep learning. The MIT Press, 2016.
2. Graupe D. Deep learning neural networks: design and case studies. World Scientific Publishing Company, 2016. https://doi.org/10.1142/10190
3. C. C. Aggarwal et al. *Neural networks and deep learning*. Cham, Springer, 2018, T. 10, № 978. https://doi.org/10.1007/978-3-319-94463-0
4. Poggio, T. Girosi F. Networks for approximation and learning, *Proceedings of the IEEE*, 1990, T. 78, №. 9. pp. 1481–1497.
5. Haykin S. Neural networks: a comprehensive foundation. Prentice Hall PTR, 2004, T. 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. Steinwart I. and Christmann A. Support Vector Machines. New York, Springer, 2008.
9. Zahirniak D. R., Chapman R., Rogers S. K., Suter B. W., Kabriski M., and Pyatti V. Pattern recognition using radial basis function network, *Aerospace Application of Artificial Intelligence*, 1990, pp. 249–260.
10. Vandewalle J., Moor B. D., Gestel T. V., Brabanter J. D., and Suykens J. A. K. Least Squares Support Vector Machines. World Scientific Publishing Company, 2003.
11. Bodyanskiy Y., Deineko, A. Brodetskyi F., and Kosmin D. Adaptive least-squares support vector machine and its online learning, *CEUR Workshop Proceedings*, Nov. 2020, vol. 2762, Art. no. 3. http://ceur-ws.org/Vol-2762/paper3.pdf
12. Xiao H., Rasul K., Vollgraf R. Fashion-mist: a novel image dataset for benchmarking machine learning algorithms. Available: https://arxiv.org/abs/1708.07747.

DOI 10.15588/1607-3274-2025-3-5

UDC 004.93, 004.8

METHOD OF PARALLEL HYBRID SEARCH FOR LARGE-SCALE CODE REPOSITORIES

Boiko V. O. – Assistant of the Department of Software Engineering, Khmelnytskyi National University, Khmelnytskyi, Ukraine.

ABSTRACT

Context. Modern software systems contain extensive and growing codebases, making code retrieval a critical task for software engineers. Traditional code search methods rely on keyword-based matching or structural analysis but often fail to capture the semantic intent of user queries or struggle with unstructured and inconsistently documented code. Recently, semantic vector search and large language models (LLMs) have shown promise in enhancing code understanding. The problem – is designing a scalable, accurate, and hybrid code search method capable of retrieving relevant code snippets based on both textual queries and semantic context, while supporting parallel processing and metadata enrichment.

Objective. The goal of the study is to develop a hybrid method for semantic code search by combining keyword-based filtering and embedding-based retrieval enhanced with LLM-generated summaries and semantic tags. The aim is to improve accuracy and efficiency in locating relevant code elements across large code repositories.

Method. A two-path search method with post-processing is proposed, where textual keyword search and embedding-based semantic search are executed in parallel. Code blocks are preprocessed using GPT-4o model to generate natural-language summaries and semantic tags.

Results. The method has been implemented and validated on a .NET codebase, demonstrating improved precision in retrieving semantically relevant methods. The combination of parallel search paths and LLM-generated metadata enhanced both result quality and responsiveness. Additionally, LLM-post-processing was applied to the top-most relevant results, enabling more precise identification of code lines matching the query within retrieved snippets. Other results can be further refined on-demand.

Conclusions. Experimental findings confirm the operability and practical applicability of the proposed hybrid code search framework. The system's modular architecture supports real-time developer workflows, and its extensibility enables future improvements through active learning and user feedback. Further research may focus on optimizing embedding selection strategies, integrating automatic query rewriting, and scaling across polyglot code environments.

KEYWORDS: hybrid code search, vector search, semantic embeddings, code summarization, LLM-generated metadata, cosine similarity, textual relevance, class and method retrieval, class-based indexing, software engineering.

ABBREVIATIONS

AST is an abstract syntax tree;
LLM is a large language model;
RAG is a retrieval augmented generation;
NLP is a natural language processing;
AI is an artificial intelligence;
GPT is a generative pre-trained transformer;
HNSW is a hierarchical navigable small world;
MRR is an average inverse rank of the first relevant search result.

NOMENCLATURE

B is a set of source code blocks;
b_i is a source code block;
q is a user-supplied natural language query;
S is a set of summaries of code blocks;
E is a set of summary embeddings;
X is input data;
Y' is output data that represents search results;
f is a general representation of a function that performs search based on input parameters;
T_i is a maximum number of input tokens provided to GPT-4o model;
$T_{\max}$ is a maximum tokens GPT-4o model can process;
T_o is a maximum amount of output tokens GPT-4o model can produce;
$C(b_i)$ is a chunking function of a particular code block;
c_{im} is a chunk of a particular code block;
$\ell(c_{ij})$ is a length of a code chunk;
$s_{i(j-m)}$ is a particular summary within a chunk c_{ij};
s_i is a particular summary of a code block;
LLM is the general representation of summarization function;
$\vec{v}_j$ is an embedding vector of the j-th token in the sequence;
k is a number of tokens obtained after tokenizing s_i;
M is an embedding model text-embedding-ada-002;
$\Re^d$ is a space of embeddings;
$\vec{e}_i$ is a raw sentence-level embedding vector;
$\hat{e}_i$ is the L2-normalized embedding vector;
t_i is a token;
$IDF(t_i)$ is an inverse document frequency function;
d is a text-based metadata;
$B(d,q)$ is a term frequency weight;

DOI 10.15588/1607-3274-2025-3-6

$BM25(d,q)$ is the lexical relevance score between the user's query q and the textual representation d of a code unit;

α is a weighting coefficient (in range [0, 1]) that balances keyword-based vs. semantic-based relevance;

$S(b_i)$ is a hybrid relevance score assigned to code block b_i;

$TopK(\cdot)$ is a function that returns the K highest-scoring elements from the input set;

Y is a set of top-K ranked results;

L is a total loss across all examples, evaluated using a binary cross-entropy loss function;

$GPTR(b_i)$ is a function that refines exact lines of code inside a code block b_i using GPT-4o model;

Y_{top} is a set of the top most relevant search results.

INTRODUCTION

Effective code retrieval is pivotal in modern software development, enabling developers to efficiently locate and reuse existing code snippets. Traditional code search methods have predominantly relied on keyword-based approaches, which, while straightforward, often fail to capture the nuanced semantics of programming languages and the intent behind code implementations. This limitation becomes particularly evident in large-scale codebases, where the sheer volume and complexity of code can hinder accurate retrieval.

Recent advancements in artificial intelligence and natural language processing have introduced semantic search techniques that utilize embeddings to represent code and queries in a continuous vector space. These embeddings facilitate the retrieval of code snippets based on semantic similarity rather than exact keyword matches, thereby enhancing search relevance. However, solely relying on semantic embeddings can overlook the precision offered by traditional keyword searches, especially when specific syntax or identifiers are involved.

The object of study is the process of searching and retrieving relevant code fragments from large-scale and semantically diverse software codebases.

The subject of study is the methods and models for hybrid code search that combine textual keyword matching, semantic embedding-based retrieval, and metadata-enriched indexing.

The known code search approaches and algorithms described by various authors and applied across different domains, including traditional keyword-based [1–2, 4] and structural analysis methods [3], are often limited in capturing the semantic context of code, especially in large and heterogeneous codebases. These methods typically rely on exact textual matches or static syntactic representations, which restrict their effectiveness in scenarios where the user query expresses intent rather than specific code tokens.

However, several recent studies [5–9] have explored semantic search or other not straightforward techniques based on neural models to enhance retrieval relevance. While these approaches demonstrate improvements in understanding code semantics, they generally do not provide a unified method that combines semantic search, keyword filtering, and metadata-based enrichment into a parallel and scalable architecture suitable for practical use in real-world software engineering environments.

The purpose of the work is to develop a method and incorporate it into an efficient and scalable hybrid model for semantic code search, which combines keyword-based filtering, embedding-based retrieval, and LLM-generated metadata. The proposed model is intended to serve as a practical framework for software engineers to search and retrieve relevant code fragments from large-scale codebases based on both natural language queries and structural context.

1 PROBLEM STATEMENT

Suppose we have a set of source code chunks $B = \{b_1, b_2, ..., b_n\}$ from a software codebase, and a user-supplied natural language query q, which represents the search intent. Each chunk $b_i \in B$ contains one or more code blocks.

The task is to develop a method that will perform an accurate hybrid code search method to retrieve a set of relevant source code sections according to the natural language query q.

This can be represented by the following model:

$$X = \{B, S, E, q\},\ S = \{s_i\}, E = \{\hat{e}_i\},\\ f : X \to Y', \tag{1}$$

where the function f from input X generates an output Y' which represents the search results.

2 REVIEW OF THE LITERATURE

A typical keyword search is often carried out using algorithms such as Rabin-Karp or Knuth-Morris-Pratt. These algorithms are commonly employed in the development of frameworks designed to detect plagiarism in text documents, as outlined in the study [1]. However, these algorithms are not effective for code search, as they can only detect specific text patterns based on explicitly defined key phrases. As a result, they are not suitable for tasks that demand more advanced search techniques.

Pattern matching search, or Regex search, is a versatile tool that enables flexible string matching by defining complex patterns. It is widely supported across various programming languages, as it is integrated into text processing libraries. In a particular publication [2], RunEx is a code search tool designed for programming instructors to easily identify patterns and mistakes in students' code. It enhances traditional search methods by incorporating runtime values and provides a user-friendly interface for

DOI 10.15588/1607-3274-2025-3-6

constructing expressive queries. RunEx outperforms baseline systems in accuracy and introduces a new approach for analyzing student code at scale. However, the industrial program code is a much more complex structure than simple text, so other methods for code search are required.

The methods mentioned earlier are useful only for textual search. However, they cannot be employed for processing difficult code structures. For this, the abstract syntax tree analysis is applicable. For example, paper [3] introduces the similarity detection technique that uses richer structural information while ensuring reasonable execution time. It generates syntax trees from program code, extracts connected *n*-gram structure tokens, and compares them using cosine correlation in the vector space model.

In general, there are a lot of AST-based methods are used and already provided in the most integrated development environments to enhance code detection, correction, syntax highlighting, and search by dependency references. However, understanding the AST is not always needed. For example, publication [4] proposes a model designed to improve code search by combining the advantages of deep learning models like DeepCS with indexing techniques for faster search. This model identifies and removes irrelevant keywords, performs fuzzy search with key query terms using Elasticsearch, and re-ranks results based on sequential token matching.

However, even using Elasticsearch database for indexing, standard search utilities don't support semantic search, which has become even more popular and effective with generative AI development in the last few years. The paper [5] introduces an annotation-based code search engine that addresses information loss by extracting features from code annotations from five perspectives. Unlike current models that treat code annotations as simple natural language, the engine preserves structural information. This approach is much better since it includes deep learning, but it still does not support search queries in a natural language despite its proximity to semantic search.

The paper [6] proposes an efficient and accurate semantic code search framework using a cascaded approach with fast and slow models. The fast model, a transformer encoder, optimizes a scalable index for quick retrieval, while the slow model re-ranks the top K results to improve accuracy. To reduce memory costs, both models are jointly trained with shared parameters. This improves accuracy and efficiency but does not integrate keyword-based filtering or metadata, and the cascaded approach adds complexity.

Another publication [7] introduces RepoRift, a code search approach that leverages RAG-powered agents to improve the accuracy of code retrieval. By enhancing user queries with relevant information from GitHub repositories, the agents provide more contextually aligned and informative inputs to embedding models. The approach also incorporates a multi-stream ensemble technique to further improve retrieval accuracy. It introduces context-awareness but relies on external repositories and augmentation agents, which may not generalize or scale in enterprise environments.

The report [8] presents a novel code retrieval system using the Dense Passage Retrieval technique, which measures functional similarity between code snippets for relevance. By leveraging large-scale pre-trained language models like CodeBERT and Starencoder, the system efficiently retrieves similar code based on natural language descriptions or source code queries. However, it uses pure embedding-based retrieval, without support for text-based filtering, tag-based classification, or enriched metadata.

Another paper [9] proposes a code semantic enrichment approach to improve deep code search by aligning the semantics of code snippets with developers' queries. Recognizing that code represents low-level implementation and queries are high-level, the approach enriches code snippets with descriptions of similar code implementations. Based on a large-scale analysis of a large amount of Java code-description pairs, the method uses syntactic similarity to retrieve similar code for each snippet, enhancing its semantic representation. The model is trained using an attention mechanism to map pairs of enriched code and query into a shared vector space. To further improve representation quality, a multi-perspective co-attention mechanism with Convolutional Neural Networks is applied to capture local correlations. This approach bridges the semantic gap, but it still does not integrate parallel keyword retrieval, nor does it utilize tags or structured metadata for boosting precision.

Chen et al. use both types of search in their retriever and feed results to an LLM. Authors in a conference paper [10] proposed a retrieval-augmented framework for improving code suggestions by combining traditional information retrieval methods and deep learning-based code search with large LLMs. Their system includes a retriever that supports multiple query types (e.g., method headers, natural language), a formulator that constructs prompts using retrieved code, and a generator based on LLMs like ChatGPT. The study demonstrates that incorporating semantically relevant code snippets significantly enhances code generation quality. However, their framework does not explicitly mention summarizing context or performing line-level GPT-based retrieval. Instead, they concatenate retrieved code snippets with the query for the LLM to consume.

While recent advancements in code search emphasize deep learning and semantic retrieval, each of the reviewed approaches addresses only part of the challenge – and none in the research area offers a unified and effective framework for a code-based search.

3 MATERIALS AND METHODS

Modern code search systems are challenged by the semantic gap between a developer's natural language query and the structure and behavior of source code. Traditional keyword-based approaches, while efficient and interpretable, often fail to capture the intent behind a query when relevant code does not share lexical similarity

DOI 10.15588/1607-3274-2025-3-6

with the input terms. Conversely, purely embedding-based semantic search can retrieve contextually aligned results but lacks explainability and may yield less precise matches when queries involve specific identifiers or domain terms.

To address these limitations, this work introduces a hybrid code search method that combines the strengths of both paradigms: the precision of keyword-based retrieval and the contextual depth of embedding-based semantic search. This hybrid architecture is further enriched through the integration of LLM-driven summarization and semantic tagging, allowing the system to index not just raw code, but also its abstracted intent and purpose. The method consists of 3 phases: indexing, retrieval, and post-processing. Each is described further.

The indexing phase serves as the preparatory stage in the hybrid code search system, where raw source code files are transformed into structured, queryable data suitable for both keyword-based and embedding-based retrieval. In information retrieval systems, indexing refers to the process of analyzing and organizing source material in a way that enables efficient and accurate search. Within the context of this work, indexing involves parsing source code files, extracting metadata, generating natural language summaries, and creating vector representations of these summaries for storage and subsequent retrieval.

Code summarization is an effective technique for improving code comprehension, maintenance, and reuse by automatically generating natural language descriptions of source code [11], so this method is used to provide descriptive text for code blocks for further search.

Each file in the codebase is processed independently. The content of the file is read and passed through a natural language model to generate a descriptive summary that captures the file's functional role and behavioral semantics. This summary is intended to reflect how a human developer might describe the purpose of the file in natural terms, which enhances its compatibility with natural language queries. For most files, especially those of modest length, the summarization is performed in a single pass using the LLM of the GPT-4o model. The entire file content is provided as input, and a concise, high-level summary is returned.

However, in the case of large files – particularly those exceeding the context window of the language model – a token-based chunking strategy is employed.

Rather than attempting to identify logical or syntactic boundaries within the file, the content is split into fixed-size chunks that fit within the model's token limit, considering space for system instructions and output. The maximum input token can be represented by the following formula (2):

$$T_i = T_{\max} - T_o. \tag{2}$$

Each chunk is summarized individually, and to maintain context coherence across the file, the summary of the preceding chunk is passed along as part of the input when processing the next chunk. The idea of coherent summarization and semantic continuity is described in the paper [12]. This sequential summarization strategy enables the aggregation of a consistent and comprehensive summary, even for files that cannot be processed in a single request. Once all chunks are processed, their summaries are merged and refined to form a single summary representing the entire file. Formula (3) represents a chunking function:

$$\begin{aligned} &C(b_i) = \{c_{i1}, c_{i2}, ..., c_{im}\}, \\ &\ell(c_{ij}) < T_{\max}. \end{aligned} \tag{3}$$

Each chunk is summarized individually using LLM (4):

$$s_{ij} = LLM(c_{ij}, s_{i(j-1)}, s_{i(j-2)}, ..., s_{i(j-m)}). \tag{4}$$

This is a chained summarization process to maintain coherence across chunks. Then, final code block summary is constructed (5):

$$s_i = LLM(s_{i1}, s_{i2}, ..., s_{im}). \tag{5}$$

The sequential summarization continues on the module or folder level and finally leads to the summary of the whole repository.

Following the natural language summary generation, the system computes a semantic embedding of the summary using a pre-trained embedding model. This embedding is a high-dimensional vector that captures the semantic meaning of the text and enables efficient similarity comparisons with query embeddings during retrieval. The selected embedding model for this process is OpenAI's text-embedding-ada-002, which has demonstrated strong performance in encoding semantic representations across diverse domains. Each vector is associated with the corresponding file and stored in a vector search database to facilitate cosine similarity queries during semantic retrieval [13]. However, we should take into account that the embedding model has a token input limit as well, which is 8191 [14]. Thus, there is a need to pass a small portion of a summary to this model. The embedding process consists of the following steps: tokenization, computing dense representation, and normalization.

Tokenization is the process of breaking down text into smaller units called tokens, which can be words, subwords, or characters. This step is foundational in NLP, enabling models to analyze and understand text data. Effective tokenization is crucial for the performance of subsequent NLP tasks. The following formula (6) represents the tokenization process [15]:

$$s_i \to \{t_1, t_2, ..., t_k\},\ k \le T_e. \tag{6}$$

DOI 10.15588/1607-3274-2025-3-6

After tokenization, each token is mapped to a dense vector, known as an embedding. These embeddings are continuous vector representations in a high-dimensional space, capturing semantic and syntactic information about the tokens. Dense embeddings allow models to discern relationships between words based on their contextual usage [16]. OpenAI uses the mean pooling across tokens to obtain the final sentence-level embedding. Formula (7) represents of how a sentence or document embedding is generated:

$$\vec{e}_i = M(s_i) = \frac{1}{k}\sum_{j=1}^{k}\vec{v}_j,\ \vec{v}_j \in \Re^d. \quad (7)$$

Normalization adjusts these dense vectors to ensure consistent scaling and distribution, which is vital for the stability and performance of neural networks. Techniques like layer normalization standardize the inputs across features, facilitating faster convergence during training and improving generalization. Formula (8) represents a process of how the normalized vector (directional embedding) is calculated:

$$\hat{e}_i = \frac{\vec{e}_i}{\|\vec{e}_i\|}. \quad (8)$$

Now, the final formula is a concise and normalized representation of how an embedding vector is computed for a text summary (8):

$$\hat{e}_i = M(s_i) = \frac{1}{k}\sum_{j=1}^{k} M(t_j). \quad (9)$$

Instead of asking the GPT-4o model to generate the overall summary, there is a need to make a prompt and point that the model should return summaries of each meaningful block of code in the file and represent it in a JSON format. Fig. 1 illustrates the template of the completion request to generate the summary of the file.

```
[
  {
    "role": "system",
    "content": "You are a code summarization assistant."
  },
  {
    "role": "user",
    "content": "Analyze the source code provided and output a
JSON array. Each item should include: type: either class, method,
or property; name: the name of the code block; tags: keywords of
a particular code block does; start_line and end_line: start_line
and end_line: estimated line numbers  end_line: estimated line
number where the block ends; summary:detailed explanation line-
by-line or by logical segment
  },
  { "role": "assistant", "content": "Provide me with a source code"},
  { "role": "user", "content": "<source_code>"}
]
```

Figure 1 – Prompts to set up the code summarization assistant

The system employs Qdrant as the underlying vector database to support high-performance semantic code search. Qdrant is selected for its efficient handling of high-dimensional vector spaces, making it particularly suitable for storing and querying dense embedding vectors derived from code summaries. Its ability to perform approximate nearest neighbor search using methods like HNSW ensures fast and scalable similarity retrieval across large codebases.

In addition to vector indexing, Qdrant offers real-time filtering and payload-based queries, allowing search results to be refined using additional metadata without post-processing overhead. This functionality is crucial for hybrid search systems that require filtering by attributes such as file names, method names, or code block positions. As a result, Qdrant supports multi-modal retrieval by combining semantic similarity scores with structured filters.

The system stores not only the normalized embeddings of summarized code chunks, but also custom metadata (payload) including the full file path, the name and type of the code block (e.g., class, method), and the corresponding line range within the file. This enables precise result mapping back to the original source files, as well as advanced use cases such as highlighting specific lines or linking results to developer tools [17]. The structure of the Quadrant data record is shown in Fig. 2.

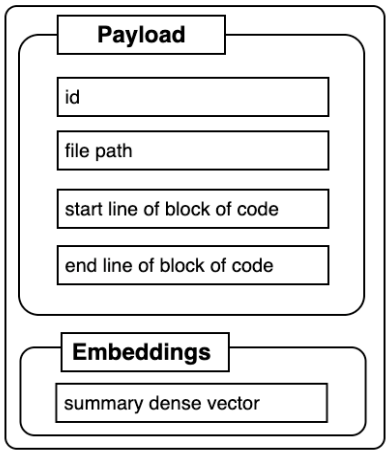


Figure 2 – A data structure for vector database Qdrant

In parallel to vector indexing, the system prepares the indexed data for keyword-based search. Summaries of code blocks, as well as additional metadata, are indexed using Elasticsearch. This database is selected due to its support for full-text search, fuzzy matching, BM25 ranking, and scalable indexing capabilities. These features make it particularly suitable for handling keyword queries where precise or partial text matches are desired [18]. The data structure is mostly the same as for vector database, but instead of embeddings – summaries and tags are saved.

By separating the vector-based search functionality and the text-based search pipeline, the indexing phase ensures that both retrieval modes can be executed inde-

DOI 10.15588/1607-3274-2025-3-6

pendently and efficiently. When the code base is changed, the system detects changes and re-generates summaries with their dense vectors only for an updated file. It can happen in the background once per some time range or while a search happens if the system detects that the code base has changed.

After this phase, the indexed codebase contains structured, searchable entries for each file, consisting of its metadata, summaries and their dense vectors, and keyword-indexed content. This structured representation supports fast and accurate retrieval in the subsequent stages of the system. Fig. 3 illustrates the activity diagram of an indexing phase.

The retrieval phase constitutes the core of the hybrid search mechanism and is responsible for identifying relevant source code blocks based on user-supplied natural language queries. This phase integrates two parallel retrieval processes: traditional keyword-based search and semantic vector-based search. Both processes operate over the indexed representation of the codebase generated during the previous phase to maximize the relevance and completeness of the search results.

Upon receiving a natural language query, the system performs keyword-based retrieval by submitting the query to a full-text search engine, such as Elasticsearch. This engine operates over the textual content indexed during the indexing phase, particularly focusing on the code block summaries and the generated semantic tags. Standard ranking techniques, including the BM25 scoring function (10), are employed to identify documents that contain direct lexical overlap with the query terms. This retrieval path provides high precision, especially for queries that include domain-specific keywords, identifiers, or terminology that match the indexed content directly.

$$BM25(d,q) = \sum_{i=1}^{n} IDF(t_i) \cdot B(d,q). \quad (10)$$

In parallel, the system conducts semantic retrieval by encoding the user query into a high-dimensional embedding using the same embedding model employed during indexing. In this case, the model used is OpenAI's text-embedding-ada-002, which produces vector representations that reflect the semantic meaning of the input text. The resulting query embedding is then compared against the stored file embeddings in a vector database using cosine similarity as the distance metric (11):

$$\cos(\hat{e}_q, \hat{e}_i) = \hat{e}_q \cdot \hat{e}_i, \quad \left\|\hat{e}_q\right\| = \left\|\hat{e}_i\right\| = 1. \quad (11)$$

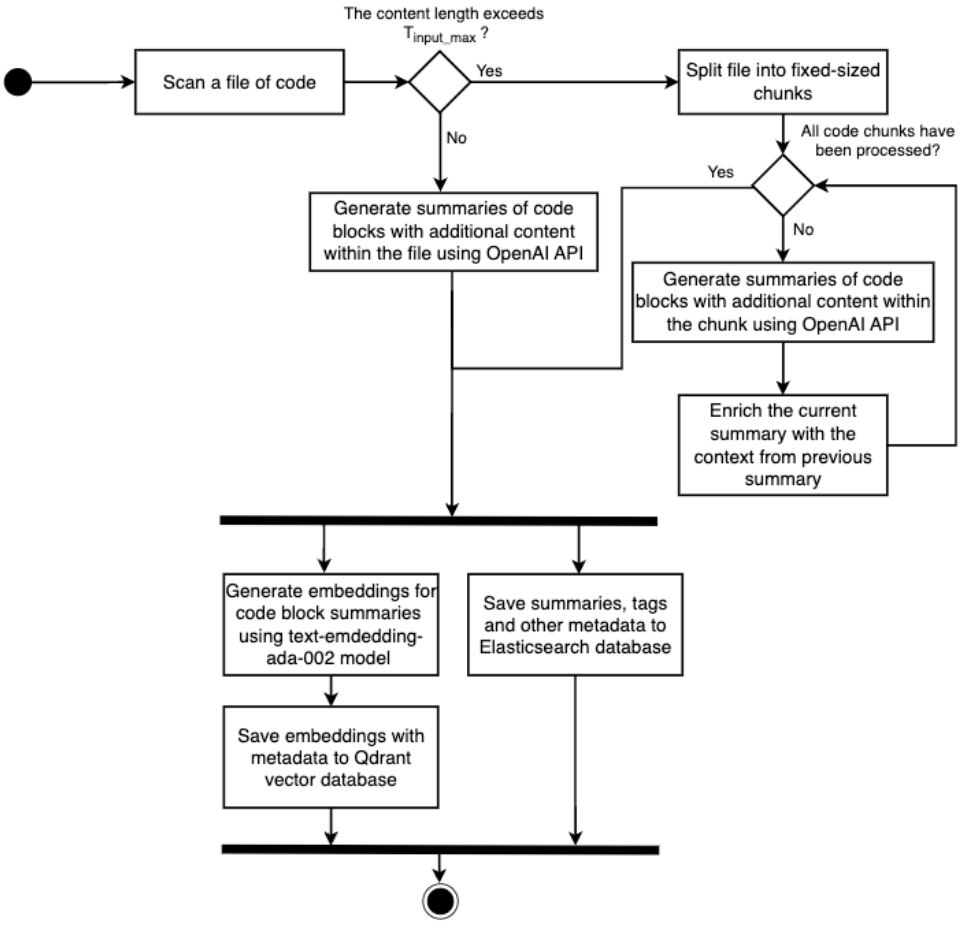


Figure 3 – Activity diagram of indexing phase

DOI 10.15588/1607-3274-2025-3-6

This retrieval path enables the system to capture conceptual and contextual similarities between the query and the code summaries, even when there is no direct lexical correspondence.

The results from both retrieval paths are processed independently and can be returned as separate ranked lists or integrated into a unified ranking. In cases where both engines yield results for the same file, these results can be merged, with ranking adjusted based on predefined weighting or scoring strategies. Each retrieved file entry includes associated metadata such as the file name, path, summary, and optionally the top-matching tags or score explanations from both retrieval methods (12):

$$\begin{aligned} &S(b_i) = \alpha \cdot BM25(q, d_i) + \\ &+ (1-\alpha) \cdot \cos(\hat{e}_q, \hat{e}_i),\ \alpha \in [0,1], \\ &Y = TopK(\{b_i \in B \mid S(b_i)\}). \end{aligned} \tag{12}$$

The dual retrieval strategy enables the system to respond effectively to a wide range of queries, from those requiring strict keyword matches to those involving abstract or concept-driven search intent. It also supports fallback mechanisms in cases where one of the retrieval paths returns no results or results of low relevance. This phase concludes with the identification of candidate files that are passed to the next stage of the pipeline for more granular analysis and line-level code identification. Fig. 4 illustrates the activity diagram of a retrieval phase.

Following the retrieval of candidate source code blocks through text-based and semantic search mechanisms, the post-processing phase is responsible for narrowing down the search results to the most relevant segments of code at a finer granularity. This stage is particularly important when user queries pertain to specific functional behavior or logic that is confined to specific code lines, rather than the code block as a whole.

To enable this refinement, each retrieved code block is further analyzed using an LLM, which operates on the full content of it, its summary, and the original user query. The objective of the model is to determine which parts of the code are most likely to satisfy the semantic intent of the query. This is achieved by prompting the model with both the query and the block-level context, asking it to identify and return the specific lines or regions of interest within the code. If training or evaluation with ground truth matches, we may define a binary relevance label $y_i \in \{0,1\}$ indicating whether code block f_i is relevant to query q, and predicted score $\hat{y}_i = S(f_i)$. We can evaluate with binary cross-entropy loss:

$$L = -\sum_{i=1}^{n} \left[y_i \log(\hat{y}_i) + (1-y_i) \log(1-\hat{y}_i) \right]. \tag{13}$$

The model receives as input the user's natural language query, the code block summary that was previously generated and indexed, and the source code in a line range.

Then, hybrid retrieval with a refinement is shown in formula (14):

$$Y' = \{GPTR(b_i) \mid b_i \in Y\}. \tag{14}$$

Thus, the entire formula expresses that every code block retrieved by the initial hybrid search is further refined using LLM, and the resulting set contains the final, context-aware, and human-readable answers that the system returns to the user. This operation ensures that the output is not just a ranked list of code blocks but a targeted extraction of meaningfully relevant code fragments.

The prompt structure, which is represented in Fig. 5, encourages the model to scan the code in context and locate logic segments that exhibit semantic alignment with the query. In some cases, the model may return an exact set of line numbers that correspond to the requested functionality.

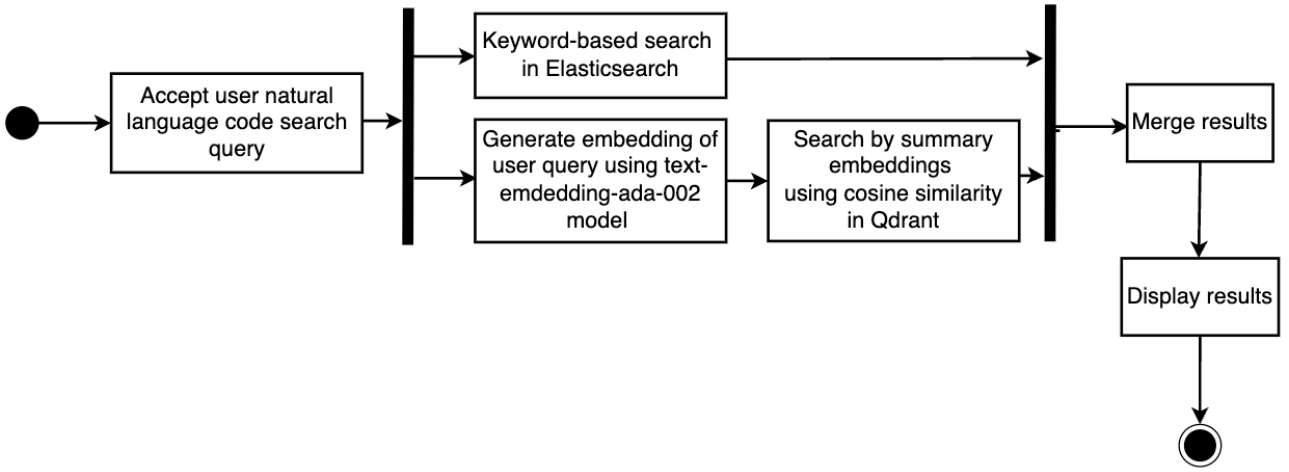


Figure 4 – Activity diagram of retrieval phase

DOI 10.15588/1607-3274-2025-3-6

```
[
  {
    "role": "system",
    "content": "You are a code analysis assistant."
  },
  {
    "role": "user",
    "content": "Ananlyze a code block to identify which parts of
the code most closely match the following user query based on
the summary and give the line numbers of possible match"
  },
  {
    "role": "assistant",
    "content": "Provide me with a summary and source code"
  },
  {
    "role": "user",
    "content": "Summary: <code_block_summary>;
               Source code: <source_code>
               Line range is: <start_line_number> -
               <end_line_number>"}
]
```

Figure 5 – Example of prompts for code lines retrieving within a post-processing stage

To optimize performance and reduce inference time, only the most relevant results (typically the top five according to the hybrid scoring function) are selected for automatic refinement. The remaining results are excluded from immediate processing and may be refined on demand, based on user interaction. This selective approach ensures scalability while still allowing detailed semantic analysis when needed. Thus, the final formula of hybrid search with refinement of top 5 the most relevant results is presented (15):

$$\begin{aligned} Y_{top} &= TopK(Y,5), \\ Y' &= \{GPTR(b_i) \mid b_i \in Y_{top}\} \cup \\ &\cup \{b_i \mid b_i \in (Y \setminus Y_{top})\}. \end{aligned} \quad (15)$$

The granularity of analysis in this phase is intended to improve the specificity and relevance of search results. While previous stages identify files likely to contain relevant content, this phase identifies and highlights the exact implementation points within those files. The output of this step may be a ranked list of method names, code excerpts, or line ranges, depending on how the model is instructed and the formatting required for downstream presentation. Fig. 6 illustrates the post-processing phase.

The proposed hybrid code retrieval method integrates structured indexing, dual-mode retrieval, and LLM-assisted post-processing to address the challenges of semantic code search in large codebases. Beginning with context-aware summarization and embedding during the indexing phase, the system enables flexible querying through parallel keyword-based and vector-based search mechanisms. The retrieval process balances lexical precision and semantic understanding, while the final line-level refinement phase leverages large language models to isolate the most relevant code fragments based on user intent. Together, these components form a scalable and context-sensitive search pipeline that supports both broad discovery and fine-grained code navigation.

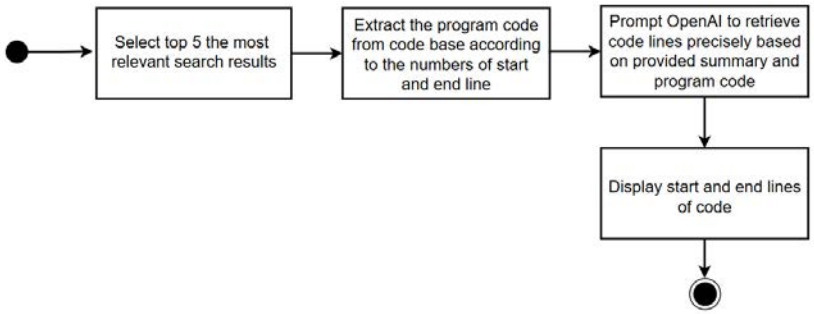


Figure 6 – Post-processing phase

4 EXPERIMENTS

To evaluate the applicability and effectiveness of the proposed hybrid code search method, a set of experiments was conducted using a real-world software codebase. The system was implemented in a .NET-based environment, incorporating components for indexing, vector storage, text search, and OpenAI API integration for both summarization and post-retrieval code refinement.

The primary goal of the experiments was to assess the ability of the system to retrieve semantically relevant code segments based on natural language queries. A test codebase in the domain of service management was selected, containing multiple classes and methods written in C#. This environment enabled the evaluation of the pipeline across different levels of abstraction – from file-level indexing to line-level code extraction.

The indexing phase was implemented as a standalone preprocessing utility that analyzed all .cs files in the selected codebase. Each file was passed through a summarization module based on OpenAI's GPT-4o model. In cases where a file exceeded the input length limitations, a recursive chunking strategy was applied, and the final summary was constructed by aggregating context-enriched chunk-level summaries.

The generated summaries were then embedded using the text-embedding-ada-002 model, and the resulting vectors were stored in a local instance of Qdrant, a high-performance vector database. Simultaneously, the summaries and metadata – including generated semantic tags – were indexed using a locally hosted Elasticsearch instance to support keyword-based retrieval.

DOI 10.15588/1607-3274-2025-3-6

Natural language queries were submitted through a simple web interface. Upon query submission, both text-based search and embedding-based retrieval were performed in parallel. Retrieved results were displayed to the user along with cosine similarity scores and matched summaries.

Following initial retrieval, the full source code of selected files was passed to GPT-4o for code analysis. The model was prompted with the original query and the code block summary and instructed to return the most relevant methods or code blocks in structured JSON format, including estimated line numbers and explanatory notes. This allowed for refined pinpointing of logic relevant to the user's intent.

A representative example query "Find places where service orders are filtered by ID" was tested on the indexed codebase. The system successfully retrieved a method named SearchOrders located in the OrderService.cs file. The post-processing phase highlighted the exact lines performing filtering based on the CustomerId field. The result was presented with the matched method, its position in the file, and a step-by-step explanation of the code logic.

5 RESULTS

To evaluate the performance and responsiveness of the implemented hybrid code search pipeline, a series of experiments were conducted using a codebase consisting of 25 C# source files. The average file size was approximately 30–40 KB, and the files varied in structural complexity, encompassing service classes, data repositories, and utility methods. The objective of the evaluation was to confirm the system's ability to index, retrieve, and refine relevant code fragments using the described keyword-based, embedding-based, and LLM-assisted methods.

During the indexing phase, all files were successfully processed without failure. Summarization of each file or chunk (for larger files) was completed using the OpenAI GPT-4o model, followed by embedding generation with the text-embedding-ada-002 model. Vector data was stored in a Qdrant instance running locally, while summaries and tags were indexed using a local Elasticsearch server.

To test retrieval performance, 10 queries were selected, each representing typical developer requests such as "Find where service orders are filtered by ID" or "Show methods that generate auth tokens." For each query, both search paths were executed in parallel. On average, the system returned relevant results within 0.90 to 1.0 seconds, including post-processing by GPT for line-level code matching only for the most relevant results.

The average time for OpenAI API calls during post-retrieval refinement was approximately 0.94 seconds, while the combined time for vector search and keyword search was under 500 milliseconds. These results indicate that the system can operate within practical response times suitable for interactive use. During testing, the standard Tier 1 subscription was used. In high-throughput scenarios, batch processing or queuing would be necessary to avoid exceeding request quotas.

To better understand the effectiveness of the proposed hybrid code search method, a benchmarking comparison (Table 1) was conducted against two baseline approaches:

– baseline A – traditional keyword-based retrieval using Elasticsearch with BM25 scoring;

– baseline B – embedding-only semantic search using text-embedding-ada-002 vectors in Qdrant without keyword filtering or refinement;

– proposed hybrid approach – parallel execution of both search strategies, followed by GPT-based post-processing for line-level code matching.

The evaluation dataset consisted of 10 developer-like queries formulated in natural language. Relevance was manually assessed by analyzing whether the retrieved code matched the intended logic or functionality described in the query. The evaluation was performed using the following metrics:

– top-1 precision – if the top result was relevant;

– top-5 recall – the proportion of relevant items among the top 5;

– MRR – average inverse rank of the first relevant result [19];

– average response time – total time to produce the final result, including post-processing.

The hybrid approach significantly outperforms both baselines in terms of retrieval quality, particularly for queries with complex or abstract semantics. The use of LLM-based summarization and refinement contributes to higher Top-1 precision and MRR scores, demonstrating the system's ability to retrieve not only relevant files but also the most accurate code segments within them.

While the hybrid method with refinement introduces additional latency due to post-processing, the average response time of approximately 0.94 seconds remains within acceptable bounds for interactive search tasks and it has been optimizing by handling a refinement of the most relevant result and the rest of results are intended to be processed on demand by user interaction. In comparison, embedding-only search is faster but occasionally less precise due to the absence of textual disambiguation and LLM refinement.

The results confirm the system's ability to produce accurate, context-sensitive, and developer-usable code search outputs across a heterogeneous codebase while maintaining acceptable execution times for all processing stages.

Table 1 – Benchmarking comparison

Method	Top-1 precision	Top-5 recall	MRR	Average time (s)
Baseline A (BM25)	0.58	0.69	0.61	0.40
Baseline B (Embeddings)	0.75	0.88	0.80	0.28
Hybrid	0.91	1.00	0.92	0.94

DOI 10.15588/1607-3274-2025-3-6

6 DISCUSSION

The proposed hybrid code retrieval method demonstrates practical applicability for source code search across large-scale and heterogeneous codebases. In comparison to prior research in code search and semantic retrieval [4–10], this approach eliminates the need for training custom models by leveraging general-purpose, pre-trained language models. Unlike domain-specific models, which often require fine-tuning on large, curated datasets, the use of GPT-based APIs allows the system to remain flexible and adaptable to a broad range of programming styles and query types without retraining.

While custom-trained models may exhibit strong performance in narrow domains, they often suffer from limited generalization when applied to unfamiliar codebases or other programming languages. The hybrid method presented in this work benefits from the broad domain coverage and general language understanding embedded in OpenAI's GPT models, enabling it to interpret developer queries more naturally and perform code summarization in a context-aware manner.

An advantage of the system lies in its layered architecture, which combines the precision of keyword-based retrieval with the semantic depth of vector-based embedding search. The addition of GPT-based post-processing further enhances the system's ability to localize relevant code fragments within files, aligning search outputs with user intent. The results of the experiments confirm that the hybrid method achieves higher precision and recall compared to standalone search methods, especially for abstract or semantically rich queries.

However, the system also inherits limitations from its reliance on external APIs. The OpenAI GPT models, while highly capable, are constrained by token-based limits and subscription-dependent rate quotas. These constraints may impact the scalability of the method in high-throughput or real-time search scenarios. To mitigate this, the system includes a chunking and recursive summarization strategy to handle large files, ensuring full coverage of the codebase even when input sizes exceed model capacity.

Moreover, while the current pipeline performs well in general software engineering contexts, future improvements may involve domain-adaptive summarization or the incorporation of static analysis techniques (e.g., AST matching or control flow analysis) to further enrich search quality. Another promising direction involves integrating the hybrid method into development environments allowing for contextual, in-line code discovery and reuse during software maintenance or refactoring tasks.

CONCLUSIONS

The hybrid code search method was developed and implemented as a solution that combines keyword-based retrieval, vector-based semantic search, and LLM-driven summarization and refinement. The system was tested on real-world codebases and evaluated using standard search effectiveness metrics to validate its practical applicability.

The scientific novelty of the obtained results lies in the integration of multiple retrieval modalities into a unified pipeline, enhanced by recursive summarization and line-level reasoning via GPT model. The proposed method introduces a structured, context-aware approach to code retrieval, enabling semantic alignment between developer queries and relevant code segments across a large-scale codebase.

The practical significance of the obtained results is reflected in the method's ability to automate code retrieval tasks without relying on rigid structures or manually crafted rules. This flexibility allows developers to search using natural language and receive highly relevant results at both the file and method levels. The modular architecture facilitates integration into software engineering workflows, development environments, and documentation systems.

Prospects for further research include exploring optimization strategies to reduce dependency on API rate limits and improve runtime performance in large-scale deployments. Additional directions may involve the use of static code analysis techniques, domain-adaptive summarization models, and the expansion of hybrid retrieval methods into other software engineering domains, including automated documentation, test generation, and intelligent code navigation tools.

REFERENCES

1. Kumar Vivek, Chinmay Bhatt, Varsha Namdeo A framework for document plagiarism detection using Rabin Karp method, *International Journal of Innovative Research in Technology and Management*, 2021, Vol. 5, pp. 17–30.
2. Zhang Ashley Ge, Chen Yan, Oney Steve RunEx: Augmenting Regular-Expression Code Search with Runtime Values, *Proceedings of the 2023 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 2023, pp. 145–155. DOI: 10.1109/VL-HCC57772.2023.00024
3. Karnalim Oscar, Simon Syntax Trees and Information Retrieval to Improve Code Similarity Detection, *Proceedings of the Twenty-Second Australasian Computing Education Conference (ACE 2020)*, 2020, pp. 48–55. DOI: 10.1145/3373165.3373171
4. Liu Chao, Xia Xin, Lo David, Liu Zhiwei, Hassan Ahmed E., Li Shanping CodeMatcher: Searching Code Based on Sequential Semantics of Important Query Words. Ithaca, arXiv, 2020, 36 p. (Preprint / arXiv; 2005.14373). DOI: 10.1145/3465403
5. Kong Xianglong, Chen Hongyu, Yu Ming, Zhang Lixiang Boosting Code Search with Structural Code Annotation, *Electronics*, 2022, Vol. 11, No. 19, P. 3053. DOI: 10.3390/electronics11193053
6. Gotmare Khilesh Deepak, Li Junnan, Joty Shafiq, Hoi Steven C. H. Cascaded Fast and Slow Models for Efficient Semantic Code Search. Ithaca: arXiv, 2021, 12 p. (Preprint / arXiv; 2110.07811). DOI: 10.48550/arXiv.2110.07811
7. Jain Sarthak, Dora Aditya, Sam Ka Seng, Singh Prabhat LLM Agents Improve Semantic Code Search. Ithaca, arXiv, 2024, 12 p. (Preprint / arXiv; 2408.11058). DOI: 10.48550/arXiv.2408.11058
8. Khan M. A. M. Development of a code search engine using natural language processing technique: Graduate thesis.

DOI 10.15588/1607-3274-2025-3-6

IUT, Department of Computer Science and Engineering, 2023, 65 p.
9. Deng Zhongyang, Xu Ling, Liu Chao, Huangfu Luwen, Yan Meng Code semantic enrichment for deep code search, *Journal of Systems and Software*, 2024, Vol. 207, P. 111856. DOI: 10.1016/j.jss.2023.111856
10. Chen Junkai, Hu Xing, Li Zhenhao, Gao Cuiyun, Xia Xin, Lo David Code Search Is All You Need? Improving Code Suggestions with Code Search, *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (ICSE 2024), Lisbon, Portugal, April 14–20, 2024*, 2024, Article No. 73, pp. 1–13. DOI: 10.1145/3597503.3639085
11. Nate Suraj, Patil Om, Medar Shreenidhi, Deshmukh Jyoti A Survey on Transformer-based Models in Code Summarization, *International Research Journal on Advanced Engineering Hub (IRJAEH)*, 2025, Vol. 3, pp. 740–745. DOI: 10.47392/IRJAEH.2025.0103
12. Parmar Mihir, Deilamsalehy Hanieh, Dernoncourt Franck, Yoon Seunghyun, Rossi Ryan A., Bui Trung Towards Enhancing Coherence in Extractive Summarization: Dataset and Experiments with LLMs, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 19810–19820. DOI: 10.18653/v1/2024.emnlp-main.1106
13. Korade Nilesh Bhikaji, Salunke Mahendra B., Bhosle Amol, Kumbharkar Prashant Babarao, Asalkar Gayatri, Khedkar Rutuja G. Strengthening Sentence Similarity Identification Through OpenAI Embeddings and Deep Learning, *International Journal of Advanced Computer Science and Applications*, 2024, Vol. 15, No. 4, pp. 821–829. DOI: 10.14569/IJACSA.2024.0150485
14. OpenAI. New and improved embedding model [Electronic resource], *OpenAI*, Mode of access: https://openai.com/index/new-and-improved-embedding-model (date of access: 09.04.2025). – Title from screen.
15. Patil Rajvardhan, Boit Sorio, Gudivada Venkat N., Nandigam Jagadeesh A Survey of Text Representation and Embedding Techniques in NLP, *IEEE Access*, 2023, Vol. 11, pp. 36120–36146. DOI: 10.1109/ACCESS.2023.3266377
16. Jiang Xue, Wang Weiren, Tian Shaohan, Wang Hao, Lookman Turab, Su Yanjing Applications of natural language processing and large language models in materials discovery, *npj Computational Materials*, 2025, Vol. 11. DOI: 10.1038/s41524-025-01554-0
17. Qdrant. Qdrant Vector Database: High-performance vector similarity search [Electronic resource], *Qdrant Documentation*. Mode of access: https://qdrant.tech/qdrant-vector-database (date of access: 09.04.2025). – Title from screen.
18. Elastic. Elasticsearch: The Official Distributed Search & Analytics Engine [Electronic resource], *Elastic*. Mode of access: https://www.elastic.co/elasticsearch (date of access: 09.04.2025). Title from screen.
19. Hoyt Charles Tapley, Berrendorf Max, Galkin Mikhail, Tresp Volker, Gyori Benjamin M. A Unified Framework for Rank-based Evaluation Metrics for Link Prediction in Knowledge Graphs. Ithaca: arXiv, 2022, 18 p. (Preprint / arXiv; 2203.07544). DOI: 10.48550/arXiv.2203.07544

Received 12.04.2025.
Accepted 21.06.2025.

УДК 004.93, 004.8

МЕТОД ПАРАЛЕЛЬНОГО ГІБРИДНОГО ПОШУКУ ДЛЯ ВЕЛИКИХ РЕПОЗИТОРІЇВ КОДУ

Бойко В. О. – асистент кафедри інженерії програмного забезпечення Хмельницького національного університету, Хмельницький, Україна.

АНОТАЦІЯ

Актуальність. Сучасні програмні системи містять великі кодові бази, що робить пошук коду критично важливим завданням для розробників програмного забезпечення. Традиційні методи пошуку коду спираються на співставлення за ключовими словами або структурний аналіз, але часто не здатні відобразити семантичний зміст запитів користувачів або мають проблеми з неструктурованим та непослідовно задокументованим кодом. Останнім часом семантичний векторний пошук і великі мовні моделі (LLM) показали перспективи в покращенні розуміння коду. Проблема полягає в розробці масштабованого, точного та гібридного методу пошуку коду, здатного знаходити відповідні фрагменти коду на основі як текстових запитів, так і семантичного контексту, при цьому підтримуючи паралельну обробку та пошуку на основі метаданих.

Мета роботи – розробка гібридного методу семантичного пошуку коду шляхом комбінування фільтрації за ключовими словами та пошуку на основі вбудованих представлень, доповненого сумаризацією та семантичними тегами, згенерованими за допомогою LLM для підвищення точності та ефективності пошуку відповідних елементів коду у великих кодових репозиторіях.

Метод. Для досягнення мети дослідження розроблено метод пошуку з двома шляхами з пост-обробкою, де пошук за текстовими ключовими словами та пошук на основі вбудовуваних семантичних представлень виконуються паралельно. Блоки коду попередньо обробляються за допомогою GPT-4o моделі для генерування сумаризації та семантичних тегів.

Результати. Метод реалізовано та перевірено на кодовій базі .NET, що продемонструвало покращену точність при знаходженні семантично релевантних методів. Комбінація паралельних шляхів пошуку та метаданих, згенерованих LLM, покращила якість результатів. Для підвищення релевантності було застосовано LLM-постобробку яка виконується над найбільш релевантними результатами, що дозволяє точніше локалізувати потрібні рядки коду в межах знайдених фрагментів. Інші результати можуть бути оброблені на вимогу користувача.

Висновки. Експериментальні результати підтвердили працездатність та практичну застосовність запропонованої гібридної системи пошуку коду. Модульна архітектура системи підтримує робочі процеси розробників в реальному часі, а її розширюваність дозволяє впроваджувати майбутні покращення через активне навчання та зворотний зв'язок від користувачів. Подальші дослідження можуть бути спрямовані на оптимізацію стратегій вибору вбудованих представлень, інтеграцію автоматичного переформатування запитів та масштабування у багатомовних кодових середовищах.

КЛЮЧОВІ СЛОВА: гібридний пошук коду, векторний пошук, семантичні вбудовування, сумаризація коду, метадані, згенеровані LLM, косинусна схожість, текстова релевантність, пошук класів та методів, індексування на основі класів, інженерія програмного забезпечення.

DOI 10.15588/1607-3274-2025-3-6

ЛІТЕРАТУРА

1. Kumar V. A framework for document plagiarism detection using Rabin Karp method / Vivek Kumar, Bhatt Chinmay, Namdeo Varsha // International Journal of Innovative Research in Technology and Management. – 2021. – Vol. 5. – P. 17–30.
2. Zhang A. G. RunEx: Augmenting Regular-Expression Code Search with Runtime Values / Ashley Ge Zhang, Yan Chen, Steve Oney // Proceedings of the 2023 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). – 2023. – P. 145–155. DOI: 10.1109/VL-HCC57772.2023.00024
3. Karnalim O. Syntax Trees and Information Retrieval to Improve Code Similarity Detection / Oscar Karnalim, Simon // Proceedings of the Twenty-Second Australasian Computing Education Conference (ACE 2020). – 2020. – P. 48–55. DOI: 10.1145/3373165.3373171
4. CodeMatcher: Searching Code Based on Sequential Semantics of Important Query Words / [Chao Liu, Xin Xia, David Lo et al.]. – Ithaca : arXiv, 2020. – 36 p. – (Preprint / arXiv; 2005.14373). DOI: 10.1145/3465403
5. Boosting Code Search with Structural Code Annotation / [Xianglong Kong, Hongyu Chen, Ming Yu, Lixiang Zhang] // Electronics. – 2022. – Vol. 11, No. 19. – P. 3053. DOI: 10.3390/electronics11193053
6. Cascaded Fast and Slow Models for Efficient Semantic Code Search / [Khilesh Deepak Gotmare, Junnan Li, Shafiq Joty, Steven C. H. Hoi]. – Ithaca: arXiv, 2021. – 12 p. – (Preprint / arXiv; 2110.07811). DOI: 10.48550/arXiv.2110.07811
7. LLM Agents Improve Semantic Code Search / [Sarthak Jain, Aditya Dora, Ka Seng Sam, Prabhat Singh]. – Ithaca: arXiv, 2024. – 12 p. – (Preprint / arXiv; 2408.11058). DOI: 10.48550/arXiv.2408.11058
8. Khan M. A. M. Development of a code search engine using natural language processing technique: Graduate thesis / Mohammad Abdullah Matin Khan. – IUT, Department of Computer Science and Engineering, 2023. – 65 p.
9. Code semantic enrichment for deep code search / [Zhongyang Deng, Ling Xu, Chao Liu et al.] // Journal of Systems and Software. – 2024. – Vol. 207. – P. 111856. DOI: 10.1016/j.jss.2023.111856
10. Code Search Is All You Need? Improving Code Suggestions with Code Search / [Junkai Chen, Xing Hu, Zhenhao Li et al.] // Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (ICSE 2024), Lisbon, Portugal, April 14–20, 2024. – 2024. – Article No. 73. – P. 1–13. DOI: 10.1145/3597503.3639085
11. A Survey on Transformer-based Models in Code Summarization / [Suraj Nate, Om Patil, Shreenidhi Medar, Jyoti Deshmukh] // International Research Journal on Advanced Engineering Hub (IRJAEH). – 2025. – Vol. 3. – P. 740–745. DOI: 10.47392/IRJAEH.2025.0103
12. Towards Enhancing Coherence in Extractive Summarization: Dataset and Experiments with LLMs / [Mihir Parmar, Hanieh Deilamsalehy, Franck Dernoncourt et al.] // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. – 2024. – P. 19810–19820. DOI: 10.18653/v1/2024.emnlp-main.1106
13. Strengthening Sentence Similarity Identification Through OpenAI Embeddings and Deep Learning / [Nilesh Bhikaji Korade, Mahendra B. Salunke, Amol Bhosle et al.] // International Journal of Advanced Computer Science and Applications. – 2024. – Vol. 15, No. 4. – P. 821–829. DOI: 10.14569/IJACSA.2024.0150485
14. OpenAI. New and improved embedding model [Electronic resource] // OpenAI. – Mode of access: https://openai.com/index/new-and-improved-embedding-model (date of access: 09.04.2025). – Title from screen.
15. A Survey of Text Representation and Embedding Techniques in NLP / [Rajvardhan Patil, Sorio Boit, Venkat N. Gudivada, Jagadeesh Nandigam] // IEEE Access. – 2023. – Vol. 11. – P. 36120–36146. DOI: 10.1109/ACCESS.2023.3266377
16. Applications of natural language processing and large language models in materials discovery / [Xue Jiang, Weiren Wang, Shaohan Tian et al.] // npj Computational Materials. – 2025. – Vol. 11. DOI: 10.1038/s41524-025-01554-0
17. Qdrant. Qdrant Vector Database: High-performance vector similarity search [Electronic resource] // Qdrant Documentation. – Mode of access: https://qdrant.tech/qdrant-vector-database (date of access: 09.04.2025). – Title from screen.
18. Elastic. Elasticsearch: The Official Distributed Search & Analytics Engine [Electronic resource] // Elastic. – Mode of access: https://www.elastic.co/elasticsearch (date of access: 09.04.2025). – Title from screen.
19. A Unified Framework for Rank-based Evaluation Metrics for Link Prediction in Knowledge Graphs / [Charles Tapley Hoyt, Max Berrendorf, Mikhail Galkin et al.]. – Ithaca: arXiv, 2022. – 18 p. – (Preprint / arXiv; 2203.07544). DOI: 10.48550/arXiv.2203.07544

DOI 10.15588/1607-3274-2025-3-6

UDC 004.93

URBAN SCENE SEGMENTATION USING HOMOGENEOUS U-NET ENSEMBLE: A STUDY ON THE CITYSCAPES DATASET

Hmyria I. O. – Post-graduate student of the Department of Software Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Kravets N. S. – Associate Professor, Associate Professor of the Department of Software Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

ABSTRACT

Context. Semantic segmentation plays a critical role in computer vision tasks such as autonomous driving and urban scene understanding. While designing new model architectures can be complex, improving performance through ensemble techniques applied to existing models has shown promising potential. This paper investigates ensemble learning as a strategy to enhance segmentation accuracy without modifying the underlying U-Net architecture.

Objective. The aim of this work is to develop and evaluate a homogeneous ensemble of U-Net models trained with distinct initialization and data augmentation techniques, and to assess the effectiveness of various ensemble aggregation strategies in improving segmentation performance on complex urban dataset.

Method. The proposed approach constructs an ensemble of five structurally identical U-Net models, each trained with unique weight initialization and augmentation schemes to ensure prediction diversity. Several ensemble strategies are examined, including softmax averaging, max voting, proportional weighting, exponential weighting, and optimized weighted voting. Evaluation is conducted on the Cityscapes dataset using a range of segmentation metrics.

Results. Experimental findings demonstrate that ensemble models outperform individual U-Net instances and the baseline in terms of accuracy, mean IoU, and specificity. The optimized weighted ensemble achieved the highest accuracy (87.56%) and mean IoU (0.6504), exceeding the best individual model by approximately 3%. However, these improvements come with a notable increase in inference time, highlighting a trade-off between accuracy and computational efficiency.

Conclusions. The ensemble-based approach effectively enhances segmentation accuracy while leveraging existing model architectures. Although the increased computational cost presents a limitation for real-time applications, the method is well-suited for high-precision tasks. Future research will focus on reducing inference time and extending the ensemble methodology to other architectures and datasets.

KEYWORDS: convolutional neural network, semantic segmentation, U-Net, ensemble learning, data augmentation techniques, model initialization, Cityscapes, urban scenes.

ABBREVIATIONS

CNN is a convolutional Neural Network;
U-Net is a U-shaped network architecture;
ELU is an Exponential Linear Unit;
ReLU is a Rectified Linear Unit;
IoU is a Intersection over Union;
Mean IoU / mIoU is a Mean Intersection over Union;
TP is a True Positive;
FP is a False Positive;
FN is a False Negative;
TN is a True Negative;
RGB is a Red Green Blue.

NOMENCLATURE

$F(X)$ is a convolutional neural network performing semantic segmentation;
X is an input image space;
Y is an output segmentation map (label space);
x is an input image;
y is a ground truth segmentation map;
H is a height of the input image;
W is a width of the input image;
C is a number of channels of the input image;
$\hat{y}$ is a predicted segmentation output;
$\arg\max_k$ is a predicted class;
c_i is a segmentation class;
$F_i(x)$ is a prediction of the U-Net model in the ensemble;
w_i is a weight assigned to the model in the ensemble;
$E(x)$ is an ensemble output;
$\hat{y}_{final}$ is a final predicted class map from the ensemble;
$Metric$ is a segmentation metric;
$accuracy$ is an accuracy of a model;
$T(F_i)$ is an inference time of the model;
$T(E)$ is a total inference time of the ensemble;
IoU is an intersection over union;
N is a normal distribution with mean;
μ is a variance;
α is a displacement intensity in elastic deformation;
σ is a standard deviation of Gaussian noise;
w is an ensemble weight vector;
fan_in is a number of input units;
fan_out is a number of output units.

INTRODUCTION

Deep learning has made impressive strides in image segmentation, especially when it comes to parsing

DOI 10.15588/1607-3274-2025-3-7

complex urban scenes. This task is essential for technologies like autonomous vehicles, smart traffic systems, and overall city infrastructure management. A widely recognized benchmark in this domain is the Cityscapes dataset [1], known for its high-resolution, finely annotated road scene images that have become standard for evaluating model performance.

Architectures like U-Net have shown strong results in semantic segmentation, yet they still struggle with generalizing across varying conditions – think changes in lighting, weather, or city layouts. To mitigate these issues, ensemble learning has emerged as a valuable approach. By merging predictions from multiple models, it boosts accuracy and stabilizes results. While traditional ensemble methods often mix different types of models [2], a homogeneous ensemble – where multiple U-Net models are trained separately – offers a balance between performance and efficiency. This diversity, introduced through unique initializations and data augmentations, helps improve outcomes without the added complexity of mixing architectures.

In this work, we investigate several homogeneous ensembling techniques – such as averaging, max pooling, and weighted voting – to see how each impacts segmentation results. We propose a U-Net-based ensemble tailored for urban image segmentation, using distinct augmentation and initialization variations, and run experiments to evaluate its effectiveness. Our analysis includes comparisons of accuracy, speed, and computational overhead to weigh the trade-offs of each method.

The object of study in this research is semantic segmentation of urban scenes, specifically focusing on the challenges posed by varying environmental conditions, such as changes in lighting, weather, and occlusions. The study is conducted using the Cityscapes dataset, which provides high-resolution images of complex urban environments with 34 semantic classes.

The subject of study is the method of constructing homogeneous U-Net ensembles to improve the accuracy and robustness of semantic segmentation in urban environments. This includes exploring different ensemble strategies to enhance generalization across diverse scenes while maintaining computational efficiency.

The purpose of this work is to **i**mprove the generalization ability of U-Net models for urban scene segmentation by leveraging networks homogeneous ensembling. The study aims to demonstrate that a homogeneous ensemble of multiple U-Net models can achieve higher segmentation accuracy and robustness compared to a single U-Net, particularly in conditions found in real-world urban environments.

1 PROBLEM STATEMENT

Formally, the semantic segmentation task can be described as a pixel-wise classification problem, where the goal is to assign each pixel of an input image to one of the predefined semantic classes.

Let the input image be denoted as $x \in X$, where $X \subset R^{H \times W \times C}$. The output segmentation map is denoted as $y \in Y$, where $Y \subset Z^{H \times W}$, and each pixel value corresponds to a class label from the set of predefined classes $C = \{c_1, c_2, ..., c_K\}$.

The model is a function $F : X \to Y$, implemented using a deep convolutional neural network architecture, particularly U-Net. The output of the model is a probability tensor $\hat{y} = F(x) \in [0,1]^{H \times W \times K}$, and the predicted class for each pixel is obtained by:

$$\hat{y}_{i,j} = \arg\max_{k \in C} F(x)_{i,j,k} .$$

Let's define a set of n trained models $\{F_1, F_2, ..., F_n\}$ each producing a prediction $\hat{y}_i = F_i(x)$. We define the ensemble function E as a weighted combination of the model outputs $E(x) = \sum_{i=1}^{n} w_i F_i(x)$, subject to $\sum_{i=1}^{n} w_i = 1$, $w_i \geq 0$. The final prediction is obtained by taking the argmax over the ensembled output:

$$\hat{y}_{final} = \arg\max_{k \in C} E(x)_{i,j,k} .$$

The objective is to find the optimal weight vector $w = (w_1, w_2, ..., w_n)$ such that the ensemble prediction maximizes a chosen segmentation quality metric, such as the mean Intersection over Union (mIoU):

$$\max_{w} Metric(E(x), y), \ \sum_{i=1}^{n} w_i = 1, \ w_i \geq 0.$$

Additionally, due to hardware limitations, inference must be performed on a CPU-based system, where parallel execution is not available, and models are evaluated sequentially. Let $T(F_i)$ denote the execution time of model F_i. The ensemble execution time is therefore:

$$T(E) = \sum_{i=1}^{n} T(F_i).$$

The constraint is to keep inference time within an acceptable range $T_{\max}$, defined based on application requirements:

$$T(E) \leq T_{\max} .$$

DOI 10.15588/1607-3274-2025-3-7

2 REVIEW OF THE LITERATURE

Image segmentation [3] plays a key role in computer vision, and in robot vision, aiming to identify and outline objects within an image. Traditional methods – like thresholding [4], region growing [5], and edge detection [6] – have largely given way to deep learning-based approaches [7–9], which excel at learning layered features and handling complex textures.

Among deep learning models, Convolutional Neural Networks (CNNs) [10] have become the backbone of many segmentation tasks. Fully Convolutional Networks (FCNs) [11] were among the earliest deep models to perform pixel-level classification, showing the potential of CNNs in segmentation. Yet, U-Net [12] – a fully convolutional encoder-decoder design with skip connections – has emerged as the preferred choice, especially in biomedical applications. Its symmetric structure allows it to retain spatial details, making it particularly suitable for use cases like autonomous driving. Enhanced versions, such as Attention U-Net [13] and Residual U-Net [14], add mechanisms for better feature focus and improved performance in complex datasets.

Despite its strengths, U-Net and other single-model architectures often face challenges in generalizing across varied datasets. Differences in image quality, noise, and structural variations can lead to inconsistent results. These issues have encouraged the adoption of ensemble learning to boost consistency and resilience.

Ensemble learning has long been explored as a way to enhance model reliability by combining multiple learners. Techniques like bagging [15], boosting [16], and stacking [17] have shown success in improving generalization through model diversity. Ensemble learning has demonstrated strong performance improvements across a variety of machine learning tasks even beyond computer vision. For instance, in time series forecasting [18], authors proposed an ensemble of adaptive predictors capable of real-time learning on multivariate non stationary sequences. In segmentation, deep learning ensembles are typically either heterogeneous or homogeneous.

Heterogeneous ensembles [19], which mix various model types, can improve accuracy by capturing different feature perspectives. However, this comes at the cost of greater computational demands and system complexity. Homogeneous ensembles [20], on the other hand, use multiple instances of the same architecture, each trained under varied conditions – such as different initializations, hyperparameters, or data augmentations. Research [21] suggests that such homogeneous setups can match or even surpass heterogeneous ensembles, all while remaining more efficient.

Several studies illustrate the promise of ensembling in segmentation. One work combined 3D CNNs for brain lesion detection [22], demonstrating reduced uncertainty through model fusion. Another leveraged a U-Net ensemble trained with diverse loss functions to improve lung nodule segmentation [23]. These examples underline the benefits of ensembles in reducing prediction variance and improving robustness.

The importance of adaptivity in visual systems has been emphasized not only in segmentation architectures but also in image preprocessing approaches. For example, Smelyakov et al. [24] developed an adaptive image enhancement model for robotic vision systems, enabling real-time responsiveness to variable environmental conditions.

The existing literature consistently shows that ensemble learning enhances both the accuracy and stability of segmentation models. While many studies have tested different ensemble strategies, few have taken a detailed look at the trade-offs between segmentation accuracy and scalability. Building on previous findings, this paper proposes a homogeneous ensemble of U-Net models, each trained with unique weight initializations and augmentation schemes, using optimized voting strategy. Various inference methods are evaluated to better understand how ensemble design choices affect segmentation performance and efficiency.

3 MATERIALS AND METHODS

The Cityscapes dataset serves as a large-scale benchmark tailored for urban scene understanding, particularly focusing on tasks such as semantic segmentation, instance segmentation, and depth estimation. It features high-resolution imagery (2048×1024 pixels) captured from a vehicle-mounted camera as it navigates through 50 cities across Germany, Switzerland, and France. These images encompass a wide range of environmental conditions, including various weather scenarios and lighting settings throughout the day, thereby providing a comprehensive dataset for evaluating and training deep learning models used in autonomous driving and urban analysis.

This dataset contains 5.000 finely annotated images, distributed across 2.975 for training, 500 for validation, and 1.525 for testing, with annotations for the test set not publicly available. Additionally, it offers 20.000 coarsely annotated images as a supplementary resource. The annotation schema spans 34 semantic categories, including classes such as roads, buildings, vegetation, vehicles, pedestrians, and traffic signs. Each pixel in the finely annotated set is labeled with a semantic class, allowing for precise pixel-wise learning. Due to its detailed labeling, high resolution, and inherent class imbalance, the Cityscapes dataset has become a gold standard for evaluating segmentation models like U-Net. Given these challenges, leveraging a homogeneous ensemble of U-Net models offers a promising approach to improving segmentation performance by reducing variance and enhancing generalization, particularly in urban environments with fine-grained structures, dynamic lighting, and frequent occlusions.

To bolster the generalization capacity of the homogeneous U-Net ensemble, a diverse range of data

DOI 10.15588/1607-3274-2025-3-7

augmentation strategies was applied, with each of the five networks in the ensemble trained using a distinct transformation method. This approach promotes the learning of unique and complementary feature representations across models, thereby reducing overfitting and enhancing robustness on the Cityscapes dataset. Augmentations were chosen to simulate real-world visual variability while preserving the fundamental structure and semantics of objects within the scene. Urban environments naturally involve variations in object distance, camera perspective, noise, occlusion, and image distortion. Therefore, the selected augmentations were designed to reflect these real-world variations while maintaining semantic integrity.

Scaling was applied by randomly resizing input images within a predefined range. This helped the model develop scale-invariant features, which are critical for segmenting objects appearing at varying distances from the camera. Rotation was used to introduce random angular transformations, enhancing the model's ability to recognize and segment objects regardless of orientation – a common challenge in dynamic urban settings. Affine transformations, including shearing, translation, and reflection, were incorporated to introduce spatial diversity without disrupting the essential spatial structure of objects, thereby encouraging the model to generalize better under changes in viewpoint or alignment. One network in the ensemble was trained using elastic deformation, a technique adapted from medical imaging applications. This method simulates local, nonlinear distortions within the image, which is particularly useful for modeling real-world deformations in classes like pedestrians or vehicles, which often exhibit variable shapes and poses. Gaussian noise was added to simulate sensor noise, compression artifacts, and environmental distortions. This augmentation made the model more resilient to unpredictable visual noise and inconsistencies present in real-world imagery.

By assigning a unique augmentation strategy to each model, the ensemble was exposed to a broad spectrum of visual conditions. This diversity in learning experiences encouraged the models to acquire distinct yet complementary internal representations. Consequently, the ensemble could capture a wider range of features and generalize more effectively across complex urban scenes with challenging visual variability.

To improve training stability and ensure convergence across the homogeneous ensemble, each U-Net model was initialized using a distinct weight initialization technique. The importance of proper initialization in deep neural networks is well-established, particularly in preventing vanishing or exploding gradients, enhancing learning efficiency, and improving generalization performance. In this work, five different initialization strategies were employed: Glorot Normal, He Uniform, Orthogonal Initialization, LeCun Normal, and Random Normal.

The Glorot Normal method [25], also referred to as Xavier Normal, initializes weights from a truncated normal distribution centered at zero, with variance scaled based on both the number of incoming and outgoing connections. Weights are initialized by sampling from a truncated normal distribution centered at 0 with a standard deviation of:

$$\sigma = \sqrt{\frac{2}{\text{fan\_in} + \text{fan\_out}}}.$$

This technique helps maintain a balanced variance of activations across layers, which is particularly beneficial when using sigmoid or tanh activation functions.

He Uniform initialization [26], designed for networks employing ReLU activations, samples weights from a uniform distribution scaled by the number of input units. This ensures that activations are well-scaled during forward propagation, improving training stability in deep architectures. Weights are initialized by sampling from a uniform distribution within [–limit, limit], where:

$$\text{limit} = \sqrt{\frac{6}{\text{fan\_in}}}.$$

Orthogonal Initialization involves generating weight matrices that form an orthogonal basis, typically achieved through QR decomposition of randomly generated matrices. This approach helps preserve information flow during both forward and backward passes, making it especially effective for deep convolutional models. Weights are initialized by generating a random matrix and applying QR decomposition to obtain an orthogonal matrix. Specifically, for a weight matrix $W = Q \times R$.

LeCun Normal initialization [27] is similar in concept to Glorot Normal but scales weights based solely on the number of input units, offering improved stability for tanh and sigmoid-based networks of moderate depth. Weights are initialized by sampling from a truncated normal distribution centered at 0 with a standard deviation of:

$$\sigma = \sqrt{\frac{1}{\text{fan\_in}}}.$$

Finally, one network was initialized using a Random Normal distribution with manually specified mean and standard deviation, providing a baseline for comparing the effectiveness of more sophisticated initializers. Weights are initialized by sampling from a normal distribution:

$$W \sim N(\mu, \sigma^2).$$

The assignment of initialization methods to specific augmentation strategies was done purposefully to enhance model diversity and learning dynamics. For

DOI 10.15588/1607-3274-2025-3-7

augmentations that alter spatial characteristics – such as scaling or affine transformations – initialization techniques like Glorot Normal and Orthogonal Initialization were chosen, as they preserve activation variance even under substantial input variation. Rotation-based augmentations, which introduce directional shifts without distorting spatial structure, were paired with He Uniform initialization due to its suitability for ReLU-based networks and its ability to facilitate rapid early learning. Elastic deformation, which applies localized and nonlinear distortions, was combined with LeCun Normal initialization, providing a low-variance starting point that helps avoid overfitting in early training phases. The combination of Gaussian noise augmentation and Random Normal initialization introduced variability both at the data and model initialization level, offering a useful control scenario for measuring the effects of structured randomness.

This strategic pairing of augmentations and initializations promoted heterogeneity in feature representations and error patterns across the ensemble, which is essential for achieving high segmentation accuracy through ensemble learning. The result is a more resilient and generalizable model, capable of handling the diverse challenges inherent in urban scene segmentation.

To establish a baseline for evaluating the effectiveness of the proposed homogeneous U-Net ensemble, a standard U-Net model was implemented. This model, widely recognized in semantic segmentation tasks, is particularly well-suited for applications involving urban scenes, such as those found in the Cityscapes dataset. The U-Net architecture (Fig. 1) adopts a symmetric encoder-decoder structure, which enables accurate pixel-level classification – an essential capability for high-resolution urban segmentation.

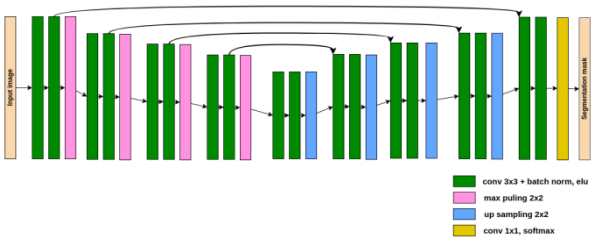


Figure 1 – Unet baseline acrhitecture

The network's architecture is composed of two main components. The encoder, also known as the contracting path, systematically reduces the spatial dimensions of the input image while extracting progressively higher-level features. Each block in the encoder includes two convolutional layers followed by Batch Normalization and ReLU activations, with Max Pooling layers applied between blocks to downsample the feature maps. At each stage of downsampling, the number of feature channels doubles, beginning from 64 and reaching up to 1024, allowing the model to learn increasingly abstract representations of the input scene.

In the decoder, or expanding path, the spatial resolution of the feature maps is gradually restored using transposed convolutions. To retain fine-grained spatial information lost during the encoding process, skip connections linking corresponding layers between the encoder and decoder. After each upsampling step, the upsampled feature map is concatenated with its encoder counterpart, followed by two convolutional layers and Batch Normalization, which further refine the segmentation outputs. The network concludes with a 1×1 convolutional layer that projects the final feature map to the desired number of segmentation classes, and a softmax activation function is applied to produce the class probabilities for each pixel.

The baseline model is configured to accept input images of size 256×256×3, with a total parameter count of approximately 35.8 million. Training was conducted using categorical cross-entropy as the loss function, optimized with the Adam algorithm and an initial learning rate of 0.001. The training set was processed in mini-batches of 16 images, and the network was trained for up to 30 epochs, with early stopping triggered based on the validation loss to prevent overfitting.

This baseline U-Net model serves as a reference point against which the ensemble approach is assessed. By comparing its performance with that of the ensemble – composed of multiple U-Net variants trained with different augmentation strategies and initialization schemes – it becomes possible to quantify the benefits of ensemble learning in enhancing segmentation accuracy and robustness.

To improve segmentation accuracy and enhance the generalization capabilities beyond what a single U-Net model can offer, an ensemble of five U-Net networks was constructed. While all five models shared the same architectural design as the baseline U-Net, they differed in their training setup through distinct combinations of weight initialization and data augmentation strategies. This intentional diversification enabled the ensemble to learn a wider array of feature representations, ultimately leading to stronger performance on complex urban segmentation tasks within the dataset.

The diversity within the ensemble was introduced through two complementary mechanisms. The first involved using different weight initialization schemes for each model, which encouraged unique learning dynamics by altering the starting conditions of training. The initializers applied – Glorot Normal, He Uniform, Orthogonal, LeCun Normal, and Random Normal – each influenced the convergence path in different ways, thereby promoting model independence and reducing the risk of all networks settling into similar local minima. The second mechanism of diversification relied on data augmentation. Each U-Net model was trained using a specific transformation technique – ranging from scaling and rotation to affine transformation, elastic deformation,

DOI 10.15588/1607-3274-2025-3-7

and Gaussian noise. These augmentations simulated a variety of real-world conditions found in urban environments, compelling each network to adapt to distinct types of variability, which in turn increased the ensemble's robustness to unseen data.

To combine the outputs from the ensemble, multiple prediction aggregation strategies were explored, each offering a different method of consolidating the networks' decisions. The first approach involved averaging the softmax probability outputs of each model on a pixel-wise basis. This method smoothed out individual inconsistencies and allowed the final segmentation map to reflect a balanced consensus across all predictions. An alternative strategy employed a maxing operation, selecting the highest softmax probability across the ensemble for each pixel. This method emphasized high-confidence predictions by giving more weight to confident outputs from any individual model.

Beyond these basic ensemble strategies, a more refined weighted voting method was developed to optimize how each model contributed to the final output. Here, the influence of each network was proportional to its validation accuracy, ensuring that more reliable models had a stronger impact on the final segmentation results. To further refine this weighting scheme, an exponential scaling mechanism was introduced, amplifying the contributions of the top-performing models while still allowing all ensemble members to participate in the decision-making process. This balance maintained the diversity benefits of ensembling while increasing the precision of the final predictions.

To optimize the distribution of weights among the models, a grid search procedure was performed. Rather than assigning equal weights, the goal was to identify the optimal weight vector w=[w1,w2,...,wN] that would yield the highest segmentation accuracy, as measured by the Dice score across the entire validation set. This optimization process ensured that the ensemble not only leveraged the strengths of individual models but also fine-tuned their contributions to achieve maximal overall performance.The ensemble prediction is computed as a weighted sum of the individual model predictions:

$$\hat{Y} = \sum_{i=1}^{N} w_i \cdot P_i \, .$$

To ensure model contributions remain meaningful and balanced, we enforce the following constraints on the weights:

$$0 \le w_i \le 1, \; \sum_{i=1}^{N} w_i \approx 1 \, .$$

This formulation prevents any single model from dominating the ensemble while allowing flexibility for weight adjustments.

The optimization process seeks to maximize the dice score:

$$Dice(Y,\hat{Y}) = \frac{2\sum_{i=1}^{N}(Y_i \cdot \hat{Y}_i) + \varepsilon}{\sum_{i=1}^{N} Y + \sum_{i=1}^{N} \hat{Y} + \varepsilon} \, .$$

We define the objective function as:

$$\max_{w} \frac{2\sum_{i=1}^{N}(Y_i \cdot \hat{Y}_i) + \varepsilon}{\sum_{i=1}^{N} Y_i + \sum_{i=1}^{N} \hat{Y}_i + \varepsilon} \, .$$

We use constrained numerical optimization or Powell's method [28] to solve for the optimal weight vector.

4 EXPERIMENTS

The experiment was conducted using the Cityscapes dataset, which was divided into 2,975 training images and 500 for validation. To ensure consistency, all input images were resized and normalized prior to training, enhancing numerical stability and model convergence. The dataset was also shuffled randomly to avoid any learning bias, and mini-batches of size 16 were used to optimize computational efficiency.

For the baseline, a standard U-Net model was deployed without explicit weight initialization – weights were set to zero by default. The Exponential Linear Unit (ELU) activation function was used throughout the network to support better gradient flow and accelerate convergence in deeper layers. An early stopping strategy was applied, halting training automatically once the validation loss ceased to improve, thus preventing overfitting and reducing computational overhead.

Performance was evaluated using a suite of metrics, including Mean Intersection over Union (Mean IoU), pixel accuracy, precision, sensitivity, and specificity. These metrics offered a comprehensive view of model performance, capturing both pixel-level accuracy and class-level segmentation effectiveness. This baseline served as a critical reference point for assessing the effectiveness of the proposed homogeneous ensemble approach.

The first ensemble model maintained the baseline architecture but introduced Glorot Normal initialization to ensure balanced activation variance across layers. The activation function was switched to sigmoid to produce smoother probability maps suitable for segmentation tasks. Additionally, scaling augmentation was applied, randomly zooming input images to simulate changes in object size and distance. These modifications aimed to improve stability, generalization, and robustness to scale

DOI 10.15588/1607-3274-2025-3-7

variance while preserving compatibility with the overall ensemble structure.

The second U-Net also retained the base architecture but employed He Normal initialization, tailored for ReLU activations, to facilitate deeper gradient flow. Rotation-based augmentation was introduced, with input images randomly rotated up to 30 degrees to simulate real-world changes in camera angle. Nearest-neighbor interpolation was used to maintain pixel quality. This configuration allowed the model to develop rotation-invariant features, enhancing its performance in dynamic urban environments.

The third model applied Orthogonal initialization to promote stable training by preserving variance throughout deep layers, in conjunction with ELU activation to support gradient propagation. Affine transformations – including translation, scaling, shearing, and minor rotations – were used as augmentations to introduce spatial diversity. This combination encouraged the model to learn features invariant to subtle spatial distortions typical in real-world imagery.

The fourth model employed LeCun Normal initialization, optimized for tanh activations, and was paired with Elastic Deformation as the augmentation technique. By introducing smooth, localized warping through parameterized displacement fields ($\alpha = 10$, $\sigma = 4$), the model became better equipped to generalize across irregular object shapes and occlusions. This setup enabled the model to develop fine-grained sensitivity to structural deformations commonly seen in urban environments.

The final U-Net used Random Normal initialization to introduce variability in early learning trajectories. Gaussian noise was added to the input during training to simulate sensor-level imperfections, using a standard deviation of $\sigma = 0.05$. The ELU activation function was retained to aid in stable convergence. This model served to improve robustness under noisy conditions, rounding out the ensemble with additional stochastic diversity.

Upon training the five U-Net models, they were integrated into a homogeneous ensemble to capitalize on their individual strengths and improve segmentation accuracy, robustness, and generalization. To achieve this, three distinct ensemble strategies were explored: averaging, maxing, and weighted voting – each offering a different method for aggregating pixel-wise predictions.

In the averaging ensemble, the probability distributions generated by each model were averaged for every pixel. This approach mitigated the noise and uncertainty present in individual model outputs, yielding smoother and more balanced segmentation maps. It was particularly effective at improving generalization by consolidating diverse prediction patterns across the ensemble.

The maxing ensemble took a different approach, selecting the highest softmax probability across all five models for each pixel. This strategy emphasized confident predictions, allowing the most certain model to determine the final class decision per pixel. While this method enhanced decisiveness, it also introduced the risk of amplifying isolated high-confidence errors, depending on the reliability of individual networks.

To further refine prediction quality, a weighted voting ensemble was implemented. Here, each model's prediction was weighted according to its validation performance. The first weighting scheme assigned weights proportional to each model's validation accuracy, allowing higher-performing models to contribute more significantly to the final segmentation output.

The second approach used exponential scaling, amplifying differences between strong and weak models by applying an exponential function to the accuracy scores. This method increased the influence of top performers while still preserving the diversity contributed by other networks. Finally, a grid search optimization was conducted to identify the optimal weight vector $w = [w_1, w_2, ..., w_N]$. This involved evaluating different weight configurations on a subset of 10 validation images. The aim was to maximize the Dice score across the ensemble, ensuring that the final weighted output delivered the highest possible segmentation accuracy.

5 RESULTS

Once the ensemble models were constructed and integrated using the proposed aggregation strategies, a comprehensive evaluation was carried out to compare their performance against the baseline U-Net. This analysis focused on measuring segmentation accuracy, generalization, and robustness across both individual and ensemble models. All models were tested on the same validation set using consistent evaluation metrics, which included Mean Intersection over Union (Mean IoU), pixel accuracy, precision, sensitivity, specificity, and execution time measured in seconds per image. This consistent methodology ensured a fair comparison and provided a granular understanding of how each configuration performed. The training accuracy graph (Fig. 2) illustrates the learning progression of multiple U-Net models compared to the baseline over 25 epochs.

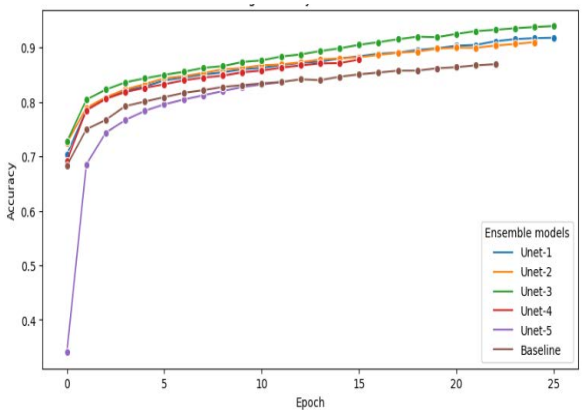


Figure 2 – Training accuracy across networks

Training accuracy trends revealed that all models experienced a rapid increase in performance within the

DOI 10.15588/1607-3274-2025-3-7

first five epochs, reflecting effective initial feature learning from the dataset. The baseline U-Net, though following a similar pattern, consistently trailed behind the other models. Among the ensemble components, U-Net-2 and U-Net-3 achieved the highest accuracy throughout training, indicating their ability to extract and generalize critical features. U-Net-5, on the other hand, consistently recorded the lowest accuracy, suggesting challenges in learning effective feature representations. By epoch 15, most models began to converge, with accuracy improvements tapering off and stabilizing near the 90% mark – except for U-Net-5, which continued to underperform. The baseline model remained consistently below the performance of all U-Net variants, reaffirming the benefits introduced by tailored augmentation and initialization strategies in the ensemble.

The graph Fig. 3 illustrates the validation accuracy of different U-Net models and the baseline over 25 epochs.

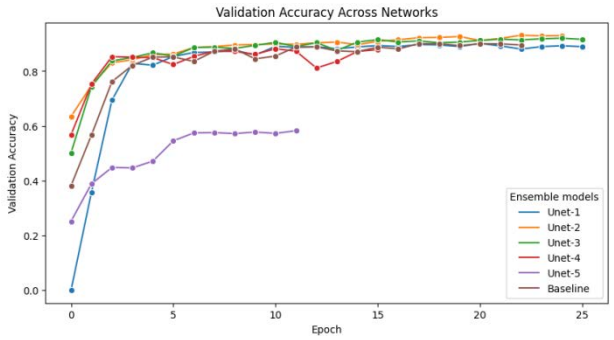


Figure 3 – Validation accuracy across networks

Validation accuracy followed a similar trajectory, offering further insight into the generalization capabilities of each model on unseen data. All models showed sharp improvements in validation accuracy during the initial training phase, mirroring their training performance. U-Net-2, U-Net-3, and U-Net-4 achieved the highest early-stage validation scores, indicating robust learning dynamics and generalization from augmented and well-initialized architectures. In contrast, U-Net-5 lagged significantly behind, maintaining a noticeably lower accuracy curve throughout training. The baseline model started with low initial accuracy but gradually improved, though by the fifth epoch, all U-Net variants had surpassed it. This confirmed the value of ensemble diversification strategies in enhancing model generalization.

By the end of training, the validation accuracy of most U-Net models converged between 82% and 84%, while the baseline plateaued slightly below this range. U-Net-5 remained a notable outlier, stabilizing around 57%, which suggests either insufficient regularization or overfitting to the training data. After epoch 10, most models displayed stable accuracy with minimal variation, indicating convergence. U-Net-2 and U-Net-3 maintained superior performance throughout, reflecting their consistency across both training and validation phases. The observed gap between training accuracy (approaching 90%) and validation accuracy (around 80%) across models points to potential overfitting – a likely result of dataset limitations and model complexity.

To assess the performance of the models, several evaluation metrics were employed, each designed to capture different aspects of segmentation accuracy and classification quality. One of the core metrics used was sparse categorical accuracy, which is particularly suitable when ground truth labels are provided as integer-encoded class indices rather than one-hot encoded vectors. This metric computes the proportion of correctly classified pixels by comparing the predicted class index – determined by the highest predicted probability – with the actual class label for each pixel:

$$accuracy = \frac{1}{\mathrm{N}}\sum_{i=1}^{N} 1(\arg\max_{\mathrm{c}} p_{i,c} = y_i) .$$

Another key metric is the Mean Intersection over Union (Mean IoU), a standard in semantic segmentation tasks. Mean IoU quantifies the average overlap between predicted and ground truth segmentation masks across all considered classes. However, given the class imbalance inherent to the Cityscapes dataset – where some classes dominate the dataset while others are infrequently represented – Mean IoU was computed over a targeted subset of six representative classes: 7 (road), 11 (building), 20 (traffic sign), 21 (vegetation), 23 (sky), and 26 (car). These selected categories encompass both large structural elements and smaller, yet semantically important, urban objects. This focused evaluation offers a more meaningful representation of model performance in real-world scenarios, rather than being skewed by rare or less relevant classes. IoU is calculated as following:

$$IoU = \frac{TP}{TP + FP + FN} .$$

Precision was also utilized to evaluate how reliable the model's positive predictions were. It measures the proportion of pixels that were correctly predicted as belonging to a particular class out of all pixels the model assigned to that class:

$$precision = \frac{TP}{TP + FP} .$$

In contrast, Sensitivity, also referred to as Recall, measures the model's ability to detect all relevant pixels that belong to a given class. This is calculated as the ratio of True Positives to the sum of True Positives and False Negatives:

$$sensitivity = \frac{TP}{TP + FN} .$$

DOI 10.15588/1607-3274-2025-3-7

Finally, Specificity was included to assess how well the model avoids false alarms. It evaluates the proportion of correctly identified negative pixels – those that do not belong to a particular class – relative to all true negatives and false positives. In this case, True Negatives refer to pixels correctly classified as not part of the target class, and False Positives indicate pixels that were incorrectly predicted as belonging to it:

$$specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}.$$

Result values of metrics are displayed in Table 1.

The results outlined in the table highlight the clear advantage of ensemble strategies over both the baseline and individual U-Net models. The most effective configuration – the ensemble with optimized weights – achieved the highest accuracy, reaching 0.8756. This marks an approximate 4.7% improvement over the baseline model, which recorded an accuracy of 0.8360. These findings are consistent with trends observed in the training and validation accuracy curves, where ensemble methods consistently surpassed the performance of individual networks, particularly in the later stages of training.

In terms of segmentation quality, the mean Intersection over Union (Mean IoU) also shows a notable boost. The optimized weight ensemble attained a Mean IoU of 0.6504, outperforming the baseline's 0.6145 by a margin of 3.6%. Beyond overall accuracy and IoU, additional evaluation metrics such as precision, sensitivity, and specificity provide deeper insight into the segmentation behavior of each model. Precision, which quantifies the correctness of positive pixel classifications, varied across configurations. The highest precision was observed in U-Net-2 at 0.4050, while the optimized ensemble closely followed with a precision of 0.3980, demonstrating its ability to maintain segmentation accuracy while effectively limiting false positives.

Specificity, which measures how accurately negative pixels are classified, remained consistently high across all configurations. The optimized ensemble achieved the highest specificity at 0.9953, indicating its strong capacity to reduce false positive classifications without compromising performance. This reliability in identifying background or non-target areas is especially valuable in high-precision segmentation tasks.

While ensemble approaches deliver substantial gains in accuracy and segmentation quality, these improvements come with increased computational demands. The baseline U-Net offered the fastest inference speed, processing an image in 0.1604 seconds. In contrast, the optimized ensemble required 0.4135 seconds per image – roughly 2.6 times longer.

Among the individual models, U-Net-4 exhibited the lowest execution time at 0.1512 seconds, making it a compelling option for applications that prioritize speed over marginal gains in accuracy. Nevertheless, the superior accuracy and segmentation fidelity achieved by ensemble configurations justify their use in domains where precision is paramount and computational cost is secondary.

Overall, the findings clearly demonstrate that ensemble methods offer meaningful improvements in both accuracy and IoU compared to standalone models. The ensemble with optimized weights emerges as the most effective approach, achieving the best overall balance: high accuracy (0.8756), strong IoU (0.6504), and leading specificity (0.9953).

Table 1 – Networks metrics

	Baseline	Unet-1	Unet-2	Unet-3	Unet-4	Unet-5	Ensemble (max)	Ensemble (avg)	Ensemble optimized weights	Ensemble proportional weights	Ensemble exponential
accuracy	0.8360	0.8265	0.8462	0.8365	0.8223	0.5754	0.8458	0.8672	0.8756	0.8620	0.8622
mean IoU	0.6145	0.6103	0.6284	0.6324	0.5898	0.3463	0.5931	0.6346	0.6504	0.6380	0.6410
precision	0.3783	0.3368	0.4050	0.3820	0.3297	0.1650	0.2922	0.2991	0.3980	0.3005	0.3003
sensitivity	0.3094	0.2893	0.3269	0.3047	0.3056	0.1623	0.2409	0.2480	0.2576	0.2485	0.2485
specificity	0.9944	0.9940	0.9947	0.9944	0.9940	0.9852	0.9946	0.9952	0.9953	0.9952	0.9952
time per image, s	0.1604	0.1524	0.1618	0.1570	0.1512	0.1627	0.3611	0.3710	0.4135	0.3947	0.4584

6 DISCUSSION

In this study, we explored the effect of ensemble methods on convolutional neural networks applied to semantic segmentation tasks. The proposed method integrates multiple U-Net networks and aggregates their outputs using an optimized weighting technique, aiming to enhance segmentation accuracy while keeping computational demands within practical limits.

Our research began with a literature review, examining established techniques for improving semantic segmentation, particularly those focused on single-model refinement and ensemble learning. While individual model optimizations can yield modest improvements, the

DOI 10.15588/1607-3274-2025-3-7

reviewed studies consistently highlight ensemble learning as a more effective approach for increasing model robustness and generalization. However, these benefits are often accompanied by a notable rise in computational cost.

To assess the proposed approach, we implemented and evaluated several U-Net models, each combined through different ensembling strategies – namely max voting, simple averaging, optimized weighting, proportional weighting, and exponential weighting. Across all configurations, the ensemble models outperformed standalone networks in terms of both accuracy and mean Intersection over Union (IoU). The ensemble using optimized weights delivered the best results, achieving an accuracy of 87.56% and a mean IoU of 0.6504, outperforming the top-performing individual U-Net by roughly 3%. These gains, however, came at the cost of increased inference time, a factor that becomes particularly relevant in time-sensitive or real-time applications, even though it stays within acceptable limits.

Our findings further underscore that ensemble performance is most effective when constituent models produce diverse yet complementary predictions. Variability among the individual U-Net models was evident, with some excelling in precision and others in sensitivity. Through ensembling, these strengths were combined, effectively balancing the trade-offs inherent in each individual model and producing a more stable and consistent segmentation output.

Despite these advantages, the study also sheds light on the limitations of ensemble learning. Running multiple networks in sequence substantially increases computational requirements, especially on systems without hardware acceleration. This poses challenges for deployment in scenarios where real-time inference is critical. Moreover, ensemble models did not show significant gains in specificity, suggesting that some segmentation errors are systemic and may persist regardless of the aggregation strategy.

Overall, the results demonstrate that ensemble techniques offer meaningful improvements in semantic segmentation performance and model generalization across diverse classes. Yet, the balance between performance gains and computational efficiency remains a key consideration. Future research should focus on optimizing ensemble methodologies to reduce overhead, potentially through model distillation, parallel inference strategies, or lightweight ensembling techniques, all while preserving segmentation quality.

CONCLUSIONS

The paper analyses the effectiveness of ensemble methods for convolutional neural networks in solving the semantic segmentation task.

The scientific novelty of the presented work lies in the development of a weighted ensemble approach based on five U-Net models sharing the same architecture, but each trained using distinct augmentation strategies and weight initialization techniques. This design improves segmentation accuracy and consistency without altering the network structure itself. By applying an optimized weighting mechanism during ensemble prediction, the proposed method achieves notable improvements in both accuracy and mean IoU when compared to individual models, while maintaining a high level of specificity. These results demonstrate that ensembling is a viable and efficient strategy for enhancing semantic segmentation performance using existing architectures.

The practical significance of the research is underscored by the fact that the ensemble models were trained and evaluated on a real-world dataset, validating their relevance for practical deployment. The findings support the recommendation of this ensemble strategy for applications that demand high segmentation accuracy, such as autonomous driving systems. However, the increased computational overhead introduced by ensemble methods should be carefully considered, particularly in scenarios requiring real-time processing.

Prospects for further research include refining the computational efficiency of the ensemble to reduce inference time while preserving segmentation quality. Future investigations may also explore the effectiveness of the proposed ensembling strategy when applied to alternative network architectures and larger, more diverse datasets, thereby broadening its applicability across different domains and use cases.

ACKNOWLEDGEMENTS

We thank the management of Kharkiv National University of Radioelectronics for the opportunity to conduct scientific research.

REFERENCES

1. Cordts M., Omran M., Ramos S. et al. The Cityscapes Dataset for Semantic Urban Scene Understanding, *In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), USA, 27–30 June 2016 : proceedings*. Las Vegas, IEEE, 2016, pp. 3213–3223. DOI: 10.1109/cvpr.2016.350.
2. A Comprehensive Survey on Ensemble Methods / Suyash Kumar, Prabhjot Kaur, Anjana Gosain, *2022 IEEE 7th International Conference for Convergence in Technology (I2CT)*. Mumbai, India, 7–9 April 2022. [S. l.], 2022. DOI: 10.1109/i2ct54291.2022.9825269.
3. Hao S., Zhou Y., Guo Y. A brief survey on semantic segmentation with deep learning, *Neurocomputing*, 2020, Vol. 406, pp. 302–321. DOI: 10.1016/j.neucom.2019.11.118.
4. Pare S. [et al.]Image Segmentation Using Multilevel Thresholding: A Research Review, *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 2019, Vol. 44, No. 1, pp. 1–29. DOI: 10.1007/s40998-019-00251-1.
5. Tang Jun A color image segmentation algorithm based on region growing, *2010 2nd International Conference on Computer Engineering and Technology*. Chengdu, China, 16–18 April 2010. [S. l.], 2010. DOI: 10.1109/iccet.2010.5486012.

DOI 10.15588/1607-3274-2025-3-7

6. Jeyalaksshmi S., Prasanna S. A Review of Edge Detection Techniques for Image Segmentation, *International Journal of Data Mining Techniques and Applications*, 2016, Vol. 5, No. 2, pp. 140–142. DOI: 10.20894/ijdmta.102.005.002.008.
7. Liu Xiangbin et al. A Review of Deep-Learning-Based Medical Image Segmentation Methods, *Sustainability*, 2021, Vol. 13, No. 3, P. 1224. DOI: 10.3390/su13031224.
8. Guo Zhe et al. Deep Learning-Based Image Segmentation on Multimodal Medical Imaging, *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2019, Vol. 3, No. 2, pp. 162–169. DOI: 10.1109/trpms.2018.2890359.
9. Mzoughi O., Mzoughi Olfa, Yahiaoui Itheri Deep learning-based segmentation for disease identification, *Ecological Informatics*, 2023, pp. 102000. DOI: 10.1016/j.ecoinf.2023.102000.
10. Sultana Farhana, Sufian Abu, Dutta Paramartha Evolution of Image Segmentation using Deep Convolutional Neural Network: A Survey, *Knowledge-Based Systems*, 2020, Vol. 201–202, P. 106062. DOI: 10.1016/j.knosys.2020.106062.
11. Shelhamer E. et al. Fully Convolutional Networks for Semantic Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, Vol. 39, No. 4, pp. 640–651. DOI: 10.1109/tpami.2016.2572683.
12. Li Xiaojin et al. Image Segmentation Based on Improved Unet, *Journal of Physics: Conference Series*, 2021, Vol. 1815, No. 1, P. 012018. DOI: 10.1088/1742-6596/1815/1/012018.
13. Mobarakol Islam et al. Brain Tumor Segmentation and Survival Prediction Using 3D Attention UNet. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham, 2020, pp. 262–272. DOI: 10.1007/978-3-030-46640-4_25.
14. Karaali A. et al. DR-VNet: Retinal Vessel Segmentation via Dense Residual UNet, *Pattern Recognition and Artificial Intelligence*. Cham, 2022, pp. 198–210. DOI: 10.1007/978-3-031-09037-0_17.
15. Ngo G. et al. Evolutionary bagging for ensemble learning, *Neurocomputing*, 2022. DOI: 10.1016/j.neucom.2022.08.055.
16. Drucker Harris et al. Boosting and Other Ensemble Methods, *Neural Computation*, 1994, Vol. 6, No. 6, pp. 1289–1301. DOI: 10.1162/neco.1994.6.6.1289.
17. Verma Anurag Kumar, Pal Saurabh Prediction of Skin Disease with Three Different Feature Selection Techniques Using Stacking Ensemble Method, *Applied Biochemistry and Biotechnology*, 2019, Vol. 191, No. 2, pp. 637–656. DOI: 10.1007/s12010-019-03222-8.
18. Bodyanskiy Y. V., Lipianina-Honcharenko K. V., Sachenko A. O. Ensemble of Adaptive Predictors for Multivariate Nonstationary Sequences and its Online Learning, *Radio Electronics, Computer Science, Control*, 2024, No. 4, P. 91. DOI: 10.15588/1607-3274-2023-4-9.
19. Ahmad Numan, Behram Wali, Khattak Asad J. Heterogeneous ensemble learning for enhanced crash forecasts – A frequentist and machine learning based stacking framework, *Journal of Safety Research*, 2022, DOI: 10.1016/j.jsr.2022.12.005.
20. Lo Hung-Yi, Wang Ju-Chiang, Wang Hsin-Min Homogeneous segmentation and classifier ensemble for audio tag annotation and retrieval, *2010 IEEE International Conference on Multimedia and Expo (ICME), Singapore*. Singapore, 19–23 July 2010. [S. l.], 2010. DOI: 10.1109/icme.2010.5583009.
21. Bian Shun, Wang Wenjia Investigation on Diversity in Homogeneous and Heterogeneous Ensembles, *The 2006 IEEE International Joint Conference on Neural Network Proceedings, Vancouver, BC, Canada, 16–21 July 2006.* [S. l.], 2006. DOI: 10.1109/ijcnn.2006.247268.
22. Kamnitsas Konstantinos et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, *Medical Image Analysis*, 2017, Vol. 36, pp. 61–78. DOI: 10.1016/j.media.2016.10.004.
23. Gautam Nandita et al. An Ensemble of UNet Frameworks for Lung Nodule Segmentation, *Current Problems in Applied Mathematics and Computer Science and Systems.* Cham, 2023, pp. 450–461. DOI: 10.1007/978-3-031-34127-4_44.
24. Smelyakov K. et al. Adaptive Image Enhancement Model for the Robot Vision System, *ENVIRONMENT. TECHNOLOGIES. RESOURCES. Proceedings of the International Scientific and Practical Conference*, 2023, Vol. 3, pp. 246–251.
25. Abdullahi A. M. et al. A comparison of weight initializers in deep learning, *2023 IEEE 21st Student Conference on Research and Development (SCOReD); 2023 Dec 13–14*. Kuala Lumpur, Malaysia, [S.l.], 2023. DOI: 10.1109/scored60679.2023.10563215.
26. Lee H. et al. Improved weight initialization for deep and narrow feedforward neural network, *Neural Networks*, 2024, P. 106362. DOI: 10.1016/j.neunet.2024.106362.
27. LeCun Y. et al. Efficient BackProp, *In: Lecture Notes in Computer Science*. Berlin, Heidelberg, 1998, pp. 9–50. DOI: 10.1007/3-540-49430-8_2.
28. Kramer O. Iterated local search with Powell's method: a memetic algorithm for continuous global optimization, *Memetic Computing*, 2010, Vol. 2, No. 1, pp. 69–83. DOI: 10.1007/s12293-010-0032-9.

Received 27.03.2025.
Accepted 21.05.2025.

DOI 10.15588/1607-3274-2025-3-7

УДК 004.93

СЕГМЕНТАЦІЯ МІСЬКИХ СЦЕН ЗА ДОПОМОГОЮ ОДНОРІДНОГО АНСАМБЛЮ U-NET: ДОСЛІДЖЕННЯ НА ДАТАСЕТІ CITYSCAPES

Гмиря І. О. – аспірант кафедри програмної інженерії, Харківський національний університет радіоелектроніки, Харків, Україна.

Кравець Н. С. – канд. техн. наук, доцент, доцент кафедри програмної інженерії, Харківський національний університет радіоелектроніки, Харків, Україна.

АНОТАЦІЯ

Актуальність. Семантична сегментація є ключовим завданням комп'ютерного зору, зокрема в таких сферах, як автономне водіння та аналіз міських сцен. Створення нових архітектур є складним і трудомістким процесом, однак поліпшення точності за допомогою ансамблевих методів на основі вже існуючих моделей показує високий потенціал. У даній роботі досліджується застосування ансамблевого навчання як стратегії підвищення точності сегментації без модифікації архітектури U-Net.

Мета роботи – розробка та оцінка однорідного ансамблю моделей U-Net, навчання яких здійснюється із використанням різних методів ініціалізації ваг та збільшення обсягу даних, а також вивчення ефективності різних стратегій агрегації ансамблю для підвищення якості сегментації на складних урбаністичних даних.

Метод. Запропоновано ансамбль з п'яти моделей U-Net з однаковою архітектурою, але різною ініціалізацією ваг та підходами до збільшення обсягу даних, що забезпечує різноманітність прогнозів. Розглянуто кілька стратегій об'єднання вихідних даних: середнє по softmax, максимум, пропорційне зважування, експоненціальне зважування та оптимізоване вагове голосування. Оцінювання виконано на датасеті Cityscapes із використанням стандартних метрик сегментації.

Результати. Результати експериментів показують, що ансамблеві моделі стабільно перевищують точність окремих моделей U-Net та базової моделі за такими показниками, як точність, середній IoU та специфічність. Ансамбль із оптимізованим зважуванням досяг найвищої точності (87,56%) та середнього IoU (0,6504), перевищивши найкращу окрему модель приблизно на 3%. Водночас покращення якості супроводжується збільшенням часу виведення результату, що вказує на необхідність компромісу між точністю та обчислювальною ефективністю.

Висновки. Запропонований підхід на основі ансамблю ефективно покращує результати сегментації без зміни архітектури моделі. Незважаючи на збільшення обчислювальних витрат, метод є придатним для задач, де критично важлива точність сегментації. Подальші дослідження будуть зосереджені на зменшенні часу виведення результату та поширенні ансамблевого підходу на інші архітектури та датасети.

КЛЮЧОВІ СЛОВА: згорткова нейронна мережа, семантична сегментація, U-Net, ансамблеве навчання, методи збільшення обсягу даних, ініціалізація ваг, Cityscapes, урбаністичні сцени.

ЛІТЕРАТУРА

1. The Cityscapes Dataset for Semantic Urban Scene Understanding / [M. Cordts, M. Omran, S. Ramos, et al.] – In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), USA, 27–30 June 2016 : proceedings. – Las Vegas : IEEE, 2016. – P. 3213–3223. DOI: 10.1109/cvpr.2016.350.
2. A Comprehensive Survey on Ensemble Methods / Suyash Kumar, Prabhjot Kaur, Anjana Gosain // 2022 IEEE 7th International Conference for Convergence in Technology (I2CT), Mumbai, India, 7–9 April 2022. – [S. l.], 2022. DOI: 10.1109/i2ct54291.2022.9825269.
3. Hao S. A brief survey on semantic segmentation with deep learning / S. Hao, Y. Zhou, Y. Guo // Neurocomputing. – 2020. – Vol. 406. – P. 302–321. DOI: 10.1016/j.neucom.2019.11.118.
4. Image Segmentation Using Multilevel Thresholding: A Research Review / S. Pare [et al.] // Iranian Journal of Science and Technology, Transactions of Electrical Engineering. – 2019. – Vol. 44, No. 1. – P. 1–29. DOI: 10.1007/s40998-019-00251-1.
5. A color image segmentation algorithm based on region growing / Jun Tang // 2010 2nd International Conference on Computer Engineering and Technology, Chengdu, China, 16–18 April 2010. – [S. l.], 2010. DOI: 10.1109/iccet.2010.5486012.
6. Jeyalaksshmi S. A Review of Edge Detection Techniques for Image Segmentation / S. Jeyalaksshmi, S. Prasanna // International Journal of Data Mining Techniques and Applications. – 2016. – Vol. 5, No. 2. – P. 140–142. DOI: 10.20894/ijdmta.102.005.002.008.
7. A Review of Deep-Learning-Based Medical Image Segmentation Methods / Xiangbin Liu [et al.] // Sustainability. – 2021. – Vol. 13, No. 3. – P. 1224. DOI: 10.3390/su13031224.
8. Deep Learning-Based Image Segmentation on Multimodal Medical Imaging / Zhe Guo [et al.] // IEEE Transactions on Radiation and Plasma Medical Sciences. – 2019. – Vol. 3, No. 2. – P. 162–169. DOI: 10.1109/trpms.2018.2890359.
9. Mzoughi O. Deep learning-based segmentation for disease identification / Mzoughi O., Mzoughi Olfa, Yahiaoui Itheri // Ecological Informatics. – 2023. – P. 102000. DOI: 10.1016/j.ecoinf.2023.102000.
10. Sultana Farhana Evolution of Image Segmentation using Deep Convolutional Neural Network: A Survey / Farhana Sultana, Abu Sufian, Paramartha Dutta // Knowledge-Based Systems. – 2020. – Vol. 201–202. – P. 106062. DOI: 10.1016/j.knosys.2020.106062.
11. Fully Convolutional Networks for Semantic Segmentation / E. Shelhamer et al. // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2017. –

DOI 10.15588/1607-3274-2025-3-7

Vol. 39, No. 4. – P. 640–651. DOI: 10.1109/tpami.2016.2572683.
12. Image Segmentation Based on Improved Unet / Xiaojin Li et al. // Journal of Physics: Conference Series. – 2021. – Vol. 1815, No. 1. – P. 012018. DOI: 10.1088/1742-6596/1815/1/012018.
13. Brain Tumor Segmentation and Survival Prediction Using 3D Attention UNet / Mobarakol Islam et al. // Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. – Cham, 2020. – P. 262–272. DOI: 10.1007/978-3-030-46640-4_25.
14. DR-VNet: Retinal Vessel Segmentation via Dense Residual UNet / A. Karaali et al. // Pattern Recognition and Artificial Intelligence. – Cham, 2022. – P. 198–210. DOI: 10.1007/978-3-031-09037-0_17.
15. Evolutionary bagging for ensemble learning / G. Ngo et al. // Neurocomputing. – 2022. DOI: 10.1016/j.neucom.2022.08.055.
16. Boosting and Other Ensemble Methods / Harris Drucker et al. // Neural Computation. – 1994. – Vol. 6, No. 6. – P. 1289–1301. DOI: 10.1162/neco.1994.6.6.1289.
17. Verma Anurag Kumar Prediction of Skin Disease with Three Different Feature Selection Techniques Using Stacking Ensemble Method / Anurag Kumar Verma, Saurabh Pal // Applied Biochemistry and Biotechnology. – 2019. – Vol. 191, No. 2. – P. 637–656. DOI: 10.1007/s12010-019-03222-8.
18. Bodyanskiy Y. V. Ensemble of Adaptive Predictors for Multivariate Nonstationary Sequences and its Online Learning / Y. V. Bodyanskiy, K. V. Lipianina-Honcharenko, A. O. Sachenko // Radio Electronics, Computer Science, Control. – 2024. – No. 4. – P. 91. DOI: 10.15588/1607-3274-2023-4-9.
19. Ahmad Numan Heterogeneous ensemble learning for enhanced crash forecasts – A frequentist and machine learning based stacking framework / Numan Ahmad, Behram Wali, Asad J. Khattak // Journal of Safety Research. – 2022. DOI: 10.1016/j.jsr.2022.12.005.
20. Lo Hung-Yi Homogeneous segmentation and classifier ensemble for audio tag annotation and retrieval / Hung-Yi Lo, Ju-Chiang Wang, Hsin-Min Wang // 2010 IEEE International Conference on Multimedia and Expo (ICME), Singapore, Singapore, 19–23 July 2010. – [S. l.], 2010. DOI: 10.1109/icme.2010.5583009.
21. Bian Shun Investigation on Diversity in Homogeneous and Heterogeneous Ensembles / Shun Bian, Wenjia Wang // The 2006 IEEE International Joint Conference on Neural Network Proceedings, Vancouver, BC, Canada, 16–21 July 2006. – [S. l.], 2006. DOI: 10.1109/ijcnn.2006.247268.
22. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation / Konstantinos Kamnitsas et al. // Medical Image Analysis. – 2017. – Vol. 36. – P. 61–78. DOI: 10.1016/j.media.2016.10.004.
23. An Ensemble of UNet Frameworks for Lung Nodule Segmentation / Nandita Gautam et al. // Current Problems in Applied Mathematics and Computer Science and Systems. – Cham, 2023. – P. 450–461. DOI: 10.1007/978-3-031-34127-4_44.
24. Adaptive Image Enhancement Model for the Robot Vision System / K. Smelyakov et al. // ENVIRONMENT. TECHNOLOGIES. RESOURCES. Proceedings of the International Scientific and Practical Conference. – 2023. – Vol. 3. – P. 246–251.
25. A comparison of weight initializers in deep learning / A. M. Abdullahi et al. // 2023 IEEE 21st Student Conference on Research and Development (SCOReD); 2023 Dec 13–14; Kuala Lumpur, Malaysia. – [S.l.]: 2023. DOI: 10.1109/scored60679.2023.10563215.
26. Improved weight initialization for deep and narrow feedforward neural network / H. Lee et al. // Neural Networks. – 2024. – P. 106362. DOI: 10.1016/j.neunet.2024.106362.
27. Efficient BackProp / Y. LeCun et al. – In: Lecture Notes in Computer Science. – Berlin, Heidelberg: 1998. – P. 9–50. DOI: 10.1007/3-540-49430-8_2.
28. Kramer O. Iterated local search with Powell's method: a memetic algorithm for continuous global optimization / O. Kramer // Memetic Computing. – 2010. – Vol. 2, No. 1. – P. 69–83. DOI: 10.1007/s12293-010-0032-9.

DOI 10.15588/1607-3274-2025-3-7

UDC 004.8

A NEURAL NETWORK APPROACH TO SEMANTIC SEGMENTATION OF VEHICLES IN VERY HIGH RESOLUTION IMAGES

Kashtan V. Yu. – PhD, Associate Professor, Associate Professor of Department of Information Technology and Computer Engineering, Dnipro University of Technology, Dnipro, Ukraine.

Hnatushenko V. V. – Dr. Sc., Professor, Head of Department of Information Technology and Computer Engineering, Dnipro University of Technology, Dnipro, Ukraine.

Udovyk I. M. – PhD, Associate Professor, Dean of Information Technologies Department, Dnipro University of Technology, Dnipro, Ukraine.

Kazymyrenko O. V. – Postgraduate student of Department of Information Technology and Computer Engineering, Dnipro University of Technology, Dnipro, Ukraine.

Radionov Y. D. – Postgraduate student of Department of Information Technology and Computer Engineering, Dnipro University of Technology, Dnipro, Ukraine.

ABSTRACT

Context. The semantic segmentation of vehicles in very high resolution aerial images is essential in developing intelligent transportation systems. It allows for the automation of real-time traffic management and the detection of congestion and emergencies.

Objective. This work aims to develop and evaluate the effectiveness of a neural network approach to semantic segmentation in very high resolution aerial images, which provides high detail and correct reproduction of object boundaries.

Method. The DeepLab architecture with ResNet-101 as a backbone is used for gradient preservation and multiscale feature analysis. We trained on DOTA data and retrained on specialized sets with classes: vehicles, green areas, buildings, and roads. A loss function based on the Dice coefficient was applied to reduce the imbalance of classes. It effectively solves the class imbalance problem and improves the accuracy of segmenting objects of different sizes. Using ResNet-101 instead of Xception in the backbone network allows us to maintain the gradient as the network depth increases.

Results. Experimental studies have confirmed the effectiveness of the proposed approach, which achieves a segmentation accuracy of more than 90%, outperforming existing analogs. The use of multiscale feature analysis allows for preserving the texture features of objects, reducing false classifications. A comparative study with U-Net, SegNet, FCN8s, and other methods confirms the higher performance of the proposed approach in terms of mIoU (82.3%) and Pixel Accuracy (95.1%).

Conclusions. The experiments confirm the effectiveness of the proposed method of semantic segmentation of vehicles in ultra-high spatial resolution images. Using DeepLab v3+ResNet-101 significantly improves the quality of vehicle segmentation in an urbanized environment. Excellent metric performance makes it promising for infrastructure monitoring and traffic planning tasks. Further research will focus on adapting the model to new datasets.

KEYWORDS: semantic segmentation, vehicles, deep neural networks, ResNet-101, DeepLab, multi-scale analysis, very high resolution images.

ABBREVIATIONS

UAVs is a Unmanned Aerial Vehicle;
RGB is a red, green, blue;
CNN is a convolutional neural network;
FCNs is a fully convolutional network;
RPN is a region proposal network;
DOTA is a dataset for object detection;
PA is a pixel accuracy;
MA is a mean accuracy;
mIoU is a mean intersection over the union;
FP is a False Positive;
FN is a False Negative;
TN is a True Negative;
TP is a True Positive.

NOMENCLATURE

X is an input image;
$f(a)$ is a function of structural features;
P is a predicate that defines the segmentation rule;
$s_{i,j}$ is a name of the region $s_{i,j} \in S$;
$P(s_{i,j})$ is an indication of the neighborhood model;
$\nabla f(a)$ is a gradient;
x_m and x_n are elements of the pixel set X;
$F(w)$ is a neural network model;
X_{norm} is a normalized image;
μ is a mean value of the image X;
σ is a standard deviation of the image X;
$F(x, \{Wi\})$ is a mapping function that represents a sequence of layers with parameters $\{Wi\}$;
y is a residual building block;
ϵ is a small positive number added to avoid division by zero in the case of no intersection between the predicted and real segments;
$D(p,q)$ is a measure of similarity between p and q;
p is a predicted segmentation;
q is a real segmentation;
L_{Dice} is a loss function;
TP is a number of correctly classified positive pixels;
FP is a number of false positive pixels;
FN is a number of false negative pixels;
TN is a number of correctly classified negative pixels;
N is a number of image pixel categories;
TP_i is a number of correctly classified pixels of class i;
FP_i is a number of false positive pixels for class i;
FN_i is a number of false negative pixels for class i.
T_i is a total number of pixels of class i;

DOI 10.15588/1607-3274-2025-3-8

X_{ii} is a total number of pixels with actual type i and prediction type i;

X_{ji} is a total number of pixels with actual type i and prediction type j.

INTRODUCTION

Uncrewed aerial vehicles (UAVs) are an effective tool for high-precision aerial surveys, providing fast and detailed ultra-high-resolution images that can reach an accuracy of several centimeters [1]. It ensures high object detail and provides operational aerial photography with minimal resource costs. One of the parameters affecting the quality of the data is the camera angle. Vertical imaging (perpendicular to the camera's optical axis) provides high accuracy but has a limited coverage area. Low-angle images (15°–30°) expand the coverage area of the scene, improve the depth of perspective, and allow for better analysis of objects in the image. Images acquired at high tilt angles (approximately 60°) provide a much wider coverage area, including horizons, making them suitable for complex analysis of traffic flows and urban environments.

UAVs combine compactness, mobility, and efficiency, which makes it possible to obtain data in real-time and adapt the research methodology depending on the specifics of the territory or object under observation. Equipped with various sensors (RGB cameras, multispectral and hyper-spectral sensors, LiDAR, and thermal imaging systems), UAVs provide multispectral information necessary for thematic image processing and environmental change analysis. Using UAVs for automated vehicle recognition and segmentation is an urgent task in security, logistics, and traffic management [2].

Traditional tracking methods based on ground-based cameras and satellite imagery have limitations associated with limited spatial coverage, high dependence on weather conditions, and delays in data updates. Using UAVs for vehicle recognition and segmentation can overcome these shortcomings by providing adaptability to the information collection process, high spatial resolution, and the ability to update data quickly.

Semantic segmentation is one of the approaches for automated analysis of UAV images. This computer vision method consists of classifying each pixel of an image according to its class [3]. Semantic segmentation allows for high-accuracy vehicle detection in complex urban and road scenes [4].

The research is relevant due to the need to develop new methods of vehicle recognition for intelligent transportation systems, including traffic monitoring, logistics process management, and road safety improvement. The use of UAV imagery in combination with deep learning architecture will increase the accuracy and speed of automated vehicle detection in real-time.

The object of study is the process of semantic segmentation of vehicles in ultra-high-resolution images.

Constructing a neural network model for semantic segmentation is complex and multi-component. It is caused by segmentation accuracy and stability of model training, which mainly depends on the amount and quality of training data, neural network architecture, choice of the loss function, and optimization strategies. In particular, it is necessary to balance computational costs and the model's generalization ability to process ultra-high-resolution images efficiently. It requires adaptation of feature extraction mechanisms and adjustment of the loss function to solve the class imbalance problem.

The subject of study is a neural network methodology for semantic vehicle segmentation based on the DeepLab + ResNet architecture with multi-scale feature extraction, loss function adaptation, and retraining on specialized datasets.

The purpose of the work is to develop and evaluate the effectiveness of a neural network approach to semantic segmentation in very high resolution aerial images, which provides high detail and correct reproduction of object boundaries.

1 PROBLEM STATEMENT

Suppose a set of image pixels $X=\{x_{i,j}\}$ is given, where each pixel $x_{i,j}$ is characterized by structural features defined by the function $f(a)$. The predicate P is also given, establishing the segmentation rule $f(a)$.

The problem of image segmentation is to find a partition of the set P into $S=\{s_{i,j}\}$, where $s_{i,j}$ is connected to non-empty subsets such that for any two pixels x_m, $x_n \in s_{i,j}$ the condition $P(x_m, x_n)$=True is fulfilled, i.e., they belong to the same segment according to the segmentation rule. The boundaries of the regions $s_{i,j}$ are determined by the contrast gradient $\nabla f(a)$ and the spatial dependencies between neighboring pixels. The background region is the set of pixels with the highest or lowest contrast relative to the segmented regions.

In general, segmentation can be considered $f(a) \rightarrow S$.

In particular, $s_{i,j}$ is the name of the region $s_{i,j} \in S$, and $P(s_{i,j})$ is an indication of the neighborhood model that characterizes the object.

2 REVIEW OF THE LITERATURE

The existing approaches to semantic vehicle segmentation can be divided into traditional methods and methods based on deep learning. Conventional methods of vehicle segmentation involve manual feature extraction and machine learning methods, such as SVM, AdaBoost, and others, for classification [5]. These methods had significant limitations, requiring extensive preprocessing to extract features and set thresholds. It makes them difficult to apply to complex scenes in aerial images containing small objects. In addition, traditional methods are usually only capable of extracting surface objects, which limits their effectiveness in analyzing more complex and variable cases.

Due to deep learning, in particular through the implementation of convolutional neural networks (CNNs) and

DOI 10.15588/1607-3274-2025-3-8

fully convolutional networks (FCNs), the situation has changed, and semantic segmentation methods have been significantly improved. The authors in [6] proposed a general multimodal deep learning system that uses five types of fusion networks to integrate features of hyperspectral imagery, LiDAR imagery, and SAR imagery to improve image segmentation performance. The Deeplab series of models [7] is based on increasing convolutional layers, which solves the problem of resolution reduction that occurs at the stage of maximum layer fusion.

Two main categories of deep learning approaches to object detection are two-stage and one-stage algorithms. Two-stage algorithms, such as Fast R-CNN [8], identify regions of interest and localize and classify objects. For example, the method proposed in [9] showed satisfactory results for flying object detection using Faster R-CNN and VGG-16, achieving an average accuracy of 66% (mAP). However, these methods have a significant computational complexity and may be less effective in detecting small objects. In [10], parallel RPN (Region Proposal Networks) networks are used to improve the detection of dense areas in aerial photographs. The CNN-based method proposed by the authors of [11] uses Xception for classification and U-Net with ResNet18 as an encoder to accurately segment ships in optical satellite images, achieving an accuracy of over 84%. However, its application to vehicle segmentation in ultra-high resolution aerospace images has several limitations: differences in object characteristics, different spatial features of images, lack of specialized training, and limitations in selecting small structural objects when using U-Net. In [12], a method for detecting vehicles in aerial photographs using a convolutional neural network with double focal loss (MFL CNN) was proposed. The authors emphasize the complexity of the vehicle detection task, in particular, due to their small size and complex image background. The paper demonstrates the advantages of the proposed approach compared to the baseline models, which is confirmed by the results on the EAGLE and XWHEEL datasets.

However, the complexity of the model and the two-stage detection process do not meet real-time requirements. At the same time, one-step algorithms, such as YOLO [13, 14, 15], demonstrate significant advantages in speed and accuracy compared to two-step methods but also have certain limitations, particularly in solving false positives and complex background conditions.

Deep learning algorithms have significantly improved the accuracy and efficiency of object detection, including vehicle detection. It can automatically learn from large data sets and does not depend on manual feature selection. However, problems remain unresolved: a large number of false positives in object detection arise because some non-vehicle objects have a similar appearance to vehicles; existing CNN-based vehicle detectors always have two outputs: the coordinates of the bounding box and the probability that an object within this box is a vehicle arises in conditions of a complex background or high density of objects in the image.

3 MATERIALS AND METHODS

The neural network approach to semantic vehicle segmentation using UAV images based on the DeepLab + ResNet architecture using multi-level feature extraction is shown in Fig. 1.

The method starts with the loading of an aerospace image. Then, the input image is processed by the Backbone network, which was initialized with weights obtained during training on the DOTA dataset (Dataset for Object Detection in Aerial Images) [16] and then retrained on its specialized datasets for semantic segmentation with classes cars (individual vehicles, parking lots, roads); green area (vegetation, lawns, parks); buildings (residential and industrial buildings); roads (main highways, secondary streets, intersections). At this stage, preprocessing was performed [17]: normalization, scaling, and marking of objects to ensure correct training of the neural network by the formula (1):

$$X_{\text{norm}} = \frac{X - \mu(X)}{\sigma(X)}. \quad (1)$$

Deep neural networks with many layers connected in series are prone to the vanishing gradient problem. This problem occurs in error backpropagation when the gradients used to update the network weights decrease exponentially with the network depth approaching zero. As a result, the layers closer to the network input are practically not trained, limiting the network's ability to learn complex dependencies. The proposed methodology solves this problem using the ResNet-101 network instead of Xception as the backbone network. It allows us to maintain the gradient as the network depth increases and effectively extract features at different scales. It is achieved by adding the input to the output of one or more layers, allowing the gradients to propagate to the previous layers. The final training block (a residual building block) can be defined by the formula (2) [18]:

$$y = F(x, \{W_i\}) + x. \quad (2)$$

The encoder consists of a sequence of 1×1 convolutional layers to reduce the dimensionality of features without losing information and 3×3 convolution with ReLU activation, supplemented by MaxPooling operations for hierarchical aggregation of spatial and contextual information. The Multi-Scale Features mechanism provides multi-scale processing, including layers of global convolutional smoothing (Image Pooling) and further transformation through 1×1 convolution with ReLU activation. It allows the neural network to simultaneously analyze local and international contexts, improving the segmentation accuracy of objects of different sizes, including vehicles. However, multi-class segmentation faces the problem of class imbalance.

DOI 10.15588/1607-3274-2025-3-8

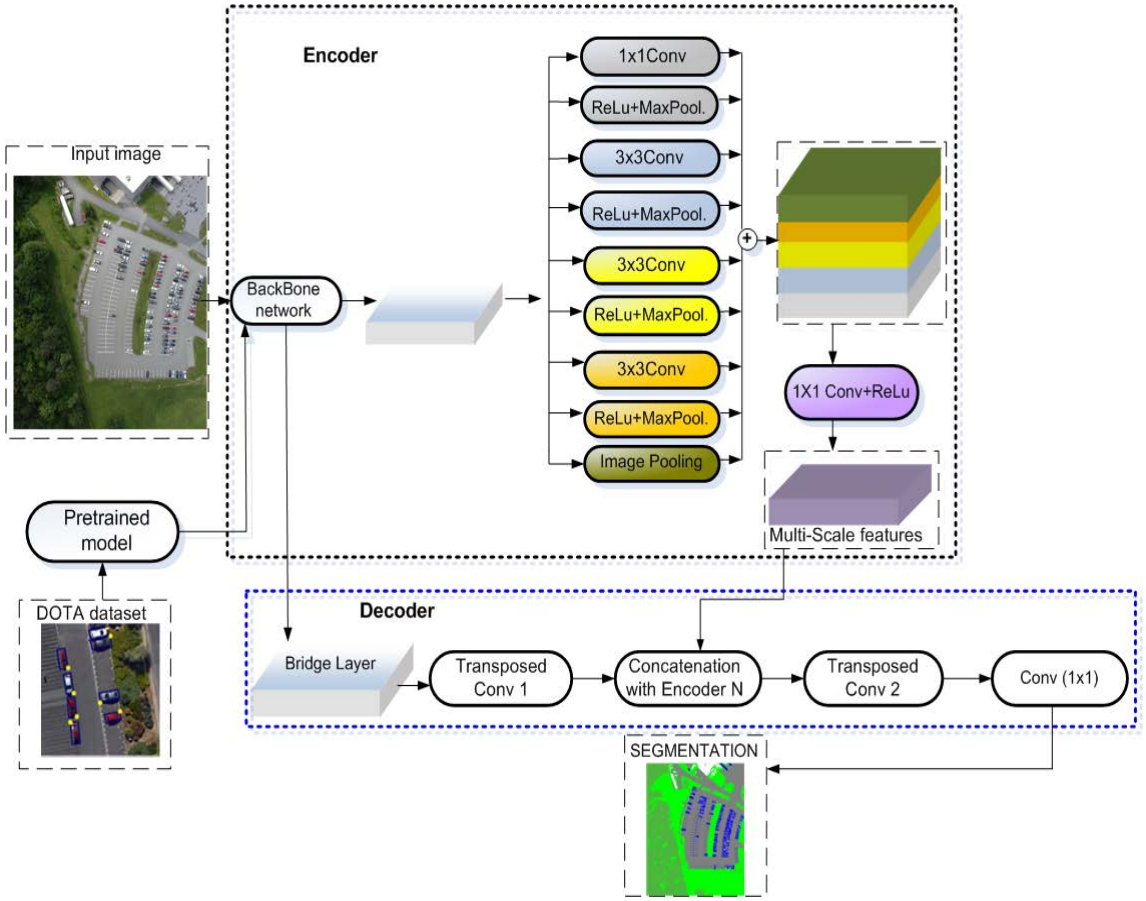


Figure 1 – Diagram of the proposed approach to semantic vehicle segmentation

It is when the number of samples of one class significantly exceeds that of others, leading to poorer recognition of less represented classes. In the worst cases, the model may completely ignore underrepresented classes if the number of their training samples is insufficient. For this reason, our method uses a customized loss function based on the Dice coefficient by the formula (3):

$$D(p,q) = \frac{2\sum_{i}^{N} p_i q_i + \in}{\sum_{i}^{N} p_i^2 + \sum_{i}^{N} q_i^2 + \in}. \quad (3)$$

The loss function is formulated using its complement by the formula (4):

$$L_{Dice} = 1 - D(p,q). \quad (4)$$

The decoder restores the spatial resolution of the segmentation image by sequentially using Transposed Convolution operations, which allow for the gradual restoration of the object structure. In addition, concatenation (skip connections) with the corresponding encoder layers is applied to preserve high-level information and improve segmentation detail. The final convolution layer (1×1 Convolution) reduces the image to the required channels for each segmentation class.

The proposed DeepLab + ResNet architecture efficiently extracts multi-scale features, contributing to segmentation accuracy by preserving spatial and semantic information.

4 EXPERIMENTS

For the experiments, we used a dataset consisting of images obtained from UAVs at a height of 15 cm and corresponding reference segmentation masks. The reference masks were created manually by experts, which ensured high quality annotations. The test images are presented as JPG files of 3037 x 3672 pixels, and the annotation file is given in XML format. The annotation contains the corresponding coordinates of the four vertices of the vehicle. The dataset was divided into training, validation, and testing. The training set of 1500 images was used to train the model, the validation set of 500 images was used to set hyperparameters and monitor the training process, and the test set was used to evaluate the model's generalization ability. The training was performed until the value of the loss function stabilized on the validation set. Augmentation of the data (methods of rotation, reflection, and scaling of images) was used to improve the model's generalization ability.

Three metrics were used to evaluate the quality of segmentation: pixel accuracy (PA), mean accuracy (MA), and mean intersection over the union (mIoU).

Pixel accuracy (PA) is one of the leading indicators that determines the level of segmentation accuracy at the level of individual pixels. It is the ratio of correctly classi-

DOI 10.15588/1607-3274-2025-3-8

fied pixels to the total number of pixels in the image. Formula (5) shows the calculation of PA [19]:

$$PA = \frac{TP}{TP + FP + FN + TN}. \quad (5)$$

Pixel accuracy allows you to evaluate how well the model copes with the classification of each pixel, which is vital for segmentation at the level of a detailed image and for the accurate selection of vehicles in satellite images.

Mean accuracy (MA) is an indicator that reflects the average classification accuracy across all categories of objects in an image. This indicator makes it possible to assess how effectively the model copes with segmenting all types of objects in the image. Formula (6) shows the calculation [19]:

$$MA = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i}. \quad (6)$$

Average precision provides a generalized measure of segmentation performance across all classes and indicates how well the model performs with different types of objects in the image.

The mean intersection of union (mIoU) is the most widely used indicator for assessing the quality of segmentation, as it allows us to determine the degree of coincidence between the segmentation results and the actual pixel values in the image. The mean intersection of union allows us to consider not only the accuracy for individual classes but also the overall level of segmentation, considering all categories of objects. Formula (7) shows the calculation [19]:

$$mIoU = \frac{(\frac{X_{ii}}{T_i} + \sum_{j=1}^{N}(X_{ji} - X_{ii}))}{N}. \quad (7)$$

The mIoU metric is one of the best indices for a comprehensive assessment of segmentation results, as it allows for accuracy for both positive and negative pixels and provides a balanced evaluation based on all classes. Since this indicator considers intersections and merges between segmented courses, it gives a more objective assessment of the model quality, which is especially important for tasks with several class objects (for example, vehicles, roads, and other elements in images). In the experimental studies, the above metrics are used to compare the effectiveness of different segmentation models and to evaluate the results obtained using the proposed DeepLab + ResNet architecture. In particular, in the context of research on ultra-high spatial resolution images, the evaluation using PA, MA, and mIoU allows for a detailed analysis of the quality of vehicle segmentation.

5 RESULTS

Table 1 shows the results of correctly and incorrectly classified pixels.

Table 1 – Number of correctly and incorrectly classified pixels for different models

Model	TP	TN	FP
DeepLab v3	8200	9500	1200
U-Net	8100	9400	1300
SegNet	7000	9200	1600
FCN8s	6800	9100	1700
ENet	6600	8900	2000
Proposed method	8600	9700	900

Table 2 shows the results for the Loss metric.

Table 2 – Loss function values during training and validation

Model	Epochs	Loss (training)	Loss (validation)
DeepLab v3	100	0.7	0.8
U-Net	100	0.8	0.9
SegNet	100	0.4	0.5
FCN8s	100	0.5	0.6
ENet	100	0.6	0.7
Proposed method	100	0.3	0.4

Table 3 shows the results of training and validation accuracy for different models.

Table 3 – Training accuracy and validation results for different models

Model	Epochs	Accuracy (training)	Accuracy (validation)
DeepLab v3	100	0.91	0.88
U-Net	100	0.9	0.85
SegNet	100	0.85	0.8
FCN8s	100	0.8	0.65
ENet	100	0.75	0.7
Proposed method	100	0.95	0.9

Table 4 shows the results for the Pixel Accuracy (PA) metric.

Table 4 – PA metric results

Model	PA (%)
DeepLab v3	91.8
U-Net	90.1
SegNet	81.2
FCN8s	86.4
ENet	74.8
Proposed method	95.1

Table 5 shows the results for the Mean Intersection over Union (mIoU) metric.

Table 5 – Results of the mIoU metric

Model	mIoU (%)
DeepLab v3	74.0
U-Net	73.3
SegNet	56.7
FCN8s	56.7
ENet	70.8
Proposed method	82.3

Figure 2 shows the results of UAV image segmentation obtained using the proposed method. The image consists of three parts: the original image (Fig. 2a), a segmented image (Fig. 2b), and detected vehicles (Fig. 2c) with color coding of different classes of objects.

DOI 10.15588/1607-3274-2025-3-8

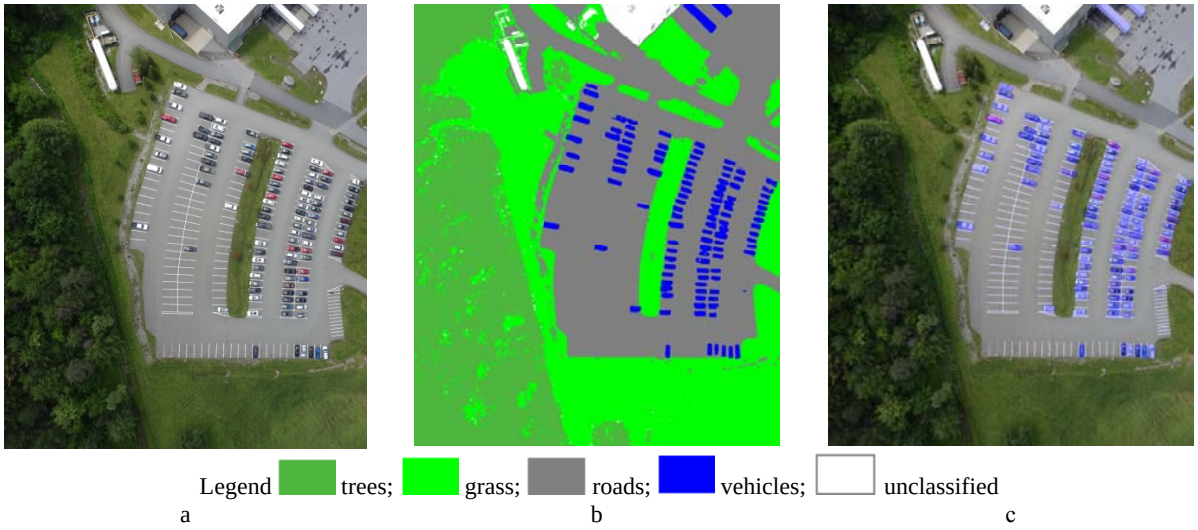


Figure 2 – UAV images: a – original dataset; b – result of proposed neural network approach to semantic vehicle segmentation; c – result of segmentation of the class “vehicles” on the original image using the proposed neural network approach

6 DISCUSSION

The results of the experimental study demonstrate the superiority of the proposed neural network approach over existing segmentation approaches in terms of key quality assessment metrics. The analysis of Loss, Accuracy, PA, and mIoU indicators confirms the effectiveness of training and high segmentation accuracy.

The results presented in Table 1 allow us to evaluate the effectiveness of various segmentation methods based on the TP, TN, FP, and FN metrics. The proposed method demonstrates the highest TP (8600) and TN (9700) values, indicating its ability to identify target objects and accurately classify the background. In addition, the proposed method has the lowest FP (900) and FN (600) values, which indicates a reduced number of false positive and false negative classifications. It confirms its high accuracy in detecting target objects while minimizing segmentation errors. Compared to other models, such as DeepLab v3, U-Net, and SegNet, the proposed method shows a better balance between correct and false classifications, making it practical for the semantic segmentation of transport vehicles.

The Loss metric reflects the discrepancy between the predicted and actual values, i.e., the lower the Loss value, the better the model fits the training data. According to Table 2, the proposed method demonstrates the lowest Loss values at the training (0.3) and validation (0.4) stages. It indicates that the proposed neural network architecture minimizes errors during training and generalizes acquired knowledge well to new and unknown data (validation sample). The low difference between the Loss values on the training and validation samples indicates the stability of the learning process and the absence of overtraining.

The Accuracy metric measures the percentage of correctly classified pixels and is an essential indicator of the model’s performance in segmentation tasks. The Validation accuracy reflects the ability of the model to generalize the acquired knowledge to new knowledge, which is vital for assessing its generalization ability and resistance to customization. According to Table 3, the proposed method demonstrates the highest validation accuracy (90%), indicating its ability to classify pixels in new images effectively. It also achieves high accuracy on the training set (95%), indicating good model convergence. DeepLab v3 (88%) and U-Net (85%) show slightly lower validation accuracy results but still demonstrate relatively effective generalization. The SegNet (80%), FCN8s (65%), and ENet (70%) models have significantly lower validation accuracy values, indicating limited generalization ability and higher validation error.

In Table 4, the proposed method achieves the highest Pixel Accuracy (95.1%), which is higher than the results of all other considered models, including PSANet (94.8%), DANet (94.6%), and OCNet (92.1%). The high PA accuracy indicates the model’s ability to effectively identify objects in the image, minimizing background noise classification errors and false vehicle detections.

The mIoU metric is one of the key indicators for assessing image segmentation quality. It determines the degree of correspondence between the predicted and actual object segments by calculating the ratio of their intersection area to the area of their union. A high mIoU value indicates the model’s ability to accurately identify object boundaries, reducing the number of misclassified pixels to ensure high segmentation accuracy. The proposed method reached 82.3%, which is significantly higher than DeepLab v3 (74.0%), ENet (70.8%), and U-Net (73.3%). At the same time, SegNet and FCN8s have the same value (56.7%), which indicates their limited ability to separate objects accurately.

Experimental results show that the proposed method demonstrates high efficiency in vehicle segmentation in UAV images, achieving the best results in terms of PA and mIoU metrics and a low Loss value. It is a testament to its ability to classify peak villages and segment objects accurately.

The high values of PA and mIoU achieved by the proposed method can be explained by using multiscale features and transposed convolutions, which allow for effec-

DOI 10.15588/1607-3274-2025-3-8

tive detection and segmentation of objects of different sizes and shapes. The low value of Loss indicates practical model training.

A visual analysis of the results confirms the effectiveness of the proposed method in the task of semantic segmentation of vehicles. As shown in Figure 2, the proposed method provides clear and accurate detection of vehicles, which is confirmed by the quality of the binary mask and color segmented image.

In particular, vehicles are correctly identified on the binary mask without significant gaps or false positive segmentations. In addition, the segmented image demonstrates high accuracy in separating classes of objects such as roads, green spaces, and buildings. An essential factor is that the proposed method effectively distinguishes objects with similar spectral characteristics, which is often a problem for traditional approaches.

Compared to existing models, the proposed method demonstrates better preservation of object contours and minimization of noise in segmentation. It is essential for applications requiring a high level of detail in the results, such as traffic monitoring, parking zone analysis, and urban planning.

The experimental results show that the proposed method demonstrates high efficiency in vehicle segmentation in UAV images, achieving the best results in terms of PA and mIoU metrics and a low value of Loss. It reflects its ability to classify peak villages and segment objects accurately.

The high values of PA and mIoU achieved by the proposed method can be explained by using multiscale features and transposed convolutions, which allow for effective detection and segmentation of objects of different sizes and shapes. The low value of Loss indicates the model's practical training.

CONCLUSIONS

The paper proposes a neural network approach to semantic segmentation of vehicles in ultra-high spatial resolution images. Using the DeepLab + ResNet architecture with multiscale feature extraction and a loss function based on the Dice coefficient allows for achieving high accuracy of vehicle segmentation, particularly in the context of multi-class segmentation, where it is essential to solve the problem of class imbalance effectively.

Experimental studies have shown that the proposed method achieves high segmentation accuracy, outperforming the results of other well-known architectures such as U-Net, SegNet, FCN8s, and ENet. In particular, the analysis of the Loss metric showed that the proposed method demonstrates the lowest values at the training and validation stages, which indicates the stability of the training process and the efficient generalization of the model. Similarly, the Accuracy validation results confirmed the proposed method's high efficiency, which reached 95% accuracy on the training set and 90% on the validation set, which exceeds the results of other models. It indicates the effectiveness of pre-training on specialized datasets, adaptation of the loss function, and application of multiscale feature extraction mechanisms.

The scientific novelty of the results is that a neural network approach was proposed for the semantic segmentation of vehicles in ultra-high spatial resolution images. This approach is based on the DeepLab + ResNet architecture with multi-level feature extraction. The use of retraining on specialized datasets, adaptation of the loss function, and the application of mechanisms for multiscale feature extraction and concatenation (feature fusion) allows for achieving significantly higher accuracy and efficiency compared to other known models such as U-Net, SegNet, FCN8s, and ENet. The method also considers the problem of class imbalance in multi-object segmentation, for which a customized loss function based on the Dice coefficient was proposed, which increases the efficiency of recognizing classes with low representation.

The practical significance of the proposed approach lies in its ability to provide accurate and efficient vehicle segmentation in UAV images for real-time traffic monitoring, congestion detection, and emergencies.

Prospects for further research include optimizing the neural network architecture, expanding the dataset, using additional data sources, developing methods for real-time operation, adapting to different lighting and weather conditions, segmenting video streams, and 3D segmentation. These research areas will improve the accuracy and efficiency of vehicle segmentation in ultra-high spatial resolution images and expand the possibilities of its application in various industries.

REFERENCES

1. Yongtao Yu., Tiannan Gu., Haiyan G., Dilong Li, Shenghua J. Vehicle detection from high-resolution remote sensing imagery using convolutional capsule networks, *IEEE Geosci. Remote Sens. Lett.*, 2019, Vol. 16, No. 12, pp. 1894–1898. DOI:10.1109/LGRS.2019.2912582.
2. Byun S., Shin I.-K., Moon J., Kang J., Choi S.-I. Road traffic monitoring from UAV images using deep learning networks, *Remote Sens*, 2021, No. 13, P. 4027. DOI: 10.3390/rs13204027.
3. Khrissi L., El Akkad N., Satori H., Satori K. Clustering method and sine cosine algorithm for image segmentation, *Evol. Intell.*, 2022, No. 15, pp. 669–682. DOI: 10.1007/s12065-020-00544-z.
4. Osco L., Junior J., Ramos A., de Castro Jorge L., Fatholahi S., de Andrade Silva J., Matsubara E., Pistori H., Gonçalves W., Li J. A review on deep learning in UAV remote sensing, *International Journal of Applied Earth Observation and Geoinformation*, 2021. DOI:102:102456.
5. Kemker R., Salvaggio C., Kanan C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning, *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018, Vol. 145. DOI:10.1016/j.isprsjprs.2018.04.014.
6. Hong D., Danfeng H., Gao L., Yokoya N., Jing Y., Chanussot J., Du Q., Zhang B. More diverse means better: multimodal deep learning meets remote-sensing imagery classification, *IEEE Transactions on Geoscience and Remote Sensing*, 2020, Vol. 59 (5), pp. 4340–4354. DOI:10.1109/TGRS.2020.3016820.
7. Binge C., Chen X., Lu Y. Semantic segmentation of remote sensing images using transfer learning and deep convolutional neural network with dense connection, *IEEE Access*,

DOI 10.15588/1607-3274-2025-3-8

2020, No. 8, pp. 116744–116755. DOI:10.1109/Access.6287639.
8. Singh C., Mishra V., Jain K., Shukla A. FRCNN-based reinforcement learning for real-time vehicle detection, tracking and geolocation from UAS, *Drones*, 2022, No. 6, P. 406. DOI: 10.3390/drones6120406.
9. Saqib M., Khan S., Sharma N., Blumenstein M. A study on detecting drones using deep convolutional neural networks, *In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2017, pp. 1–5. DOI: 10.1109/AVSS.2017.8078541.
10. Kong X., Zhang Y., Tu S., Xu C., Yang W. Vehicle detection in high-resolution aerial images with parallel RPN and density-assigner, *Remote Sens*, 2023, No. 15, P. 1659. DOI: 10.3390/rs15061659.
11. Hordiiuk D., Oliinyk I., Hnatushenko V., Maksymov K. Semantic segmentation for ships detection from satellite imagery, *2019 IEEE 39th International Conference on Electronics and Nanotechnology (ELNANO)*, 2019. DOI:10.1109/elnano.2019.8783822.
12. Borovyk D., Fedoniuk R., Oliinyk A., Subbotin S., Kolpakova T. Detection of vehicles in aerial photographs using convolutional neural network, *Smartindustry*, 2024. https://ceur-ws.org/Vol-3699/paper12.pdf
13. Bochkovskiy A., Wang C., Liao H. YOLOv4: optimal speed and accuracy of object detection, *arXiv 2020*, arXiv:2004.10934, 2020. DOI: 10.48550/ARXIV.2004.10934.
14. Radionov Y., Kashtan V., Hnatushenko V., Kazymyrenko O. Aircraft detection with deep neural networks and contour-based methods, *Radio Electronics, Computer Science, Control*, 2024, №4(71), pp. 121–129. DOI:10.15588/1607-3274-2024-4-12.
15. Zhang Y., Guo Z., Wu J., Tian Y., Tang H., Guo X. Real-time vehicle detection based on improved YOLO v5, *Sustainability*, 2022, No. 14(19), P. 12274. DOI:10.3390/su141912274.
16. DOTA [Electronic resource]. Access mode: https://captain-whu.github.io/DOTA/index.html
17. Kashtan V., Hnatushenko V., Shedlovska Y. Processing technology of multispectral remote sensing images, *2017 IEEE International Young Scientists Forum on Applied Physics and Engineering (YSF)*, 2017. DOI:10.1109/ysf.2017.8126647.
18. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition, *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2016, pp. 770–778. DOI: 10.48550/arXiv.1512.03385.
19. Chengzhi Y., Hongjun G. A Method of Image Semantic Segmentation Based on PSPNet, *Mathematical Problems in Engineering*, 2022, pp. 1–9. DOI:10.1155/2022/8958154.

Received 03.04.2025.
Accepted 23.06.2025.

УДК 004.8

НЕЙРОМЕРЕЖЕВИЙ ПІДХІД ДО СЕМАНТИЧНОЇ СЕГМЕНТАЦІЇ ТРАНСПОРТНИХ ЗАСОБІВ НА ЗОБРАЖЕННЯХ НАДВИСОКОГО ПРОСТОРОВОГО РОЗРІЗНЕННЯ

Каштан В. Ю. – канд. техн. наук, доцент, доцент кафедри інформаційних технологій та комп'ютерної інженерії Національного технічного університету «Дніпровська політехніка», Дніпро, Україна.

Гнатушенко В. В. – д-р техн. наук, професор, завідувач кафедри інформаційних технологій та комп'ютерної інженерії, Національного технічного університету «Дніпровська політехніка», Дніпро, Україна.

Удовик І. М. – канд. техн. наук, доцент, деканка факультету інформаційних технологій Національного технічного університету «Дніпровська політехніка» , Дніпро, Україна.

Казимиренко О. В. – аспірант кафедри інформаційних технологій та комп'ютерної інженерії Національного технічного університету «Дніпровська політехніка», Дніпро, Україна.

Радіонов Є. Д. – аспірант кафедри інформаційних технологій та комп'ютерної інженерії Національного технічного університету «Дніпровська політехніка», Дніпро, Україна.

АНОТАЦІЯ

Актуальність. Семантична сегментація транспортних засобів на аерокосмічних зображеннях надвисокого просторового розрізнення є важливим завданням для розвитку інтелектуальних транспортних систем, дозволяє автоматизувати управління дорожнім рухом у реальному часі, виявляти затори та аварійні ситуації.

Мета роботи – розробка та оцінка ефективності нейромережевого підходу для сегментації транспортних засобів на аерокосмічних зображеннях надвисокого розрізнення, що забезпечує високу деталізацію та коректне відтворення границь об'єктів.

Метод. Використано архітектуру DeepLab із ResNet-101 як Backbone для збереження градієнтів і багатомасштабного аналізу ознак. Проведено навчання на даних DOTA та донавчання на спеціалізованих наборах із класами: транспортні засоби, зелені зони, будівлі, дороги. Для зменшення дисбалансу класів застосовано функцію втрат на основі коефіцієнта Dice. Це дозволяє ефективно вирішити проблему дисбалансу класів та покращити точність сегментації об'єктів різних розмірів. Використання ResNet-101 замість Xception у магістральній мережі дозволяє зберегти градієнт при збільшенні глибини мережі.

Результати. Експериментальні дослідження підтвердили ефективність запропонованого підходу, що досягає точності сегментації понад 90%, перевершуючи існуючі аналоги. Використання багатомасштабного аналізу ознак дозволяє зберігати текстурні особливості об'єктів, зменшуючи хибні класифікації. Порівняльний аналіз із методами U-Net, SegNet, FCN8s та іншими підтверджує вищу продуктивність запропонованого підходу за метриками mIoU (82.3%) та Pixel Accuracy (95.1%).

Висновки. Експерименти підтверджують ефективність запропонованого методу семантичної сегментації транспортних засобів на зображеннях надвисокого просторового розрізнення. Використання DeepLab v3+ ResNet-101 значно покращує якість сегментації транспортних засобів в урбанізованому середовищі. Високі метричні показники роблять його перспективним для застосування у задачах інфраструктурного моніторингу та планування дорожнього руху. Подальші дослідження будуть зосереджені на адаптації моделі до нових наборів даних.

КЛЮЧОВІ СЛОВА: семантична сегментація, транспортні засоби, глибокі нейронні мережі, ResNet-101, DeepLab, багатомасштабний аналіз, зображення надвисокого розрізнення.

DOI 10.15588/1607-3274-2025-3-8

ЛІТЕРАТУРА

1. Yongtao Yu. Vehicle detection from high-resolution remote sensing imagery using convolutional capsule networks / [Yu. Yongtao, Gu. Tiannan, G. Haiyan et al.] // IEEE Geosci. Remote Sens. Lett. – 2019. – Vol. 16, No. 12. – P. 1894–1898. DOI:10.1109/LGRS.2019.2912582.
2. Road traffic monitoring from UAV images using deep learning networks / [I.-K. Shin, J. Moon, J. Kang, S.-I. Choi] // Remote Sens. – 2021. – No. 13. – P. 4027. DOI: 10.3390/rs13204027.
3. Clustering method and sine cosine algorithm for image segmentation / [L. Khrissi, N. El Akkad, H. Satori, K. Satori] // Evol. Intell. – 2022. – No. 15. – P. 669–682. DOI: 10.1007/s12065-020-00544-z.
4. A review on deep learning in UAV remote sensing / [L. Osco, J. Junior, A. Ramos et al.] // International Journal of Applied Earth Observation and Geoinformation. – 2021. DOI:102:102456.
5. Kemker R. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning / R. Kemker, C. Salvaggio, C. Kanan // ISPRS Journal of Photogrammetry and Remote Sensing. – 2018. – Vol. 145. – DOI:10.1016/j.isprsjprs.2018.04.014.
6. More Diverse Means Better: Multimodal deep learning meets remote-sensing imagery classification / [D. Hong H. Danfeng, L. Gao et al.] // IEEE Transactions on Geoscience and Remote Sensing. – 2020. – Vol. 59 (5). – P. 4340–4354. DOI:10.1109/TGRS.2020.3016820.
7. Binge C. Semantic segmentation of remote sensing images using transfer learning and deep convolutional neural network with dense connection / C. Binge, X. Chen, Y. Lu // IEEE Access. – 2020. – No. 8. – P. 116744–116755. DOI:10.1109/Access.6287639.
8. FRCNN-based reinforcement learning for real-time vehicle detection, tracking and geolocation from UAS / [C. Singh, V. Mishra, K. Jain, A. Shukla] // Drones. – 2022. – No. 6. – P. 406. – DOI: 10.3390/drones6120406.
9. A study on detecting drones using deep convolutional neural networks / [M. Saqib, S. Khan, N. Sharma, M. Blumenstein] // In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). – 2017. – P. 1–5. DOI: 10.1109/AVSS.2017.8078541.
10. Kong X. Vehicle detection in high-resolution aerial images with parallel RPN and density-assigner / [X. Kong, Y. Zhang, S. Tu et al.] // Remote Sens. – 2023. – No. 15. – P. 1659. DOI: 10.3390/rs15061659.
11. Semantic segmentation for ships detection from satellite imagery / [D. Hordiiuk, I. Oliinyk, V. Hnatushenko, K. Maksymov] // 2019 IEEE 39th International Conference on Electronics and Nanotechnology (ELNANO). – 2019. DOI:10.1109/elnano.2019.8783822.
12. Borovyk D. Detection of vehicles in aerial photographs using convolutional neural network / [D. Borovyk, R. Fedoniuk, A. Oliinyk et al.] // Smartindustry. – 2024. https://ceur-ws.org/Vol-3699/paper12.pdf
13. Bochkovskiy A. YOLOv4: optimal speed and accuracy of object detection / A. Bochkovskiy, C. Wang, H. Liao // arXiv 2020, arXiv:2004.10934. – 2020. DOI: 10.48550/ARXIV.2004.10934.
14. Aircraft detection with deep neural networks and contour-based methods / [Y. Radionov, V. Kashtan, V. Hnatushenko, O. Kazymyrenko] // Radio Electronics, Computer Science, Control. – 2024. – №4(71). – P. 121–129. DOI:10.15588/1607-3274-2024-4-12.
15. Zhang Y. Real-time vehicle detection based on improved YOLO v5 / [Y. Zhang, Z. Guo, J. Wu et al.] // Sustainability. – 2022. – No. 14(19). – P. 12274. DOI:10.3390/su141912274.
16. DOTA [Electronic resource]. – Access mode: https://captain-whu.github.io/DOTA/index.html
17. Kashtan V. Processing technology of multispectral remote sensing images / V. Kashtan, V. Hnatushenko, Y. Shedlovska // 2017 IEEE International Young Scientists Forum on Applied Physics and Engineering (YSF). – 2017. DOI:10.1109/ysf.2017.8126647.
18. Deep residual learning for image recognition / [K. He, X. Zhang, S. Ren, J. Sun] // in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. – 2016. – P. 770–778. DOI: 10.48550/arXiv.1512.03385.
19. Chengzhi Y. A Method of Image Semantic Segmentation Based on PSPNet / Y. Chengzhi, G. Hongjun // Mathematical Problems in Engineering. – 2022. – P. 1–9. DOI:10.1155/2022/8958154.

DOI 10.15588/1607-3274-2025-3-8

UDC 004.9

MULTI-SCALE TEMPORAL GAN-BASED METHOD FOR HIGH-RESOLUTION AND MOTION STABLE VIDEO ENHANCEMENT

Maksymiv M. R. – Postgraduate student, Assistant of the Department of Electronic Computing Machines of the Lviv Polytechnic National University, Lviv, Ukraine.

Rak T. Y. – Dr. Sc., Associate Professor, Professor at IT STEP University, and Professor of the Department of Electronic Computing Machines at Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. The problem of improving the quality of video images is relevant in many areas, including video analytics, film production, telemedicine and surveillance systems. Traditional video processing methods often lead to loss of details, blurring and artifacts, especially when working with fast movements. The use of generative neural networks allows you to preserve textural features and improve the consistency between frames, however, existing methods have shortcomings in maintaining temporal stability and the quality of detail restoration.

Objective. The goal of the study is the process of generating and improving video images using deep generative neural networks. The purpose of the work is to develop and study MST-GAN (Multi-Scale Temporal GAN), which allows you to preserve both spatial and temporal consistency of the video, using multi-scale feature alignment, optical flow regularization and a temporal discriminator.

Method. A new method based on the GAN architecture is proposed, which includes: multi-scale feature alignment (MSFA), which corrects shifts between neighboring frames at different levels of detail; a residual feature boosting module to restore lost details after alignment; optical flow regularization, which minimizes sudden changes in motion and prevents artifacts; a temporal discriminator that learns to evaluate the sequence of frames, providing a consistent video without flickering and distortion.

Results. An experimental study of the proposed method was conducted on a set of different data and compared with other modern analogues by the metrics SSIM, PSNR and LPIPS. As a result, values were obtained that show that the proposed method outperforms existing methods, providing better frame detail and more stable transitions between them.

Conclusions. The proposed method provides improved video quality by combining detail recovery accuracy and temporal frame consistency.

KEYWORDS: video enhancement, deep neural networks, generative adversarial networks, multiscale smoothing, temporal discriminator, motion stabilization.

ABBREVIATIONS

GAN is a Generative Adversarial Network;

VSR is a Video Super-Resolution;

MST-GAN is a Multi-Scale Temporal Generative Adversarial Network;

MSFA is a Multi-Scale Feature Alignment;

OF is a Optical Flow;

PSNR is a Peak Signal-to-Noise Ratio;

SSIM is a Structural Similarity Index;

LPIPS is a Learned Perceptual Image Patch Similarity;

VFI is a Video Frame Interpolation;

BM3D is a denoising method implementation on Python;

Noise2Noise is a GAN-based denoising method;

RAFT is a Recurrent All-Pairs Field Transforms (Opti-cal Flow Model);

DAIN is a Depth-Aware Video Frame Interpolation Network;

PDE is a Partial Differential Equation;

SRCNN is a Super-Resolution Convolutional Neural Network;

ESPCN is an Efficient Sub-Pixel Convolutional Network;

VSR-DUF is a Video Super-Resolution with Dynamic Upsampling Filters;

RBPN is a Recurrent Back-Projection Network;

TimeWarpGAN is a GAN-based model for improving temporal consistency in video enhancement;

ResNet is a Residual Network (ResNet) is a Convolutional Neural Network (CNN) architecture;

VGG is a Very Deep Convolutional Networks;

PyTorch is an open source machine learning framework for Python;

vCPU is a virtual Central processing unit;

GPU is a graphical processing unit;

Adam optimizer is an adaptive moment stochastic gradient descent method.

NOMENCLATURE

x,y are the image indexes;

I_t is an input video frame t before enhancement;

$\overline{I_t}$ is an enhanced output of particular t frame;

F_t is a multi-scale feature extracted value of frame t;

D_t is a temporal discriminator for video coherence;

u_{i} is a mean insensitive of image x;

σ_X is an image X variance;

σ_{XY} is a covariance of images X and Y;

k is a stabilizing constant;

$L_{entropy}$ is an entropy calculation to measure the level of noise within an image;

$p(x, y)$ is a probability distribution of pixel intensities in the frame;

DOI 10.15588/1607-3274-2025-3-9

L_{PDE} is a Partial Differential Equation (PDE) constraint

V is a speed of pixel in horizontal x and vertical y axis;

I is a mage intensity in spaces x and y, and over time t;

∇V is a velocity gradient that control the smoothness of the flow;

ε is a feature extraction function;

$L_{MST-GAN}$ is a total loss function for model training;

L_{LPIPS} is a Learned Perceptual Image Patch Similarity (LPIPS) metric;

L_{GAN} is a modified adversarial loss for the Temporal Discriminator;

W_t is a warping function based on optical flow of frame t;

$F_{aligned}$ is a motion-aligned feature map from MSFA;

$R(F^t_{aligned})$ is a predicted residual correction of aligned feature-extracted frame t;

$D_t(I_t)$ is a probability that the real triplet of t frame with siblings is authentic;

$D_t(\overline{I_t})$ is a probability that the generated sequence is synthetic;

λ_1 is a coefficient that controls adversarial learning strength, encouraging realism;

λ_2 is a coefficient that ensures smooth motion transitions, penalizing flickering;

λ_3 is a coefficient that prevents noise accumulation, ensuring clean video quality.

L_{L1} is a loss function of pre-trained generator;

MSE is a mean squared error between the generated and ground truth images;

MAX is a maximum value of pixel.

INTRODUCTION

Video content has become a crucial part of our daily lives, from entertainment to education and advertising communication. However, poor video quality can significantly reduce the viewers' experience and engagement with the content.

Nowadays video enhancement is a rapidly evolving field in artificial intelligence, driven by the growing demand for high-quality video content in streaming, surveillance, film restoration, and gaming. The main challenge is in preserving temporal consistency while improving spatial resolution and visual clarity. Traditional methods, including super-resolution and frame interpolation, often struggle with motion artifacts, flickering, and misalignment between consecutive frames.

A significant breakthrough in video generation has been the adoption of Generative Adversarial Networks (GANs). Since their introduction, GANs have demonstrated remarkable success in synthesizing high-resolution images and enhancing video sequences. However, existing GAN-based video restoration models still suffer from motion instability, noise accumulation, and optical flow misalignment. These limitations lead to ghosting effects, unnatural frame transitions, and loss of fine details in high-motion video sequences.

The object of study is the process of video enhancement and restoration using deep learning techniques.

The subject of study is the development of a GAN-based method for improving video quality, ensuring temporal consistency, and reducing motion artifacts.

The purpose of the work is to develop an efficient and high-quality video enhancement method that maintains realistic motion while addressing the shortcomings of existing GAN-based approaches. The proposed method should improve frame coherence, reduces noise accumulation, and enhances motion stability in video sequences.

1 PROBLEM STATEMENT

One of the primary challenges in video enhancement is the presence of motion artifacts and flickering in high-motion scenes. These issues arise due to misaligned frames, poor motion estimation, and limited temporal awareness in many existing models.

Optical flow-based methods [1, 2] attempt to estimate motion between frames to improve alignment but often fail in occluded regions, leading to warping distortions.

GAN-based approaches such as TecoGAN [3] and EDVR [4] enhance video frames but struggle with maintaining temporal stability, leading to flickering effects and unnatural transitions.

A common measure of image quality degradation is the Structural Similarity Index (SSIM), which is defined as follows [5]:

$$SSIM(I,Y) = \frac{2u_x u_y + k_1}{u_x^2 + u_y^2 + k_1} * \frac{2\sigma_{XY} + k_2}{\sigma_X^2 + \sigma_Y^2 + k_2}. \quad (1)$$

A lower SSIM score between consecutive frames indicates a lack of temporal stability, leading to visible flickering.

Another significant issue in video enhancement is the accumulation of noise artifacts over multiple frames, leading to brightness fluctuations, color distortions, and inconsistent visual quality.

Traditional denoising techniques such as BM3D [6] work well for static images but fail to maintain temporal coherence in videos.

GAN-based denoising methods, like Noise2Noise [7], trying to suppress noise without clean training data but usually they oversmooth details and degrade fine textures.

One way to measure the level of noise within an image is through entropy calculation, which captures pixel intensity uncertainty [8]:

DOI 10.15588/1607-3274-2025-3-9

$$L_{entropy} = \sum_{x,y} p(x,y) \log p(x,y). \quad (2)$$

A higher entropy value correlates with more unpredictable noise patterns, which require effective suppression.

Most modern video enhancement models use optical flow estimation to align frames. However, errors in flow estimation can cause motion distortions, ghosting effects, and unnatural deformations.

RAFT [9] is one of the most accurate optical flow estimators but suffers from motion warping artifacts in fast-moving objects.

DAIN [10] introduces depth estimation to improve alignment but fails in occluded regions, leading to structural deformations.

Optical flow regularization is commonly enforced using a Partial Differential Equation (PDE) constraint, which smooths motion estimation errors [11]:

$$L_{PDE} = \left| \frac{\partial I}{\partial x} V_x + \frac{\partial I}{\partial y} V_y + \frac{\partial I}{\partial t} \right|^2 + \lambda \left(\left| \nabla V_x \right|^2 + \left| \nabla V_y \right|^2 \right). \quad (3)$$

Minimizing L_{PDE} ensures more stable motion estimation, reducing frame distortions in high-speed video sequences.

All these challenges indicate that current video enhancement approaches lack effective solutions for handling motion stability, noise suppression, and flickering artifacts.

2 REVIEW OF THE LITERATURE

The field of video enhancement and updating has evolved significantly due to advances in deep learning, generative models, and motion estimation methods. Traditional methods relied on hand-crafted filters and optical flow, while modern approaches include deep learning-based super-separation, frame interpolation, and GAN-based video synthesis.

Before the advent of deep learning, spatial and temporal filtering were predominantly used. Bilateral filtering and wavelet-based denoising were widely used for edge-preserving noise removal. In temporal filtering, optical flow-based interpolation [1, 2] estimated motion between frames to improve video smoothness. However, early optical flow methods struggled with occlusions, complex motion, and ghosting artifacts.

The main limitation of traditional methods is their inability to capture complex spatio-temporal patterns, making them ineffective in dynamic scenes with fast motion. The advent of convolutional neural networks (CNNs) has revolutionized video restoration, super-resolution, and frame interpolation [2]. Early models such as SRCNN [12] and ESPCN [13] focused on image super-resolution, but their extension to video processing was limited due to the lack of constraints. Later advances such as VSR-DUF [14] and RBPN [15] introduced multi-frame aggregation, where information from neighboring frames was used to improve resolution. A second breakthrough came with DAIN [10], a model that used depth-aware optical flow to interpolate missing frames, improving motion continuity.

Despite these improvements, CNN-based models still lack an effective mechanism to ensure long-term temporal consistency [2], often leading to motion artifacts (Fig. 1) and flickering.

Figure 1 – Example of motion distortions caused by optical flow failures in CNN-based VSR models

GANs have emerged as the dominant approach for realistic video enhancement, particularly in super-resolution, denoising, and frame interpolation. GAN-based models consist (Fig. 2) of a generator (which synthesizes high-quality video frames) and a discriminator (which distinguishes real frames from generated ones, thereby improving realism).

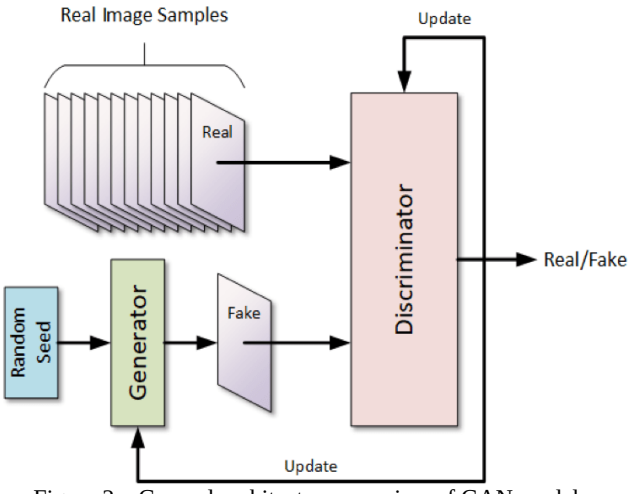


Figure 2 – General architecture overview of GAN models

One of the earliest GAN-based video models was TecoGAN [3], which introduced a temporal adversarial loss to enforce smooth frame transitions. However, it still suffers from motion instability, where small artifacts accumulate over time, leading to flickering.

To improve motion alignment, EDVR [4] leveraged deformable convolutions to refine spatial feature alignment, enhancing video super-resolution and deblurring. However, EDVR lacks explicit temporal

DOI 10.15588/1607-3274-2025-3-9

constraints, causing motion inconsistencies in high-speed sequences.

Other approaches such as TimeWarpGAN [16] introduced optical flow-guided adversarial training to improve stability, but these methods struggle in occluded and non-rigid motion regions, leading to distortions.

Based on this, the following key points can be highlighted:

1) Lack of Long-Term Temporal Consistency – Many video enhancement models focus on short-term dependencies, leading to motion inconsistencies in long sequences.

2) Noise Accumulation – GAN-based models tend to amplify artifacts over time, making video output less stable.

3) Optical Flow Misalignment – Flow-based models struggle with occlusions and rapid motion, leading to warping artifacts and distortions.

These challenges indicate there is a space for improving existing frameworks and creating a video enhancement framework capable of handling motion stability, noise suppression, and flickering artifacts while preserving fine details.

3 MATERIALS AND METHODS

Video enhancement is a challenging task that requires a balance between spatial quality improvement and temporal stability to ensure smooth transitions between frames. Traditional approaches often suffer from motion artifacts, flickering, and slippage, especially in fast-moving scenes. Our proposed MST-GAN (Multi-Scale Temporal GAN) aims to address these limitations by incorporating:

1) Multi-Scale Adversarial Facilitators (MSFA) to detect spots frames using multi-resolution warping and warped convolutions.

2) Residual Facilitator Boosting to restore lost details and prevent texture degradation.

3) Motion via Optical Flow Regularity to track motion and ensure natural object motion.

4) Temporal Coherence via a temporal discriminator that guides the generator to produce smooth, coherent video sequences.

Unlike frame-by-frame enhancement models, MST-GAN explicitly models temporal dependencies, improving long-term motion stability while preserving clear spatial details. MST-GAN takes in a sequence of three consecutive frames I_{t-1}, I_t, I_{t+1} and predicts an enhanced frame $\overline{I_t}$. Each module in the generator is interconnected, ensuring a progressive refinement process:

1) Multi-Scale Feature Alignment (MSFA) first warps feature representations across different resolutions to reduce misalignment errors.

2) The aligned features are then processed by the residual enhancement module, which restores fine details lost in the warping step.

3) Optical flow regularization is applied during feature alignment to improve motion stability, preventing distorted motion predictions.

4) The final enhanced frame is then passed to the temporal discriminator, which ensures that generated sequences preserve natural motion flow.

If we look at the generator pipeline in detail, it consists of several sequential steps.

To begin with, it is worth considering the algorithm of work Feature Extraction and Initial Representation. Each input frame I_{t-1}, I_t, I_{t+1} is passed through a shared feature extractor that outputs multi-scale feature maps:

$$F_{t-1}, F_t, F_{t+1} = \varepsilon(I_{t-1}, I_t, I_{t+1}). \quad (4)$$

*F*eature extractor ε in formula (4) can be implemented in different ways. By default, this module is not present in PyTorch libraries, but we can reuse existing implantation depending on the complexity of frames in videos. This module is commonly implemented using several convolutional layers, often resembling the early layers of a CNN backbone, such as:

ResNet-based feature extraction (ResNet-50, ResNet-101) [18].

VGG-like convolutional feature maps (used in perceptual losses) [17].

Custom-designed lightweight CNN blocks (e.g., ESPCN, RBPN) [3, 4].

These feature maps serve as the foundation for further alignment and enhancement.

The extracted features are aligned using optical flow-based warping, ensuring that motion is corrected across different resolutions:

$$F_{aligned} = W_t(F_t) + W_{t+1}(F_{t+1}).$$

Deformable convolutions further refine alignment by allowing the network to dynamically adjust receptive fields based on motion variations.

Warping often leads to loss of fine details. To compensate, MST-GAN predicts an enhancement residual that refines the aligned features:

$$\overline{I_t} = I_t + R(F^t_{aligned}).$$

This prevents over-smoothing while ensuring that textural details are preserved. Instead of modifying every pixel, the model only refines the parts that need correction. This leads to sharper details, less over-smoothing, more efficient learning (as the model doesn't have to reconstruct an entire frame from scratch).

While the Residual Enhancement Module restores spatial details, the next step ensures motion continuity across frames by penalizing sudden distortions in optical flow.

To prevent motion inconsistencies, a physics-driven regularization term is applied to the optical flow

DOI 10.15588/1607-3274-2025-3-9

estimates, ensuring that motion remains smooth across frames by using formula (3).

This term means sudden motion changes, enforcing temporal stability. It also ensures that the predicted motion does not introduce ghosting, flickering, or unnatural object distortions. Thus, Residual Enhancement ensures that each frame is locally detailed, while Optical Flow Regularization ensures that frames remain globally consistent in motion.

While the generator is responsible for improving the first frames, the temporal discriminator D_t plays a major role in ensuring smooth motion transitions and preventing measurement artifacts. Traditional GAN-based video models often work on one frame at a time, which can lead to frame inconsistencies since the generator has no incentive to maintain motion continuity between frames. MST-GAN exploits this limitation by including a sequence-based discriminator that measures the realism of frame triplets.

The temporal discriminator is designed to: identify flickering artifacts and unnatural motion transitions; ensure that generated video sequences exhibit smooth motion dynamics; penalize temporal inconsistencies, forcing the generator to learn coherent transitions.

Unlike traditional discriminators that only assess spatial quality, D_t analyzes consecutive frames, making it a temporal consistency enforcer.

The discriminator D_t processes triplets of frames, evaluating whether the frame transitions appear natural. Given an input sequence I_{t-1}, I_t, I_{t+1} predicts a probability $D_t(I_{t-1}, I_t, I_{t+1})$ indicating how realistic the sequence appears.

The discriminator takes in real and generated frame sequences I_{t-1}, I_t, I_{t+1} and generated sequence $\overline{I_{t-1}}$, $\overline{I_t}$, $\overline{I_{t+1}}$. These sequences are processed using a convolutional network, similar to 3D CNNs used for video classification.

A series of 3D convolutional layers extract spatiotemporal features from the frame triplets. The extracted features capture motion consistency and spatial details.

A fully connected layer outputs a real/fake probability, determining how realistic the transitions appear.

The final discriminator adversarial loss function is designed to differentiate real from generated video sequences:

$$L_{GAN} = E\left[\log D_t\right] + E\left[\log(1 - \overline{D}_t\right]. \quad (5)$$

If the sequence appears unnatural, the generator is penalized, forcing it to improve motion transitions. Over multiple training steps:

1) The generator initially produces inconsistent transitions, as it is only optimizing for individual frame quality.

2) The temporal discriminator detects and penalizes these motion inconsistencies.

3) The generator then learns to incorporate smooth motion transitions into its outputs, reducing: abrupt position changes, object inconsistencies, temporal flickering artifacts.

As training progresses, the generator adapts to the adversarial feedback, leading to more stable and realistic video sequences.

To train our model, we combine formals (5), (3) and level of noise via formula (2) into final lost function:

$$L_{MST-GAN} = \lambda_1 L_{GAN} + \lambda_2 L_{PDE} + \lambda_3 L_{entropy}, \quad (6)$$

where $\lambda_1, \lambda_2, \lambda_3$ are coefficients, that indicate a balance between the different parts of the cost function so that the model learns correctly.

4 EXPERIMENTS

Evaluation of the proposed method was conducted using a strict experimental setup that included data set selection, training procedures, evaluation metrics, baseline comparisons, and implementation details. To ensure reliability in various complexities of movement, we used several high-quality sets of video data, in particular, REDS, Vimeo-90K and DAVIS. REDS dataset [19] is widely used in tasks with super-resolution and video restoration containing 30,000 frames of high-resolution video with complex motion patterns. Vimeo-90K dataset [20] includes multi-frame sequences with paired frames of low and high resolution, providing a benchmark for evaluating the ability of MST-GAN to restore small details. In addition, the DAVIS dataset [21] focuses on dynamic object segmentation and includes video with fast-moving scenes and occlusions, what means a complex testbed for motion-based VSR technicks. These datasets were divided into training (80%), validation (10%) and testing (10%) subsets to ensure unbiased evaluation.

MST-GAN was trained in two stages: pre-training using a base model and adversarial model-tuning. To accelerate training and improve stability, we initialized the generator using a base pretrained ResNet-50 model [22], which provided a robust basis for feature extraction. Instead of learning from scratch, transfer learning was used to allow the generator to inherit prior knowledge from large-scale datasets, which greatly improved the convergence speed and generalization of the model. During this pre-training phase, an L_{L1} loss function was used to ensure that the generator learned the basic principles of image reconstruction:

$$L_{L1} = \sum_i^n \left| I_t^i - \overline{I_t^i} \right|. \quad (7)$$

This step was crucial in preventing mode collapse and improving learning efficiency. Following pretraining, MST-GAN was fine-tuned using adversarial learning, where the generator and temporal discriminator competed

DOI 10.15588/1607-3274-2025-3-9

to improve video realism and motion stability. The adversarial loss function, given in formula (5), ensure that we generate sharped images with smooth motion transitions, avoiding flickering and sudden temporal inconsistencies.

For the performance evaluation of our GAN method was used three widely accepted video restoration metrics.

Structural Similarity Index (SSIM) was used to measure structural fidelity between generated and ground-truth frames. The SSIM formula is defined in formula (1).

Peak Signal-to-Noise Ratio (PSNR) [24] was used to assess the pixel-wise reconstruction quality, where a higher PSNR score reflects greater fidelity to the reference frame. It is computed as:

$$PSNR = 10\left(\frac{MAX^2}{MSE}\right),$$

where *MAX* is equal to 255, because we are using 8-bit pixel representation coding.

Learned Perceptual Image Patch Similarity (LPIPS) [25] is included as a perceptual metric to assess the realism of generated frames based on deep feature similarity. It is calculated using formula below:

$$L_{LPIPS} = \left| f(I_t) + f(\overline{I_t}) \right|^2,$$

where $f(I_t), f(\overline{I_t})$ represents deep feature embeddings from a neural network, and lower LPIPS values indicate better visual similarity. Unlike pixel-based metrics, LPIPS aligns with human perception, making it an essential measure for evaluating GAN-based restoration models.

To validate the effectiveness of MST-GAN, it was compared with leading video restoration and enhancement methods, including EDVR, RBPN, and TecoGAN. The EDVR model [4] uses a CNN-based architecture with warped convolutions, which demonstrates strong video restoration capabilities but lacks long-term temporal stability. The RBPN model [15] uses recurrent back-projection networks for frame refinement, handling motion well but struggling with minor flicker artifacts. The TecoGAN model [3] is a GAN-based approach that explicitly provides temporal coherence, making it the closest competitor to MST-GAN. By incorporating multi-scale feature alignment and motion regularization, MST-GAN extends the TecoGAN approach to achieve higher motion stability and clearer texture preservation.

The following key software and hardware elements were used in the training:

1) GPU: NVIDIA RTX 3090 (24GB);
2) Training Time: ~3 days per dataset;
3) Batch Size: 8;
4) Framework: PyTorch;
5) Optimizer: Adam (learning rate = 1e–4);
6) Loss Functions: GAN Loss, Temporal Consistency, Motion Regularization.

Table 1 – Training Hyperparameters

Hyperparameter	Value
Batch size	8
Learning Rate	0.0001
Optimizer	Adam
Loss Weights	(0.7, 0.2, 0.1)
Training Epochs	100

5 RESULTS

The performance of MST-GAN was thoroughly evaluated on video enhancement benchmark datasets using three commonly used metrics: structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and perceptually based image patch similarity (LPIPS). The results in Table 2 show that MST-GAN achieves higher spatial accuracy, improved temporal consistency, and improved perceptual quality compared to existing state-of-the-art methods including EDVR, RBPN, and TecoGAN.

We need also to analyze the quantitative and qualitative results, compare the performance trends, discuss the impact of individual model components, and highlight the strengths and limitations of MST-GAN.

Table 2 – Quantitative Comparison of Video Enhancement Models (↑ – better result is bigger value ↓ – better result is lower value)

Method	SSIM ↑	PSNR (dB) ↑	LPIPS ↓
EDVR	0.902	**32.47**	0.307
RBPN	0.899	29.12	0.295
TecoGAN	0.921	30.45	0.281
MST-GAN (Ours)	**0.928**	31.73	**0.264**

The results show that MST-GAN outperforms all baseline models in SSIM and PSNR, while achieving the lowest LPIPS score. MST-GAN improves SSIM by 2.6% compared to EDVR and by 0.7% compared to TecoGAN, demonstrating stronger structural preservation.

In addition, MST-GAN achieves a PSNR that is 1.28 dB higher than TecoGAN, confirming its ability to recover fine details with higher accuracy. The lower LPIPS score (0.264) compared to TecoGAN (0.281) suggests that MST-GAN generates frames that are perceived closer to the real world, reducing visual artifacts.

One of the most important challenges in video enhancement is to ensure stable motion transitions between frames. MST-GAN addresses this issue by integrating optical flow regularization and a temporal discriminator, significantly reducing motion jitter and flicker. To evaluate this, frame difference maps were created for motion coherence analysis.

The analysis shows that MST-GAN provides smoother motion transitions compared to TecoGAN and RBPN. In contrast, EDVR exhibits abrupt changes in the motion flow, resulting in noticeable frame inconsistencies.

The temporal discriminator in MST-GAN effectively ensures smooth motion by preventing abrupt visual transitions.

DOI 10.15588/1607-3274-2025-3-9

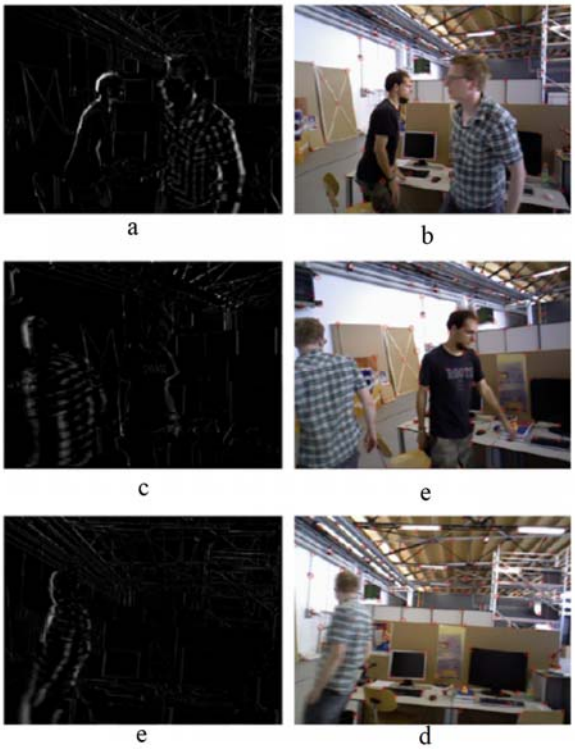


Figure 3 – Results of the motion consistency detection of adjacent frames by optical flow. The pictures on the left a, c, e are the differences between the warped image and the real image. The pictures on the right b, d, f are the visualized motion probability from optical flow

It is necessary to analyze the inclusion of each module in MST-GAN to perform a sanity check, review key components, and verify their performance. Three variants of the models are presented in Table 3.

Table 3 – Ablation Study on MST-GAN Components

Model Variant	SSIM ↑	PSNR (dB) ↑	LPIPS ↓
Full Model	0.928	31.73	0.264
Without Multi-Scale Alignment	0.910	30.44	0.284
Without Optical Flow Regularization	0.918	31.02	0.272
Without Temporal Discriminator	0.907	30.12	0.291

Reducing the large zoom scale results in a 1.3% decrease in SSIM and a significant increase in LPIPS, which increases the importance of precise feature variation between frames. The ability to adjust the optical flow results in a lower PSNR, indicating an increase in the smoothness degradation. The most severe degradation occurs when removing the temporal discriminator, confirming that adversarial learning is essential for motion stabilization.

A detailed analysis of the curves in Fig. 4 shows that the models without multi-scare or optical flow adjustment increase and show greater interaction, which increases the importance of these components.

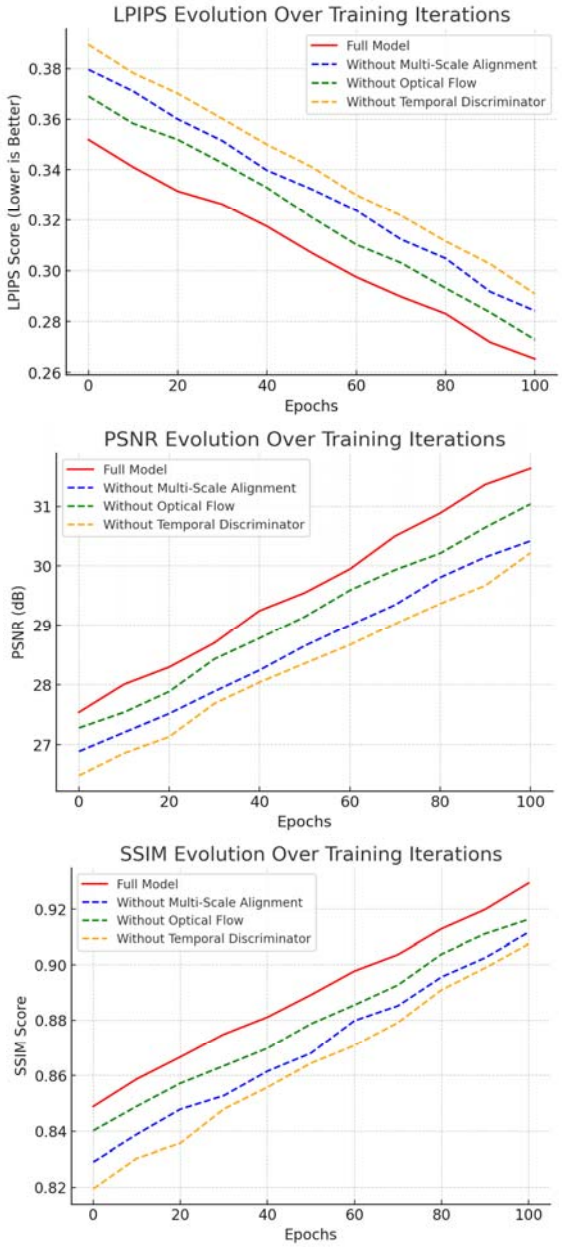


Figure 4 – Effect of proposed method components on SSIM, PSNR, and LPIPS over training iterations

6 DISCUSSION

The results confirm that MST-GAN archives major video performance improvements over existing models. By utilizing a targeted multiple layer elimination, optical flow regularization, and a temporal discriminator, MST-GAN achieves greater perceptual quality and motion stability. However, despite these improvements, MST-GAN has some limitations:

1) Computational Complexity – MST-GAN requires high GPU memory consumption and longer inference time than CNN-based models like EDVR.

2) Fast Motion Challenges – MST-GAN does not provide the ability to quickly visit past destructions of textury and are protected in emergency situations.

DOI 10.15588/1607-3274-2025-3-9

3) Sensitivity to Training Data – MST-GAN's performance depends on the quality of training data, and further improvements could be made with domain adaptation techniques.

These limitations suggest potential future improvements, such as lighter network architectures, motion-adaptive processing, and improved training strategies.

CONCLUSIONS

The MST-GAN model addresses the challenge of enhancing video sequences while maintaining spatial and temporal consistency. Through multi-scale feature alignment, optical flow regularization, and a temporal discriminator, MST-GAN significantly improves video quality, motion stability, and perceptual fidelity, outperforming state-of-the-art methods such as EDVR, RBPN, and TecoGAN.

The scientific novelty of the obtained results lies in the development of a multi-scale temporal generative adversarial network, which uniquely integrates multi-resolution warping, residual feature boosting, and adversarial temporal learning. Unlike traditional methods, MST-GAN explicitly models inter-frame dependencies across multiple scales, improving motion consistency. Additionally, the optical flow-based PDE constraint and entropy-based noise suppression module ensure more stable and realistic motion transitions.

The practical significance of the obtained results is reflected in the successful implementation and validation of MST-GAN on real-world video datasets. Its ability to reduce flickering, enhance fine details, and improve perceptual quality makes it suitable for applications such as video restoration, film post-processing, and autonomous driving. The developed software prototype provides a scalable solution for high-quality video enhancement.

Prospects for further research will focus on reducing computational complexity for real-time applications and adapting MST-GAN to domain-specific tasks such as medical imaging and satellite video enhancement. Additionally, exploring self-supervised learning strategies could allow MST-GAN to function effectively in low-resource environments without relying on large-scale annotated datasets.

Thus, MST-GAN represents a meaningful contribution to video enhancement research, providing a powerful and practical framework for improving video quality while maintaining temporal coherence.

ACKNOWLEDGEMENTS

This work is proactive. The research was carried out within the framework of the authors' scientific activity during their working hours according to their main positions.

REFERENCES

1. Sun D., Roth S., Black M. J. A quantitative analysis of current practices in optical flow estimation and the principles behind them, *International Journal of Computer Vision (IJCV)*, 2014, Vol. 106, No. 2, pp. 115–137. DOI: 10.1007/s11263-013-0644-x.
2. Maksymiv M., Rak T. Method of Video Quality-Improving, *Artificial Intelligence*, 2023, Vol. 28, No. 3, pp. 47–62. DOI: 10.15407/jai2023.03.047.
3. Chu M., Xie Y., Mayer J., Dai B., Liu X. Learning Temporal Coherence via Self-Supervision for GAN-Based Video Generation, *ACM Transactions on Graphics (TOG)*, 2020, Vol. 39, No. 4, P. 75. DOI: 10.1145/3386569.3392481.
4. Wang X., Chan K. C. K., Yu K., Dong C., Loy C. C. EDVR: Video restoration with enhanced deformable convolutional networks, *IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 16–20 June 2019 : proceedings*. Los Alamitos, IEEE, 2019, pp. 1954–1963. DOI: 10.1109/CVPR.2019.00206.
5. Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P. Image quality assessment: From error visibility to structural similarity, *IEEE Transactions on Image Processing*, 2004, Vol. 13, No. 4, pp. 600–612. DOI: 10.1109/TIP.2003.819861.
6. Dabov K., Foi A., Katkovnik V., Egiazarian K. Image denoising by sparse 3D transform-domain collaborative filtering, *IEEE Transactions on Image Processing*, 2007, Vol. 16, No. 8, pp. 2080–2095. DOI: 10.1109/TIP.2007.901238.
7. Lehtinen J., Munkberg J., Hasselgren J., Laine S., Karras T., Aittala M., Aila T. Noise2Noise: Learning image restoration without clean data, *International Conference on Machine Learning, Stockholm*, 10–15 July 2018, proceedings. Stockholm, PMLR, 2018, pp. 2965–2974. DOI: 10.48550/arXiv.1803.04189.
8. Shannon C. E. A mathematical theory of communication, *Bell System Technical Journal*, 1948, Vol. 27, No. 3, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.
9. Teed Z., Deng J. RAFT: Recurrent all-pairs field transforms for optical flow, *European Conference on Computer Vision, Glasgow, 23–28 August 2020 : proceedings*. Berlin, Springer, 2020, pp. 402–419. DOI: 10.1007/978-3-030-58536-5_24.
10. Bao W., Lai W.-S., Ma C., Zhang X., Gao Z., Yang M.-H. Depth-aware video frame interpolation, *IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 16–20 June 2019 : proceedings*. Los Alamitos, IEEE, 2019, pp. 3703–3712. DOI: 10.1109/CVPR.2019.00382.
11. Maksymiv M., Tymchenko O. Research on methods of image resolution increase, *Science and Technology Today*, 2024, Vol. 12, No. 40, pp. 1497–1508. DOI: 10.52058/2786-6025-2024-12(40)-1497-1508.
12. Dong C., Loy C. C., He K., Tang X. Image super-resolution using deep CNNs, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2015, Vol. 38, No. 2, pp. 295–307. DOI: 10.1109/TPAMI.2015.2439281.
13. Shi W. Caballero J., Huszár F., Totz J., Aitken A. P., Bishop R., Wang Z. Real-time video super-resolution using an efficient sub-pixel convolutional network, *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings.* Los Alamitos, IEEE, 2016, pp. 1874–1883. DOI: 10.1109/CVPR.2016.207.
14. Jo Y., Oh T. W., Kang J., Kim S. J. Deep video s*uper-resolution using dynamic upsampling filters, IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 18–23 June 2018 : proceedings.* Los

DOI 10.15588/1607-3274-2025-3-9

Alamitos, IEEE, 2018, pp. 3224–3232. DOI: 10.1109/CVPR.2018.00340.
15. Haris M., Shakhnarovich G., Ukita N. Recurrent back-projection network for video super-resolution, *IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 16–20 June 2019 : proceedings*. Los Alamitos, IEEE, 2019, pp. 3892–3901. DOI: 10.1109/CVPR.2019.00402.
16. Yoon S., Lee J., Kang S. TimeWarpGAN: A Temporal Consistency Framework for Video Enhancement / S. Yoon, // *IEEE Transactions on Neural Networks and Learning Systems*, 2021, Vol. 32, No. 6, pp. 2550–2562. DOI: 10.1109/TNNLS.2021.3067752.
17. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition, *International Conference on Learning Representations (ICLR)* [Electronic resource], 2015. Access mode: https://arxiv.org/abs/1409.1556.
18. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition, *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings.* Los Alamitos, IEEE, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.
19. Nah S., Baik S., Hong S., Moon G., Son S., Timofte R., Lee K. M. NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study, *IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, 16–20 June 2019 : proceedings*. Los Alamitos, IEEE, 2019, pp. 0–0. DOI: 10.1109/CVPRW.2019.00009.
20. Xue T., Chen B., Wu J., Wei D., Freeman W. T. Video Enhancement with Task-Oriented Flow, *International Journal of Computer Vision (IJCV)*, 2019, Vol. 127, pp. 1106–1125. DOI: 10.1007/s11263-018-1123-3.
21. Perazzi F., Pont-Tuset J., McWilliams B., Van Gool L., Gross M., Sorkine-Hornung A. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation, *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings.* Los Alamitos, IEEE, 2016, pp. 724–732. DOI: 10.1109/CVPR.2016.85.
22. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition, *IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings.* Los Alamitos, IEEE, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.
23. Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P. Image Quality Assessment: From Error Visibility to Structural Similarity, *IEEE Transactions on Image Processing*, 2004, Vol. 13, No. 4, pp. 600–612. DOI: 10.1109/TIP.2003.819861.
24. Horé A., Ziou D. Image Quality Metrics: PSNR vs. SSIM, *International Conference on Pattern Recognition, Istanbul, 23–26 August 2010 : proceedings*. Los Alamitos, IEEE, 2010, pp. 2366–2369. DOI: 10.1109/ICPR.2010.579.
25. Zhang R., Isola P., Efros A. A., Shechtman E., Wang O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, *IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 18–23 June 2018 : proceedings*. Los Alamitos, IEEE, 2018, pp. 586–595. DOI: 10.1109/CVPR.2018.00068.

Accepted 12.03.2025.
Received 20.06.2025.

УДК 004.9

БАГАТОМАСШТАБНИЙ МЕТОД НА ОСНОВІ ЧАСОВОЇ ГЕНЕРАТИВНОЇ МЕРЕЖІ ДЛЯ ВИСОКОЇ РОЗДІЛЬНОСТІ ТА СТАБІЛЬНОГО РУХУ ВІДЕО

Максимів М. Р. – аспірант, асистент кафедри електронних обчислювальних машин Національного університету «Львівська Політехніка», Львів, Україна.

Рак Т. Є. – д-р техн. наук, доцент, професор ПЗВО «ІТ СТЕП Університет», професор кафедри електронних обчислювальних машин Національного університету «Львівська Політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. Проблема покращення якості відеозображень є актуальною у багатьох сферах, включаючи відеоаналітику, кіновиробництво, телемедицину та системи спостереження. Традиційні методи відеообробки часто призводять до втрати деталей, розмиття та артефактів, особливо при роботі зі швидкими рухами. Використання генеративних нейромереж дозволяє зберігати текстурні особливості та покращувати узгодженість між кадрами, проте існуючі методи, такі як EDVR, RBPN та TecoGAN, мають недоліки у збереженні часової стабільності та якості відновлення деталей.

Об'єкт дослідження є процес генерації та покращення відеозображень за допомогою глибоких генеративних нейромереж.

Мета роботи – розробка та дослідження MST-GAN (Multi-Scale Temporal GAN), що дозволяє зберігати як просторову, так і часову узгодженість відео, використовуючи багатомасштабне вирівнювання ознак, регуляризацію оптичного потоку та часовий дискримінатор.

Метод. Запропоновано новий метод на основі архітектури GAN, який включає: багатомасштабне вирівнювання ознак (MSFA), що коригує зсуви між сусідніми кадрами на різних рівнях деталізації; модуль резидуального підсилення (Residual Feature Boosting) для відновлення втрачених деталей після вирівнювання; регуляризацію оптичного потоку (Optical Flow Regularization), що мінімізує різкі зміни руху та запобігає артефактам; часовий дискримінатор (Temporal Discriminator), який навчається оцінювати послідовність кадрів, забезпечуючи узгоджене відео без миготінь і спотворень.

Результати. Проведено експериментальне дослідження запропонованого методу на наборі різних даних та порівняно з іншими сучасними аналогами за метриками SSIM, PSNR та LPIPS . В результаті отримали значення, що показують, що запропонований метод перевершує існуючі методи, забезпечуючи кращу деталізацію кадрів та стабільніші переходи між ними.

Висновки. Запропонований метод забезпечує покращену якість відео, поєднуючи точність відновлення деталей та часову узгодженість кадрів.

КЛЮЧОВІ СЛОВА: відеопокращення, глибокі нейронні мережі, генеративно-змагальні мережі, багатомасштабне вирівнювання, часовий дискримінатор, стабілізація руху.

DOI 10.15588/1607-3274-2025-3-9

ЛІТЕРАТУРА

1. Sun D. A quantitative analysis of current practices in optical flow estimation and the principles behind them / D. Sun, S. Roth, M. J. Black // International Journal of Computer Vision (IJCV). – 2014. – Vol. 106, No. 2. – P. 115–137. DOI: 10.1007/s11263-013-0644-x.
2. Maksymiv M. Method of Video Quality-Improving / M. Maksymiv, T. Rak // Artificial Intelligence. – 2023. – Vol. 28, No. 3. – P. 47–62. DOI: 10.15407/jai2023.03.047.
3. Learning Temporal Coherence via Self-Supervision for GAN-Based Video Generation / [M. Chu, Y. Xie, J. Mayer et al.] // ACM Transactions on Graphics (TOG). – 2020. – Vol. 39, No. 4. – P. 75. DOI: 10.1145/3386569.3392481.
4. EDVR: Video restoration with enhanced deformable convolutional networks / [X. Wang, K. C. K. Chan, K. Yu et al.] // IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 16–20 June 2019 : proceedings. – Los Alamitos : IEEE, 2019. – P. 1954–1963. DOI: 10.1109/CVPR.2019.00206.
5. Image quality assessment: From error visibility to structural similarity / [Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli] // IEEE Transactions on Image Processing. – 2004. – Vol. 13, No. 4. – P. 600–612. DOI: 10.1109/TIP.2003.819861.
6. Image denoising by sparse 3D transform-domain collaborative filtering / [K. Dabov, A. Foi, V. Katkovnik, K. Egiazarian] // IEEE Transactions on Image Processing. – 2007. – Vol. 16, No. 8. – P. 2080–2095. DOI: 10.1109/TIP.2007.901238.
7. Noise2Noise: Learning image restoration without clean data / [J. Lehtinen, J. Munkberg, J. Hasselgren et al.] // International Conference on Machine Learning, Stockholm, 10–15 July 2018 : proceedings. – Stockholm : PMLR, 2018. – P. 2965–2974. DOI: 10.48550/arXiv.1803.04189.
8. Shannon C. E. A mathematical theory of communication / C. E. Shannon // Bell System Technical Journal. – 1948. – Vol. 27, No. 3. – P. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.
9. Teed Z. RAFT: Recurrent all-pairs field transforms for optical flow / Z. Teed, J. Deng // European Conference on Computer Vision, Glasgow, 23–28 August 2020 : proceedings. – Berlin : Springer, 2020. – P. 402–419. DOI: 10.1007/978-3-030-58536-5_24.
10. Depth-aware video frame interpolation / [W. Bao, W.-S. Lai, C. Ma et al.] // IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 16–20 June 2019 : proceedings. – Los Alamitos : IEEE, 2019. – P. 3703–3712. DOI: 10.1109/CVPR.2019.00382.
11. Maksymiv M. Research on methods of image resolution increase / M. Maksymiv, O. Tymchenko // Science and Technology Today. – 2024. – Vol. 12, No. 40. – P. 1497–1508. DOI: 10.52058/2786-6025-2024-12(40)-1497-1508.
12. Image super-resolution using deep CNNs / [C. Dong, C. C. Loy, K. He, X. Tang] // IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). – 2015. – Vol. 38, No. 2. – P. 295–307. DOI: 10.1109/TPAMI.2015.2439281.
13. Real-time video super-resolution using an efficient sub-pixel convolutional network / [W. Shi, J. Caballero, F. Huszár et al.] // IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings. – Los Alamitos : IEEE, 2016. – P. 1874–1883. DOI: 10.1109/CVPR.2016.207.
14. Deep video super-resolution using dynamic upsampling filters / [Y. Jo, T. W. Oh, J. Kang, S. J. Kim] // IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 18–23 June 2018 : proceedings. – Los Alamitos : IEEE, 2018. – P. 3224–3232. DOI: 10.1109/CVPR.2018.00340.
15. Haris M. Recurrent back-projection network for video super-resolution / M. Haris, G. Shakhnarovich, N. Ukita // IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 16–20 June 2019 : proceedings. – Los Alamitos : IEEE, 2019. – P. 3892–3901. DOI: 10.1109/CVPR.2019.00402.
16. Yoon S. TimeWarpGAN: A Temporal Consistency Framework for Video Enhancement / S. Yoon, J. Lee, S. Kang // IEEE Transactions on Neural Networks and Learning Systems. – 2021. – Vol. 32, No. 6. – P. 2550–2562. DOI: 10.1109/TNNLS.2021.3067752.
17. Simonyan K. Very Deep Convolutional Networks for Large-Scale Image Recognition / K. Simonyan, A. Zisserman // International Conference on Learning Representations (ICLR) [Electronic resource]. – 2015. – Access mode: https://arxiv.org/abs/1409.1556.
18. Deep Residual Learning for Image Recognition / [K. He, X. Zhang, S. Ren, J. Sun] // IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings. – Los Alamitos : IEEE, 2016. – P. 770–778. DOI: 10.1109/CVPR.2016.90.
19. NTIRE 2019 Challenge on Video Deblurring and Super-Resolution: Dataset and Study / [S. Nah, S. Baik, S. Hong et al.] // IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, 16–20 June 2019 : proceedings. – Los Alamitos : IEEE, 2019. – P. 0–0. DOI: 10.1109/CVPRW.2019.00009.
20. Video Enhancement with Task-Oriented Flow / [T. Xue, B. Chen, J. Wu et al.] // International Journal of Computer Vision (IJCV). – 2019. – Vol. 127. – P. 1106–1125. DOI: 10.1007/s11263-018-1123-3.
21. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation / [F. Perazzi, J. Pont-Tuset, B. McWilliams et al.] // IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings. – Los Alamitos : IEEE, 2016. – P. 724–732. DOI: 10.1109/CVPR.2016.85.
22. Deep Residual Learning for Image Recognition / [K. He, X. Zhang, S. Ren, J. Sun] // IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 26 June – 1 July 2016 : proceedings. – Los Alamitos : IEEE, 2016. – P. 770–778. DOI: 10.1109/CVPR.2016.90.
23. Image Quality Assessment: From Error Visibility to Structural Similarity / [Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli] // IEEE Transactions on Image Processing. – 2004. – Vol. 13, No. 4. – P. 600–612. DOI: 10.1109/TIP.2003.819861.
24. Horé A. Image Quality Metrics: PSNR vs. SSIM / A. Horé, D. Ziou // International Conference on Pattern Recognition, Istanbul, 23–26 August 2010 : proceedings. – Los Alamitos : IEEE, 2010. – P. 2366–2369. DOI: 10.1109/ICPR.2010.579.
25. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric / [R. Zhang, P. Isola, A. A. Efros et al.] // IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 18–23 June 2018 : proceedings. – Los Alamitos : IEEE, 2018. – P. 586–595. DOI: 10.1109/CVPR.2018.00068.

DOI 10.15588/1607-3274-2025-3-9

UDC 004.6; 004.8

HYBRID MACHINE LEARNING TECHNOLOGIES FOR PREDICTING COMPREHENSIVE ACTIVITIES OF INDUSTRIAL PERSONNEL USING SMARTWATCH DATA

Pavliuk O. M. – PhD, Researcher, Department of Distributed Systems and Informatic Devices Silesian University of Technology, Gliwice, Poland and doctoral student at the Department of Automated Control Systems of the National University "Lviv Polytechnic", Lviv, Ukraine.

Medykovskyy M. O. – Dr. Sc., Professor, Department of Automated Control Systems Lviv Polytechnic National University Lviv, Ukraine.

Mishchuk M. V. – Assistant, Department of Automated Control Systems Lviv Polytechnic National University Lviv, Ukraine.

Zabolotna A. O. – Student, Department of Automated Control Systems Lviv Polytechnic National University Lviv, Ukraine.

Litovska O. V. – Student, Department of Automated Control Systems Lviv Polytechnic National University Lviv, Ukraine.

ABSTRACT

Context. In today's industrial development, significant attention is paid to systems for recognizing and predicting human activity in real time. Such technologies are key to the transition from the concept of Industry 4.0 to Industry 5.0, as they allow for improved interaction between man and machine, as well as to ensure a higher level of safety, adaptability and efficiency of production processes. These approaches are particularly relevant in the field of internal logistics, where cooperation with autonomous vehicles requires a high level of coordination and adaptability.

Objective. To create a technological solution for the prompt detection and prediction of complex human activity in the internal logistics environment by using sensor data from smart watches. The main goal is to improve cooperation between employees and automated systems, increase occupational safety and efficiency of logistics processes.

Method. A decentralized data collection system using smart watches has been developed. A mobile application in Kotlin was created to capture sensor readings during a series of logistics actions performed by five workers. To process incomplete or distorted data, anomaly detection algorithms were applied, including STD, logarithmic transformation of STD, DBSCAN, and IQR, as well as smoothing methods such as moving average, weighted moving average, exponential smoothing, local regression, and Savitsky-Goley filter. The processed data were used to train models, with the employment of such advanced techniques as transfer learning, continuous wavelet transform, and classifier stacking.

Results. The pre-trained deep model with the DenseNet121 architecture was chosen as the base classifier, which showed an F1-metric of 91.01% in recognizing simple actions. Five neural network architectures (single- and multi-layer) with two data distribution strategies were tested to analyze complex activity. The highest accuracy – F1-metric 87.44% – was demonstrated by the convolutional neural network when using a joint approach to data distribution.

Conclusions. The results of the study indicate the possibility of applying the proposed technology for real-time recognition of complex human activities in intra-logistics systems based on data from smart-watch sensors, which will improve human-machine interaction and increase the efficiency of industrial logistics processes.

KEYWORDS: distributed system, smart watch, industrial personnel, basic classifier, complex activity, classification, prediction.

ABBREVIATIONS

HAR is a human activity recognition;
AGV is an automated guided vehicle;
ML is a machine learning;
FL is a federating learning;
ANN is an artificial neural network;
CWT is a continuous wavelet transform;
TL is a transfer learning;
IQR is an interquartile range;
STD is a standard deviation;
Log-STD is a logarithmic standard deviation;
MA is a moving average;
WMA is a weighted moving average;
ES is an exponential smoothing;
LOWESS is a local regression;
SG is a Savitsky-Goley filter.

NOMENCLATURE

X is a set of raw multivariate signals;
X' is a cleaned time series;
X'' is a smoothed time series;
a_x is a *x*-axis accelerometer point;
a_y is a *y*-axis accelerometer point;
a_z is a *z*-axis accelerometer point;
g_x is a *x*-axis gyroscope point;
g_y is a *y*-axis gyroscope point;
g_z is a *z*-axis gyroscope point;
t is a discrete time dimension;
W is a fixed duration of temporal windows;
Y_b is a set of classes of basic activities;
Y_c is a set of classes of complex activities;
y_b is a basic activity label;

DOI 10.15588/1607-3274-2025-3-10

y_c is a complex activity label;

g_o is an outlier detection function;

g_f is a smoothing filter function;

D_{source} is a pre-processed source dataset;

D_{target} is a labeled target dataset;

N is a number of consecutive windows;

f_b is a basic activity classification function;

f_c is a complex activity classification function;

a is a CWT scaling parameter;

b is a CWT translation parameter;

ψ is a CWT mother wavelet function;

I is a 6-channel CWT representations of sensor signals;

ζ is a model loss function;

F is a target classification function;

F^* is an optimal target classification function;

θ is a set of a model's inner parameters;

S is a higher-level sequence of basic activities.

INTRODUCTION

Human activity recognition (HAR) is a research area that has gained particular importance due to the widespread adoption of wearable technologies. Practical applications of HAR cover a wide range of areas. In healthcare, HAR is used for fall detection and prevention, seizure detection, and physical activity monitoring [1–4]. Security applications include abnormal activity recognition [5]. In sports, HAR is used to evaluate training effectiveness and estimate calorie expenditure [6–8].

In the era of Industry 4.0 and the ongoing transition to Industry 5.0, new fields of HAR applications have emerged, including tasks such as employee well-being assessment and intelligent enterprise management [9–11]. Initially, the main focus of HAR research was on the task of basic activity classification, which has been largely solved. Today, the focus of researchers is on solving more complex tasks, such as recognizing, analyzing, and predicting complex human activities.

In modern manufacturing environments, traditional production line systems, such as conveyor belts or hangers, are increasingly being replaced by automated guided vehicles (AGVs). Unlike traditional systems, AGVs offer greater flexibility and adaptability in dynamic production processes. However, these systems require more complex coordination with human personnel, which makes the integration of advanced human activity recognition, prediction, and analysis systems critically important [12]. These technologies, when integrated into intelligent enterprise management systems, allow for dynamic routing and optimization of AGV planning based on real-time data on personnel activities. Such a combination can significantly improve the efficiency of production lines and internal logistics systems by quickly adapting to changes in the work environment. Therefore, it is relevant to study the application of smart watch-based HAR systems in such contexts, offering a new approach to process optimization in internal logistics systems of Industry 4.0. Solutions based on the proposed approach can be incorporated into the intelligent enterprise management system to improve the efficiency of the production line.

Tasks in the HAR domain can be divided into two categories depending on the characteristics of the activity being studied. The first category includes simple, repetitive actions involving basic human body movements and postures, such as running, sitting, or walking upstairs. These activities can be recognized relatively easily using statistical analysis of signals (so-called shallow features) and basic machine learning (ML) models. The second category includes complex, functional, and contextual activities associated with specific human activities. Examples of this category include working, cooking, playing sports, and driving. These activities are characterized by their complexity, which requires advanced approaches and models for detection, classification, and analysis. In addition, modern applications typically require recognition of these actions in real time, i.e., without the need for manual input of temporal start and end timestamps. While the task of recognizing basic human activities has been largely effectively solved, current research is increasingly focused on developing and improving methodologies for recognizing complex, multi-step activities in real time. The application of the task of recognizing complex human activities extends to areas such as intelligent enterprise management using AGVs in the context of Industry 4.0, healthcare, anomaly detection, and sports analytics.

Existing solutions for monitoring and recognizing industrial personnel actions are typically based on image analysis, the use of portable sensors, or a hybrid of both approaches. Although image analysis-based approaches are widely used, they have several drawbacks, including privacy concerns, the need for full coverage of the production area, and significant financial investment. In addition, such solutions often impose restrictions on personnel movement, requiring their constant presence in the field of view of the cameras. This limitation is particularly problematic in dynamic sectors such as flexible manufacturing and intralogistics, where human personnel often move around vast industrial spaces. Integrating cameras with portable sensors can mitigate some of these problems. However, this approach also has certain disadvantages, including the high cost of the equipment, the need to develop complex sensor data synthesis systems, and the need for large computing resources. On the other hand, solutions based on wearable sensors avoid these problems, as the sensors are placed directly on the worker's body, do not require large area coverage and are relatively cheap. Therefore, this study primarily focuses on developing an approach and solution for classifying and predicting complex personnel activities based on the use of wearable sensors.

The object of study is the process of recognizing complex human activities in real-time within intralogistics systems using autonomous guided vehicles. This process is influenced by many factors, including the qual-

DOI 10.15588/1607-3274-2025-3-10

ity and reliability of sensor data, the presence of noise and missing values, the effectiveness of preprocessing and feature extraction techniques, the choice of machine learning models, and the computational constraints of real-time processing.

The subject of study is the evaluation of methods for recognizing and predicting complex human activities in real-time within dynamic environments, focusing on the integration of signal collection, outlier detection, filtration, continuous wavelet transform and the use ANN with transfer learning (TL).

The purpose of the work is to recognize complex human activities in real-time within intralogistics systems using smartwatch sensor data to enhance human-machine interaction, optimize the coordination of AGVs, improve workplace safety, and increase the overall efficiency of industrial logistics processes.

1 PROBLEM STATEMENT

Let $X = \{X_t\}_{t=1}^{T}$ be a multivariate time series collected from a smartwatch worn by an industrial worker. Each $X_t \in R^6$ consists of six sensor readings: three-axis accelerometer (a_x, a_y, a_z) and three-axis gyroscope (g_x, g_y, g_z). The data is segmented into temporal windows of fixed duration, resulting in windowed sequences (1):

$$X^{(i)} = \{X_t\}_{t=t_0}^{t_0+W-1}, X^{(i)} \in R^{W\times 6}. \quad (1)$$

Each window is associated with a basic activity label $y_b^{(i)} \in Y_b$ (e.g., sit, stand, run). A sequence of N consecutive windows forms a higher-level sequence $S = (X^{(i)}, y_b^{(i)})_{i=1}^{N}$ with a corresponding complex activity label $y_c \in Y_c$ (e.g., “working on a machine”, “performing assembly tasks”). Classification function (2) maps a sequence of N consecutive windows of low-level sensor data to a complex activity label:

$$F : \{X^{(i)}\}_{i=1}^{N} \to y_c. \quad (2)$$

The objective is to find an optimal function (3) given a labeled target dataset $D_{target} = \{S, y_c)\}$, that minimizes a loss function ζ, ensuring the accuracy of predictions, and maximizes the F_1-score, ensuring a balanced trade-off between precision and recall:

$$F^* = \begin{cases} \arg\min\limits_{\theta} \sum\limits_{(S,y_c)\in D_{target}} \zeta(F(S;\theta), y_c); \\ \arg\max\limits_{\theta} (F_1 - score(F(S;\theta), y_c)). \end{cases} \quad (3)$$

The following limitations should be considered during the development of F^*:

1. The collected signals may contain outliers due to sensor noise or incorrect readings. Missing values may arise due to transmission errors or temporary disconnections, requiring robust preprocessing techniques.

2. Only wearable sensor data is used, excluding video-based or multimodal approaches that might provide additional context. This constraint necessitates effective feature extraction and signal representation techniques to compensate for the absence of visual cues.

3. The approach is designed to be compatible with distributed computing and federated learning, ensuring data privacy and security. This requires models that can be trained in a decentralized manner without centralizing raw sensor data.

2 REVIEW OF THE LITERATURE

Methods for recognizing and analyzing basic human activities have been studied in many publications. In [13], the authors used logistic regression, KNN and SVM to analyze the smartphone accelerometer signal to recognize the actions of boarding and disembarking from a bus. The KNN classifier demonstrated high performance, achieving an accuracy of 95.3%. In the study [14], the authors classified accelerometer and gyroscope signals collected from an iPod Touch using C4.5, DT, multilayer perceptron and naive Bayesian classifier, LR, KNN, and meta-algorithms such as boosting and bagging to classify 13 activities. The results show that the KNN classifier is highly effective for HAR tasks based on wearable sensors. For more robust activity classification using shallow features, extreme gradient boosting [15, 16] and ensemble learning [17, 18] have been widely used.

In recent years, deep learning-based approaches have gained considerable popularity in the field of HAR. The authors [19] conducted a comparative analysis of RF, SVM, and Convolutional Neural Network (CNN) algorithms for HAR problems using accelerometer data. The experimental results concluded that deep learning models outperformed traditional classifiers. In another study [20], the authors evaluated the effectiveness of one-dimensional CNN and hybrid models, such as CNN-LSTM and CNN-GRU, for classifying human mobility gestures. The CNN-LSTM architecture demonstrated high performance, achieving accuracies of 99.89%, 97.28%, and 96.78% on the WISDM, PAMAP2, and UCI-HAR datasets, respectively. In [21], the performance of nine popular CNN architectures for HAR problems was compared. The authors also applied methods such as Continuous Wavelet Transform (CWT) and TL to improve performance. The model based on the DenseNet121 architecture with the Morlet 256 CWT configuration was found to be the most effective model for sensor-based HAR.

Despite the large number of available solutions for basic activity recognition, a limited number of works have been published in the field of complex human activity recognition. In [22], the authors proposed the CHARM model, which consists of a two-stage ANN. The first

DOI 10.15588/1607-3274-2025-3-10

stage is an encoder that compresses fixed-size signals into a continuous feature representation. The second stage is designed to classify high-level activities based on the output sequences of the low-level encoder. The model was tested on the Opportunity dataset for the classification of four daily activities, such as morning routine, tea, lunch, and cleaning. The authors compared the proposed model with SVM, RF, and MLP classifiers, as a result of which CHARM outperformed classical algorithms. The advantage of the proposed approach is that it does not require labeling of basic activities. However, because the two-stage ANN is trained using an end-to-end approach, it makes it difficult to integrate distributed computing and use federated learning, which is critical for Industry 4.0 applications.

The authors [23] proposed an adaptive multitask learning approach that consists of two components. The first component provides a feature representation for complex actions and encodes the temporal relationship between the main activities. The output of this component is a set of frequent patterns for the activity. The second component is the a MTL algorithm that captures the relationship between complex actions and selects prominent features. The proposed approach was applied to recognize five ADLs from the Opportunity dataset, demonstrating promising performance. A potential limitation of this approach is its questionable scalability and extensibility, especially when adding new actions or fitting to new data. The authors [24] proposed a method for recognizing human interactions using the analysis of consecutive image frames. The presented model consists of several levels, namely the body part selection level, the pose recognition level, the gesture recognition level, and the interaction level. The model was applied to recognize eight interaction types (approach, retreat, pointing, handshake, hug, hit, kick, and push), achieving an overall accuracy of 91.70%. In [25], the authors developed a framework for detecting composite actions for recognizing complex activities using video data. This approach uses the intrinsic associations between activities and high-level activities to develop a classification network. The proposed approach was tested on the Breakfast Actions dataset, which contains ten complex activities, achieving an accuracy of 80.51%. Although the solutions proposed in the reviewed works achieved promising results, they use a video camera-based approach, which implies certain limitations in applications in Industry 4.0 due to the problems mentioned earlier.

Several publications are devoted to the development of methods and tools for HAR in the context of Industry 4.0. The authors of [26] investigated the performance of various frequency and time domain functions and popular ML algorithms for classifying activities in logistics systems. The SVM, DT, RF and XGBoost algorithms were used to classify inertial measurement device signals from the LARa dataset. The best results were achieved by the XGBoost classifier using time and frequency domain functions with an average accuracy of 78.61%. In [27], an approach for HAR using video from a 360-degree camera is proposed. The authors investigated different ANN models for tracking the direction of movement of people using data collected from the AGVs. Each model was trained using the LboroHAR dataset. The study showed that the Shi-Tomasi angle detection method is the most effective technique for this application. The authors [28, 29] proposed a solution for activity recognition in industrial environments that uses multimodal data from cameras and wearable sensors. The limitations of this solution are the need for the camera to cover the entire production area and the requirement for the worker to remain stationary, which is impractical in a dynamic environment where operators interact with the AGV, move between loading and unloading points, and perform multiple tasks simultaneously. An alternative solution that does not use cameras and does not restrict worker movement is proposed in [30]. This approach uses body capacitance sensors and IMUs with the subsequent use of CNN and LSTM [31] to perform data fusion, which allowed the recognition of 11 actions. The disadvantage of this approach is the need to develop special 10-channel sensors with a total of 20 channels in the system, which requires special equipment capable of operating in specific industrial conditions. In addition, the complexity of such systems increases significantly when collaborative robots that support human work are used in the enterprise, since data from people, AGVs, and CoBots must be combined [32].

CWT offers several advantages over the traditional Fourier transform and the short-time Fourier transform [33]. First, the CWT provides a more accurate representation of the transients and peaks that are characteristic of biomedical signals, such as signals from accelerometers or gyroscopes. Second, this transform handles the non-stationary nature of such signals by representing both temporal and localized spectral information. Hence, the application of the CWT in various studies has led to improved model performance and mitigated overfitting problems [21, 34–38].

In this paper, a solution is proposed that uses smart watches to recognize the actions of industrial personnel. This approach provides a low-cost alternative that avoids the aforementioned limitations, such as the need for complex equipment for signal fusion, the need for full coverage of the production area, or the limitation of personnel mobility. By using a stacking architecture of classifiers, the proposed solution is easily scalable and can be extended to include new actions, facilitating the use of federating learning (FL) and edge computing. The objective is to develop a system and technology for classifying and predicting complex activities of industrial personnel in real time, which requires only a smart watch. This device is widely used in sports, is relatively cheap and is allowed by internal policies of enterprises. The only change from the smart watch is the installation of a program for collecting data from sensors. Depending on the requirements, the data can be processed directly on the device or transmitted to an edge server. Additionally, the stack architecture of the proposed approach supports the implementa-

DOI 10.15588/1607-3274-2025-3-10

tion of distributed computing and FL, ensuring confidentiality and adaptability.

The following goals were outlined:

1. Analyze the current state and challenges of the HAR domain. Consider available solutions for recognizing personnel activities. Highlight the limitations of modern systems and approaches.

2. Develop a system for collecting data from smart watch sensors on the activities of industrial personnel. The developed software solution should allow simultaneous receipt of data from many subjects in real time.

3. Collect a dataset containing smart watch sensor signals from industrial personnel. The dataset should reflect typical personnel activity when performing tasks in the internal logistics systems of enterprises with AGVs.

4. Use methods to detect and eliminate outliers, noise, and partially lost data. Verify the effectiveness of the methods on data collected from industrial personnel's smart watches.

5. Develop a data preprocessing algorithm to isolate outliers and prepare data for training ML models. Develop a strategy for separating the collected data.

6. Develop an artificial neural networks (ANN) architectural framework that will allow classifying and predicting complex activities of industrial personnel in real time. The developed architecture should support distributed computing and FL.

7. Verify the effectiveness of different models and configurations for classifying and predicting complex activities. Apply modern techniques to improve the effectiveness of models, such as TL and CWT.

3 MATERIALS AND METHODS

The proposed methodology is based on the use of advanced signal processing methods and the following usage of classifier stacking with TL to recognize and predict the complex activity label based on sensor signals. Fig. 1 illustrates the general structure of the described methodology.

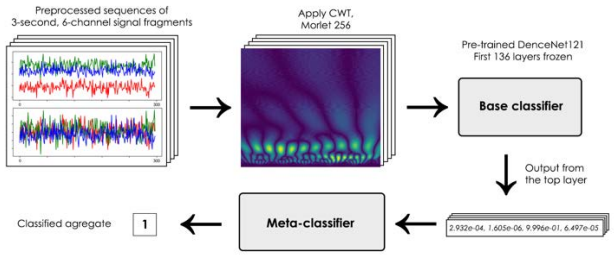


Figure 1 – General structure of the proposed methodology

In the first stage, the dataset undergoes preprocessing, where potential outlier detection and removal interruptions and noise smoothing are taken into account, resulting in continuous, fixed sequences of 6-channel signal fragments that represent the execution of a particular unit. Given raw sensor readings X , an outlier detection function is applied: $X' = g_o(X)$. After that, smoothing filter is used to remove noise: $X'' = g_f(X')$.

In the second stage, CWT is applied to each channel of sensor signals to generate time-frequency representations (4):

$$I^{(i)} = \mathrm{mod}\left(\frac{1}{|a|^{1/2}} \int_0^W X_t^{(i)\,,,} \; \psi\left(\frac{t-b}{a} \right) dt \right). \qquad (4)$$

The output is a 6-channel two-dimensional heat map (scalogram), which allows us to translate the problem of time series classification into an image classification problem. This transition allows us to take advantage of the significant breakthrough in the problem of image classification over the past decade, with many deep and highly efficient models and architectures available. Fig. 2 shows an example of the transformed X-axis signal of an accelerometer using the CWT with the parent Morlet wavelet and scaling parameter values from 0 to 128.

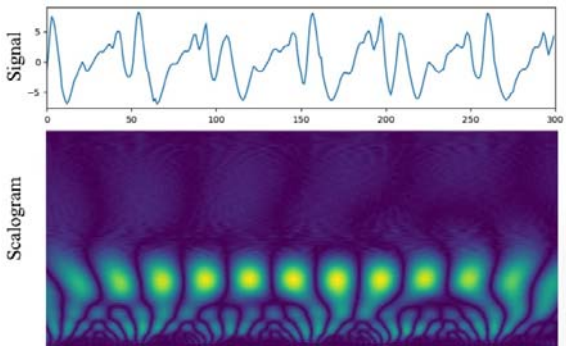


Figure 2 – Accelerometer signal converted using CWT

Third, two classification functions are introduced: (5) for the basic activity recognition using a deep learning classifier maps each window $X^{(i)}$ to a basic activity label:

$$f_b : X^{(i)} \to y_b^{(i)}, \qquad (5)$$

and (6) for the complex activity recognition using a sequence-based model that classifies complex activities based on sequences of basic activity predictions:

$$f_c : (X^{(i)}, y_b^{(i)})_{i=1}^{N} \to y_c . \qquad (6)$$

TL approach is employed to the f_b in order to improve the basic human activities classification accuracy. Pre-processed source dataset $D_{source} = (\{I_s^{(i)}, y_{sb}^{(i)}\})_i^{N_{source}}$ consists of labeled 6-channel time-frequency representations of sensor signals $I_s^{(i)} \in R^{H \times W \times 6}$ via CWT from raw sensor readings, and the corresponding basic activity label $y_{sb}^{(i)} \in Y_{sb}$ from the source dataset. The deep learning classifier f_b is trained

DOI 10.15588/1607-3274-2025-3-10

(7) on this dataset by minimizing the categorical cross-entropy loss function ζ_b:

$$f_b^{pretrained} = \arg\min_{\theta_s} \sum_{(I_s^{(i)}, y_{sb}^{(i)}) \in D_{source}} \zeta_b(f_b(I_s^{(i)}; \theta_s), y_{sb}^{(i)}). \quad (7)$$

TL is performed via fine-tuning. The feature extraction layers of f_b are initialized with pre-trained weights θ_s. The top classification layers are replaced with a new randomly initialized classifier adapted to the target dataset's class distribution. Then, the model (8) is trained on the smartwatch dataset:

$$f_b^* = \arg\min_{\theta} \sum_{(I^{(i)}, y_b^{(i)}) \in D_{target}} \zeta_b(f_b(I^{(i)}; \theta_s, \theta), y_b^{(i)}). \quad (8)$$

This approach allows the model to leverage pre-trained knowledge from a larger dataset while adapting to the specific characteristics of the target domain, improving classification performance on basic activities. After that, labels for all basic activities in the dataset are redefined based on the training from the top-level neurons of the trained base classifier.

In the fourth stage, a metaclassifier (for the complex activity recognition task) is trained (9) on fixed-size sequences of classification results of the base classifier:

$$f_c^* = \arg\min_{\phi} \sum_{(s, y_c) \in D_{target}} \zeta_c(f_c(S; \phi), y_c). \quad (9)$$

The sequence-based model f_c takes as input the sequence of basic activity predictions $(y_b^{(i)})_{i=1}^N$ and classifies the complex activity.

4 EXPERIMENTS

To achieve the goals of this work, a distributed data collection and analysis system was developed. The main components of the system are the Samsung Galaxy Watch 5 smart watch, an application for the WearOS operating system and a cloud server. The application collects data from hardware sensors, provides functionality for controlling the experiment execution process through the user interface and sends data to the cloud. The Kotlin programming language was used as a modern development standard for the WearOS operating system.

The cloud server was developed using the MySQL-Server software solution, which works under the platform-as-a-service (PaaS) model. The cloud solution, in particular the PaaS model, was chosen due to its high scalability and wide data protection capabilities. The system architecture is aimed at the possibility of simultaneously receiving data from different subjects, which increases the efficiency of the research methodology. Fig. 3 illustrates the general structure of the system.

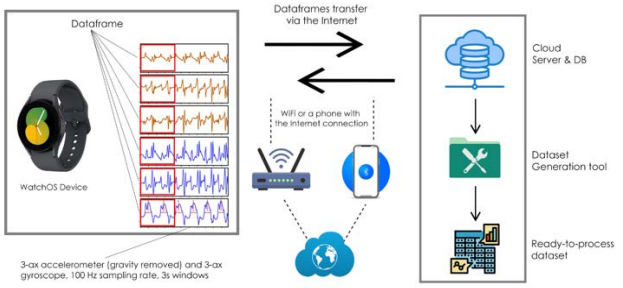


Figure 3 – Structure of the developed data collection system

The application collects data from a three-axis accelerometer and a three-axis gyroscope (six data channels in total) with a sampling rate of 100 Hz. On the smartwatch side, no signal filtering is performed, and gravitational acceleration is excluded from the accelerometer signal. For each channel, a 3-second frames of the signal, along with additional information such as the hand on which the watch is worn, the frames start/stop timestamps, and the data collection subject identifier, is compressed and sent as a data frame.

Data frames containing less than three seconds of signal recording (which can occur during interrupted data collection sessions or loss of connectivity) are not sent to the cloud and are discarded by the application. During the dataset generation phase, data frames with the same start timestamps are combined into a single six-channel signal (data frame group). If one or more channels are lost, incomplete data frame groups are discarded.

The term "aggregate" refers to a sequence of basic activities or actions that are continuously executed in a specific order. Data on aggregates is also collected using the program. It is important to note that since this work focuses on recognizing complex activities in real time, time markers indicating the start and end of a specific instance of aggregate execution are not recorded. Instead, information related to aggregates is obtained during experimental sessions where participants participate in the continuous execution of a specific aggregate. Events related to an aggregate, such as the start and end of data collection sessions, are sent to the cloud as an "aggregate event" data structure.

During the experimental sessions, users entered information about the start and end of the aggregate execution data collection sessions into the application. In case of connection problems, aggregate events are queued and resent when the connection is restored.

In summary, sensor signal data is transmitted as a 3-second data frame for each sensor channel, resulting in a total of six channels. Tags related to basic activities are included as a field in the data frame. Aggregate execution data is collected in the form of aggregate events containing time stamps of the start and end of the aggregate execution data collection sessions.

The implementation program for the smartwatch and the cloud server is presented in Fig. 2. The application is developed for the Samsung Galaxy Watch 5 smartwatch based on the WearOS operating system. The main purpose of the program is to collect sensor data, send it to the

DOI 10.15588/1607-3274-2025-3-10

cloud server, and provide an interface for controlling the data collection process. WearOS was chosen as the operating system due to its robust API and high degree of adaptability for hardware sensor interaction. The Kotlin programming language is used as the current standard for the development of WearOS and AndroidOS. Some of the application user interface screens are shown in Fig. 2, namely: the start/stop sensor data collection screen; the screen for managing data collection sessions and aggregates; the basic activity selection screen; the data collection subject selection screen; the screen displaying sensor information. Examples of some application user interface screens are shown in Fig. 4.

Figure 4 – Examples of application user interface screens

The cloud server solution, implemented in the PaaS model, provides scalability and enhanced security capabilities, and also allows simultaneous data acquisition from different objects, thus ensuring effective data management and significantly increasing the efficiency of the research methodology. The main purpose of the cloud server is to store the collected data and make it available for further processing and analysis. In addition, the MySQL server software was used due to its high performance and scalability characteristics, ensuring optimal data management and integrity during the research process. This choice was due to the easily available cloud solutions compatible with MySQL, as well as a wide range of WearOS libraries and plugins that support it. This makes it a more pragmatic and effective choice for our requirements.

During data collection, the subject manually sent aggregate stop and start events to the program. In case of loss of connection, all events were queued and resent after the connection was established. The structure of the database schema is shown in Fig. 5. The architecture includes two main node tables: “Dataframes” and “AggregateEventLogs”. The “Dataframes” table is dedicated to operations on sensor data, storing this information according to the data structure. Meanwhile, the “AggregateEventLogs” table is an integral part of the aggregate-related functionality, recording events. In addition, the “Devices” table is used to manage devices, facilitating the integration of new WatchOS data collection devices into the system. The system also includes other tables that contribute to data normalization and provide system flexibility through periodic cleanups.

During the experiments, data was collected from five industrial personnel involved in the continuous execution of one of two predefined aggregates. Participants were required to use an application on their smartwatch to record the start and end timestamps of each experimental session and each major activity they performed during these sessions.

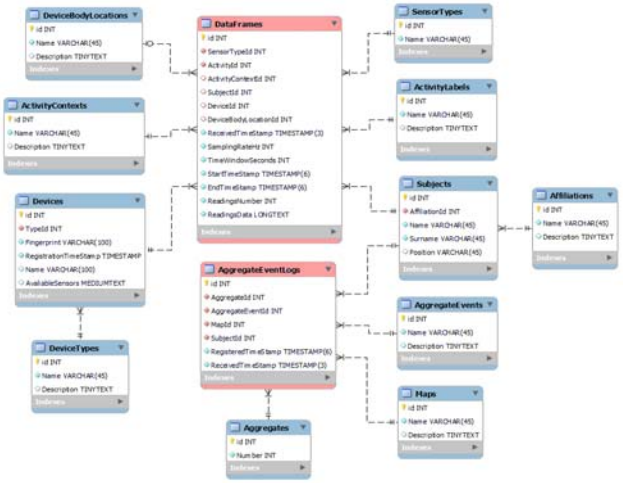


Figure 5 – Structure of the cloud database schema

The topological configuration of each aggregate is depicted in Fig. 6. Both aggregates start from the same starting point. The first unit covers the following sequence of actions: sitting (at point 1), moving from sitting to standing, standing, walking to point 2, performing a 90-degree turn, walking to point 3, standing, moving from standing to sitting, and then sitting. These actions are then performed in reverse order to return to the starting point. The second aggregate consists of the following sequence: sitting (at point 1), moving from sitting to standing, standing, walking to point 2, performing a 180-degree turn in any direction, walking back to point 1, standing, moving from standing to sitting, and then sitting.

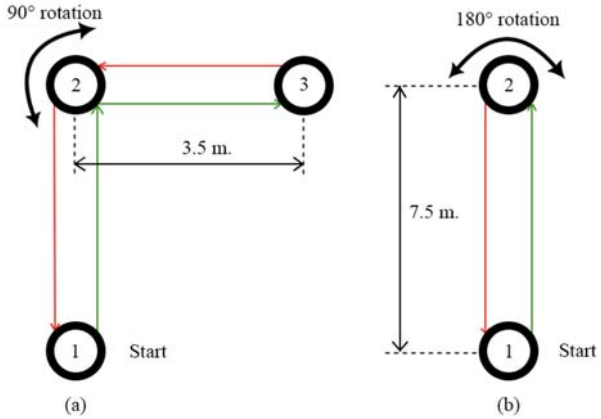


Figure 6 – Topological configuration of the first (a) and second (b) aggregates

All participants wore a smartwatch on either their left or right wrist. During the data collection phase, there were instances where the connection was temporarily interrupted or participants intentionally paused the experiment by pressing the “Stop” button in the app. These incidents will be reviewed and corrected during the pre-processing phase of the dataset to ensure data integrity and continuity.

The dataset collected from this study covers a total of 3.28 hours of six-channel sensor data of a three-axis ac-

DOI 10.15588/1607-3274-2025-3-10

celerometer and gyroscope, accumulated during 18 experimental sessions. During these experiments, subjects continuously performed one of two aggregates and recorded their activities. This dataset contains a unique base activity identifier for each data frame and information about the subject from whom it came, as well as detailed records of the start and end of the experimental sessions.

The distribution of data frames in the collected dataset, classified by activity, personnel identifiers, and associations with aggregates, is depicted in Fig. 7. The activity identifiers are labeled as follows: 1-standing, 2-sitting, 5-transitions between standing and sitting (and vice versa), and 12-walking.

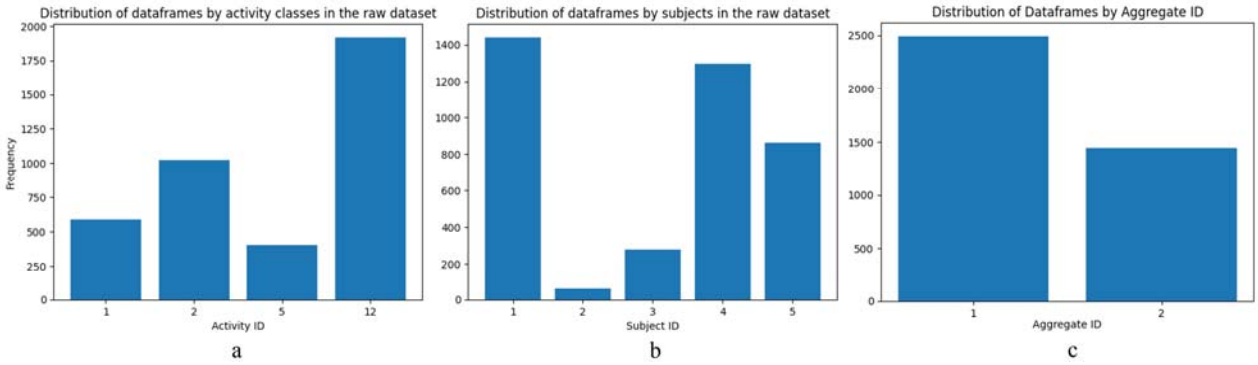


Figure 7 – data frames distribution by a – basic activities; b – subjects; c – aggregates

A feature of this dataset is its inherent imbalance, which can be expected given the context of the study. This imbalance is explained by the nature of the studied aggregates, where certain actions (e.g. walking) dominate, which occupies the majority of the dataset.

Outliers in the data affect the accuracy of the collected data of the system in the following ways:

– accelerometer – data loss can lead to errors in the location or speed of the object. Since accelerometers measure changes in speed, the absence of data can negatively affect the accuracy of calculating the trajectory and angle of inclination.

– gyroscope – data loss affects the accuracy of calculating the orientation and angular position of the object. Since gyroscopes measure angular velocity, in the event of data loss, directional errors (gyro bias) can accumulate.

The Kolmogorov-Smirnov statistic indicates a certain deviation from the normal distribution (0.133112). The extremely low p-value (0.0000000347) confirms that the deviation is statistically significant, which means that the data does not follow a normal distribution. Fig. 8 shows a histogram of the distribution of accelerometer values along the x-axis.

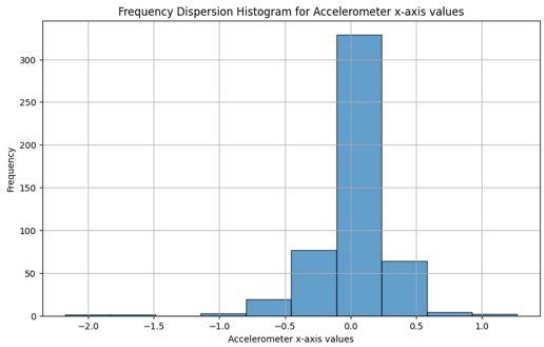


Figure 8 – Histogram of the distribution of accelerometer values along the x-axis

Most of the values are centred around zero, with some deviations in either direction. This indicates that the bulk of the data is concentrated in the central part, but there are some outliers. Figure 9a shows a histogram with outlier thresholds determined using the standard deviation (STD). The red vertical lines show the limits at which values are considered outliers. Most values are within these thresholds, but there are some values that fall outside the limits, indicating the presence of outliers. Figure 9b shows a histogram with outlier thresholds determined using the interquartile range (IQR). The red vertical lines also show the outlier limits. As in the previous case, most values are within these thresholds, but there are some outliers. Figure 9c shows a histogram of log-transformed data with outlier thresholds determined using the STD. It helped reduce the impact of large values, but there are still some outliers.

Figure 10 shows a histogram with outliers determined using the DBSCAN algorithm with parameters eps=0.01, min_samples=10. The percentage of outliers for each method is: STD- 10.80%, IQR- 12.00%, Log-STD- 8.60%, DBSCAN: 26.00%. The STD and IQR methods detect approximately the same number of outliers, indicating their similarity in determining outlier thresholds.

Logarithmic standard deviation (Log-STD) reduces the number of outliers, which can be useful for data with large deviations. The DBSCAN method detects the largest number of outliers, which may indicate its sensitivity to anomalies in the data. Therefore, for further analysis, it is recommended to use a combination of outlier detection methods to obtain more accurate results. In general, logarithmic transformation can be useful for reducing the impact of large values, but it should be noted that it can change the structure of the data. Using DBSCAN can be useful for detecting more anomalies, but caution should be exercised with its sensitivity.

DOI 10.15588/1607-3274-2025-3-10

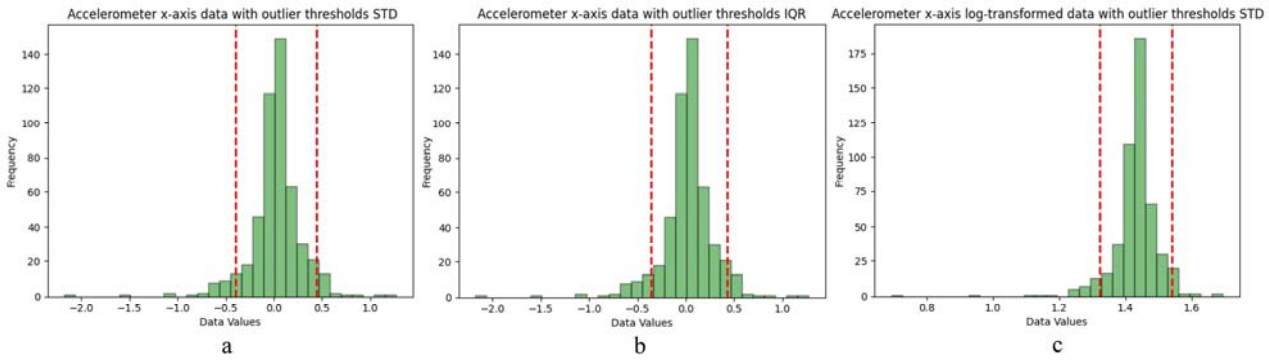


Figure 9 – Histogram with outlier thresholds: a – using standard deviation; b – using interquartile range; c – using log-transformed data with outlier thresholds determined using standard deviation

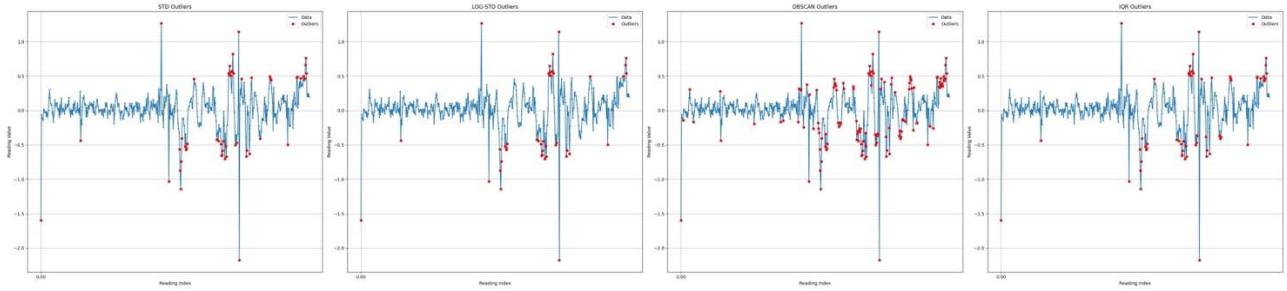


Figure 10 – Outlier detection using STD, Log-STD, DBSCAN, IQR methods

Detected outliers in sensor data in the control system can lead to incorrect analysis of industrial personnel movements, incorrect behaviour of automated devices, or a decrease in the overall efficiency of the production process. Assignment, interpolation, filtering, and smoothing methods are used to minimize the impact of noise on partially lost and distorted data. The filter helps to smooth out noise and eliminate gaps. Smoothing methods are effectively used to restore distorted or partially lost data based on adjacent values. Fig. 11 shows the results of using such filters as moving average, weighted moving average, exponential smoothing, local regression, and the Savitsky-Goley filter.

The results of calculating the deviations of all methods are presented in Table 1. According to the results, it is advisable to smooth the noise by local regression. Since it is used to smooth the data by constructing a local polynomial regression with small intervals between the data. Therefore, it effectively processes nonlinear data by adjusting the degree of the polynomial in each interval, locally adapting it to the shape of the trend in each interval. But this requires a sufficient amount of data in each interval.

In this work, the following parameters of the CWT were chosen: the Morlet mother wavelet, the value of the parameter a from 0 to 256 and the value of the parameter b from 0 to 300. This choice was based on studies [3, 21], where these parameters were determined to be the best for HAR problems based on wearable sensors when used in combination with the DenceNet121 model.

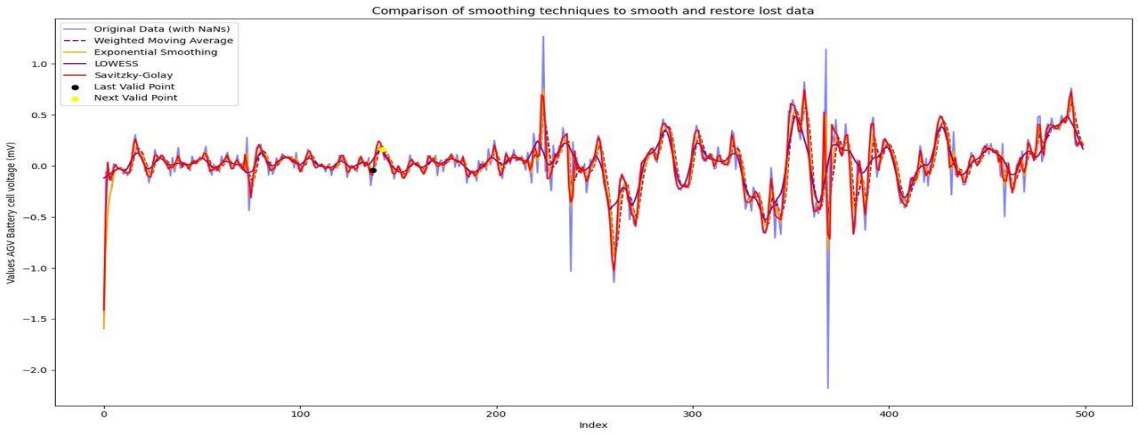


Figure 11 – Comparison of smoothing methods using filters

DOI 10.15588/1607-3274-2025-3-10

Table 1 – Deviations of smoothing filters

Deviation	**Orig. Data**	**MA**	**WMA**	**ES**	**LOWESS**	**SG**
X, axe	0.2824	0.2021	0.2086	0.2291	0.1795	0.2477
Y, axe	0.5635	0.4323	0.4472	0.4618	0.4284	0.5160
Z, axe	1.0584	0.5361	0.5248	1.1328	0.4430	0.9993
X, hyr	0.2244	0.2040	0.2073	0.2074	0.2143	0.2225
Y, hyr	0.1189	0.1076	0.1103	0.1095	0.1130	0.1182
Z, hyr	0.1962	0.1929	0.1934	0.1897	0.1913	0.1964
Dispersion	**Orig. Data**	**MA**	**WMA**	**ES**	**LOWESS**	**SG**
X, axe	0.0798	0.0408	0.0435	0.0525	0.0322	0.0613
Y, axe	0.3175	0.1869	0.2000	0.2133	0.1835	0.2662
Z, axe	1.1203	0.2874	0.2754	1.2832	0.1962	0.9987
X, hyr	0.0504	0.0416	0.0430	0.0430	0.0459	0.0495
Y, hyr	0.0141	0.0116	0.0122	0.0120	0.0128	0.0140
Z, hyr	0.0385	0.0372	0.0374	0.0360	0.0366	0.0386

This study proposes a six-step dataset preprocessing pipeline, shown in Fig. 12. It receives a set of data frames collected during experimental sessions as input, and produces datasets with fixed-size continuous sequences as output. In the first step, data frames recorded outside the experimental sessions are deleted based on their timestamps to eliminate possible outliers. In the second step, gaps in the data frame sequences are identified and highlighted using the time delta criterion. To do this, the timestamp of the end of one data frame is compared with the timestamp of the start of the next frame. If the interval exceeds 500 milliseconds, this indicates a possible pause in the experimental session or a hardware failure. In this case, this marks the end of one continuous sequence and the beginning of a new one. In the third step, the continuous data frame sequences are reduced to a fixed size of 20 frames (equivalent to 60 seconds). This size is chosen based on the fact that the initial activities of both units are the same. Hence, it is expected that a minute will be enough for the subject to perform some basic activities and the metaclassifier and predictor will have enough information to distinguish them. In the fourth stage, 50% overlap between fixed sequences is performed to expand the dataset.

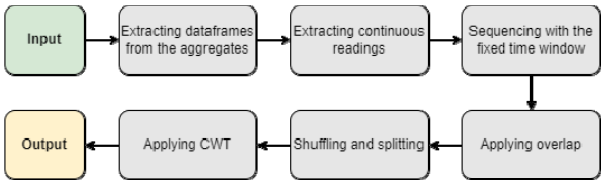


Figure 12 – Dataset preprocessing pipeline

In the fifth stage, shuffling is performed, after which the dataset is divided into subsets for training, testing, and validation. In this study, two strategies are used to divide the dataset, which are illustrated in Fig. 13. In the first partitioning strategy, 40% of the dataset is used to train the base classifier and validate the meta-classifier, the other 40% is used to train the meta-classifier and validate the base classifier, and the last 20% is used for testing. This strategy provides unique data for training the models at each level, which is the "ideal" scenario, but potentially provides insufficient data for training the meta-classifier because it does not use TL. The second strategy allocates 40% to train both classifiers, another 40% to validate the base classifier and further train the meta-classifier, and the remaining 20% for testing. The second strategy provides more data for the meta-classifier, but may result in it not capturing errors from the base classifier on new data. Finally, in the sixth step, the CWT is applied to each of the six channels of the data frames using the Morlet mother wavelet and scaling parameter values from 0 to 256.

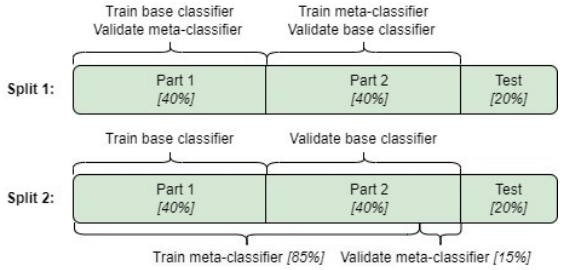


Figure 13 – The dataset partitioning strategies

In this work, the model proposed in [21, 39] was used as the baseline classifier. It is based on the DenseNet121 architecture and is specifically designed for HAR tasks. This model is pre-trained on the KU-HAR dataset [40], and CWT was used to improve performance. The proposed model achieved an F1 score of 97.52% on the KU-HAR dataset, which outperformed state-of-the-art works and demonstrated improved performance on small datasets when using layer freezing.

The original KU-HAR dataset contains 20,750 non-overlapping samples with three-axis accelerometer and three-axis gyroscope signals collected using a smartphone for 18 different activity classes. Fig. 14 illustrates the methodology used to apply knowledge transfer from the KU-HAR dataset to the collected dataset.

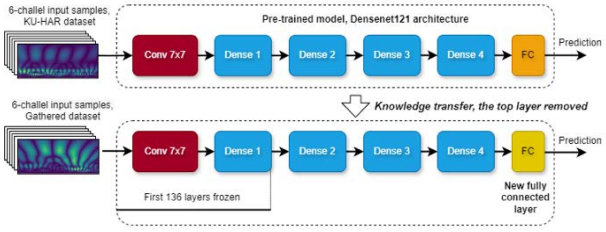


Figure 14 – Knowledge transfer approach used for the base classifier

To fit the DenseNet121 model pre-trained on the KU-HAR dataset, several manipulations are made. First, the top fully connected layer of the pre-trained model is removed and replaced with a new one initialized using the Xavier scheme. The new layer contains four neurons, which corresponds to the number of activity classes in the collected dataset. According to the study [39], freezing the layers of the pre-trained DenseNet121 can improve the performance of the model on small HAR datasets, with the optimal number being 136 layers. Accordingly,

DOI 10.15588/1607-3274-2025-3-10

the same configuration is used in this work. Additionally, appropriate class weights are used when training the base classifier to mitigate the problem of imbalance in the dataset.

The hyperparameters for training the base classifier were chosen experimentally and include the Adam optimizer, 100 training epochs, and a batch size of 32. Callbacks such as “Model checkpoint”, “Early stop”, and “Reduce training intensity at plateau” were used during training. The model was trained 10 times, and the results of the best performing instance are presented in this study.

In this study, LSTM, BiLSTM, GRU, BiGRU, and CNN architectures were used as meta-classifiers. These models were chosen because of their ability to capture temporal dependencies in fixed-size sequential data, which is important in our case. Regarding the CNN-based model, the architecture of the meta-classifier used is illustrated in Fig. 15. The input to the meta-classifier is a matrix (20×4) representing a sequence of 20 classification results from the top-level neurons of the base classifier. The architecture of the meta-classifier based on the CNN includes two convolutional blocks with pooling and batch normalization layers, as well as two fully connected blocks. The Leaky ReLU activation function is used, which prevents the problem of “dying neurons”. Dropout layers were enabled during training to improve generalization.

For the LSTM, BiLSTM, GRU and BiGRU models, experiments were conducted with both single-layer and multi-layer configurations. Each layer consists of 64 neurons, and the models include a fully connected layer with two neurons and a softmax activation function. In multilayer configurations, two consecutive layers were used, each containing the same number of neurons.

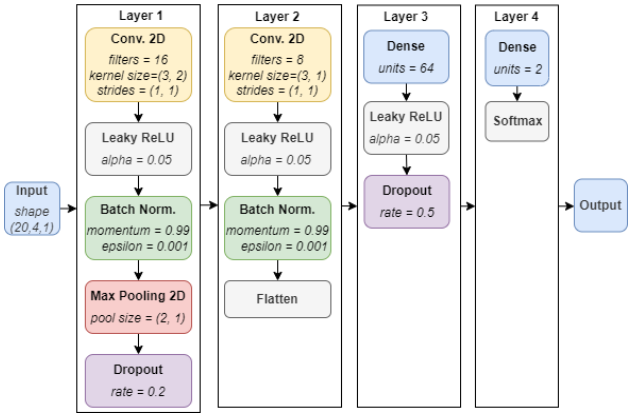


Figure 15 – Architecture of a meta-classifier based on CNN

The hyperparameters used in training the meta-classifiers include a simple gradient descent optimizer, 500 training epochs, and a batch size of 16. As with the base classifier, the same training callbacks were enabled and class weights were set. Each model configuration was trained 100 times, and the results of the best models are reported in this study.

5 RESULTS

Table 2 presents the performance metrics of the base classifier on the test subset [41]. The model generalization result from the base classifier training data is high. Furthermore, the validation accuracy and loss rates show minimal changes with increasing epochs, indicating that the model converges relatively early due to the pre-training. The results demonstrate extremely high accuracy and significant F1 scores. Given that the F1 score metric is insensitive to dataset imbalance, this alignment suggests that the model generalizes well and maintains unbias across activities.

Table 2 – Classification results of the base classifier on a subset of tests

Accuracy	Precision	Recall	AUC	F1-score
90.90%	91.33%	90.69%	97.26%	91.01%

After training and evaluating the base classifier, the dataset labels for all samples were updated based on the output from the top fully connected layer neurons of the trained base classifier.

The performance metrics of the metaclassifiers trained using the first dataset partitioning strategy [41] are shown in Table 3. The precision, recall, and F1 scores were calculated using a “weighted” approach. In this method, the metrics are calculated for each class and then the average is weighted by the support (the number of true instances for each class), which is a valid approach in the case of class imbalance. The results show that the CNN-based model showed a rougher performance, achieving an F1 score of 79.07%. Another model with satisfactory accuracy is the single-layer BiLSTM network, which achieved an F1 score of 73.89%. The remaining models showed comparable performance, with F1 scores of approximately 76%. The multilayer BiGRU model was the least efficient with an F1 score of 73.89%.

Table 3 – Classification results of meta-classifiers on the test subset (first partitioning strategy).

Classifier	Accuracy	Precision	Recall	AUC	F1-score
CNN	79.17%	79.01%	79.17%	84.34%	79.07%
Single-layer LSTM	75.00%	76.11%	75.00%	78.03%	75.32%
Multi-layer LSTM	76.39%	77.15%	76.39%	77.41%	76.63%
Single-layer BiLSTM	77.78%	78.84%	77.78%	79.13%	78.07%
Multi-layer BiLSTM	75.00%	75.49%	75.00%	79.91%	75.19%
Single-layer GRU	76.39%	76.21%	76.39%	84.51%	76.28%
Multi-layer GRU	75.00%	75.00%	75.00%	79.92%	75.00%
Single layer BiGRU	75.00%	74.45%	75,00%	78.13%	74.48%
Multi-layer BiGRU	73.61%	74.42%	73.61%	79.68%	73.89%

DOI 10.15588/1607-3274-2025-3-10

Interestingly, increasing the number of layers in the BiLSTM, GRU, and BiGRU-based architectures did not lead to an increase in performance, but rather to a decrease in it. This may be a case of overfitting with insufficient training data to utilize the additional layers in these architectures. Furthermore, the close correspondence between accuracy and F1 score in all models indicates that the model was not disproportionately affected by the more prevalent class. This was achieved by including class weights during training.

Table 4 presents the classification results of the meta-classifiers for the second partitioning strategy [41]. As can be seen, the CNN-based model also showed good performance, achieving an F1 score of 87.44%. Furthermore, when applying the second partitioning strategy, this model showed higher performance compared to the first. This indicates that the model using the second strategy benefited from the extended knowledge obtained from the shared training data, while effectively using the information from the second subset to mitigate the inaccuracies inherent in the base classifier. Interestingly, applying the second partitioning strategy resulted in a decrease in performance for the other models, indicating their inability to adapt both the extended knowledge and the errors of the base classifier. Among all the models, the single-layer LSTM network showed the lowest performance in terms of precision, granularity, recall, and F1 score. Notably, the inclusion of the second splitting strategy shows that the introduction of multiple levels in the LSTM, BiLSTM, and BiGRU models leads to an overall performance improvement. This suggests that the additional complexity of these models is an advantage when using the second splitting strategy.

Table 4 – Classification results of meta-classifiers on the test subset (second partitioning strategy).

Classifier	Accuracy	Precision	Recall	AUC	F1-score
CNN	87.50%	87.43%	87.50%	92.40%	87.44%
Single-layer LSTM	69.44%	71.42%	69.44%	76.93%	69.96%
Multi-layer LSTM	72.22%	72.75%	72.22%	78.70%	72.43%
Single-layer BiLSTM	72.22%	75.04%	72.22%	78.43%	72.75%
Multi-layer BiLSTM	73.61%	74.42%	73.61%	78.76%	73.89%
Single-layer GRU	72.22%	76.08%	72.22%	79.74%	72.77%
Multi-layer GRU	72.22%	72.22%	72.22%	80.84%	72.22%
Single-layer BiGRU	72.22%	75.04%	72.22%	77.93%	72.75%
Multi-layer BiGRU	73.61%	76.92%	73.61%	80.84%	74.13%

In light of the observed results, we propose a CNN-based metaclassifier with a second partitioning strategy as the optimal configuration among the tested ones. Considering the challenges encountered, including the similarity of the aggregates, different execution speeds, and the possibility of overlap between the main activity labels in the data frames when subjects choose actions during data collection, we evaluate the performance of the metaclassifier as satisfactory. It is important to acknowledge the potential limitations associated with the proposed approach, in particular with regard to the generation of scalograms. The computational intensity of the CWT may make it difficult to directly implement our method on wearable devices such as smartwatches or smartphones. However, this limitation can be mitigated by using edge computing and FLs, which provide decentralized data processing and model training, thereby reducing the computational constraints of individual devices.

6 DISCUSSION

This study proposes a real-time, multi-stage, complex HAR approach that is applicable to, but not limited to, intralogistics systems using AGVs. The proposed approach uses a smartwatch and techniques such as classifier stacking, CWT, and TL. In the context of this study, a distributed data collection system based on a smartwatch was developed. A dataset containing readings from five industrial personnel performing continuous sequences of actions representing typical intralogistics tasks was also collected and published.

A HAR-specific pre-trained DenseNet121 model using CWT was used as the base classifier, achieving an F1 score of 91.01% for the base activity classification. For the multi-stage activity classification task, metaclassifiers based on convolutional neural networks (CNN), long-short-term memory (LSTM), bidirectional LSTM, recurrent gating unit (GRU), and bidirectional GRU were compared. Two strategies for using the dataset were tested to optimize metaclassifier training. The most effective model using CNN and shared training data between classifiers resulted in the metaclassifier obtaining an F1 value of 87.44%. It is important to note that the temporal resolution of the data for the baseline activities is limited by the duration of the data frame, which is 3 seconds. This limitation creates a potential problem, since baseline activities with a duration shorter than this interval (e.g., a subject walk for 1 second) may overlap with the next baseline activity. Similarly, if a data frame contains data for two different baseline activities (e.g., 2 seconds of walking followed by 1 second of standing), the label corresponding to the last activity (standing in this case) will be assigned. This behavior is a potential problem that could affect the performance of classifiers and will be addressed in future research.

We hypothesize that the overall performance of the model can be improved by expanding the dataset, including data from more subjects, and including additional baseline activities such as rotation. Furthermore, the problem of overlapping activities in a time window can be

DOI 10.15588/1607-3274-2025-3-10

addressed by assigning labels based on the majority duration within a particular activity class rather than relying on the last activity. Furthermore, hybrid architectures combining CNNs with LSTMs or GRUs can yield superior results, suggesting promising directions for future research.

With appropriate modifications, the proposed approach can be integrated into an intelligent enterprise management system using CWT, improving the productivity of human-machine interaction and increasing the overall efficiency of the production line.

CONCLUSIONS

The current problem of developing an innovative approach for recognizing complex human actions in real time, focused on internal logistics systems using AGVs, is being solved.

The scientific novelty of the results is the creation of an innovative system for recognizing and predicting complex human activities in industrial intralogistics of enterprises in real time. For this purpose, a data collection system based on a smart watch was developed. This approach combines advanced data preprocessing methods and state-of-the-art machine learning models, including hybrid machine learning technologies based on DenseNet121 and CNN architectures, to achieve high accuracy of classification and prediction of activities.

The practical significance of the study in making industrial environments safer and more efficient by recognizing and predicting worker activities in real time. The system can be integrated into workplaces to streamline processes and support smarter decision-making in fast-paced conditions. By fostering smoother collaboration between humans and machines, it not only enhances productivity but also prioritizes the well-being and comfort of employees, aligning with the principles of Industry 5.0.

Prospects for further research are to focus on expanding the dataset to include more subjects, units, and major activities, and using hybrid models to improve model accuracy. Other promising directions include integrating FL technology and using the proposed architectural framework to predict worker activity.

ACKNOWLEDGEMENTS

The work is supported by the Norway Grants 2014–2023, which the National Centre operates for Research and Development under the project “Automated Guided Vehicles integrated with Collaborative Robots for Smart Industry Perspective” (Project Contract no.: NOR/POLNOR/CoBotAGV/0027/2019 00).

REFERENCES

1. Li Haobo et al. Bi-LSTM Network for Multimodal Continuous Human Activity Recognition and Fall Detection [Electronic resource], *IEEE Sensors Journal*, 2020, Vol. 20, No. 3, pp. 1191–1201 DOI: 10.1109/JSEN.2019.2946095
2. Jaafar S. T., Mohammad M. Epileptic Seizure Detection using Deep Learning Approach [Electronic resource], *UHD Journal of Science and Technology*, 2019, Vol. 3, No. 2, P. 41. DOI: 10.21928/uhdjst.v3n2y2019.pp41–50
3. Butt Fatima Sajid et al. Fall Detection from Electrocardiogram (ECG) Signals and Classification by Deep Transfer Learning [Electronic resource], *Information*, 2021, Vol. 12, No. 2, P. 63. DOI: 10.3390/info12020063
4. Tzallas A. T., Tsipouras M. G., Fotiadis D. I. Epileptic Seizure Detection in EEGs Using Time-Frequency Analysis [Electronic resource], *IEEE Transactions on Information Technology in Biomedicine*, 2009, Vol. 13, No. 5, pp. 703–710. DOI: 10.1109/titb.2009.2017939
5. Dhiman C., Vishwakarma D. K. A review of state-of-the-art techniques for abnormal human activity recognition [Electronic resource], *Engineering Applications of Artificial Intelligence*, 2019, Vol. 77, pp. 21–45. DOI: 10.1016/j.engappai.2018.08.014
6. Nadeem A. Jalal A., Kim K. Automatic human posture estimation for sport activity recognition with robust body parts detection and entropy markov model [Electronic resource], *Multimedia Tools and Applications*, 2021,Vol. 80, No. 14, pp. 21465–21498. DOI: 10.1007/s11042-021-10687-5
7. Zhuang Z., Xue Y. Sport-Related Human Activity Detection and Recognition Using a Smartwatch [Electronic resource], *Sensors*, 2019, Vol. 19, No. 22, P. 5001. DOI: 10.3390/s19225001
8. Kalpesh Jadhav et al. Human Physical Activities Based Calorie Burn Calculator Using LSTM [Electronic resource], *Intelligent Cyber Physical Systems and Internet of Things*. Cham, 2023, pp. 405–424. DOI: 10.1007/978-3-031-18497-0_31
9. Castro-García J. A. [et al.]Towards Human Stress and Activity Recognition: A Review and a First Approach Based on Low-Cost Wearables [Electronic resource], *Electronics*, 2022, Vol. 11, No. 1, P. 155. DOI: 10.3390/electronics11010155
10. Mohsen S., Elkaseer A., Scholz S. G. Industry 4.0-Oriented Deep Learning Models for Human Activity Recognition [Electronic resource], *IEEE Access*, 2021, Vol. 9, pp. 150508–150521. DOI: 10.1109/access.2021.3125733
11. Niemann Friedrich et al. Context-Aware Human Activity Recognition in Industrial Processes [Electronic resource], *Sensors*, 2021, Vol. 22, No. 1, P. 134. DOI: 10.3390/s22010134
12. Autonomous Guided Vehicles for Smart Industries – The State-of-the-Art and Research Challenges [Electronic resource] / Rafal Cupek [et al.] // Lecture Notes in Computer Science. – Cham, 2020. – P. 330–343. DOI: 10.1007/978-3-030-50426-7_25
13. Fang L., Yishui S., Wei Ch. Up and down buses activity recognition using smartphone accelerometer [Electronic resource], *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing*. China, 20–22 May 2016, [S. l.], 2016. DOI: 10.1109/itnec.2016.7560464
14. Setiaji B. R., Utama D. Q., Adiwijaya A. Smartphone Purchase Recommendation System Using the K-Nearest Neighbor (KNN) Algorithm [Electronic resource], *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 2022, Vol. 6, No. 4, P. 2180. DOI: 10.30865/mib.v6i4.4753
15. Li Kenan et al. Applying Multivariate Segmentation Methods to Human Activity Recognition From Wearable Sensors’ Data [Electronic resource], *JMIR mHealth and uHealth*, 2019, Vol. 7, No. 2, P. e11201. DOI: 10.2196/11201

DOI 10.15588/1607-3274-2025-3-10

16. Zhang W., Zhao X., Li Z. A Comprehensive Study of Smartphone-Based Indoor Activity Recognition via Xgboost [Electronic resource], IEEE Access, 2019, Vol. 7, P. 80027–80042 DOI: 10.1109/access.2019.2922974
17. Garcia-Ceja E., Galván-Tejada C. E., Brena R. Multi-view stacking for activity recognition with sound and accelerometer data [Electronic resource], *Information Fusion*, 2018, Vol. 40, pp. 45–56. DOI: 10.1016/j.inffus.2017.06.004
18. Tawosi V., Soufineyestani M., Sajedi H. Human activity recognition based on mobile phone sensor data using stacking machine learning classifiers [Electronic resource], *International Journal of Digital Signals and Smart Systems*, 2019, Vol. 3, No. 4, P. 204. DOI: 0.1504/ijdsss.2019.10027378
19. Alema Khatun M., Yousuf M. A. Human Activity Recognition Using Smartphone Sensor Based on Selective Classifiers [Electronic resource], *2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI)*. Dhaka, Bangladesh, 19–20 December 2020, [S. l.], 2020. DOI: 10.1109/sti50764.2020.9350486
20. Gaud N., Rathore M., Suman U. Hybrid Deep Learning-Based Human Activity Recognition (HAR) Using Wearable Sensors: An Edge Computing Approach [Electronic resource], *Proceedings of Data Analytics and Management*. Singapore, 2024, pp. 399–410. DOI: 10.1007/978-981-99-6544-1_30
21. Pavliuk O., Mishchuk M., Strauss C. Transfer Learning Approach for Human Activity Recognition Based on Continuous Wavelet Transform [Electronic resource], *Algorithms*, 2023, Vol. 16, No. 2, P. 77. DOI: 10.3390/a16020077
22. Khan Y. A. et al. Classification of Human Motion Activities using Mobile Phone Sensors and Deep Learning Model [Electronic resource], *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*. Coimbatore, India, 25–26 March 2022, [S. l.], 2022. DOI: 10.1109/icaccs54159.2022.9785009
23. Shah Zainudin M. N. et al. Recognizing Complex Human Activities using Hybrid Feature Selections based on an Accelerometer Sensor [Electronic resource], *International Journal of Technology*, 2017, Vol. 8, No. 5, P. 968. DOI: 10.14716/ijtech.v8i5.879
24. Ryoo M. S., Aggarwal J. K. Recognition of Composite Human Activities through Context-Free Grammar Based Representation [Electronic resource], *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 2 (CVPR'06). New York, NY, USA, [S. l.]. DOI: 10.1109/cvpr.2006.242
25. Ding G., Yao A. Temporal Action Segmentation with High-level Complex Activity Labels [Electronic resource], *IEEE Transactions on Multimedia*, 2022, pp. 1–12. DOI: 10.1109/tmm.2022.3231099
26. Li F. et al. Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors [Electronic resource], *Sensors*, 2018, Vol. 18, No. 3, P. 679. DOI: 10.3390/s18020679
27. Tammvee M., Anbarjafari G. Human activity recognition-based path planning for autonomous vehicles [Electronic resource] / Martin Tammvee, Gholamreza Anbarjafari, *Signal, Image and Video Processing*, 2020. DOI: 10.1007/s11760-020-01800-6
28. Al-Amin Md et al. Action Recognition in Manufacturing Assembly using Multimodal Sensor Fusion [Electronic resource], *Procedia Manufacturing*, 2019, Vol. 39, pp. 158–167. DOI: 10.1016/j.promfg.2020.01.288
29. Tao W., Leu M. C., Yin Z. Multi-modal recognition of worker activity for human-centered intelligent manufacturing [Electronic resource], *Engineering Applications of Artificial Intelligence*, 2020, Vol. 95, P. 103868. DOI: 10.1016/j.engappai.2020.103868
30. Suh S. et al. Worker Activity Recognition in Manufacturing Line Using Near-body Electric Field [Electronic resource], *IEEE Internet of Things Journal*, 2023, P. 1. DOI: 10.1109/jiot.2023.3330372
31. Abdel-Basset M. et al. ST-DeepHAR: Deep Learning Model for Human Activity Recognition in IoHT Applications [Electronic resource], *IEEE Internet of Things Journal*, 2020, P. 1. DOI: 10.1109/jiot.2020.3033430
32. Ziebinski Adam et al. Challenges Associated with Sensors and Data Fusion for AGV-Driven Smart Manufacturing [Electronic resource] / Adam Ziebinski [et al.] // Computational Science – ICCS 2021. – Cham, 2021. – P. 595–608. DOI: 10.1007/978-3-030-77970-2_45
33. Prots'ko I., Mishchuk M. Block-Cyclic Structuring of the Basis of Fourier Transforms Based on Cyclic Substitution [Electronic resource], *Cybernetics and Systems Analysis*, 2021, Vol. 57, No. 6, pp. 1008–1016. DOI: 10.1007/s10559-021-00426-x
34. Vuong T. H., Doan T., Takasu A. Deep Wavelet Convolutional Neural Networks for Multimodal Human Activity Recognition Using Wearable Inertial Sensors [Electronic resource], *Sensors*, 2023, Vol. 23, No. 24, P. 9721. DOI: 10.3390/s23249721
35. Jalal L., Peer A. Emotion Recognition from Physiological Signals Using Continuous Wavelet Transform and Deep Learning [Electronic resource], *HCI International 2022 – Late Breaking Papers. Multimodality in Advanced Interaction Environments*. Cham, 2022, pp. 88–99. DOI: 10.1007/978-3-031-17618-0_8
36. Tavakkoli M., Nazerfard E., Amirmazlaghani M. Wavelet-domain human activity recognition utilizing convolutional neural networks [Electronic resource], *Journal of Ambient Intelligence and Smart Environments*, 2023, pp. 1–14. DOI: 10.3233/ais-230174
37. Lu X., Ling Y., Liu S. Temporal Convolutional Network with Wavelet Transform for Fall Detection [Electronic resource], *Journal of Sensors*, 2022, Vol. 2022, pp. 1–19. DOI: 10.1155/2022/7267099
38. Izonin Ivan et al. A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain [Electronic resource], *Mathematics*, 2022, Vol. 10, No. 11, P. 1942. DOI: 10.3390/math10111942
39. Pavliuk O., Mishchuk M. A novel deep-learning model for human activity recognition based on continuous wavelet transform, *CEUR Workshop Proceedings*, 2022, Vol. 3302. pp. 236–245. https://ceur-ws.org/Vol-3302/paper14.pdf
40. Sikder N., Nahid Abdullah-Al KU-HAR: An open dataset for heterogeneous human activity recognition [Electronic resource], *Pattern Recognition Letters*, 2021, Vol. 146, pp. 46–54. DOI: 10.1016/j.patrec.2021.02.024
41. Pavliuk O., Mishchuk M. Smartwatch-Based Human Staff Activity Classification: A Use-Case Study in Internal Logistics Systems Utilizing AGVs [Electronic resource], *2024 IEEE International Conference on Big Data (BigData)*. Washington, DC, USA, 15–18 December 2024. – [S. l.], 2024, pp. 8198–8207. DOI: 10.1109/bigdata62323.2024.10825909

Received 03.03.2025.
Accepted 25.06.2025.

DOI 10.15588/1607-3274-2025-3-10

УДК 004.6; 004.8

ГІБРИДНІ ТЕХНОЛОГІЇ МАШИННОГО НАВЧАННЯ ДЛЯ ПРОГНОЗУВАННЯ КОМПЛЕКСНОЇ ДІЯЛЬНОСТІ ПРОМИСЛОВОГО ПЕРСОНАЛУ ЗА ДАНИМИ СМАРТ-ГОДИННИКІВ

Павлюк О. М. – канд. техн. наук, дослідник кафедри розподілених систем та інформаційних пристроїв Сілезької політехніки, Глівїце, Польща та докторант кафедри автоматизованих систем управління Національного університету «Львівська політехніка», Львів, Україна.

Медиковський М. О. – д-р техн. наук, професор кафедри автоматизованих систем управління Національного університету «Львівська політехніка», Львів, Україна.

Міщук М. В. – студент кафедри автоматизованих систем управління Національного університету «Львівська політехніка», Львів, Україна.

Заболотна А. О. – студентка кафедри автоматизованих систем управління Національного університету «Львівська політехніка», Львів, Україна.

Літовська О. В. – студентка кафедри автоматизованих систем управління Національного університету «Львівська політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. У сучасному промисловому виробництві значна увага приділяється системам розпізнавання та прогнозування людської активності в реальному часі. Такі технології є ключовими для переходу від Індустрії 4.0 до Індустрії 5.0, оскільки вони забезпечують покращену взаємодію між людиною і машиною, а також вищий рівень безпеки, адаптивності та ефективності виробничих процесів. Ці підходи особливо актуальні в галузі внутрішньої логістики, де співпраця з автоматизованими транспортними засобами вимагає високого рівня координації та гнучкості.

Мета. Створити технологічне рішення для оперативного виявлення та прогнозування складної поведінки людини у системах внутрішньої логістики шляхом використання сенсорних даних зі розумних годинників. Основна ціль – підвищити рівень взаємодії між працівниками та автоматизованими системами, збільшити безпеку праці й ефективність логістичних процесів.

Метод. Розроблено децентралізовану систему збору даних із використанням розумних годинників. У мобільному додатку, написаному мовою Kotlin, фіксувалися показники сенсорів під час виконання серії логістичних активностей п’ятьма працівниками. Для обробки неповних або спотворених даних застосовано алгоритми виявлення аномалій, зокрема STD, логарифмічне перетворення STD, DBSCAN та IQR, а також методи згладжування, такі як ковзне середнє, зважене ковзне середнє, експоненційне згладжування, локальна регресія й фільтр Савіцького-Голея. Оброблені дані використовувалися для навчання моделей із застосуванням таких сучасних підходів, як передавальне навчання, неперервне вейвлет-перетворення та стекінг класифікаторів.

Результати. У ролі базового класифікатора обрано попередньо натреновану глибоку модель з архітектурою DenseNet121, яка показала F1-метрику 91,01 % при розпізнаванні простих дій. Для аналізу складних активностей випробувано п’ять архітектур нейронних мереж (однашарових і багатошарових) з двома стратегіями розподілу даних. Найвищу точність – F1-метрику 87,44 % – продемонструвала згорткова нейронна мережа при використанні об’єднаного підходу до розподілу даних.

Висновки. Результати дослідження свідчать про можливість застосування запропонованої технології розпізнавання складної людської діяльності в режимі реального часу в інтралогістичних системах на основі даних з сенсорів смартгодинника яка покращить взаємодію людини та машини та підвищить ефективність промислових логістичних процесів.

КЛЮЧОВІ СЛОВА: вибірка, фрактальна розмірність, метрика якості, кластер, формування вибірок.

ЛІТЕРАТУРА

1. Bi-LSTM Network for Multimodal Continuous Human Activity Recognition and Fall Detection [Electronic resource] / Haobo Li [et al.] // IEEE Sensors Journal. – 2020. – Vol. 20, No. 3. – P. 1191–1201 DOI: 10.1109/JSEN.2019.2946095
2. Jaafar S. T. Epileptic Seizure Detection using Deep Learning Approach [Electronic resource] / Sirwan Tofiq Jaafar, Mokhtar Mohammadi // UHD Journal of Science and Technology. – 2019. – Vol. 3, No. 2. – P. 41. DOI: 10.21928/uhdjst.v3n2y2019.pp41–50
3. Fall Detection from Electrocardiogram (ECG) Signals and Classification by Deep Transfer Learning [Electronic resource] / Fatima Sajid Butt [et al.] // Information. – 2021. – Vol. 12, No. 2. – P. 63. DOI: 10.3390/info12020063
4. Tzallas A. T. Epileptic Seizure Detection in EEGs Using Time–Frequency Analysis [Electronic resource] / A. T. Tzallas, M. G. Tsipouras, D. I. Fotiadis // IEEE Transactions on Information Technology in Biomedicine. – 2009. – Vol. 13, No. 5. – P. 703–710. DOI: 10.1109/titb.2009.2017939
5. Dhiman C. A review of state-of-the-art techniques for abnormal human activity recognition [Electronic resource] / Chhavi Dhiman, Dinesh Kumar Vishwakarma // Engineering Applications of Artificial Intelligence. – 2019. – Vol. 77. – P. 21–45. DOI: 10.1016/j.engappai.2018.08.014
6. Nadeem A. Automatic human posture estimation for sport activity recognition with robust body parts detection and entropy markov model [Electronic resource] / Amir Nadeem, Ahmad Jalal, Kibum Kim // Multimedia Tools and Applications. – 2021. – Vol. 80, no. 14. – P. 21465–21498. DOI: 10.1007/s11042-021-10687-5
7. Zhuang Z. Sport-Related Human Activity Detection and Recognition Using a Smartwatch [Electronic resource] / Zhendong Zhuang, Yang Xue // Sensors. – 2019. – Vol. 19, no. 22. – P. 5001. DOI: 10.3390/s19225001
8. Human Physical Activities Based Calorie Burn Calculator Using LSTM [Electronic resource] / Jadhav Kalpesh [et al.] // Intelligent Cyber Physical Systems and Internet of Things. – Cham, 2023. – P. 405–424. DOI: 10.1007/978-3-031-18497-0_31
9. Towards Human Stress and Activity Recognition: A Review and a First Approach Based on Low-Cost Wearables [Electronic resource] / Juan Antonio Castro-García [et al.] // Electronics. – 2022. – Vol. 11, No. 1. – P. 155. DOI: 10.3390/electronics11010155
10. Mohsen S. Industry 4.0-Oriented Deep Learning Models for Human Activity Recognition [Electronic resource] / Saeed Mohsen, Ahmed Elkaseer, Steffen G. Scholz // IEEE Access. – 2021. – Vol. 9. – P. 150508–150521. DOI: 10.1109/access.2021.3125733
11. Context-Aware Human Activity Recognition in Industrial Processes [Electronic resource] / Friedrich Niemann [et al.] // Sensors. – 2021. – Vol. 22, No. 1. – P. 134. – DOI: 10.3390/s22010134
12. Autonomous Guided Vehicles for Smart Industries – The State-of-the-Art and Research Challenges [Electronic resource] / Rafal Cupek [et al.] // Lecture Notes in Computer Science. – Cham, 2020. – P. 330–343. DOI: 10.1007/978-3-030-50426-7_25
13. Fang L. Up and down buses activity recognition using smartphone accelerometer [Electronic resource] / Li Fang, Shui Yishui, Chen Wei // 2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 20–22 May 2016. – [S. l.], 2016. DOI: 10.1109/itnec.2016.7560464

DOI 10.15588/1607-3274-2025-3-10

14. Setiaji B. R. Smartphone Purchase Recommendation System Using the K-Nearest Neighbor (KNN) Algorithm [Electronic resource] / Bayu Rahmat Setiaji, Dody Qori Utama, Adiwijaya Adiwijaya // JURNAL MEDIA INFORMATIKA BUDIDARMA. – 2022. – Vol. 6, No. 4. – P. 2180. DOI: 10.30865/mib.v6i4.4753
15. Applying Multivariate Segmentation Methods to Human Activity Recognition From Wearable Sensors' Data [Electronic resource] / Kenan Li [et al.] // JMIR mHealth and uHealth. – 2019. – Vol. 7, no. 2. – P. e11201. DOI: 10.2196/11201
16. Zhang W. A Comprehensive Study of Smartphone-Based Indoor Activity Recognition via Xgboost [Electronic resource] / Wenting Zhang, Xiaohui Zhao, Zan Li // IEEE Access. – 2019. – Vol. 7. – P. 80027–80042 DOI: 10.1109/access.2019.2922974
17. Garcia-Ceja E. Multi-view stacking for activity recognition with sound and accelerometer data [Electronic resource] / Enrique Garcia-Ceja, Carlos E. Galván-Tejada, Ramon Brena // Information Fusion. – 2018. – Vol. 40. – P. 45–56. DOI: 10.1016/j.inffus.2017.06.004
18. Tawosi V. Human activity recognition based on mobile phone sensor data using stacking machine learning classifiers [Electronic resource] / Vali Tawosi, Mahsa Soufineyestani, Hedieh Sajedi // International Journal of Digital Signals and Smart Systems. – 2019. – Vol. 3, No. 4. – P. 204. DOI: 0.1504/ijdsss.2019.10027378
19. Alema Khatun M. Human Activity Recognition Using Smartphone Sensor Based on Selective Classifiers [Electronic resource] / Mst Alema Khatun, Mohammad Abu Yousuf // 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), Dhaka, Bangladesh, 19–20 December 2020. – [S. l.], 2020. DOI: 10.1109/sti50764.2020.9350486
20. Gaud N. Hybrid Deep Learning-Based Human Activity Recognition (HAR) Using Wearable Sensors: An Edge Computing Approach [Electronic resource] / Neha Gaud, Maya Rathore, Ugrasen Suman // Proceedings of Data Analytics and Management. – Singapore, 2024. – P. 399–410. DOI: 10.1007/978-981-99-6544-1_30
21. Pavliuk O. Transfer Learning Approach for Human Activity Recognition Based on Continuous Wavelet Transform [Electronic resource] / Olena Pavliuk, Myroslav Mishchuk, Christine Strauss // Algorithms. – 2023. – Vol. 16, No. 2. – P. 77. DOI: 10.3390/a16020077
22. Classification of Human Motion Activities using Mobile Phone Sensors and Deep Learning Model [Electronic resource] / Yusuf Ahmed Khan [et al.] // 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 25–26 March 2022. – [S. l.], 2022. DOI: 10.1109/icaccs54159.2022.9785009
23. Recognizing Complex Human Activities using Hybrid Feature Selections based on an Accelerometer Sensor [Electronic resource] / M. N. Shah Zainudin [et al.] // International Journal of Technology. – 2017. – Vol. 8, No. 5. – P. 968. DOI: 10.14716/ijtech.v8i5.879
24. Ryoo M. S. Recognition of Composite Human Activities through Context-Free Grammar Based Representation [Electronic resource] / M. S. Ryoo, J. K. Aggarwal // 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Volume 2 (CVPR'06), New York, NY, USA. – [S. l.]. DOI: 10.1109/cvpr.2006.242
25. Ding G. Temporal Action Segmentation with High-level Complex Activity Labels [Electronic resource] / Guodong Ding, Angela Yao // IEEE Transactions on Multimedia. – 2022. – P. 1–12. DOI: 10.1109/tmm.2022.3231099
26. Comparison of Feature Learning Methods for Human Activity Recognition Using Wearable Sensors [Electronic resource] / Frédéric Li [et al.] // Sensors. – 2018. – Vol. 18, No. 3. – P. 679. DOI: 10.3390/s18020679
27. Tammvee M. Human activity recognition-based path planning for autonomous vehicles [Electronic resource] / Martin Tammvee, Gholamreza Anbarjafari // Signal, Image and Video Processing. – 2020. DOI: 10.1007/s11760-020-01800-6
28. Action Recognition in Manufacturing Assembly using Multimodal Sensor Fusion [Electronic resource] / Md Al-Amin [et al.] // Procedia Manufacturing. – 2019. – Vol. 39. – P. 158–167. DOI: 10.1016/j.promfg.2020.01.288
29. Tao W. Multi-modal recognition of worker activity for human-centered intelligent manufacturing [Electronic resource] / Wenjin Tao, Ming C. Leu, Zhaozheng Yin // Engineering Applications of Artificial Intelligence. – 2020. – Vol. 95. – P. 103868. DOI: 10.1016/j.engappai.2020.103868
30. Worker Activity Recognition in Manufacturing Line Using Near-body Electric Field [Electronic resource] / Sungho Suh [et al.] // IEEE Internet of Things Journal. – 2023. – P. 1. DOI: 10.1109/jiot.2023.3330372
31. ST-DeepHAR: Deep Learning Model for Human Activity Recognition in IoHT Applications [Electronic resource] / Mohamed Abdel-Basset [et al.] // IEEE Internet of Things Journal. – 2020. – P. 1. DOI: 10.1109/jiot.2020.3033430
32. Challenges Associated with Sensors and Data Fusion for AGV-Driven Smart Manufacturing [Electronic resource] / Adam Ziebinski [et al.] // Computational Science – ICCS 2021. – Cham, 2021. – P. 595–608. DOI: 10.1007/978-3-030-77970-2_45
33. Prots'ko I. Block-Cyclic Structuring of the Basis of Fourier Transforms Based on Cyclic Substitution [Electronic resource] / I. Prots'ko, M. Mishchuk // Cybernetics and Systems Analysis. – 2021. – Vol. 57, No. 6. – P. 1008–1016. DOI: 10.1007/s10559-021-00426-x
34. Vuong T. H. Deep Wavelet Convolutional Neural Networks for Multimodal Human Activity Recognition Using Wearable Inertial Sensors [Electronic resource] / Thi Hong Vuong, Tung Doan, Atsuhiro Takasu // Sensors. – 2023. – Vol. 23, No. 24. – P. 9721. DOI: 10.3390/s23249721
35. Jalal L. Emotion Recognition from Physiological Signals Using Continuous Wavelet Transform and Deep Learning [Electronic resource] / Lana Jalal, Angelika Peer // HCI International 2022 – Late Breaking Papers. Multimodality in Advanced Interaction Environments. – Cham, 2022. – P. 88–99. DOI: 10.1007/978-3-031-17618-0_8
36. Tavakkoli M. Wavelet-domain human activity recognition utilizing convolutional neural networks [Electronic resource] / Mohammad Tavakkoli, Ehsan Nazerfard, Maryam Amirmazlaghani // Journal of Ambient Intelligence and Smart Environments. – 2023. – P. 1–14. DOI: 10.3233/ais-230174
37. Lu X. Temporal Convolutional Network with Wavelet Transform for Fall Detection [Electronic resource] / Xilin Lu, Yuanxiang Ling, Shuzhi Liu // Journal of Sensors. – 2022. – Vol. 2022. – P. 1–19. DOI: 10.1155/2022/7267099
38. A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain [Electronic resource] / Ivan Izonin [et al.] // Mathematics. – 2022. – Vol. 10, no. 11. – P. 1942. DOI: 10.3390/math10111942
39. Pavliuk O. A novel deep-learning model for human activity recognition based on continuous wavelet transform / Olena Pavliuk, Myroslav Mishchuk // CEUR Workshop Proceedings. – 2022. – Vol. 3302. – P. 236–245. https://ceur-ws.org/Vol-3302/paper14.pdf
40. Sikder N. KU-HAR: An open dataset for heterogeneous human activity recognition [Electronic resource] / Niloy Sikder, Abdullah-Al Nahid // Pattern Recognition Letters. – 2021. – Vol. 146. – P. 46–54. DOI: 10.1016/j.patrec.2021.02.024
41. Pavliuk O. Smartwatch-Based Human Staff Activity Classification: A Use-Case Study in Internal Logistics Systems Utilizing AGVs [Electronic resource] / Olena Pavliuk, My-roslav Mishchuk // 2024 IEEE International Conference on Big Data (BigData), Washington, DC, USA, 15–18 Decem-ber 2024. – [S. l.], 2024. – P. 8198–8207. DOI: 10.1109/bigdata62323.2024.10825909

DOI 10.15588/1607-3274-2025-3-10

UDC 681.518:004.93

HIERARCHICAL MACHINE LEARNING SYSTEM FOR FUNCTIONAL DIAGNOSIS OF EYE PATHOLOGIES BASED ON THE INFORMATION-EXTREMAL APPROACH

Shelehov I. V. – PhD, Associate Professor of the Department of Cybernetics and Informatics Department of Sumy National Agrarian University, Sumy, Ukraine; PhD, Associate Professor at the Computer Science Department of Sumy State University, Sumy, Ukraine.

Prylepa D. V. – PhD, Assistant at the Computer Sciences Department of Sumy State University, Sumy, Ukraine.

Khibovska Y. O. – Postgraduate student at the Department of Computer Science, Sumy State University, Sumy, Ukraine.

Tymchenko O. A. – Postgraduate student at the Department of Computer Science, Sumy State University, Sumy, Ukraine.

ABSTRACT

Context. The task of information-extremal machine learning for the diagnosis of eye pathologies based on the characteristic signs of diseases is considered. The object of the study is the process of hierarchical machine learning in the system for diagnosing ophthalmological diseases. The aging population and the increasing prevalence of eye diseases, such as glaucoma, optic nerve atrophy, retinal detachment, and diabetic retinopathy, necessitate effective methods for early diagnosis to prevent vision loss. Traditional diagnostic methods largely rely on the experience of the physician, which can lead to errors. The use of artificial intelligence (AI) and machine learning (ML) can significantly improve the accuracy and speed of diagnosis, making this topic highly relevant.

Objective. To enhance the functional efficiency of a computerized system for diagnosing eye pathologies based on image data.

Method. A method of information-extremal hierarchical machine learning for a system of eye pathology diagnosis based on the characteristic signs of diseases is proposed. The method is based on a functional approach to modeling cognitive processes of natural intelligence, ensuring the adaptability of the diagnostic system under any initial conditions for the formation of pathology images and allowing flexible retraining of the system when the recognition class alphabet expands. The foundation of the method is the principle of maximizing the criterion of functional efficiency based on a modified Kullback information measure, which is a functional of the diagnostic rule precision characteristics. The learning process is considered as an iterative procedure for optimizing the parameters of the diagnostic system's operation according to this information criterion. Based on the proposed categorical functional model, an information-extremal machine learning algorithm with a hierarchical data structure in the form of a binary recursive tree is developed. This data structure enables the division of a large number of recognition classes into pairs of nearest neighbors, for which the machine learning parameters are optimized using a linear algorithm of the necessary depth.

Results. An intelligent technology for diagnosing eye pathologies has been developed, which includes a comprehensive set of information, algorithmic, and software components. A comparative analysis of the effectiveness of different methods for organizing decision rules during system training has been conducted. It was found that the use of recursive hierarchical classifier structures allows achieving higher diagnostic accuracy compared to binary classifiers.

Conclusions. The developed intelligent computer-based diagnostic system for eye pathologies demonstrates high efficiency and accuracy. The implementation of such a system in medical practice could significantly improve the quality of eye disease diagnostics, reduce the workload on physicians, and minimize the risk of misdiagnosis. Further research could focus on refining algorithms and expanding their application to other types of medical images.

KEYWORDS: computer diagnosis of eye pathologies, artificial intelligence, machine learning, image processing, pattern recognition, information-extremal technology, hierarchical classifier structure.

ABBREVIATIONS

IEI-technology – information-extremal intelligent technology;

SCT – system of control tolerances;

AI – artificial intelligence;

DL – deep learning;

ML – machine learning;

ODS – ophthalmological diagnostic system.

NOMENCLATURE

M is a power of the alphabet of diagnostic classes;

m is a number of the current classes of ophthalmic diagnostics;

N is a power of the dictionary of diagnostic features;

i is a number of the diagnostic feature;

n is a volume of the training matrix of diagnostic classes;

j is a number of the structured vector of diagnostic feature values in the training matrix;

H is a set of strata in the de-recursive tree;

h is a number of the stratum in the de-recursive tree;

S is a set of strata in the de-recursive tree;

s is a number of the stratum in the de-recursive tree;

$x_{0,i}^{(h,s)}$ is an averaged feature vector of the base diagnostic class $x_0^{(h,s)}$ at the h-th level of the s-th stratum of the de-recursive tree;

$x_{1,i}^{(h,s)}$ is an averaged feature vector of the class nearest to the base diagnostic class $x_1^{(h,s)}$ at the h-th level of the s-th stratum of the de-recursive tree;

DOI 10.15588/1607-3274-2025-3-11

$d_0^{(h,s)}$ is a radius of the hyperspherical container of the base diagnostic class $x_0^{(h,s)}$;

$d_1^{(h,s)}$ is a radius of the hyperspherical container of the class nearest to the base diagnostic class $x_1^{(h,s)}$;

K is a set of machine learning steps;

$\delta_{K,i}^{(h,s)}$ is a parameter equal to half of the control tolerance field for the features of the diagnostic classes at the h-th level of the s-th stratum;

G_E is a working (permissible) domain for the definition of the information optimization criterion function;

$G_{d_0^{(h,s)}}$ is a permissible domain of the radius values for the container of the base diagnostic class $x_0^{(h,s)}$ at the h-th level of the s-th stratum;

$G_{d_1^{(h,s)}}$ is a permissible domain of the radius values for the container of the class nearest to the base diagnostic $x_1^{(h,s)}$ class at the h-th level of the s-th stratum;

$E_0^{(h,s)}$ is an information criterion for optimizing the machine learning parameters for the base diagnostic class $x_0^{(h,s)}$;

$E_1^{(h,s)}$ is an information criterion for optimizing the machine learning parameters for the class nearest to the base diagnostic class $x_1^{(h,s)}$;

G is a set of input factors that influence the ODS;

T is a set of moments in time for reading the information;

Ω is a feature space for diagnostics;

Z is a set of technical states of the diagnostic object;

$Y^{(h,s)}$ is a input training matrix of the diagnostic classes at the h-th level of the s-th stratum of the de-recursive tree;

$X^{(h,s)}$ is a binary training matrix of the diagnostic classes at the h-th level of the s-th stratum of the de-recursive tree;

f_0 is an operator for forming the de-recursive binary trees H;

f_1 is an operator for forming the training matrix $Y^{(h,s)}$;

f_2 is an operator for forming the binary training matrices $X^{(h,s)}$;

$d_0^{*(h,s)}$ is an optimal radius of the hyperspherical container of the base diagnostic class $x_0^{(h,s)}$;

$d_1^{*(h,s)}$ is an optimal radius of the hyperspherical container of the class nearest to the base diagnostic class $x_1^{(h,s)}$;

$X_m^{o(h,s)}$ is a container of the diagnostic class at the h-th level of the s-th stratum of the de-recursive tree;

$I_{0,m}^{(h,s)}$ is a set of statistical hypotheses for the decision rule of the base diagnostic class $x_m^{(h,s)}$;

$I_{1,m}^{(h,s)}$ is a set of alternative hypotheses for the decision rule of the class nearest to the base diagnostic class $x_m^{(h,s)}$;

ψ is an operator for testing the main statistical hypothesis about the assignment of the vector $x_m^{(h,s)}$ to the diagnostic class $X_m^{o(h,s)}$;

γ is an operator for forming the set of accuracy characteristics;

ζ is an operator for forming the set of reference vector values and optimal radii;

$D_1^{h,s}$ is an extreme value of the first reliability at the h-th level of the s-th stratum of the de-recursive tree;

$D_2^{h,s}$ is an extreme value of the second reliability at the h-th level of the s-th stratum of the de-recursive tree;

$\alpha^{h,s}$ is a first-type error, calculated at the h-th level of the s-th stratum of the de-recursive tree;

$\beta^{h,s}$ is a second-type error, calculated at the h-th level of the s-th stratum of the de-recursive tree;

φ_1 is an operator for forming the value of the optimization criterion $E_1^{(h,s)}$ and $E_0^{(h,s)}$;

φ_2 is an operator for forming the value of the total optimization criterion $\overline{E}^{(h,s)}$;

u is an operator that regulates the machine learning process;

r is an operator for partitioning Ω the diagnostic feature space into classes;

f_H is an operator that regulates the process of forming and evaluating the functional efficiency of the strata in the de-recursive tree;

$\rho_{m,i}$ is a selection level;

$\otimes$ is a symbol for the repetition operation;

d is a parameter that characterizes the radius values of the diagnostic class containers in code units;

10^{-p} is a sufficiently small number introduced to avoid division by zero (in practice, it is taken as $p = 2$);

$x^{(j)}$ is a structured vector of diagnostic feature values, formed during the stage of diagnostic decision-making;

μ_m is a membership function of vector $x^{(j)}$ for the diagnostic class $X_m^{o,(h,s)}$.

DOI 10.15588/1607-3274-2025-3-11

INTRODUCTION

Modern medicine faces numerous challenges, among which the tasks of diagnosing and treating eye diseases stand out [1]. With the growing number of patients with ophthalmic problems such as optic nerve atrophy, glaucoma, retinal detachment, and diabetic retinopathy, there is a need for the development of new, more effective diagnostic methods [2, 3]. Traditional diagnostic methods, based on visual assessment of fundus images and other examinations, largely depend on the subjective evaluation of the physician, which can lead to errors and inaccuracies.

Currently, the widespread use of AI and ML technologies in computerized diagnostic systems allows for a significant acceleration of medical image processing and an increase in the accuracy of pathology detection, providing effective support for medical professionals in clinical decision-making. Priority is given to artificial neural network technology, as it is believed to be capable of effectively processing large volumes of medical data, ensuring high accuracy in detecting patterns and anomalies in medical images [4–6]. An alternative to the use of artificial neural networks is information-extreme intellectual technologies, the effectiveness of which has been proven in solving many practical problems across various tasks [7–9].

The object of the research is the process of hierarchical machine learning in the system of functional diagnosis of ophthalmic diseases.

The subject of the research is the methods for building and optimizing the system of information-extreme hierarchical machine learning for diagnosing eye pathologies based on images.

The purpose of the work is to develop an IEI technology for computerized diagnosis of eye pathologies. Such a technology should include modern image processing and pattern recognition methods, as well as utilize machine learning algorithms to improve diagnostic accuracy.

The article discusses the main stages of developing and implementing this technology, starting from the analysis of existing methods and ending with the creation of software and its testing on real medical data. It is expected that the results of this research will contribute to improving the quality of medical care and serve as a foundation for further research in the field of medical diagnostics.

1 PROBLEM STATEMENT

Let us consider the formalized formulation of the information synthesis task for a ODS capable of learning based on images of human eye pathologies.

Let the alphabet $\{X_m^o \mid m=\overline{1,M}\}$, of diagnostic classes be given, which is formed according to the main diseases of the human visual organs. The peculiarities of ophthalmic diagnostics, which include visual examination and fundus scanning, allow the use of its images to form a training matrix $\| y_{m,i}^{(j)} \mid i=\overline{1,N};\ j=\overline{1,n} \|$. In this case $\left\{ y_{m,i}^{(j)} \mid i=\overline{1,N} \right\}$, a row of the matrix represents the *j*-th realization, and a column $\left\{ y_{m,i}^{(j)} \mid j=\overline{1,n} \right\}$ represents the training sample of values for the *i*-th diagnostic feature.

According to the concept of IEI-technology, the input training matrix is transformed during deep machine learning into a set of diagnostic decision rules, the parameters of which are optimized (in the informational sense) by maximizing the functional efficiency of the ODS. Let the depth of machine learning be two levels. At the first level, the optimal phenotypic parameters of the ODS are determined, namely the geometric parameters of the hyperspherical containers of diagnostic classes, and at the second level, the genotypic parameters, namely the system of control tolerances for diagnostic features, are determined. The structured vector of parameters influencing the functional efficiency of deep machine learning in the ODS is as follows:

$$g^{(h,s)} = \left\langle \left\{ x_{0,i}^{(h,s)} \right\}, d_0^{(h,s)}, \left\{ x_{1,i}^{(h,s)} \right\}, d_1^{(h,s)}, \left\{ \delta_{K,i}^{(h,s)} \right\} \right\rangle \qquad (1)$$

with the corresponding constraints [7] to the strata of the recursive hierarchical structure of pairs of nearest neighboring diagnostic classes.

During the machine learning process in the ODS, it is necessary to:

1) optimize the parameters of the vector (1) for each stratum of the hierarchical structure of diagnostic classes:

$$E^{(h,s)} = \frac{\max\limits_{G_E \cap G_{d_0^{(h,s)}}} E_0^{(h,s)}\left(d_0^{(h,s)}\right) + \max\limits_{G_E \cap G_{d_1^{(h,s)}}} E_1^{(h,s)}\left(d_1^{(h,s)}\right)}{2}. \qquad (2)$$

2) Based on the optimal geometric parameters of the container classes obtained through machine learning, we will construct decision rules for each stratum of the hierarchical structure, ensuring a high probability of making correct diagnostic decisions.

3) At the examination stage, it is necessary to make a diagnostic decision about the assignment of the structured vector of diagnostic feature values to one of the classes in the formed alphabet of the corresponding final stratum.

Thus, the task of information-extreme synthesis of the learnable ODS is to optimize the parameters of its machine learning by approximating the global maximum of the information criterion (2) to its maximum limiting value.

2 REVIEW OF THE LITERATURE

The implementation of AI in the field of medical diagnostics is one of the key trends in the development of modern science. Ophthalmology, as one of the branches of medicine, actively utilizes the potential of AI to improve the accuracy and effectiveness of diagnosing

DOI 10.15588/1607-3274-2025-3-11

ophthalmological diseases, as evidenced by numerous scientific studies.

In connection with the global demographic trend of an aging population, a significant increase in the number of patients suffering from ophthalmological diseases is predicted [1–3]. Timely diagnosis and appropriate treatment are crucial for preventing the progression of ophthalmological diseases and vision loss. Traditional diagnostic methods largely depend on the professional experience of doctors, which can lead to a high frequency of misdiagnoses and loss of medical data. The deep synergy between ophthalmology and artificial intelligence contributes to the creation of innovative methods for processing and analyzing medical data, providing ophthalmologists with powerful tools to enhance the accuracy and speed of diagnostics [4–6].

AI, first proposed by John McCarthy in 1956, became a general term for technologies that mimic intelligent behavior. However, the real breakthrough in the application of AI occurred only recently, thanks to the emergence of new algorithms, specialized hardware, and large volumes of data. ML, as a subfield of AI, encompasses methods for automatically detecting patterns in data and using them to predict future events under conditions of uncertainty [10, 11].

DL, which emerged in the early 21st century, became a catalyst for revolutionary changes in the field of AI. This technology forms the foundation of many modern systems, particularly in tasks such as image recognition, automatic translation, and intelligent control. In healthcare, DL is applied to histopathological analysis, skin cancer classification, cardiovascular disease risk prediction, and lung cancer detection [12–14].

In ophthalmology, AI is actively used for the diagnosis of retinal diseases, glaucoma, diabetic retinopathy, and other pathologies. For example, probabilistic neural networks have been used for analyzing the blood vessels of the retina, while a three-layer artificial neural network and support vector machine methods have been applied for classifying retinal diseases based on fundus images [6, 15, 16].

Intelligent diagnostic systems provide high accuracy, reduce computational costs, and shorten working time, making them indispensable in medical practice. For example, in [17], it was shown that machine learning algorithms can be used for accurately determining the condition of the retina and predicting the development of diseases.

The main advantages of using AI in ophthalmology are the ability to process large volumes of data, automate the diagnostic process, and achieve high accuracy in results. Furthermore, images obtained through slit-lamp examination, visual acuity testing, fundus images, ultrasound imaging, and optical coherence tomography can be stored for further analysis and monitoring [16–18].

Thus, the development of artificial intelligence technologies opens up new opportunities for ophthalmology, providing more accurate and timely diagnosis of eye diseases, which contributes to improving the quality of life of patients and reducing the risks of vision loss.

3 MATERIALS AND METHODS

The method of information synthesis for the ODS will be considered within the framework of the IEI-technology, which is based on maximizing the functional efficiency of the system during deep machine learning. This approach enables the use of both linear and hierarchical structures of diagnostic classes. An essential task within this process is the automation of forming optimal hierarchical structures in the form of a binary de-recursive information tree. Unlike recursive structures, in this case, the attribute from the node of the upper tier is passed to the node of its corresponding stratum in the lower tier. In our approach, the attributes of the nodes are represented by training matrices corresponding to the respective diagnostic classes. The de-recursive hierarchical structure is partitioned into strata, each consisting of two classes with the closest Hamming feature distance in the binary space. This structure enables the application of a linear algorithm of information-extreme machine learning with the required depth for classification. In contrast to neural-like structures, the depth of information-extreme machine learning is determined not by the number of hidden layers, but by the number of machine learning parameters optimized according to an information criterion [7, 9, 19].

The input mathematical description of the ODS is considered as a set-theoretic structure

$$I_B =< G, T, \Omega, Z, H, Y^{(h,s)}, X^{(h,s)}, f_0, f_1, f_2 >,$$

and the functional categorical model of information-extreme machine learning based on the hierarchical data structure is represented as a diagram of mappings between these sets by machine learning operators.

The categorical model of information-extreme machine learning for the ODS in stratum *s* at level *h* of the de-recursive hierarchical data structure is shown in Fig. 1 [8].

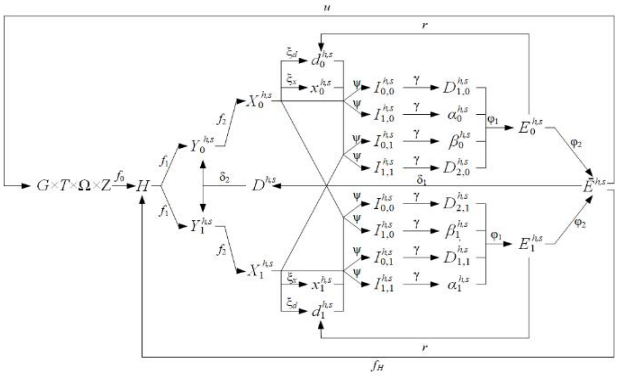


Figure 1 – Categorical Model of Machine Learning for the ODS

DOI 10.15588/1607-3274-2025-3-11

The operator f_0, shown in Fig. 1, originating from the information source defined by the Cartesian product of sets $G \times T \times \Omega \times Z$, generates the de-recursive binary tree H, while the operator f_1 forms the input training matrices for the corresponding strata $Y^{(h,s)}$. The operator f_2, by comparing the values of diagnostic features with their specified tolerance limits, forms the corresponding set $X^{(h,s)}$ of binary working matrices, which, during the machine learning process, are adapted through permissible transformations to achieve the maximum overall probability of making correct classification decisions. The term-set E, whose elements are the values of the information criterion computed at each step of the learning process, is common to all optimization loops of the learning parameters, in accordance with the principle of complete composition. The operator ζ computes the set of reference vector values $x_0^{(h,s)}$, $x_1^{(h,s)}$ and the optimal radii $d_0^{*(h,s)}$, $d_1^{*(h,s)}$ of the diagnostic class containers $X_m^{o(h,s)}$. The operator $\psi : X_0^{(j)} \in X_0^{(h,s)} \to I_{0,0}^{(h,s)}$, tests the main statistical hypothesis $I_{0,0}^{(h,s)}$ (or $I_{0,1}^{(h,s)}$) and the alternative hypothesis $I_{1,0}^{(h,s)}$ (or $I_{1,1}^{(h,s)}$). The operator γ forms the set of accuracy characteristics $D_1^{h,s}, \alpha^{h,s}, \beta^{h,s}, D_2^{h,s}$, while the operator φ_1 calculates the values of the optimization criterion $E_1^{(h,s)}$ and $E_0^{(h,s)}$ for the neighboring diagnostic states of the human eye. At the same time, the operator φ_2 computes the overall value of the optimization criterion $\overline{E}^{(h,s)}$. The operator $u : \overline{E}^{(h,s)} \to G \times T \times \Omega \times Z$ regulates the machine learning process during the ophthalmological diagnosis of the human eye.

Thus, the proposed categorical model of information-extreme machine learning enables the automatic formation of a de-recursive hierarchical structure of diagnostic classes in real-time.

The implementation of information-extreme machine learning using a hierarchical data structure represented by a binary de-recursive tree is carried out according to the following scheme:

1. Formation of the tolerance field for diagnostic features of stratum s at hierarchical level h:

1.1 Determination of the averaged vector of structured diagnostic features for class $X_0^{o(h,s)}$

$$y_{0,i}^{(h,s)} = \frac{1}{n}\sum_{j=1}^{n} Y_{0,i}^{j(h,s)};$$

1.2 Determination of the upper control tolerance

$$A_{BK,i}^{(h,s)} = y_{0,i}^{(h,s)} + \delta_{K,i}^{(h,s)};$$

1.3 Determination of the lower control tolerance

$$A_{HK,i}^{(h,s)} = y_{0,i}^{(h,s)} - \delta_{K,i}^{(h,s)};$$

2. Formation of the binary training matrix $X^{(h,s)}$

$$X_{m,i}^{j(h,s)} = \begin{cases} 1, \text{ if } A_{HKm,i}^{(h,s)} < Y_{m,i}^{j(h,s)} < A_{BKm,i}^{(h,s)}; \\ 0, \text{ if otherwise;} \end{cases}$$

3. A set $\{x_m^{(h,s)}\}$ of binary averaged vectors of diagnostic features is formed according to the following rule

$$x_{m,i}^{(h,s)} = \begin{cases} 1, \text{ if } \frac{1}{n}\sum_{j=1}^{n} X_{m,i}^{j(h,s)} \geq \rho_{m,i}; \\ 0, \text{ if otherwise;} \end{cases}$$

4. Ranking of $\{x_m^{(h,s)}\}$ by code distance from the zero binary vector and determination of the composition of the two branches of stratum s at level h of the binary de-recursive tree by dividing the set of classes into two approximately equal and non-overlapping groups.

5. Optimization (in the informational sense) of the values of phenotypic and genotypic learning parameters, and derivation of decision rules for the two classes with the smallest code distance $\{x_m^{(h,s)}\}$, belonging to different branches of stratum s at level h.

6. The branching continues until the formation of so-called final strata, the branches of which contain only one diagnostic class each.

Thus, during the formation of strata in the binary de-recursive tree, the optimal set of phenotypic and genotypic learning parameters is obtained for all pairs of the nearest neighboring classes, which is a necessary condition for the pairwise partitioning of the diagnostic feature space by means of information-extreme machine learning using a linear algorithm [20, 21].

According to the categorical model (Fig. 1), the information-extreme machine learning algorithm of the ODS based on a hierarchical data structure is presented as a procedure regulated by operator f_H for searching the global maximum of the criterion (2) averaged over the alphabet $\{X^{o(h,s)}\}$ of the corresponding diagnostic classes of the stratum:

$$\delta_K^{*(h,s)} = \arg\max_{G_\delta^{(h,s)}} \left\{ \max_{G_E \cap G_d} \overline{E}^{(h,s)}(d) \right\}. \qquad (3)$$

Thus, unlike the linear algorithm, in which the optimal value of the parameter δ is determined for the entire

DOI 10.15588/1607-3274-2025-3-11

diagnostic class alphabet, in information-extreme machine learning based on a hierarchical de-recursive data structure, the parameter $\left\{\delta_{K,i}^{(h,s)}\right\}$ is determined separately for each stratum.

The internal loop of procedure (3) implements the basic algorithm, whose functions include calculating the criterion (2) at each step of the machine learning process, searching for its global maximum, and determining the optimal geometric parameters of the diagnostic class containers [21, 22].

$$d_m^{*(h,s)} = \arg \max_{G_E \cap G_d} \overline{E}^{(h,s)}\left(d_m^{(h,s)}\right), m = \overline{1, M^{(h,s)}}. \quad (4)$$

In the outer loop of procedure (3), the operator for adjusting the parameter $\left\{\delta_{K,i}^{(h,s)}\right\}$ of the control tolerance field is executed until the value of the information criterion for optimizing the machine learning parameters reaches its maximum [9, 19].

As the optimization criterion for machine learning parameters of the ODS within each stratum of the de-recursive hierarchical data structure, a modified Kullback information measure was used [7, 23], which, for two equally probable alternative hypotheses, takes the following form:

$$E_m^{(k,h,s)} = \frac{1}{2}\{2 - [\alpha_m^{(k,h,s)}(d) + \beta_m^{(k,h,s)}(d)]\} \times \times \log_2 \frac{2 - [\alpha_m^{(k,h,s)}(d) + \beta_m^{(k,h,s)}(d)] + 10^{-p}}{\alpha_m^{(k,h,s)}(d) + \beta_m^{(k,h,s)}(d)] + 10^{-p}}. \quad (5)$$

The decision rules are constructed in the form of an implication based on the optimal geometric parameters of the hyperspherical containers of the diagnostic classes.

$$(\forall h,s,m)\left(\forall x^{(j)} \in \Omega\right)\left\{if\ d\left(x^{(j)} \oplus \left\{x_{m,i}^{(h,s)}\right\}\right) \le d_m^{(h,s)}\ \&\ \left(m^* = \arg\max_{\{m\}}\left(1 - \frac{d\left(x^{(j)} \oplus \left\{x_{m,i}^{(h,s)}\right\}\right)}{d_m^{(h,s)}}\right)\right)\right. \quad (6)$$
$$\left. then\ x^{(j)} \in X_{m^*}^{o(h,s)}\ else\ x^{(j)} \notin X_{m^*}^{o(h,s)}\right\}.$$

Thus, the feature vector $x^{(j)}$ is assigned to the class from the given alphabet of the corresponding stratum for which the membership function (6) is positive and maximal. Moreover, the decision rules (6), developed within the framework of the geometric approach, enable diagnostic decisions to be made in real time.

4 EXPERIMENTS

This study is dedicated to exploring the application of information-extreme machine learning in ophthalmological diagnostic systems. The training process utilized images of six common ocular pathologies, each representing a corresponding class for recognition purposes [19, 24].

These classes were ordered according to the proposed method of forming a variational series, the visualization of which is presented in Fig. 2. The specified set of images served as a test dataset to evaluate the effectiveness of the developed machine learning approach in the context of computer-aided diagnosis of visual system diseases.

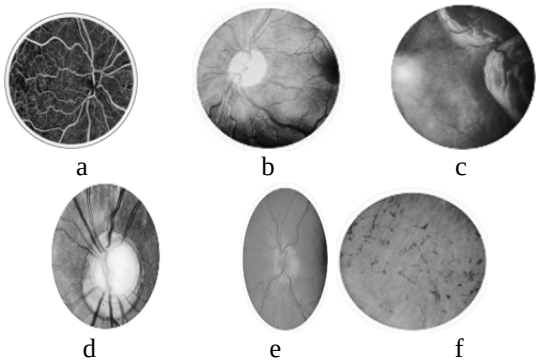


Figure 2 – Photographs of human eye pathologies: a – class X_1^o; b – class X_2^o; c – class X_3^o; d – class X_4^o; e – class X_5^o; f – class X_6^o

Below are the diagnostic classes that form a defined classification system:

1) normal eye condition (recognition class X_1^o);

2) optic nerve atrophy (recognition class X_2^o);

3) retinal detachment (recognition class X_3^o);

4) glaucoma (recognition class X_4^o);

5) optic neuritis (recognition class s X_5^o);

6) pigmentary retinopathy (recognition class X_6^o);

In the process of creating the training dataset for the ODS targeting visual system diseases, image fragments of ocular pathologies with a resolution of 100 × 100 pixels were used. The value of each pixel, expressed as a numerical code ranging from 0 to 255, represented the brightness level of that pixel in the image and served as a diagnostic feature. Assuming invariance of the brightness characteristics, the images were digitized using a Cartesian coordinate system. To enhance the informational content of the input data, the transposed version of the primary training matrix was added to it, effectively doubling the number of diagnostic features. This approach increases the volume of input information and, according to the maximum-distance principle of pattern recognition theory, contributes to an increase in the average interclass distance between the code representations of different image classes [9].

DOI 10.15588/1607-3274-2025-3-11

5 RESULTS

The analysis in works [9, 24] showed that using a hierarchical binary structure for storing diagnostic data in the context of information-extreme machine learning, exemplified by the development of the ODS for eye diseases based on pathology images, was inefficient due to the significant complexity of search, insertion, and deletion operations for diagnostic class elements. Specifically, with the increasing volume of images and their processing, the linear processing of the binary structure leads to a slowdown in processes, negatively affecting the speed of analysis and the accuracy of predictions. Since information-extreme machine learning involves processing large amounts of data for accurate pathology detection, it is advisable to switch to a de-recursive hierarchical structure. This structure will significantly speed up the search and classification processes, thereby enhancing the functional efficiency of eye disease diagnosis and reducing the time required for model training, which, in turn, will provide more accurate and faster results in ophthalmological systems.

To improve the functional efficiency of the ODS, a de-recursive hierarchical structure has been applied (Table 1).

Table 1 – Results of Machine Learning for the ODS

Class	Distance from the Zero Vector	Neighbor Class	Distance to Neighbor Class
X_1^o	15	X_3^o	40
X_2^o	79	X_4^o	30
X_3^o	37	X_5^o	31
X_4^o	59	X_5^o	11
X_5^o	54	X_4^o	11
X_6^o	62	X_4^o	27

The graphical representation of the results presented in Table 1, illustrating the distances from the zero vector and the distances to the neighbor class for each class as points on the plane, is shown in Figure 3.

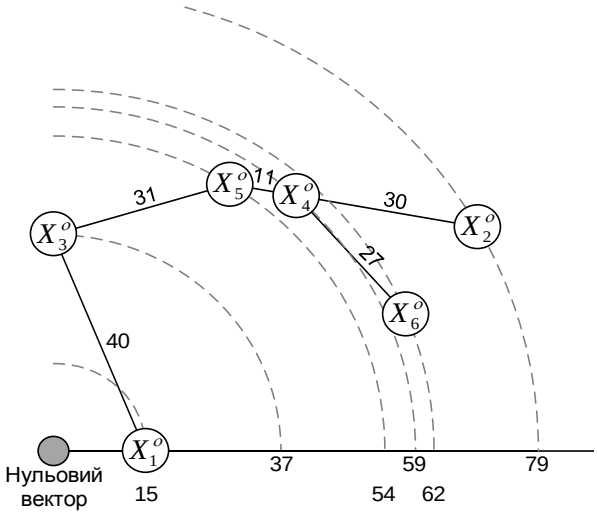


Figure 3 – Graphical Representation of Table 1

According to the information-extreme machine learning scheme using a hierarchical data structure, a binary de-recursive tree was formed, as shown in Figure 4.

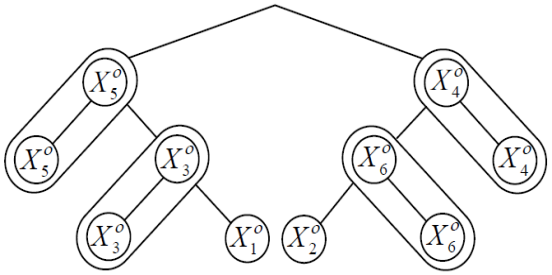


Figure 4 – Three-Level De-Recursive Hierarchical Structure

The analysis of the data presented in Figure 4 indicates the distribution of the alphabet, consisting of six recognition classes, into four final strata. Each of these strata contains two adjacent classes with the highest degree of similarity. In cases where one class belongs to two final strata, the membership function (7) is applied when forming the decision rules described by formula (6). In this case, the optimal geometric parameters of the class are selected, for which the radius, calculated according to procedure (4), takes the minimum value.

For the initial set of recognition classes, a classifier was formed for the first-level classes of the hierarchy $X_0^{o(h=1,s=1)} = X_4^o$ and $X_1^{o(h=1,s=1)} = X_5^o$. During the development, parallel optimization algorithms for the diagnostic class decision (Fig. 5) and optimization of the geometric parameters of decision rules (Fig. 6) were used.

Figure 5 shows the graphical representation of the functional dependence of the averaged information criterion, calculated according to formula (5), on the parameter of the control tolerance field for diagnostic features. This dependence was obtained by applying procedure (3), which involves parallel optimization of the tolerance limits for diagnostic features.

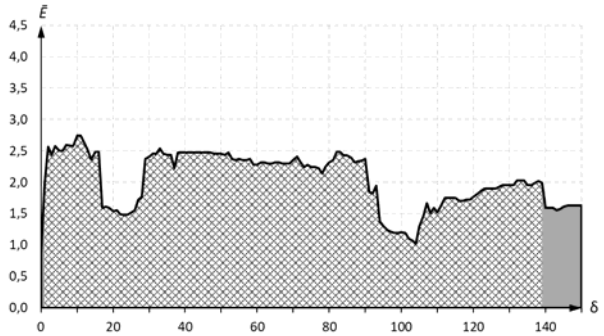


Figure 5 – Graph of the dependence of the information criterion on the parameter of the control tolerance field for the first-level hierarchy

In Figure 5 and subsequent graphical representations, the working domain for the definition of the criterion function (5) is marked by a double dashed line. This

DOI 10.15588/1607-3274-2025-3-11

domain is characterized by values of the first reliability exceeding 0.5 and a second-type error less than 0.5. The analysis of the graph presented in Figure 5 shows that the optimal value of the control tolerance field parameter is $\delta^*_{1,1}=10$ (measured in brightness gradations, which is also used for subsequent measurements). In this case, the maximum value of the information criterion $\overline{E}^*_{1,1}$=2.74 is achieved.

The formation of decision rules, described by formula (6), requires the determination of the optimal geometric parameters of the recognition class containers. Figure 6 illustrates the functional dependencies of the information criterion, calculated according to formula (5), on the radii of the hyperspherical containers of the recognition classes.

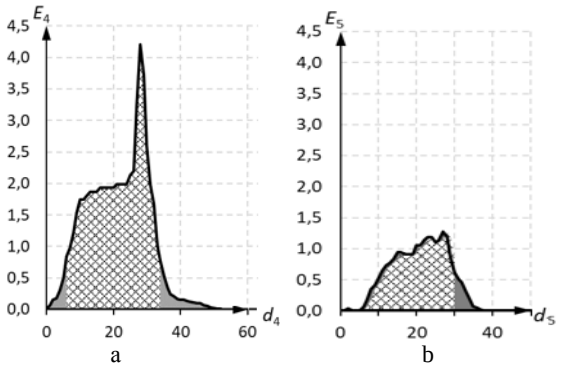


Figure 6 – Graph of the dependence of the information criterion on the radii of the recognition class containers for the first-level hierarchy: a – class X_4^o; b– class X_5^o

Based on the analysis presented in Figure 6, the optimal radius values for the recognition class containers are: $d_4^*=28$ (in code units) for recognition class X_4^o and $d_5^*=27$ for recognition class X_5^o. The maximum values of the Kullback information measure (5) are E_4^*=4.39 and E_5^*=1.27, respectively. The accuracy characteristics, specifically the first reliability and the second-type error for recognition class X_4^o, are as follows: D_{14}=1, β_4 = 0, while for recognition class X_5^o, they are D_{15}=0.97, β_5 = 0.32.

To improve the system's efficiency, two key algorithmic approaches were implemented: sequential optimization of SCT and optimization of the geometric parameters of decision rules. The graphical representation of these methods is shown in Figures 7 and 8, respectively. Figure 7 illustrates the functional dependence of the averaged information criterion, defined by formula (5), on the parameter of the control tolerance field for diagnostic features.

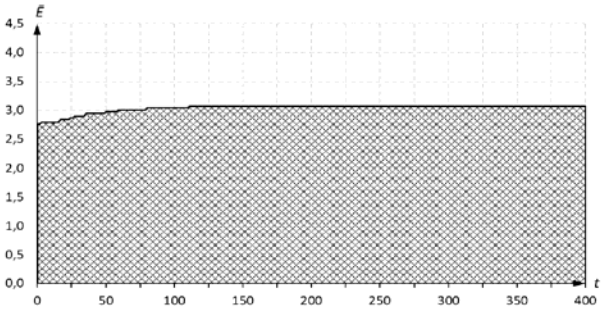


Figure 7 – Graph of the change in the information criterion during the sequential optimization of SCT for the first-level hierarchy

Based on the analysis of Figure 7, the maximum value of the averaged information criterion was reached at the 103rd iteration and amounted to 3.066, which is higher than the value obtained using parallel optimization. Figure 8 demonstrates the results of optimizing the geometric parameters of the recognition class containers obtained during the machine learning process.

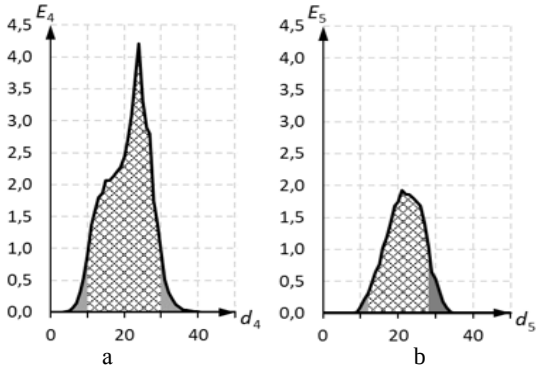


Figure 8 – Graph of the dependence of the information criterion on the radii of the recognition class containers for the first-level hierarchy: a – class X_4^o; b – class X_5^o

Based on the analysis presented in Figure 8, the optimal radius values for the recognition class containers are: $d_4^*=24$ for the recognition class X_4^o and $d_5^*=21$ for the recognition class X_5^o. At the same time, the maximum values of the Kullback information measure (5) are E_4^*=4.39 and E_5^*=1.93, respectively. The accuracy characteristics, specifically the first reliability and the second-type error for recognition class X_4^o, are as follows: D_{14}=1, β_4 =0, while for recognition class X_5^o, they are D_{15}=0.99, β_5 =0.22. A comparison with the previous results shows a significant improvement in these indicators.

For the initial set of recognition classes, a classifier was developed for the second-level classes of the first stratum, X_3^o and X_5^o. In this process, as at the previous level, parallel optimization algorithms for SCT (Fig. 9) and optimization of the geometric parameters of decision rules (Fig. 10) were used.

DOI 10.15588/1607-3274-2025-3-11

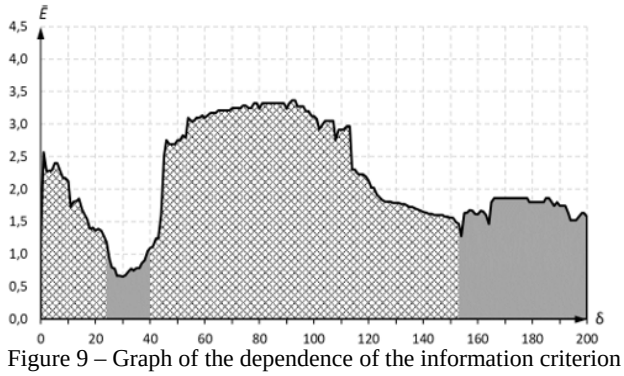
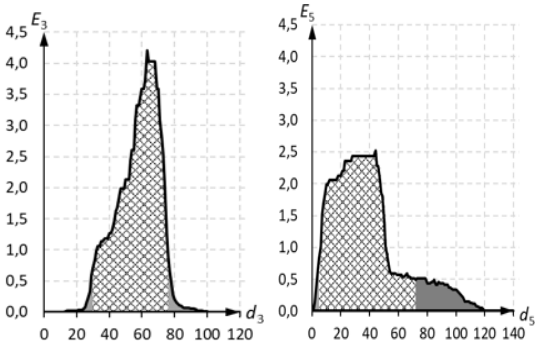


Figure 9 – Graph of the dependence of the information criterion on the parameter of the control tolerance field for the second-level hierarchy of the first stratum

The analysis of the graphical dependency presented in Figure 9 shows the achievement of the maximum value of the averaged Kullback information criterion (5) at the 92nd iteration of the process. The numerical value of this maximum is 3.362. Figure 10 demonstrates the results of optimizing the geometric parameters, obtained through the use of the optimal SCT, which was determined at the previous stage of optimization.



a b

Figure 10 – Graph of the dependence of the information criterion on the radii of the recognition class containers for the second-level hierarchy of the first stratum: a – class X_3^o; b – class X_5^o

The analysis of Figure 10 shows that the optimal radii of the recognition class containers are: $d_3^* = 63$ for recognition class X_3^o and $d_5^* = 44$ for recognition class X_5^o. The maximum values of the Kullback information measure (5) are E_3^*=4.39 and E_5^*=2.52, respectively. The accuracy characteristics, specifically the first reliability and the second-type error for diagnostic class X_3^o, are as follows: D_{13}=1, β_3=0, while for diagnostic class X_5^o, they are D_{15}=0.86, β_5=0.01.

To improve the system's efficiency, sequential optimization algorithms for SCT (Fig. 11) and optimization of the geometric parameters of decision rules (Fig. 12) were used.

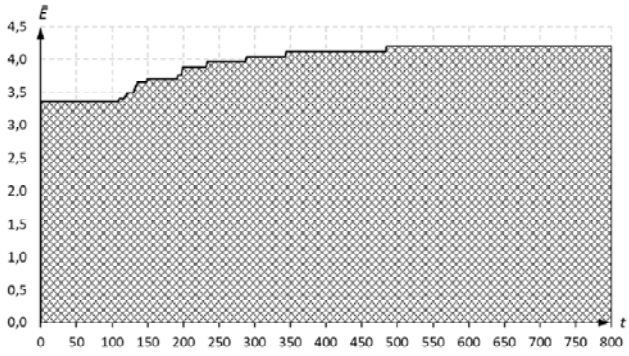
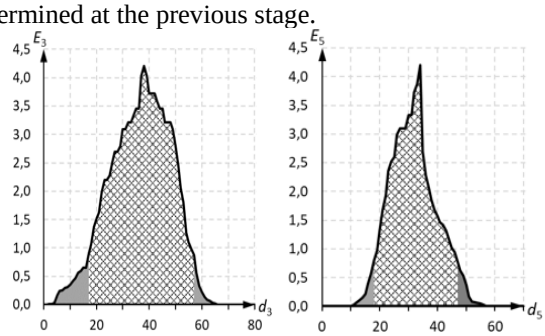


Figure 11 – Graph of the change in the information criterion during the sequential optimization of SCT for the second-level hierarchy of the first stratum

The analysis of the graphical dependence shown in Figure 11 demonstrates that the maximum value of the averaged Kullback information criterion is achieved at the 490th step of the iterative optimization process. The numerical value of this maximum is 4.206. Figure 12 illustrates the results of further optimization of the geometric parameters, carried out using the optimal SCT determined at the previous stage.



a b

Figure 12 – Graph of the dependence of the information criterion on the radii of the recognition class containers for the second-level hierarchy of the first stratum: a – class X_3^o; b – class X_5^o

The analysis of Figure 12 shows that the optimal radii of the recognition class containers are: $d_3^* = 38$ for recognition class X_3^o and $d_5^o = 34$ for recognition class X_5^o. At the same time, the maximum values of the Kullback information measure (5) are E_3^*=4.39 and E_5^*=4.39, respectively. The accuracy characteristics, specifically the first reliability and the second-type error for diagnostic class X_3^o, are as follows: D_{13}=1, β_3, while for diagnostic class X_5^o, they are D_{15}=1, β_5=0. A comparison with previous results shows a significant improvement in these indicators.

For the initial diagnostic class alphabet, a classifier was formed for the second-level classes of the second stratum, X_4^o and X_6^o. As in the previous stages of the study, parallel optimization algorithms for SCT (Fig. 12) and optimization of the geometric parameters of decision rules (Fig. 13) were used.

DOI 10.15588/1607-3274-2025-3-11

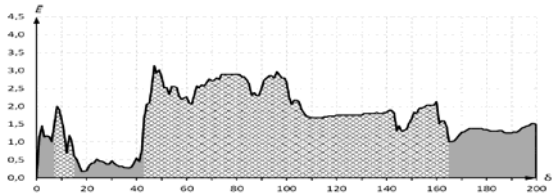


Figure 13 – Graph of the dependence of the information criterion on the parameter of the control tolerance field for the second-level hierarchy of the second stratum

The analysis of the graphical dependency presented in Figure 13 shows that the maximum value of the averaged Kullback information criterion (5) is achieved at the 47th step of the iterative process. The numerical value of this maximum is 3.133. Figure 14 displays the results of optimizing the geometric parameters, obtained through the use of the optimal SCT, determined during the previous optimization stage.

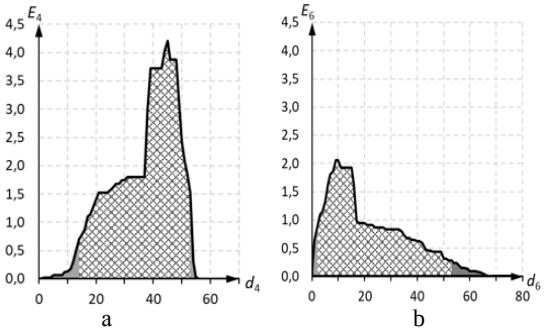


Figure 14 – Graph of the dependence of the information criterion on the radii of the diagnostic class containers for the second-level hierarchy of the second stratum: a – class X_4^o; b – class X_6^o

The analysis of Figure 14 shows that the optimal radii of the recognition class containers are: $d_4^* = 45$ for recognition class X_4^o and $d_6^* = 10$ for recognition class X_6^o. The maximum values of the Kullback information measure (5) are E_4^*=4.39 and E_6^*=2.06, respectively. The accuracy characteristics, specifically the first reliability and the second-type error for diagnostic class X_4^o, are as follows: D_{14}=1, β_4=0, while for recognition class X_6^o, they are D_{16}=0.79, β_6=0.0.

To improve the system's efficiency, sequential optimization algorithms for SCT (Fig. 15) and optimization of the geometric parameters of decision rules (Fig. 16) were used.

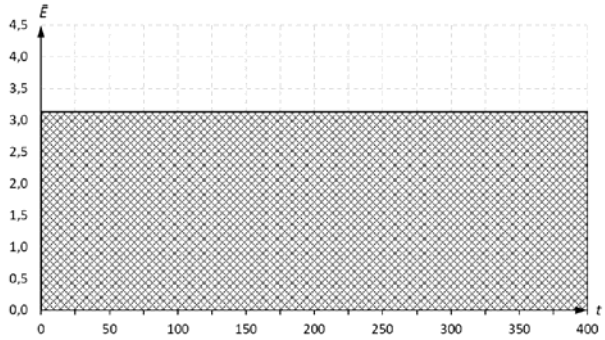


Figure 15 – Graph of the change in the information criterion during the sequential optimization of SCT for the second-level hierarchy of the second stratum

The analysis of the graphical dependency presented in Figure 15 shows that the maximum value of the averaged Kullback information criterion (5) is achieved at the 1st step of the iterative process. The numerical value of this maximum is 3.133. Figure 16 displays the results of optimizing the geometric parameters, obtained through the use of the optimal SCT, determined during the previous optimization stage.

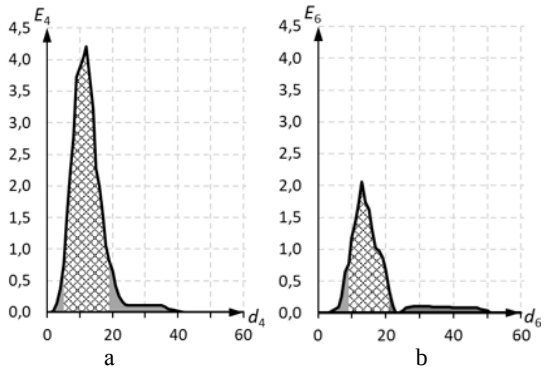


Figure 16 – Graph of the dependence of the information criterion on the radii of the recognition class containers for the second-level hierarchy of the second stratum: a – class X_4^o; b – class X_6^o

The analysis of Figure 16 shows that the optimal radii of the recognition class containers are: $d_4^* = 12$ for diagnostic class X_4^o and $d_6^* = 13$ for diagnostic class X_6^o. The maximum values of the Kullback information measure (5) are E_3^*=4.39 and E_5^*=2.06, respectively. The accuracy characteristics, specifically the first reliability and the second-type error for recognition class X_4^o, are as follows: D_{14}=1, β_4=0, while for recognition class X_6^o, they are D_{16}=0.79, β_6=0.

At the next step, a classifier was developed for the third-level classes of the first stratum, X_1^o and X_3^o. As in the previous stages of the study, parallel optimization algorithms for SCT (Fig. 17) were used, as well as algorithms for optimizing the geometric parameters of decision rules (Fig. 18).

DOI 10.15588/1607-3274-2025-3-11

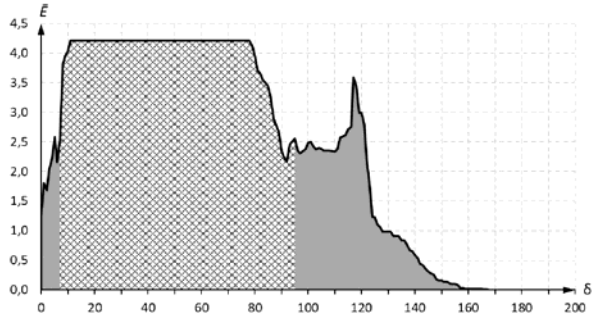
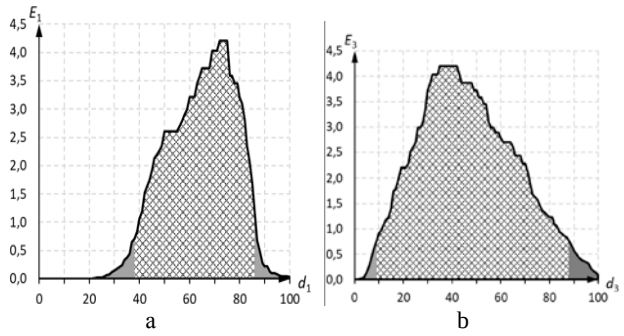


Figure 17 – Graph of the dependence of the information criterion on the parameter of the control tolerance field for the third-level hierarchy of the first stratum

The analysis of the graphical dependency presented in Figure 17 shows that the maximum value of the averaged Kullback information criterion (5) is achieved at the 12th step of the iterative process. The numerical value of this maximum is 4.205.

Figure 18 displays the results of optimizing the geometric parameters, obtained through the use of the optimal SCT, determined during the previous optimization stage.



a b

Figure 18 – Graph of the dependence of the information criterion on the radii of the recognition class containers for the third-level hierarchy of the first stratum: a – class X_1^o; b – class X_3^o

The analysis of Figure 18 shows that the optimal radii of the recognition class containers are: $d_1^* = 75$ for recognition class X_1^o and $d_3^* = 42$ for recognition class X_3^o. The maximum values of the Kullback information measure (5) are E_1^*=4.39 and E_3^*=4.39, respectively. The accuracy characteristics, specifically the first reliability and the second-type error for recognition class X_1^o, are as follows: D_{11}=1, β_1=0, while for recognition class X_3^o, they are D_{13}=1, β_3=0. Since an error-free classifier has been built, the use of sequential optimization algorithms for SCT for the first stratum classes of the third-level hierarchy is unnecessary.

At the next step, a classifier was developed for the third-level classes of the second stratum, X_2^o and X_6^o. As in the previous stages, parallel optimization algorithms for SCT (Fig. 19) were used, as well as algorithms for optimizing the geometric parameters of decision rules (Fig. 20) in the classification process.

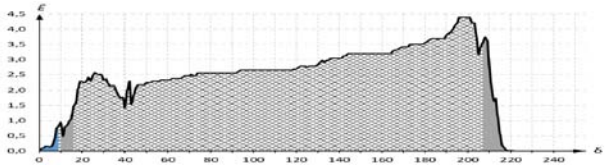
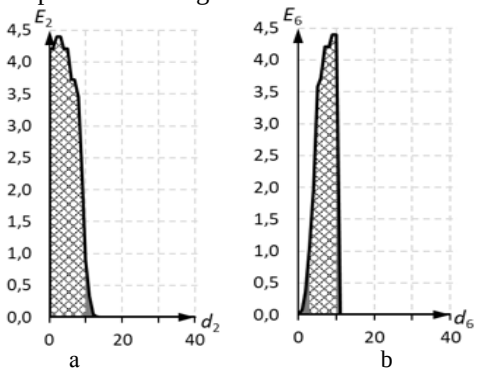


Figure 19 – Graph of the dependence of the information criterion on the parameter of the control tolerance field for the third-level hierarchy of the second stratum

The analysis of the graphical dependency presented in Figure 19 shows that the maximum value of the averaged Kullback information criterion (5) is achieved at the 192nd step of the iterative process. The numerical value of this maximum is 4.478. Figure 20 displays the results of optimizing the geometric parameters, obtained through the use of the optimal SCT, determined during the previous optimization stage.



a b

Figure 20 – Graph of the dependence of the information criterion on the radii of the recognition class containers for the third-level hierarchy of the second stratum: a – class X_2^o; b – class X_6^o

The analysis of Figure 20 shows that the optimal radii of the recognition class containers are: $d_2^* = 3$ for recognition class X_2^o and $d_3^* = 10$ for recognition class X_6^o. The maximum values of the Kullback information measure (5) are E_2^*=4.39 and E_3^*=4.39, respectively. The accuracy characteristics, specifically the first reliability and the second-type error for recognition class X_2^o, are as follows: D_{12}=1, β_2=0, while for recognition class X_6^o, they are D_{16}=1, β_6=0. Since an error-free classifier has been built for the first stratum classes of the third-level hierarchy, the use of sequential optimization algorithms for SCT is also unnecessary.

Thus, the comparative analysis of the training results of the computerized ODS, which uses binary and de-recursive hierarchical structures of decision rules, confirms the high effectiveness of these approaches in the task of classifying six functional states of the human eye based on images. Both structures demonstrate the ability to ensure the accurate formation of a classifier that eliminates errors during the processing of the training matrix.

DOI 10.15588/1607-3274-2025-3-11

6 DISCUSSION

The results of information-extremal machine learning based on a hierarchical data structure in the form of a recursive tree open new possibilities for solving the problem of enhancing the functional efficiency of the ODS. Using the example of information synthesis in the eye disease diagnosis system based on the characteristic signs of pathologies, the possibility of forming highly accurate diagnostic rules in the form of a three-layer recursive tree is demonstrated. Unlike a linear rule structure in a fixed diagnostic feature space, this method operates with optimal tolerance systems for each diagnostic class. Each layer of the tree uses strata consisting of pairs of nearest neighbor classes. The formation of decision rules is carried out using information-extremal machine learning methods with a depth of two levels for each such pair. It was established that, at the first level, it is advisable to apply the standard iterative optimization procedure for genotype parameters, while at the second level, a parallel-sequential optimization of phenotype functional parameters is applied, specifically the control tolerances for diagnostic features. This improves both the recognition accuracy and the efficiency of the machine learning process.

CONCLUSIONS

An important task of information analysis and synthesis of the intellectual component of the ODS, capable of information-extreme machine learning, is solved.

The scientific novelty of the obtained results lies in the fact that, for the first time, a methodology for selecting the training sample has been proposed. It determines the weighting coefficients that characterize the term and utility of the function for a given initial set of precedents and a specified division of the function space. It characterizes the individual absolute and relative informativeness of instances relative to the centers and boundaries of feature intervals based on the weighting values. This allows automating the sample analysis and its division into subsets, which, in turn, reduces the dimensionality of the training data. This, in turn, shortens the time and ensures acceptable accuracy for training the neural model.

The practical significance of the obtained results lies in the fact that software has been developed to implement the proposed indicators, and experiments have been conducted to study their properties. The results of the experiment allow recommending the proposed indicators for practical use, as well as determining the effective conditions for applying the proposed indicators.

Prospects for further research lie in studying the proposed set of indicators for a wide range of practical tasks.

REFERENCES

1. Lírio L. R. de, Malheiros É. F. R., Stabile G. et al. Retinoblastoma e a radiologia intervencionista: papel no tratamento da população infantil, *Brazilian Journal of Implantology and Health Sciences*, 2024, Vol. 8, № 6, pp. 3380–3399. DOI: 10.36557/2674-8169.2024v6n8p3380-3399.
2. Stitt A. W., Lois N., Medina R. J. et al. Advances in Our Understanding of Diabetic Retinopathy, *Clin Sci (Lond)*, 2013, Vol. 125, №1, pp. 1–17. DOI: 10.1042/CS20120588.
3. Cho N. H., Shaw J. E., Karuranga S. et al. IDF Diabetes Atlas: Global Estimates of Diabetes Prevalence for 2017 and Projections for 2045, *Diabetes Research and Clinical Practice*, 2018, Vol. 138, pp. 271–281. DOI: 10.1016/j.diabres.2018.02.023.
4. Oshika T. Artificial Intelligence Applications in Ophthalmology, *JMAJ*, 2025, Vol. 8, № 1, pp. 66–75. DOI: 10.31662/jmaj.2024-0139.
5. Li Z., Wang L., Wu X. et al. Artificial Intelligence in Ophthalmology: The Path to the Real-World Clinic, *Cell Reports Medicine*, 2023, Vol. 4, № 7, Article number: 101095. DOI: 10.1016/j.xcrm.2023.101095.
6. Srivastava O., Tennant M., Grewal P. et al. Artificial Intelligence and Machine Learning in Ophthalmology: A Review, *Indian Journal of Ophthalmology*, 2023, Vol. 71, №1, pp. 11–17. DOI: 10.4103/ijo.IJO_1569_22.
7. Shelehov I. V., Prylepa D. V., Khibovska Y. O. et al. Machine learning decision support systems for adaptation of educational content to the labor market requirements, *Radio Electronics, Computer Science, Control*, 2023, Vol. 1, pp. 62–72. DOI: 10.15588/1607-3274-2023-1-6.
8. Prylepa D.V. Informatsiyno-ekstremal'na intelektual'na tekhnolohiya diahnostuvannya emotsiyno-psykhichnoho stanu lyudyny. Dys. c.t.n. [Information-extreme intellectual technology for diagnosing the emotional and mental state of a person. Candidate of Technical Sciences diss.]. Kharkiv, 2024. 188 p.
9. Shelehov I. V., Barchenko N. L., Prylepa D. V. et al. Information-extreme machine training system of functional diagnosis system with hierarchical data structure, *Radio Electronics, Computer Science, Control*, 2022, Vol. 2, pp. 189–200. DOI: 10.15588/1607-3274-2022-18.
10. Chandra A., Romano M. R., Chao D. L. Implementing the New Normal in Ophthalmology Care Beyond COVID-19, *European Journal of Ophthalmology*, 2020, Vol. 31, № 2. DOI: 10.1177/1120672120975331.
11. Bejnordi B. E., Zuidhof G., Balkenhol M. et al. Context-Aware Stacked Convolutional Neural Networks for Classification of Breast Carcinomas in Whole-Slide Histopathology Images, *Journal of Medical Imaging*, 2017, Vol. 4, № 4, pp. 1. DOI: 10.1117/1.jmi.4.4.044504.
12. Gu H., Gu Y., Wei A. et al. Deep Learning for Identifying Corneal Diseases from Ocular Surface Slit-Lamp Photographs, *Scientific Reports*, 2020, Vol. 10, Article number: 17851. DOI: 10.1038/s41598-020-75027-3.
13. Kermany D. S., Goldbaum M., Cai W. et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning, *Cell*, 2018, Vol. 172, №5, pp. 1122–1131.e9. DOI: 10.1016/j.cell.2018.02.010.
14. Gulshan V., Peng L., Coram M. et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, *JAMA*, 2016, Vol. 316, №22, pp. 2402–2410. DOI: 10.1001/jama.2016.17216.
15. Woreta F. A., Gordon L. K., Pérez-González C. E. Enhancing Diversity in the Ophthalmology Workforce, *Ophthalmology*, 2022, Vol. 129, №10, pp. 127–136. DOI: 10.1016/j.ophtha.2022.06.033.
16. Lu W., Tong Y., Yu Y. et al. Applications of Artificial Intelligence in Ophthalmology: General Overview, *Journal of Ophthalmology*, 2018, Article number: 30581604, pp. 1–15. DOI: 10.1155/2018/5278196.

DOI 10.15588/1607-3274-2025-3-11

17. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2019, *Diabetes Care*, 2019, Vol. 42(Supplement 1), pp. 13–28. DOI: 10.2337/dc19-s002.
18. Kermany D. S., Goldbaum M., Cai W. et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning, *Cell*, 2018, Vol. 172, №5, pp. 1122–1131.e9. DOI: 10.1016/j.cell.2018.02.010.
19. Putivets', A.V. Intelektual'na systema vyyavlennya urazhen' sitkivky oka: robota na zdobuttya kvalifikatsiynoho stupenya bakalavra; spets.: 122 – komp"yuterni nauky (informatyka) [Elektron. resurs] / A.V. Putivets'; nauk. kerivnyk I.V. Shelekhov. – Sumy: SumDU, 2020. – 72 s. – Rezhym dostupu: http://essuir.sumdu.edu.ua:8080/handle/123456789/79779.
20. Suprunenko M. K., Zborshchyk O. P., Sokolov O. Information-extreme machine learning of wrist prosthesis control system based on the sparse training matrix, *Journal of Engineering Sciences*, 2022, Vol. 9, iss. 2, pp. 28–35. DOI: 10.21272/jes.2022.9(2).e4.
21. Dovbysh A., Zimovets V. Hierarchical Algorithm of the Machine Learning for the System of Functional Diagnostics of the Electric Drive, *Advanced Information Systems and Technologies, VI International Conference, Sumy, 16–18 May 2018*: *proceedings*, Sumy, Sumy State University, 2018, pp. 85–88.
22. Moskalenko V. V., Moskalenko A. S., Korobov A. G. Models and methods of intellectual information technology of autonomous navigation for compact drones, *Radio Electronics, Computer Science, Control*, 2018, № 3, pp. 68–77. DOI: 10.15588/1607-3274-2018-3-8.
23. Dovbysh A., Shelehov I., Romaniuk A., et al. Decision-making support system for diagnosis of oncopathologies by histological images, *Journal of Pathology Informatics*, 2023, Article number: 100193. DOI: 10.1016/j.jpi.2023.100193.
24. Shelehov I., Prylepa D., Khibovska Y. et al. Information-Extreme Machine Learning of an Ophthalmic Diagnostic System with a Hierarchical Class Structure, *Artificial Intelligence*, 2024, № 3, pp. 114–125. DOI: 10.15407/jai2024.03.114.

Received 24.03.2025.
Accepted 25.06.2025.

УДК 681.518:004.93

ІЄРАРХІЧНЕ МАШИННЕ НАВЧАННЯ СИСТЕМИ ФУНКЦІОНАЛЬНОГО ДІАГНОСТУВАННЯ ПАТОЛОГІЙ ОКА НА ОСНОВІ ІНФОРМАЦІЙНО-ЕКСТРЕМАЛЬНОГО ПІДХОДУ

Шелехов І. В. – канд. техн. наук, доцент, доцент каф. кібернетики та інформатики, Сумський національний аграрний університет, Суми, Україна; канд. техн. наук, доцент, доцент каф. комп'ютерних наук, Сумський державний університет, Суми, Україна.

Прилепа Д. В. – канд. техн. наук, асистент кафедри комп'ютерних наук Сумського державного університету, Суми, Україна.

Хібовська Ю. О. – аспірант кафедри комп'ютерних наук Сумського державного університету, Суми, Україна.

Тимченко О. А. – аспірант кафедри комп'ютерних наук Сумського державного університету, Суми, Україна.

АНОТАЦІЯ

Актуальність. Розглянуто задачу інформаційно-екстремального машинного навчання системи діагностування патологій ока за характерними ознаками захворювань. Об'єктом дослідження є процес ієрархічного машинного навчання системи діагностування офтальмологічних захворювань. Старіння населення та поширення захворювань очей, таких як глаукома, атрофія зорового нерва, відшарування сітківки та діабетична ретинопатія, вимагають ефективних методів ранньої діагностики для запобігання втрати зору. Традиційні методи діагностики значною мірою залежать від досвіду лікаря, що може призводити до помилок. Використання штучного інтелекту (ШІ) та машинного навчання (МН) може суттєво покращити точність і швидкість діагностування, що робить цю тему надзвичайно актуальною.

Мета. Підвищення функціональної ефективності комп'ютеризованої системи діагностування патологій ока на основі зображень.

Метод. Запропоновано метод інформаційно-екстремального ієрархічного машинного навчання для системи діагностування патологій ока на основі характерних ознак захворювань. Метод базується на функціональному підході до моделювання когнітивних процесів природного інтелекту, що забезпечує адаптивність системи діагностування за будь-яких початкових умов формування зображень патологій і дозволяє гнучко перенавчати систему при збільшення потужності алфавіту класів розпізнавання. Основою методу є принцип максимізації критерію функціональної ефективності на базі модифікованої інформаційної міри Кульбака, яка є функціоналом від точносних харатеристик діагростичних правил. Процес навчання розглядається як ітераційна процедура оптимізації параметрів роботи системи діагностування за цим інформаційним критерієм. На основі запропонованої категорійної функціональної моделі розроблено алгоритм інформаційно-екстремального машинного навчання з ієрархічною структурою даних у вигляді бінарного декурсивного дерева. Така структура даних дозволяє розділяти велику кількість класів розпізнавання на пари найближчих сусідів, для яких параметри машинного навчання оптимізуються за лінійним алгоритмом необхідної глибини.

Результати. Розроблено інтелектуальну технологію діагностики патологій ока, яка включає комплекс інформаційного, алгоритмічного та програмного забезпечення. Проведено порівняльний аналіз ефективності різних методів організації вирішальних правил у процесі навчання системи. Виявлено, що використання декурсивних ієрархічних структур класифікаторів дозволяє досягти вищої точності діагностики у порівнянні з бінарними класифікаторами.

Висновки. Розроблена інтелектуальна система комп'ютерного діагностування патологій ока демонструє високу ефективність та точність. Впровадження такої системи у медичну практику може суттєво підвищити якість діагностики очних захворювань, знизити навантаження на лікарів та мінімізувати ризик помилкових діагнозів. Подальші дослідження можуть бути спрямовані на вдосконалення алгоритмів та розширення їх застосування на інші типи медичних зображень.

КЛЮЧОВІ СЛОВА: комп'ютерна діагностика патології ока, штучний інтелект, машинне навчання, обробка зображень, розпізнавання образів, інформаційно-екстремальна технологія, ієрархічна структура класифікаторів.

DOI 10.15588/1607-3274-2025-3-11

ЛІТЕРАТУРА

1. Lírio L. R. de. Retinoblastoma and interventional radiology: role in the treatment of children / [L. R. de Lírio, É. F. R. Malheiros, G. Stabile, L. C. V. et al.] // Brazilian Journal of Implantology and Health Sciences. – 2024. – №8(6). – P. 3380–3399. DOI:10.36557/2674-8169.2024v6n8p3380-3399.
2. Advances in Our Understanding of Diabetic Retinopathy / [A. W. Stitt, N. Lois, R. J. Medina et al.] // Clin Sci (Lond). – 2013. – Vol. 125(1). – P. 1–17. DOI: 10.1042/CS20120588.
3. IDF Diabetes Atlas: Global Estimates of Diabetes Prevalence for 2017 and Projections for 2045 / [N. H. Cho, J. E. Shaw, S. Karuranga et al.] // Diabetes Research and Clinical Practice. – 2018. – Vol. 138. – P. 271–281. DOI: 10.1016/j.diabres.2018.02.023.
4. Oshika T. Artificial Intelligence Applications in Ophthalmology / T. Oshika // JMAJ. – 2025. – №8(1). – P. 66–75. DOI: 10.31662/jmaj.2024-0139.
5. Artificial Intelligence in Ophthalmology: The Path to the Real-World Clinic / [Z. Li, L. Wang, X. Wu et al.] // Cell Reports Medicine. – 2023. – Vol. 4, Issue 7, Article number: 101095. DOI: 10.1016/j.xcrm.2023.101095.
6. Artificial Intelligence and Machine Learning in Ophthalmology: A Review / [O. Srivastava, M. Tennant, P. Grewal et al.] // Indian Journal of Ophthalmology. – 2023. – Vol. 71(1). – P. 11–17. DOI: 10.4103/ijo.IJO_1569_22.
7. Machine Learning Decision Support Systems for Adaptation of Educational Content to the Labor Market Requirements / [I. V. Shelehov, D. V. Prylepa, Y. O. Khibovska et al.] // Radio Electronics, Computer Science, Control. – 2023. – Vol. 1. – P. 62–72. DOI: 10.15588/1607-3274-2023-1-6.
8. Прилепа Д. В. Інформаційно-екстремальна інтелектуальна технологія діагностування емоційно-психічного стану людини : дис. … канд. техн. наук : 05.13.06 / Прилепа Дмитро Вікторович. – Харків, 2024. 188 с.
9. Information-extreme machine training system of functional diagnosis system with hierarchical data structure / [I. V. Shelehov, N. L. Barchenko, D. V. Prylepa et al.] // Radio Electronics, Computer Science, Control. – 2022. – №2. – P. 189–200. DOI:10.15588/1607-3274-2022-18.
10. Chandra A. Implementing the New Normal in Ophthalmology Care Beyond COVID-19 / A. Chandra, M. R. Romano, D. L. Chao // European Journal of Ophthalmology. – 2020. – Vol. 31, Issue 2. – DOI: 10.1177/1120672120975331.
11. Bejnordi B. E. Context-Aware Stacked Convolutional Neural Networks for Classification of Breast Carcinomas in Whole-Slide Histopathology Images / [B. E. Bejnordi, G. Zuidhof, M. Balkenhol et al.] // Journal of Medical Imaging. – 2017. – Vol. 4(4). – P. 1. DOI: 10.1117/1.jmi.4.4.044504.
12. Deep Learning for Identifying Corneal Diseases from Ocular Surface Slit-Lamp Photographs / [H. Gu, Y. Gu, A. Wei et al.] // Scientific Reports. – 2020. – Vol. 10, Article number: 17851. DOI: 10.1038/s41598-020-75027-3.
13. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning / [D. S. Kermany, M. Goldbaum, W. Cai et al.] // Cell. – 2018. – Vol. 172(5). – P. 1122–1131.e9. DOI: 10.1016/j.cell.2018.02.010.
14. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs / [V. Gulshan, L. Peng, M. Coram et al.] // JAMA. – 2016. – Vol. 316(22). – P. 2402–2410. DOI: 10.1001/jama.2016.17216.
15. Woreta F. A. Enhancing Diversity in the Ophthalmology Workforce / F. A. Woreta, L. K. Gordon, C. E. Pérez-González // Ophthalmology. – 2022. – Vol. 129(10). – P. 127–136. DOI: 10.1016/j.ophtha.2022.06.033.
16. Applications of Artificial Intelligence in Ophthalmology: General Overview / [W. Lu, Y. Tong, Y. Yu et al.] // Journal of Ophthalmology. – 2018. – P. 1–15. DOI: 10.1155/2018/5278196.
17. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2019 / American Diabetes Association // Diabetes Care. – 2018. – Vol. 42(Supplement 1). – P. S13–S28. DOI: 10.2337/dc19-s002.
18. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning / [D. S. Kermany, M. Goldbaum, W. Cai et al.] // Cell. – 2018. – Vol. 172(5). – P. 1122–1131.e9. DOI: 10.1016/j.cell.2018.02.010.
19. Путівець А.В. Інтелектуальна система виявлення уражень сітківки ока: робота на здобуття кваліфікаційного ступеня бакалавра; спец.: 122 – комп'ютерні науки (інформатика) [Електрон. ресурс] / А. В. Путівець; наук. керівник І. В. Шелехов. – Суми: СумДУ, 2020. – 72 с. – Режим доступу: http://essuir.sumdu.edu.ua:8080/handle/123456789/79779.
20. Suprunenko M. K. Information-extreme machine learning of wrist prosthesis control system based on the sparse training matrix / M. K. Suprunenko, O. P. Zborshchyk, O. Sokolov // Journal of Engineering Sciences. – 2022. – Vol. 9, Issue 2. – P. 28–35. – DOI: 10.21272/jes.2022.9(2).e4.
21. Dovbysh A. Hierarchical Algorithm of the Machine Learning for the System of Functional Diagnostics of the Electric Drive / A. Dovbysh, V. Zimovets // Advanced Information Systems and Technologies : VI International Conference, Sumy, 16–18 May 2018 : proceedings. – Sumy: Sumy State University, 2018. – P. 85–88.
22. Moskalenko V. V. Models and methods of intellectual information technology of autonomous navigation for compact drones / V. V. Moskalenko, A. S. Moskalenko, A. G. Korobov // Radio Electronics, Computer Science, Control. – 2018. – № 3. – P. 8. – DOI: 10.15588/1607-3274-2018-3-8.
23. Dovbysh A. Decision-making support system for diagnosis of oncopathologies by histological images / [A. Dovbysh, I. Shelehov, A. Romaniuk et al.] // Journal of Pathology Informatics. – 2023. – P. 100193. – DOI: 10.1016/j.jpi.2023.100193.
24. Shelehov I. Information-Extreme Machine Learning of an Ophthalmic Diagnostic System with a Hierarchical Class Structure / I. Shelehov, D. Prylepa, Y. Khibovska // Artificial Intelligence. – 2024. – №3. – P. 114–125. DOI: 10.15407/jai2024.03.114.

DOI 10.15588/1607-3274-2025-3-11

p-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2025. № 3
e-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2025. № 3

ПРОГРЕСИВНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

PROGRESSIVE INFORMATION TECHNOLOGIES

UDC 004.852

SIMPLE, FAST AND SCALABLE RECOMMENDATION SYSTEMS VIA EXTERNAL KNOWLEDGE DISTILLATION

Androsov D. V. – Post-graduate student, Institute for Applied System Analysis, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.

Nedashkovskaya N. I. – Dr. Sc., Professor, Department of Mathematical Methods of System Analysis, Institute for Applied Systems Analysis, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Associate Professor, Kyiv, Ukraine.

ABSTRACT

Context. Recommendation systems are important tools for modern businesses to generate more income via proposing relevant goods to clients and achieve more loyal attendees. With deep learning emergence and hardware capabilities evolution it became possible to grasp customer behavioral patterns in data-driven way. However, accuracy of prediction is dependent on complexity of system, and these factors lead to increased delay in model's output. The object of the study is the task of issuing sequential recommendations, namely the next most relevant product, subject to restrictions on system response time.

Objective. The goal of the research is the synthesis of a deep neural network that can retrieve relevant items for a large portion of users with minimal delay.

Method. The proposed method of obtaining recommendation systems that leverages a mixture of Attention-based deep learning model architectures with application of knowledge graphs for prediction quality enhancement via explicit enrichment of recommendation candidate pool, demonstrates the benefits of decoder-only models and distillation learning framework. The latter approach was proven to demonstrate outstanding performance in solving recommendation retrieval task while responding fast for large user batch processing.

Results. A model of a recommender system and a method for its training are proposed, combining the knowledge distillation paradigm and learning on knowledge graphs. The proposed method was implemented via two-tower deep neural network to solve recommendation retrieval problem. A system for predicting the most relevant proposals for the user has been built, which includes the proposed model and its training method, as well as ranking indicators MAP@k and NDCG@k to assess the quality of the models. A program has been developed that implements the proposed architecture of the recommendation system, with the help of which the problem of issuing the most relevant proposals has been studied. When conducting experiments on a large amount of real data from user visits to an online retail store, it was found that the proposed method for designing recommender systems guarantees high relevance of the recommendations issued, is fast and unpretentious to computing resources at the stage of receiving responses from the system.

Conclusions. Series of conducted experiments confirmed that the proposed system effectively solves the problem in a short period of time, which is a strong argument in favor of its use in real conditions for large businesses that operate millions of visits per month and thousands of products. Prospects for further research within the given research topic include the use of other knowledge distillation methods, such as internal or self-distillation, the use of deep learning architectures other than the attention mechanism, and optimization of embedding vector storage.

KEYWORDS: knowledge distillation, knowledge graphs, decoder-only models, node embeddings, transformer models, attention mechanism, recurrent neural networks, long short-term memory networks, deep neural networks, personalized sequential recommendations, predicting the next most relevant product, user modeling.

ABBREVIATIONS

CBF is a content-based filtering;
CF is a collaborative filtering;
CNN is a convolutional neural network;
DNN is a deep neural network;
KD is a knowledge distillation process;
KG is a knowledge graph;
KLD is a Kullback-Leibler divergence;
LSTM is a long short-term memory network;
MAP is a mean average precision;
NDCG is a normalized discounted cumulative gain;
MHA is a multi-head attention network;
NBO is a next-best offer problem;
NBA is a next-best action problem;
NDCG is a normalized discounted cumulative gain;
PMI is a pointwise mutual information;
RS is a recommender system.

DOI 10.15588/1607-3274-2025-3-12

NOMENCLATURE

s is a session data;
I is an item data;
t is a (discrete) time point;
$i_j^{(t)}$ is an item at time point t;
s_t is a sample of time series;
$R(\cdot|\cdot)$ is a conditional probability mass function;
G is a weighted graph, i.e. KG;
V is a set of users and items – the vertices of G;
$\aleph(v)$ is a set of neighbors of node v in a graph G;
P is a projection operator over graph G;
$\Pr(\cdot)$ is a probability mass function in PMI definition;
$\Pr(\cdot,\cdot)$ is a joint probability mass function in PMI;
σ is a sigmoid activation function;
T is a softmax temperature;
L is a RS loss function;
h_t is a hidden state of neural network at time t, $h_t \in \mathrm{h}$;
b_h is a bias vector for hidden state of a recurrent neural network;
b_u is a bias vector for output state of a recurrent neural network;
$W_{\cdot h}$ is a weight matrix for hidden state of a recurrent neural network;
$W_{\cdot u}$ is a weight matrix for output state of a recurrent neural network;
α is a PMI threshold, $\alpha \in \mathbb{R}$;
v_i is a logit for the i-th score produced by the student model;
z_i is a logit by teacher model;
q_i is a soft target output for i-th score produced by the student model;
$\hat{q}_i$ is a soft target output for i-th score produced by the teacher model;
$||$ is a vector concatenation operator;
Q_A is an attention query weight matrix;
K_A is an attention key weight matrix;
V_A is an attention value weight matrix;
d_K is a number of columns in matrix K.

INTRODUCTION

Users' purchase decisions are significantly influenced not only by their general preferences but also by their most recent interactions with a given platform or marketplace. Understanding user behavior patterns is crucial for any customer-oriented business, as this obtained knowledge allow to propose the most relevant items to a given customer base, increasing revenue in both short- and long-term perspective. Such item proposal systems are called recommendation systems.

A recommendation system (RS) consists of a set of statistical models that analyze a user's interaction history, along with knowledge about the user and the items available, to generate relevant content recommendations [1]. Relevance, in this context, refers to the likelihood of a user engaging with the items presented. Consequently, there exists a broad spectrum of recommendation approaches, including non-personalized, semi-personalized, and personalized methods [1]. This work focuses specifically on the development of personalized recommendation systems, and thus, the terms "recommendation system" and "personalized recommendation system" are used interchangeably.

The content filtering (CBF) approach for recommendation systems construction is based on idea that the user is interested in items that are similar to items that were already interesting to this user earlier. Unlike collaborative filtering (CF) models, the similarity of items is determined not by a set of user actions, but based on the internal characteristics of the items themselves. To address the problem of items' feature descriptions extraction, deep learning methods are often used in the process of content filtering systems construction.

In recent years, RS have achieved substantial success across various real-world applications, including e-commerce platforms, streaming services, and online retail. A particularly notable application of recommendation systems is the next best offer (NBO) task, which involves predicting the items a user is likely to view or purchase after interacting with a platform.

NBO, also referred to as next best action (NBA) [2], or more broadly as next-basket recommendation (NBR) [3], is a prevalent use case for any enterprise engaged in business-to-consumer (B2C) operations. Marketing teams in these enterprises have been implementing NBO/NBA projects for many years, though many of these initiatives have failed to meet expectations [2]. Several factors contribute to this underperformance, including reliance on traditional methods, failure to update NBO models with new features (resulting in underutilization of both the breadth and depth of available data), inadequate campaign validation methods, technological shortcomings, and more.

The advent of machine learning and, consequently, deep neural networks (DNN) has introduced new opportunities for NBO/NBA by enabling the utilization of advanced technologies and large data sets to improve and optimize basket recommendations more effectively than ever before.

For instance, by leveraging deep learning techniques, the delivery of personalized offers and recommendations has been significantly enhanced, leading to notable improvements in customer engagement. These advancements can increase customer satisfaction and loyalty, ultimately driving higher sales and revenue for businesses [3, 4].

DOI 10.15588/1607-3274-2025-3-12

The object of study is the next-best offer (NBO) recommendation problem. NBO is a difficult task, since most session-based models' prediction pool is too narrow to accomplish the goal of grasping long-term inter-item dependencies and user behavior patterns. On the other hand, leveraging ubiquitously used collaborative filtering (CF) models do not capture short-term dependencies between items, what may be unsuitable for marketing campaigns design. Therefore, it is proposed to construct a new model, based on multi-head attention (MHA) mechanism, knowledge graphs (KG) and knowledge distillation (KD) techniques.

The subject of study is methods for sequential recommendation retrieval.

The purpose of the work is to create fast and scalable RS to solve NBO/NBA task for a large number of users.

1 PROBLEM STATEMENT

For a given multiset $s = \{i_j^{(t)} \mid j \in \mathbb{N}, j \le |I|, t \in \mathbb{N}\}$ of items in some set of available items (goods) I and t is a (discrete) time, called session, it is desired to model a likelihood function R such that:

$$\hat{i}^{(t)} = \arg\max_i P(i \mid s). \quad (1)$$

Suppose the items and users are described by many categorical and numerical (continuous) features. Each categorical feature is presented by an embedding vector, thus generalizing the concept of latent variables in matrix factorization.

The main difficulty of such task is that it should be approached both by CBF and CF methods, since solving (1) solely with respect to the given user's session may diminish the explorative capabilities of RS, while applying only collaborative filtering, considering the demand of such models to be trained on large historical interactions datasets, may result in a system which cannot adapt to drift in the user behavior and is feasible to use for recognition of general preferences of users. The second challenge arises on the inference step – it is preferrable to update the recommendations on-line, adapting them to the newest user actions, thus limiting the complexity of the obtained RS. Current research addresses these challenges by developing a hybrid method of building RS.

2 REVIEW OF THE LITERATURE

Considering the variability in user session lengths, it becomes essential to capture both short- and long-term dependencies that exist between items within a session and the potential future items that a user might interact with. This challenge has led to the emergence of models based on high-order Markov chains, which offer a sophisticated approach to understanding and predicting user behavior. Among these models, context tree models (CT) [5, 6] and Markov chain similarity models [7] have proven particularly effective.

Context tree models function by first constructing a partition tree that represents each user session. This partition tree is then traversed to define a high-order Markov chain, allowing the model to encapsulate the user session [6]. The hierarchical structure of the partition tree provides a powerful framework for modeling the sequential nature of user interactions, enabling the recommendation system to account for complex patterns and long-term dependencies that simpler models might overlook.

In addition to context tree models, another promising approach involves integrating high-order Markov chains with similarity-based methods, such as sparse linear methods (SLIM) and factored item similarity models (FISM). This hybrid approach leverages the strengths of both Markov chains and similarity measures to capture a comprehensive range of relationships within the data. By combining these methodologies, the model is capable of simultaneously addressing short-term and long-term dependencies between users and items, as well as item-to-item relationships, thereby offering a more nuanced and accurate prediction of user preferences [7]. The integration of similarity-based techniques with high-order Markov chains enhances the model's ability to generalize across different users and sessions, ultimately leading to more personalized and effective recommendations.

Aside from Markov chain-based models, deep learning techniques have increasingly gained traction in addressing the challenges imposed by sequential recommendation tasks. Among the various deep neural networks (DNN) architectures, recurrent neural networks (RNNs) have emerged as a leading choice due to their capacity to model sequences of data, capturing both the short-term and long-term dependencies that characterize user interactions over time.

Consider the sample of time series $s_t \subset s$ and the remaining time series s_{T-t}. RNN in this case is a mapping function $f : s_t \to s_{T-s}$, and that function is a chain of non-linear transformations over affine transformations that are provided by state-space modeling of s_{T-t} [8]. Vanilla RNN models these chains in a following way:

$$h_t = \sigma(W_{xh} x_t + W_{hh} h_t + b_h),$$
$$s_{T-t} = u_t = \sigma(W_{hu} h_t + b_u).$$

RNN is aimed to maximize logarithmic likelihood $\log P(s_{T-t} \mid s_t, U, W, V, c)$ [8]. However, despite its ability to model non-stationary time series, RNN has a couple of significant drawbacks – disability to parallelize hidden states computations, and consequently gradient vanishing, which is directly caused by its architecture [8].

Long short-term memory networks (LSTMs), firstly introduced in 1997 [9] have become the most widely adopted variant of RNNs for sequential recommendation tasks. LSTMs are designed to overcome the limitations of traditional RNNs, particularly the issues of vanishing and exploding gradients, by introducing memory cells that can

DOI 10.15588/1607-3274-2025-3-12

maintain and update information over long sequences. This makes LSTMs particularly effective at modeling the temporal dependencies within user sessions, allowing them to predict future interactions with a high degree of accuracy. Moreover, modifications of LSTMs, such as bi-directional LSTMs, further enhance the model's capability by enabling it to consider both past and future context when making predictions [10, 11]. Bi-directional LSTMs, contrary to "classic" ones, estimate parameters by traversing input sequences in both in forward and backward directions, i.e, at each time step, the outputs of the forward and backward LSTMs are concatenated (or combined) to form the final output. This allows the network to have access to both past and future context when making predictions. However, due to their architecture, these models are prone to violate causality requirements on sequential data, and the requirement to have all sequence available to perform backward pass make them unsuitable for online recommendation engines.

As an alternative to recurrent neural networks, relatively new family of deep learning approaches, called attention networks, have recently become ubiquitous choice for analyzing sequential data. Attention networks are based on attention mechanism, introduced in 2017 [12]. It allows the model to focus on specific parts of the input sequence when producing an output, enabling it to handle long-range dependencies more effectively than LSTMs or RNNs.

In traditional sequence-to-sequence models, such as those used in machine translation, the encoder processes the input sequence into a fixed-length context vector, which is then used by the decoder to generate the output sequence. However, this fixed-length context vector can be a bottleneck, especially for long sequences, as it forces the model to compress all information into a single vector.

The attention mechanism addresses this issue by allowing the decoder to access different parts of the encoder's output sequence directly, enabling it to focus on the most relevant parts of the input when generating each element of the output sequence [12].

Several variants of the attention mechanism exist, depending on the application and architecture. Self-attention used in transformer models, where the attention mechanism is applied to the same sequence, allowing each element to attend to all other elements in the sequence. The formula for scaled dot-product attention is:

$$\text{Attention}(Q_A, K_A, V_A) = \varsigma\left(\frac{Q_A{}^{\mathrm{T}} K_A}{\sqrt{d_K}}\right) V_A,$$

where d_K is a number of columns in key matrix, $\varsigma(\cdot)$ is a softmax function.

The other popular modification of attention mechanism is multi-head attention (MHA). It extends self-attention by applying multiple attention mechanisms (heads) in parallel, each with different learned parameters, and then concatenating their outputs. This allows the model to focus on different parts of the input sequence.

Attention-based networks have become increasingly prevalent in recommendation retrieval tasks due to their ability to effectively model complex relationships in data. These networks, such as hierarchical attention networks, are designed to process and analyze inputs that capture both user-item and item-item interactions. By considering these interactions simultaneously, hierarchical attention networks can more accurately predict subsequent user actions, leading to more personalized and relevant recommendations [13].

Moreover, stochastic self-attention networks represent an another advancement in this domain. These networks leverage the self-attention mechanism to dynamically assess the importance of different elements within the input sequence, thereby generating candidate recommendations with enhanced precision. The stochastic nature of these models introduces an element of randomness, which can help in exploring a broader range of potential recommendations, thereby improving the diversity and relevance of the suggested items [14].

In summary, attention networks play a critical role in the evolution of recommendation systems. Their ability to incorporate complex interactions and adapt to various input dynamics makes them indispensable tools for enhancing the accuracy and diversity of recommendations in modern retrieval tasks.

3 MATERIALS AND METHODS

Since the main drawback of sequential recommendation is in narrow candidate pool, it is crucial to enrich recommendation proposals beyond trending items and short-term user-item and item-item relationships. To overcome this challenge, it is proposed to change the structure of received data and augment the given RS with some external context.

The relationships between users and items, as well as between users themselves and items themselves, can be naturally represented by a graph $G = G(V, E, w, f)$, where $V = \langle U, I \rangle$ denotes the set of users and items – the vertices of the graph – and $E = \{(u,i) \mid u \in V, i \in V\}$ represents the set of edges connecting users with items, items with items and users with users. The edge weights $w \in \mathbb{R}$, are assigned through a mapping f. Since this graph represents stable relationships between items and users, it is proposed to name this structure a knowledge graph (KG).

Given this graph-based representation, geometric deep learning frameworks, such as graph neural networks (GNNs), are well-suited for addressing recommendation retrieval tasks.

GNNs are a class of deep learning models specifically designed to operate on data that is represented as graphs [15, 16]. These networks excel in tasks that require inference over graph-structured data by iteratively updating the representations of vertices through the aggregation of information from their neighboring vertices.

DOI 10.15588/1607-3274-2025-3-12

During the training process, each vertex $v \in V$ within the graph $G = G(V, E)$ refines its feature representation by incorporating features from its adjacent vertices. This iterative process of feature aggregation and representation updating can be expressed as:

$$h_v^{(k+1)} = \sigma\left(\sum_{u \in \aleph(v)} \left(W^{(k)} h_u^{(k)} + b^{(k)} \right) \right), \quad (2)$$

where $h_v^{(k)}$ denotes the feature vector of vertex v at the k-th layer, $\aleph(v)$ represents the set of neighboring vertices of v. The parameters $W^{(k)}$ and $b^{(k)}$ are the weights and biases specific to k-th layer, respectively. The function σ is a non-linear activation function, for example, sigmoid.

The iterative aggregation (2) allows GNNs to effectively capture and propagate local structural information throughout the network, leading to a more comprehensive understanding of the graph's global structure. As a result, GNNs are well-suited for a variety of tasks, including node classification, link prediction, and graph classification, where the relationships and dependencies between entities are naturally modeled as graphs. The ability of GNNs to leverage the inherent graph structure makes them particularly powerful for applications in domains such as social network analysis, molecular chemistry, recommendation systems, and more.

To process categorical data, embeddings are used that form a dense representation of each category.

For instance, [17] demonstrates an approach that integrates the attention mechanism with graph convolutional networks [15] to effectively learn embeddings from the user-item graph. This combined model is then leveraged to generate recommendations for the next item a user is likely to interact with.

In the current work it is proposed to modify MHA by referencing not only input sequence (i.e. self-attention heads), but to aggregate first-order neighbors of each input of a sequence with respect to the given knowledge graph.

More specifically, consider the heterogenous graph:

$$G = G(V, E, w, f, P), \quad (3)$$

where P is an edge properties set. Let's define such projection operator over (3) as follows:

$$\begin{aligned} P : G, p &\to G_p, \\ G_p &= G(V, E, w, f, P = p). \end{aligned} \quad (4)$$

It is obvious that graph G_p is weighted graph where only those edges preserved that share same property, e.g. user-movie graph that contain only US-based users.

The proposed attention modification, named as Mixed Attention, applies self-attention over the given session multiset s and operator (4) over the KG (3), performing the following computation:

$$\begin{aligned} &\text{MixedAttention}(Q_A, K_A, V_A) = \\ &= \Big\| \underset{p \in P \cup s}{} \text{Attention}\left(Q_p, K_p, V_p\right). \end{aligned} \quad (5)$$

This application of graph-based learning methods is called to enhance the potential of Attention Networks in capturing the intricate relational structures inherent in recommendation systems, improving the accuracy of the recommendations produced.

However, this computation implies traversing two structures – knowledge graph $G = G(V, E, w, f, P)$ (3), (4) and session s simultaneously to find each item nearest neighbors, hence the same computations are required on each inference step, which may increase complexity and latency of proposed RS. This problem could be mitigated by knowledge distillation (KD) techniques.

Knowledge distillation is a model compression technique that is designed to transfer the knowledge encapsulated within a large, highly complex model, known as the teacher, to a smaller and more computationally efficient model, referred to as the student [18, 19]. The principal idea underlying knowledge distillation is to enable the student model to mimic the behavior of the teacher model. This is achieved by training the student model to replicate the output distributions produced by the teacher model, in addition to the conventional training on labeled data. To facilitate learning, the concepts of learning on logits is introduce.

There exist multiple ways to perform KD, but the chosen one in the current proposal is performed via teacher's output distribution temperature scaling.

By transferring knowledge from teacher to student in a classification problem, we minimize the loss function of the class distribution predicted by the teacher model. Let us consider the case of accurate model, when the prediction of the probability of one of the classes (the correct one) is close to unity, and all others are close to zero. Such data is usually of little help for the student model, since it practically does not differ from the original labels. Therefore, a softmax temperature (normally set to 1) is introduced [18], which helps the student model to repeat not the classification labeled data, but the probability distribution, and allows the student model to better adopt the teacher's behavior. Let v_i denote the logits or pre-softmax outputs for the i-th score produced by the student model. The corresponding student soft target output q_i for i-th score is computed as follows:

$$q_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)}.$$

DOI 10.15588/1607-3274-2025-3-12

Higher values of a temperature parameter T produce softer probability distributions, which contain more information about the relative confidence levels across classes.

Suppose the teacher model has logits z_i, which produce soft target probabilities $\hat{q}_i$, and distillation is performed at temperature T. Then, the gradient of the cross-entropy function L_{CE} with respect to each logit v_i of the student model is given by [18]:

$$\frac{\partial L_{CE}}{\partial v_i} = \frac{1}{T}\left(q_i - \hat{q}_i\right) = \frac{1}{T}\left(\frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)} - \frac{\exp\left(\frac{v_i}{T}\right)}{\sum_j \exp\left(\frac{v_j}{T}\right)}\right).$$

If T is high in comparison with the logits v_i, the gradient of loss L_{CE} can be approximated as follows:

$$\frac{\partial L_{CE}}{\partial v_i} \approx \frac{1}{T}\left(\frac{1+\frac{z_i}{T}}{N+\frac{\sum_j z_j}{T}} - \frac{1+\frac{v_i}{T}}{N+\frac{\sum_j v_j}{T}}\right). \quad (6)$$

Let the logits v_i and z_i have a zero mean separately for each transfer case. Then, equation (6) is simplified to the following:

$$\frac{\partial L_{CE}}{\partial v_i} \approx \frac{1}{NT^2}\left(z_i - v_i\right),$$

and distillation is equivalent to minimizing $\frac{1}{2}\left(z_i - v_i\right)^2$ under the above conditions.

If T is relatively low, distillation practically ignores large negative logits (which are much more negative than the average). On the one hand, this is an advantage, since such logits could be very noisy. It has been shown in [18] that intermediate temperatures T work best when the student model is much too small in comparison with the teacher model.

Thus, in the paper, in process of training the student model, it is proposed to minimize Kullback-Leibler divergence (KLD) measure between student and teacher models, defined as:

$$KL(q \mid \hat{q}) = \int_{\infty}^{+\infty} q(x) \log \frac{q(x)}{\hat{q}(x)} dx. \quad (7)$$

By learning from the teacher's soft targets, the student model can generalize better on unseen data, often leading to improved performance compared to directly training the smaller model from scratch.

The proposed method consists of the following stages:

1. To construct the teacher model, which consists of the following elements:

– knowledge graph (3), (4);

– mixed attention, that consumes both user session and traverses knowledge graph to compute the operation (5);

– multi-layer perceptron with several hidden layers, which takes results of the mixed attention operations an input and has the softmax output layer to produce some probability distribution.

2. To construct the student model, which consists of the following elements:

– multi-head attention that takes as an input the user session sequence;

– convolutional neural network with several filters, consuming the attention scores, obtained at the previous stage along outputting softmax-mapped vector of the same size as the teacher model output.

3. To perform student model learning: the KL divergence (7) between model outputs is minimized.

4 EXPERIMENTS

It is proposed to solve the problem of NBO/NBA recommendation leveraging information retrieved from user interaction history and item properties.

More precisely, consider the dataset retrieved from an anonymous multi-brand and multi-category e-commerce store, which schema is provided in Table 1.

The dataset contains historical data from October 2019 to November 2019, overall storing approximately 6.5 million user sessions, or 69 million records.

Let us predict the final item for each user session leveraging the proposed method.

As a baseline model for solving the problem (1), an LSTM-based architecture is chosen (Fig. 1).

As a teacher model, the proposed extension of attention mechanism (5) is introduced instead of LSTM module, thus allowing the model to capture long-term relationships from a given KG (Fig. 2).

It could be observed, that the proposed architecture (Fig. 2) combines both content-based RS (via considering past interaction history) and collaborative filtering (via considering user-item model branching).

Table 1 – Dataset fields description

Field Name	Data Type	Description
Session Id	Base64	Unique identifier of user visit
Product Id	Integer	Stock keeping unit (SKU) of an item
Product Description	String	Description of item
Brand	String	Brand name
Category	String	Category of item, e.g. furniture
Price	Integer	Price in cents
Action	String	User action over item, e.g. add to cart, view
Timestamp	Timestamp	Date and time of interaction

DOI 10.15588/1607-3274-2025-3-12

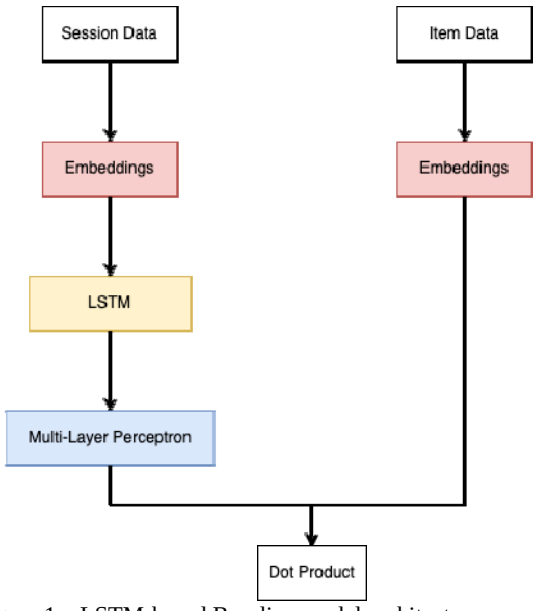


Figure 1 – LSTM-based Baseline model architecture

The general purpose of model is to generate embeddings for NBO/NBA candidate items. Thus, for solving task (1) it is proposed to minimize cross-entropy loss L between obtained embedding e_{obtained} and ground truth e_{true}:

$$L(e_{\text{obtained}}, e_{\text{true}}) = -e_{\text{obtained}}{}^{\text{T}} \log(e_{\text{true}}).$$

It is worth mentioning that the construction of KG is done in data-driven way by thresholding pointwise mutual information (PMI) between item pairs.

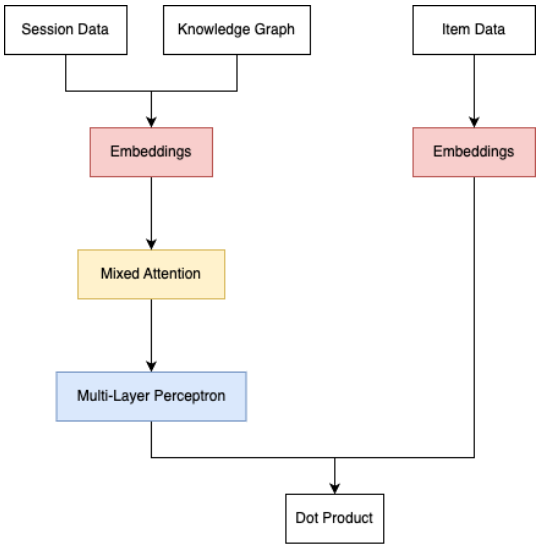


Figure 2 – The proposed Mixed Attention-based model architecture

KG is obtained by choosing only pairs of items where PMI value is greater than α. The selected threshold is defined at the 25-th percentile of all PMI values.

PMI between two items i and j is defined as:

$$\text{PMI}(i, j) = \log\left(\frac{\Pr(i, j)}{\Pr(i)\Pr(j)}\right),$$

where $\Pr(i, j)$ is the joint probability of items i and j co-occurring, $\Pr(i)$ and $\Pr(j)$ are the individual probabilities of items i and j occurring independently [20].

Item embeddings could be chosen in two ways – random initialization or leveraging pre-trained embeddings. To achieve the latter from the obtained KG, it is proposed to apply Node2Vec algorithm [21].

The main purpose of Node2Vec is to capture both the local and global network structure of a given graph. It does this by performing biased random walks on the graph, by balancing between breadth-first search (BFS) and depth-first search (DFS). This allows Node2Vec to generate node embeddings that reflect both the community structure (via BFS) and functional similarity (via DFS) within the graph [21]. For the experiment purposes, the only adjusted hyper-parameter is embedding dimension, which should align with the corresponding hyper-parameter in all proposed architectures.

The next step is to define the student model, parameters of which will be optimized via temperature scaling. This model is utilizing MHA module, thus eliminating the need for constantly traversing KG for each recommendation suggestion. The schematic representation of proposed model is shown in Fig. 3.

The student model learns the probability distribution of the teacher; thus, it is proposed to minimize KL divergence (7) between student model and teacher model.

Concluding, the following hyper-parameters are set for baseline model:

1. Embedding dimension – 64.
2. Multi-Layer Perceptron layer number – 2.
3. LSTM cell size – 64.

Consequently, the following hyper-parameters are set for teacher model:

1. Embedding dimension – 64.
2. Multi-Layer Perceptron layer number – 2.
3. Query, Key and Value matrix size – 64.
4. Causal mask – applied to guarantee that current decisions don't affect previous ones.

On the other side, since student model utilizes convolutional neural networks (CNN) instead of MLP, the following hyper-parameters are picked:

1. Stride size – 1.
2. Padding – "same".
3. Dropout rate – 50%.
4. Number of heads in MHA – 2.
5. Embedding dimension – 64.
6. Query, Key and Value matrix size – 64.
7. Temperature T – [0.5, 1, 2, 5].

DOI 10.15588/1607-3274-2025-3-12

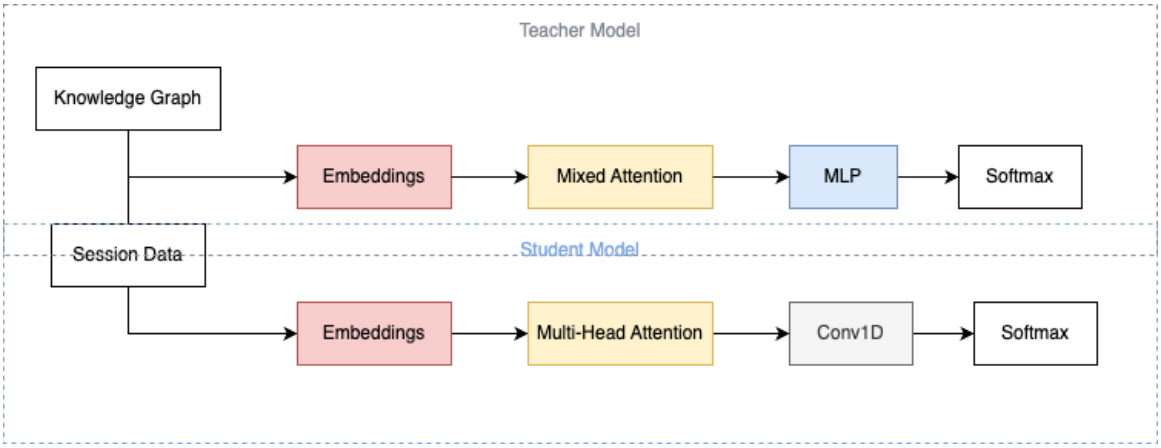


Figure 3 – The proposed teacher-student architecture

The choice of padding type and stride size is influenced by the constraint on output of NN to be an embedding of next item, given users' previous interactions data.

In order to examine the ability of models to retrieve relevant items, the following ranking metrics were chosen:

1. Mean average precision @ k (MAP @ k) – ranking metric used to evaluate the accuracy of a ranked list of items up to a cutoff rank k.

2. Normalized discounted cumulative gain @ k (NDCG @ k) – measure of ranking quality that considers the position of relevant items in the ranked list up to a cutoff rank k and applies penalty to relevant items that appear lower in the list by applying a logarithmic discount.

For both metrics, k values 1, 10 and 100 are considered.

Also, since the complexity of system affects recommendation candidate calculation time, mean retrieval time is proposed as the auxiliary metric to consider along with ranking ones.

5 RESULTS

In the Tables 2–3 MAP@k and NDCG@k metrics for given k ranking cut-off for LSTM baseline and Mixed Attention model statistics are provided. The best Mixed Attention model is selected as the teacher model.

Table 2 – Results of Baseline and Mixed Attention models benchmarking by MAP metric

Model	MAP@1	MAP@10	MAP@100
Baseline	0.1492	0.2766	0.2859
Baseline + Node2Vec	0.1453	0.2740	0.2824
Mixed Attention	0.1769	0.3003	0.3082
Mixed Attention + Node2Vec	0.2	0.3316	0.3378

Table 3 – Results of Baseline and Mixed Attention models benchmarking by NDCG metric

Model	NDCG@1	NDCG@10	NDCG@100
Baseline	0.1529	0.3296	0.3767
Baseline + Node2Vec	0.1509	0.3276	0.374
Mixed Attention	0.1755	0.3487	0.3865
Mixed Attention + Node2Vec	0.2025	0.3794	0.417

On the other hand, Table 4 and Table 5 reflect the values of MAP@k and NDCG@k results of KD for different temperature values, respectively. Table 6 summarizes models time performance. Figures 4–7 depict the evolution of ranking metrics with each epoch.

Table 4 – Results of KD benchmarking by MAP metric

T	MAP@1	MAP@10	MAP@100
0.5	0.1977	0.2727	0.2773
1	0.1952	0.325	0.3315
2	0.0832	0.1441	0.1487
5	0.076	0.1352	0.1789

Table 5 – Results of KD benchmarking by NDCG metric

T	NDCG@1	NDCG@10	NDCG@100
0.5	0.198	0.3041	0.33
1	0.1948	0.3745	0.374
2	0.0831	0.1708	0.1956
5	0.0368	0.1180	0.1417

Table 6 – Average time per 1000 requests per model

Model	Average time per 1000 requests, s
Baseline	1.01
Mixed Attention + Node2Vec	2.67
Proposed student model	0.189

DOI 10.15588/1607-3274-2025-3-12

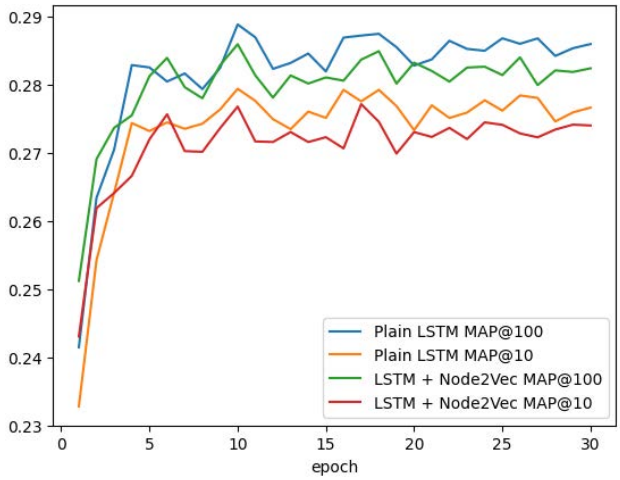


Figure 4 – MAP@k per epoch for baseline models

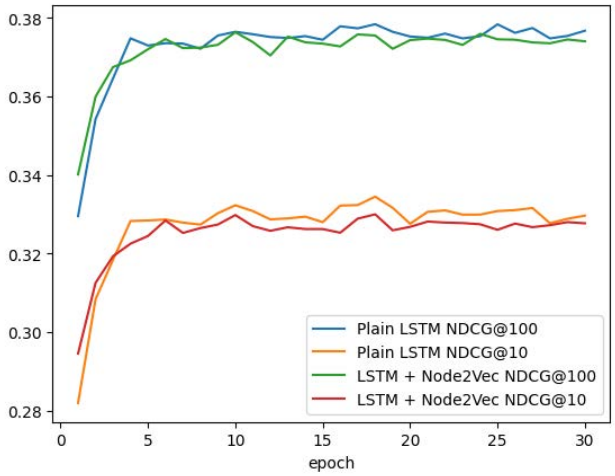


Figure 5 – NDCG@k per epoch for baseline models

6 DISCUSSION

As follows from Table 2 and Table 3, proposed Mixed Attention approach strongly outperforms LSTM-based baseline models. However, the fact that Node2Vec pre-trained embeddings cause little-to-no impact on ranking metrics for non-graph-based model but significantly enhances predictive capabilities of models that utilize KGs, is quite surprising and contradicts the initial suggestion that "implicit" knowledge, reflected solely in pre-trained embeddings could enhance sequential models.

It is worth noticing that this behavior persists for each epoch, as shown on Fig. 4 and 5 for Baseline models and Fig. 6 and 7 for Mixed Attention models, respectively.

Since the best model by all ranking metrics is Mixed Attention model with Node2Vec pre-trained embedding, this model is used as a teacher model.

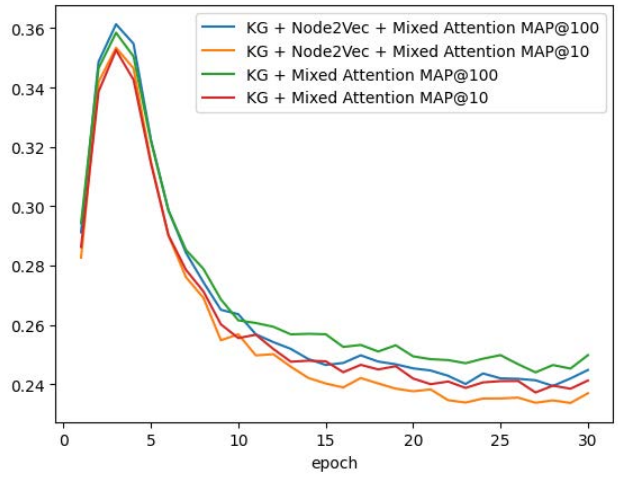


Figure 6 – MAP@k per epoch for Mixed Attention models

After performing temperature scaling with different values of parameter *T*, various student models were obtained. Since the bigger temperature, the more output distribution is uniform-like. Whilst very low value can introduce overconfidence to model decisions, it was predictable that both high and low *T* values could decrease predictive capabilities of model. The overall dependencies between temperature scaled outputs of teacher model and student model performance are listed in Tables 4 and 5.

The best model was obtained without performing temperature scaling of teacher model outputs. It is also worth noticing that results are only slightly worse than teacher's model ones, namely Mixed Attention + Node2Vec models in Tables 2 and 3.

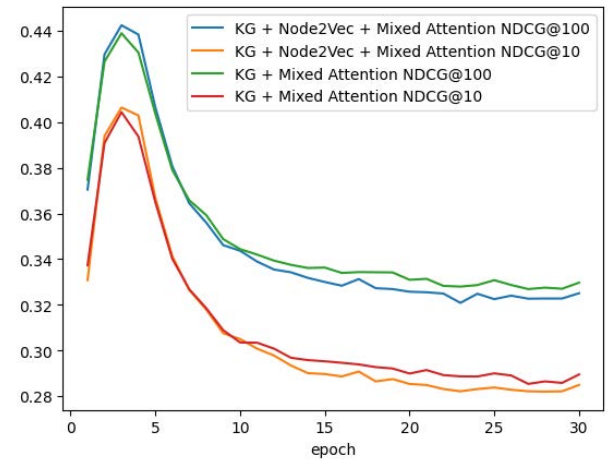


Figure 7 – NDCG@k per epoch for Mixed Attention models

It is also worth noticing the graph representation quality of the recommender system, obtained with the proposed KG and distillation method (Fig. 3). On Fig. 8 one can see the top-5 recommendations given that user has put Nike shoes to the basket or purchased this item. As one can see, the proposed model captures associations from the KG with a decent accuracy, grasping relationships between sport shoes and fitness vehicles and equipment, although the model itself gives irrelevant recommendation to buy desktop.

DOI 10.15588/1607-3274-2025-3-12

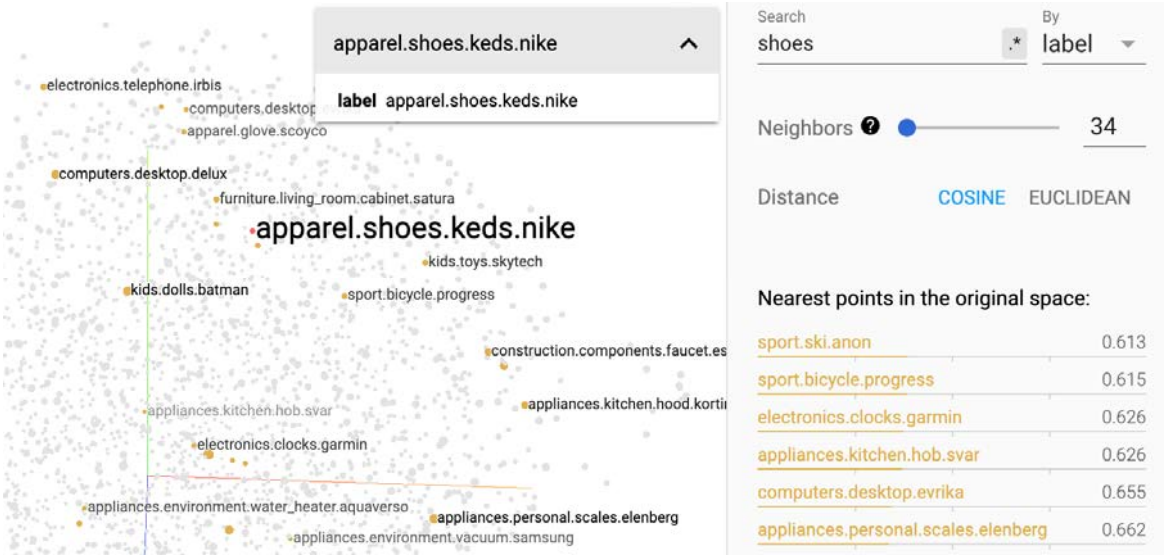


Figure 8 – Top-5 recommendations, obtained using the proposed model

The final assessment is conducted between best models in each category – LSTM, Mixed Attention and distilled student model. The main objective is to measure average time per 1000 requests for retrieving 100 most relevant items, given user sessions data. The received measures, recorded into Table 6, show that the proposed architecture (Fig. 3) and method significantly outperforms baseline solution by offering the main benefit of transformer-like architecture over RNNs – high degree of computations parallelization.

CONCLUSIONS

The problem of next best offer prediction is solved in this work using multiple deep learning-based approaches.

The scientific novelty of obtained results shows that by combining learning on graphs and knowledge distillation it is feasible to build scalable, fast and precise recommendations systems.

The practical significance of current work and its results is that implemented models could be applied to forecast users next interactions on the enterprise scale.

Prospects for further research are to examine other architectural approaches, different from decoder-only models, and propose alternatives to Attention networks.

ACKNOWLEDGEMENTS

This study was funded and supported by National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute” (NTUU KPI) in Kyiv (Ukraine), and also financed in part of the NTUU KPI Science-Research Work by the National Academy of Sciences of Ukraine “Development of models and methods for solving predictive problems based on large amounts of poorly structured information in conditions of uncertainty” (State Reg. No. 0122U000671).

REFERENCES

1. Falk K. Practical Recommender Systems. Shelter Island, Manning, 2019, 432 p.
2. Rasool A. Next Best Offer (NBO) / Next Best Action (NBA) – why it requires a fresh perspective? [Electronic resource]. Access mode: https://www.linkedin.com/pulse/next-best-offer-nbo-action-nba-why-requires-fresh-azaz-rasool/
3. Wang S., Wang Y., Hu L. et al. Modeling User Demand Evolution for Next-Basket Prediction, *IEEE Transactions on Knowledge and Data Engineering*, 2023, Vol. 35, Issue 11, pp. 11585–11598. DOI: 10.1109/TKDE.2022.3231018.
4. Eliyahu K. A. Achieving Commercial Excellence through Next Best Offer models. [Electronic resource]. Access mode: https://www.linkedin.com/pulse/achieving-commercial-excellence-through-next-best-offer-kisliuk/
5. Wang S., Hu L., Wang Y. et al. Sequential Recommender Systems: Challenges, Progress and Prospects, *International Joint Conference on Artificial Intelligence : Twenty-eighth international joint conference, IJCAI 2019, Macao, 10–16 August 2019 : proceedings*. Macao: International Joint Conference on Artificial Intelligence, 2019, pp. 6332–6338. DOI: 10.24963/ijcai.2019/883.
6. Garcin F., Dimitrakakis C., Faltings B. Personalized News Recommendation with Context Trees, *Recommender systems : Seventh ACM conference, RecSys'13, Hong-Kong, 12–16 October 2013 : proceedings*. New York, Association for Computing Machinery, 2013, pp. 105–112. DOI: 10.1145/2507157.2507166.
7. He R., McAuley J. Fusing Similarity Models with Markov Chains for Sparse Sequential Recommendation, *ArXiv*, 2016. DOI: 1609.09152v1.
8. Geron A. Hands-On Machine Learning with Scikit-Learn and TensorFlow. Sebastopol, O’Reilly Media Inc., 2017, 760 p.
9. Hochreiter S., Schmidhuber J. Long short-term memory, *Neural computation*, 1997, Vol. 9, № 8, pp. 1735–1780.
10. Xia Q., Jiang P., Sun F. et al. Modeling Consumer Buying Decision for Recommendation Based on Multi-Task Deep Learning, *Information and Knowledge Management : Twenty-seventh ACM international conference, CIKM '18, Torino, 22–26 October 2018 : proceedings.* New York, Association for Computing Machinery, 2018, pp. 1703–1706. DOI: 10.1145/3269206.3269285.
11. Zhao C., You J., Wen X., Li X. Deep Bi-LSTM Networks for Sequential Recommendation, *Entropy (Basel)*, 2020, Vol. 22, Issue 8, P. 870. DOI: 10.3390/e22080870.

DOI 10.15588/1607-3274-2025-3-12

12. Vaswani A., Shazeer N., Parmar N. et al. Attention is all you need, *Neural Information Processing Systems : Thirty-first international conference, NIPS '17, Long Beach, California, 04–09 December 2017 : proceedings*. New York: Curran Associates Inc., 2017, pp. 6000–6010.
13. Ying H., Zhuang F., Zhang F. et al. Sequential Recommender System based on Hierarchical Attention Network, *International Joint Conference on Artificial Intelligence : Twenty-seventh international joint conference, IJCAI '18, Stockholm, 13–19 July 2018 : proceedings*. Menlo Park, AAAI Press, 2018, pp. 3926–3932. DOI: 10.24963/ijcai.2018/546.
14. Fan Z., Liu Z., Wang Y. et al. Sequential Recommendation via Stochastic Self-Attention, *ACM Web Conference 2022, WWW '22, Lyon, 25–29 April 2022 : proceedings*. New York, Association for Computing Machinery, 2022, pp. 2036–2047. DOI: 10.1145/3485447.3512077.
15. Kipf T. N., Welling M. Semi-Supervised Classification with Graph Convolutional Networks, *International Conference on Learning Representations : Fifth international conference, ICLR 2017, Toulon, 24–26 April 2017 : proceedings*. New York, Curran Associates Inc., 2017. DOI: 10.48550/arXiv.1609.02907.
16. Wu Z., Pan S., Chen F. et al. A Comprehensive Survey of Graph Neural Networks for Knowledge Graphs, *IEEE Transactions on Neural Networks and Learning Systems*, 2022, Vol. 32, № 1, pp. 4–24. DOI: 10.1109/TNNLS.2020.2978386.
17. Hekmatfar T., Haratizadeh S., Razban P., Goliaei S.]Attention-Based Recommendation On Graphs, *ArXiv*, 2022. DOI: 2201.05499.
18. Hinton G., Vinyals O., Dean J. Distilling the knowledge in a neural network, *ArXiv*, 2015. DOI: 1503.02531.
19. Ba L. J., Caruana R. Do Deep Nets Really Need to be Deep? *Advances in Neural Information Processing Systems*, 2014, Vol. 27, pp. 2654–2662. DOI: 1312.6184.
20. Church K. W., Hanks P. Word association norms, mutual information, and lexicography, *Computational Linguistics*, 1990, Vol. 16, № 1, pp. 22–29.
21. Grover A., Leskovec J. Node2vec: Scalable Feature Learning for Networks, *ArXiv*, 2016. DOI: 1607.00653.

Received 11.05.2025.
Accepted 04.07.2025.

УДК 004.852

ПРОСТІ, ШВИДКІ ТА МАСШТАБОВАНІ РЕКОМЕНДАЦІЙНІ СИСТЕМИ ЗАСНОВАНІ НА ФІЛЬТРАЦІЇ ЗНАНЬ ВІД ВЧИТЕЛЯ

Андросов Д. В. – аспірант кафедри штучного інтелекту Навчально-наукового Інституту прикладного системного аналізу Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.

Недашківська Н. І. – д-р техн. наук, професор кафедри математичних методів системного аналізу Навчально-наукового Інституту прикладного системного аналізу Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», доцент, Київ, Україна.

АНОТАЦІЯ

Актуальність. Системи рекомендацій – важливі інструменти для сучасного бізнесу, які дають змогу отримувати більший дохід за рахунок пропозиції клієнтам відповідних товарів та залучення більш лояльних відвідувачів. З появою глибокого навчання та розвитком апаратних можливостей стало можливим уловлювати моделі поведінки клієнтів на основі даних. Однак точність прогнозу залежить від складності системи, і ці фактори призводять до збільшення затримки виведення на основі моделі. Об’єктом дослідження є задача видачі послідовних рекомендацій, а саме – наступного найбільш релевантного товару в умовах наявності обмежень по часу відповіді системи.

Ціль. Метою дослідження є синтез глибокої нейронної мережі, яка з мінімальною затримкою може отримувати релевантні елементи для більшості користувачів.

Метод. Пропонований метод отримання систем рекомендацій, який використовує поєднання архітектур моделей глибокого навчання на основі уваги із застосуванням графів знань для підвищення якості прогнозування за допомогою явного збагачення пулу кандидатів для рекомендацій, демонструє переваги моделей декодування та структури дистиляційного навчання. Було доведено, що підхід дистиляції знань є надзвичайно продуктивним під час вирішення завдань пошуку рекомендацій, одночасно швидко реагуючи на пакетну обробку великих обсягів даних користувачів.

Результати. Запропоновано модель рекомендаційної системи та метод її навчання, що поєднує парадигму дистиляції знань та навчання на графах знань. Запропонований метод був реалізований через двобаштову глибоку нейронну мережу для вирішення проблеми пошуку рекомендацій. Побудовано систему прогнозування найбільш релевантних наступних пропозицій для користувача, яка включає пропоновану модель та метод її навчання, а також показники ранжування MAP@k та NDCG@k для оцінки якості роботи моделей. Розроблено програму, яка реалізує пропоновану архітектуру рекомендаційної системи, за допомогою якої досліджена проблема видачі найрелевантніших наступних пропозицій. Під час проведення експериментів на великій кількості реальних даних візитів користувачів до онлайн магазину роздрібної торгівлі було встановлено, що пропонований метод конструкції рекомендаційних систем гарантує високу релевантність виданих рекомендацій, є швидким та невибагливим до обчислювальних ресурсів на етапі отримання відповідей від системи.

Висновки. Проведені експерименти підтвердили, що запропонована система ефективно вирішує поставлену задачу за малий проміжок часу, що є вагомим аргументом на користь її застосування в реальних умовах для великих бізнесів, що оперують мільйонами візитів на місяць та тисячами товарів. Перспективи подальших досліджень в рамках заданої теми дослідження включають в себе використання інших методів дистиляції знань, таких як внутрішня або само-дистиляція, використання відмінних від механізму уваги архітектур глибинного навчання, а також оптимізація сховища векторів вкладень.

КЛЮЧОВІ СЛОВА: дистиляція знань, графи знань, декодувальні моделі, вкладення вершин графів, архітектури типу «трансформер», механізм уваги, рекурентні нейронні мережі, мережі довгострокової короткої пам’яті, глибинні нейронні мережі, персоналізовані послідовні рекомендації, прогнозування наступного найбільш релевантного товару, моделювання користувача.

DOI 10.15588/1607-3274-2025-3-12

ЛІТЕРАТУРА

1. Falk K. Practical Recommender Systems / K. Falk. – Shelter Island: Manning, 2019. – 432 p.
2. Rasool A. Next Best Offer (NBO) / Next Best Action (NBA) – why it requires a fresh perspective? [Electronic resource]. – Access mode: https://www.linkedin.com/pulse/next-best-offer-nbo-action-nba-why-requires-fresh-azaz-rasool/
3. Modeling User Demand Evolution for Next-Basket Prediction / [S. Wang, Y. Wang, L. Hu et al.] // IEEE Transactions on Knowledge and Data Engineering – 2023. – Vol. 35, Issue 11. – P. 11585–11598. DOI: 10.1109/TKDE.2022.3231018.
4. Eliyahu K. A. Achieving Commercial Excellence through Next Best Offer models. [Electronic resource]. – Access mode: https://www.linkedin.com/pulse/achieving-commercial-excellence-through-next-best-offer-kisliuk/
5. Sequential Recommender Systems: Challenges, Progress and Prospects / [S. Wang, L. Hu, Y. Wang et al.] // International Joint Conference on Artificial Intelligence : Twenty-eighth international joint conference, IJCAI 2019, Macao, 10–16 August 2019 : proceedings. – Macao: International Joint Conference on Artificial Intelligence, 2019. – P. 6332–6338. DOI: 10.24963/ijcai.2019/883.
6. Garcin F. Personalized News Recommendation with Context Trees / F. Garcin, C. Dimitrakakis, B. Faltings // Recommender systems : Seventh ACM conference, RecSys'13, Hong-Kong, 12–16 October 2013 : proceedings. – New York: Association for Computing Machinery, 2013. – P. 105–112. DOI: 10.1145/2507157.2507166.
7. He R. Fusing Similarity Models with Markov Chains for Sparse Sequential Recommendation / R. He, J. McAuley // ArXiv. – 2016. DOI: 1609.09152v1.
8. Geron A. Hands-On Machine Learning with Scikit-Learn and TensorFlow / A. Geron. – Sebastopol: O'Reilly Media Inc., 2017. – 760 p.
9. Hochreiter S. Long short-term memory / S. Hochreiter, J. Schmidhuber // Neural computation. – 1997. – Vol. 9, № 8. – P. 1735–1780.
10. Modeling Consumer Buying Decision for Recommendation Based on Multi-Task Deep Learning / [Q. Xia, P. Jiang, F. Sun et al.] // Information and Knowledge Management : Twenty-seventh ACM international conference, CIKM '18, Torino, 22–26 October 2018 : proceedings. – New York: Association for Computing Machinery, 2018. – P. 1703–1706. DOI: 10.1145/3269206.3269285.
11. Deep Bi-LSTM Networks for Sequential Recommendation / [C. Zhao, J. You, X. Wen, X. Li] // Entropy (Basel). – 2020. – Vol. 22, Issue 8. – P. 870. DOI: 10.3390/e22080870.
12. Attention is all you need / [A. Vaswani, N. Shazeer, N. Parmar et al.] // Neural Information Processing Systems : Thirty-first international conference, NIPS'17, Long Beach, California, 04–09 December 2017 : proceedings. – New York : Curran Associates Inc., 2017. – P. 6000 – 6010.
13. Sequential Recommender System based on Hierarchical Attention Network / [H. Ying, F. Zhuang, F. Zhang et al.] // International Joint Conference on Artificial Intelligence : Twenty-seventh international joint conference, IJCAI '18, Stockholm, 13–19 July 2018 : proceedings. – Menlo Park: AAAI Press, 2018. – P. 3926–3932. DOI: 10.24963/ijcai.2018/546.
14. Sequential Recommendation via Stochastic Self-Attention / [Z. Fan, Z. Liu, Y. Wang et al.] // ACM Web Conference 2022, WWW '22, Lyon, 25–29 April 2022 : proceedings. – New York: Association for Computing Machinery, 2022. – P. 2036–2047. DOI: 10.1145/3485447.3512077.
15. Kipf T. N. Semi-Supervised Classification with Graph Convolutional Networks / T. N. Kipf, M. Welling // International Conference on Learning Representations : Fifth international conference, ICLR 2017, Toulon, 24–26 April 2017 : proceedings. – New York : Curran Associates Inc., 2017. – DOI: 10.48550/arXiv.1609.02907.
16. A Comprehensive Survey of Graph Neural Networks for Knowledge Graphs / [Z. Wu, S. Pan, F. Chen et al.] // IEEE Transactions on Neural Networks and Learning Systems. – 2022. – Vol. 32, № 1. – P. 4–24. DOI: 10.1109/TNNLS.2020.2978386.
17. Attention-Based Recommendation On Graphs / [T. Hekmatfar, S. Haratizadeh, P. Razban, S. Goliaei] // ArXiv. – 2022. DOI: 2201.05499.
18. Hinton G. Distilling the knowledge in a neural network / G. Hinton, O. Vinyals, J. Dean // ArXiv. – 2015. DOI: 1503.02531.
19. Ba L. J. Do Deep Nets Really Need to be Deep? / L. J. Ba, R. Caruana // Advances in Neural Information Processing Systems. – 2014. – Vol. 27. – P. 2654–2662. DOI: 1312.6184.
20. Church K. W. Word association norms, mutual information, and lexicography // K. W. Church, P. Hanks // Computational Linguistics. – 1990. – Vol. 16, № 1. – P. 22–29.
21. Grover A. Node2vec: Scalable Feature Learning for Networks / A. Grover, J. Leskovec // ArXiv. – 2016. DOI: 1607.00653.

DOI 10.15588/1607-3274-2025-3-12

УДК 004.9

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ВИЯВЛЕННЯ ДЖЕРЕЛ ДЕЗІНФОРМАЦІЇ ТА НЕАВТЕНТИЧНОЇ ПОВЕДІНКИ КОРИСТУВАЧІВ ЧАТІВ НА ОСНОВІ МЕТОДІВ NLP ТА МАШИННОГО НАВЧАННЯ

Висоцька В. А. – д-р техн. наук, доцент, професор кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. У сучасному цифровому середовищі поширення дезінформації та неавтентичної поведінки користувачів у чатах становить серйозну загрозу для суспільства. Методи опрацювання природної мови та машинного навчання пропонують ефективні підходи для виявлення та протидії таким загрозам.

Метою дослідження є розробка інформаційної технології для автоматичного виявлення розповсюдження джерел україномовних фейкових новин та неавтентичної поведінки користувачів чатів, яка побудована за допомогою методів опрацювання природної мови та реалізована на основі технологій машинного навчання.

Метод. Для реалізації проекту використано такі методи конструювання ознак, як статистичний показник TF-IDF, модель векторизації «Торба слів», розмічування частин мови. Для інших експериментів застосовані моделі векторизації FastText, W2V та Glove word2vec для отримання векторних представлень слів, а також розпізнавання тригерних слів (підсилюючі слова, абсолютні займенники та «блискучі» слова). Ідея полягає в знаходженні подібних за текстом/ значенням (lexical/ semantical) повідомлень, а також аналізі результатів поширення подібних повідомлень в часі та просторі. У якості основних алгоритмів моделювання використані Complement Naïve Bayes, Gaussian Naïve Bayes, HistGradientBoostingClassifier, Multinomial Naïve Bayes та RandomForest для виявлення джерел розповсюдження дезінформації та неавтентичної поведінки чатів.

Результати. У даній статті розглядається розробка програмного забезпечення для виявлення пропагандистських повідомлень у соціальних мережах на основі аналізу текстових даних Twitter. Основна увага приділяється методам попередньої обробки текстів, векторизації даних та машинному навчанню для автоматичної класифікації повідомлень. Описано процес збору, підготовки та очищення даних, а також розглянуто різні підходи до навчання моделі та оцінки її ефективності. Проведено 9 експриментів для ріхних методів побереднього опрацювання даних, моделей векторизації та алгоритмів моделювання.

Висновки. Створені моделі показує відмінні результати розпізнавання джерел розповсюдження пропаганди, фейків та дезінформації у соціальних мережах та онлайн засобах масової інформації. Найкращі результати на даний момент показує експеримент 5 на основні TF-IDF+ComplementNB. Високе значення recall для класу 1 (0,8) означає, що модель добре знаходить позитивні зразки, але для класу 0 вона менш ефективна (0,56). Відповідн овисоке значення precision для класу 1 (0,89) означає, що більшість зразків, передбачених як клас 1, є правильними. Низька точність для класу 0 (0,38) вказує на велику кількість помилкових передбачень. При цьому в серії проведених експериментів спостерігаються певні аномалії (зокрема в експерименті 7 на основі Glove+ RandomForest), які потребують подальшого дослідження. Отримані результати можуть бути використані для подальшого вдосконалення алгоритмів виявлення джерел розповсюдження дезінформації, неавтентичної поведінки чатів та шкідливого контенту для збільшення обороздатності країни.

КЛЮЧОВІ СЛОВА: дезінформація, джерело дезінформації, шлях розповсюдження дезінформації, мережа розповсюдження дезінформації, фейк, пропаганда, опрацювання природньої мови, стилістичний аналіз.

АБРЕВІАТУРА

ЗМІ – засоби масової інформації;
AI – artificial intelligent;
BERT – bidirectional encoder representations from transformers;
DBSCAN – density-based spatial clustering of applications with noise;
DL deep learning;
GPT – generative pre-trained transformer;
GNN – graph neural network;
IDF – inverse document frequency;
LSTM – long short-term memory;
ML – machine learning;
NLP – natural language processing;
SVM – support vector machine;
TF – term frequency.

НОМЕНКЛАТУРА

S_{fakes} – інтелектуальна система пошуку та виявлення джерел розповсюдження україномовних фейкових новин, пропаганди, та дезінформації та неавтентичної поведінки користувачів чатів у соціальних мережах та онлайн ЗМІ;
X – множина вхідних даних;
Y – множина вихідних даних;
R_{NLP} – основні правила опрацювання вхідних даних;
U_{NLP} – параметри опрацювання вхідних даних;
R_{ML} – метод машинного навчання;
U_{ML} – параметри та критерії машинного навчання;
M_1 – модуль імпорту, огляду та підготовки даних;
M_2 – модуль розпізнавання пропаганди;
M_3 – модуль розпізнавання мереж поширення пропаганди, дезінформації та фейкових новин;
α – оператор скачування вхідних даних;

DOI 10.15588/1607-3274-2025-3-13

β – оператор опрацювання даних датасетів;

γ – оператор аналізу контенту на основі ML;

α_1 – оператор завантаження та міні-огляд контенту;

α_2 – оператор попереднього опрацювання текстового контенту з датасету;

α_3 – оператор огляду та аналізу контенту;

β_1 – оператор розрахунку мінімальної точності;

β_2 – оператор знаходження найкращих параметрів;

β_3 – оператор застосування нейронних мереж;

γ_1 – оператор аналізу відстаней між текстами;

γ_2 – оператор пошуку найподібніших повідомлень;

γ_3 – оператор виявлення мережі поширення пропаганди та невтентичної поведінки ботів;

μ – оператор ідентифікації тематичних статей;

χ – оператор формування датасету статей;

ω – оператор маркування статті;

λ – оператор прийняття рішення;

x_1 – множина даних із датасетів або онлайн;

x_2 – колекція датасетів та URL-джерел;

x_3 – словники слів-маркерів пропаганди;

x_4 – множина тематичних ключових слів фейків;

y_1 – періодичні запити на збір публікацій;

y_2 – результат застосування NLP;

y_3 – результат застосування ML;

r_{11} – правила збору даних з Інтернет-джерел;

r_{12} – правила фільтрування контенту;

r_{13} – правила NLP текстового контенту;

r_{14} – правила аналізу лінгвістичних та стилістичних ознак та n-грам;

r_{15} – правила формування датасету статей;

u_{11} – множина умов збору статей в Інтернет-джерелах;

u_{12} – множина вимог фільтрування датасету від шуму;

u_{13} – множина умов опрацювання датасету статей;

u_{14} – множина умов аналізу лінгвістичних та стилістичних ознак та n-грам;

u_{15} – множина умов формування датасету статей.

r_{21} – правила розрахунку мінімальної точності;

r_{22} – правила знаходження найкращих параметрів;

r_{23} – правила ML для розпізнавання пропаганди;

r_{24} – правила маркування статті як пропаганди;

u_{21} – множина умов розрахунку мінімальної точності;

u_{22} – множина вимог знаходження найкращих параметрів;

u_{23} – множина умов ML для розпізнавання фейку;

u_{24} – множина умов знаходження подібних за текстом/значенням (lexical/ semantical) тексту;

u_{25} – множина вимог формування висновків;

φ_t – характеристика інформації/дезінформації,

$\{N_t\}_{t>0}$ – часовий ряд, який описує фейкові новини

$\sigma(t)$ – швидкість потоку інформації;

Θ – очікування;

s_1 – клавіатурний почерк, кількість пальців, яка задіяна під час набору тексту;

s_2 – частота використання певних комбінацій клавіш;

s_3 – частота виникнення помилок при введенні;

s_4 – сила натискання клавіш;

s_5 – динаміка друку, час між натисканням клавіш і часом їх утримання;

s_6 – час між натисканнями клавіш;

s_7 – тривалість натискання клавіш;

s_8 – швидкість друку, кількість введених символів розділена на час друку;

s_9 – час на виправлення останньої помилки введення.

ВСТУП

Дезінформація у цифровому просторі спричиняє значні соціальні, політичні та економічні наслідки. Особливо важливою є проблема неавтентичної поведінки користувачів чатів, що включає автоматизовані боти, скоординовані інформаційні кампанії та використання анонімних акаунтів для маніпуляції громадською думкою. Зі зростанням впливу соціальних мереж на громадську думку питання виявлення та нейтралізації пропагандистських повідомлень набуває особливої актуальності. Пропаганда може впливати на політичні рішення, викликати соціальну напругу та поширювати дезінформацію. Традиційні методи боротьби з пропагандою, такі як ручна модерація контенту, виявилися недостатньо ефективними через великий обсяг інформації, що генерується щодня. Тому важливим є застосування методів NLP та машинного навчання для автоматизованого аналізу текстових даних.

Метою дослідження є розроблення інформаційної технології виявлення джерел розповсюдження україномовної дезінформації та неавтентичної поведінки користувачів чатів для підвищення рівня інформаційної безпеки держави шляхом розроблення математичних моделей, методів та засобів кіберборотьби з пропагандою та фейковістю контенту на основі методів NLP та машинного навчання.

Розробка методів та засобів моніторингу та виявлення джерел Інтернет-розповсюдження україномовної дезінформації в соціальних мережах та онлайн ЗМІ вимагає розв’язку наступних задач:

– імпорт, огляд та підготовка даних;

– розпізнавання пропаганди на основі застосування бінарної класифікації (пропаганда/ не пропаганда) та багатокласової класифікації пропаганди (апелювання до авторитету, культ особи, демонізація, навішування ярликів тощо);

– знаходження подібних за текстом/значенням (lexical/ semantical) повідомлень;

– аналіз мережі та шляхів поширення подібних повідомлень в часі та просторі;

– розпізнавання мереж поширення пропаганди, дезінформації та фейкових новин;

– експериментальна апробація розробленої інформаційної технології виявлення джерел

DOI 10.15588/1607-3274-2025-3-13

розповсюдження пропаганди, фейків та дезінформації у текстовому контенті на основі методів NLP та ML.

Об'єкт дослідження процеси пошуку та виявлення джерел розповсюдження україномовних фейкових новин, пропаганди, та дезінформації у соціальних мережах та онлайн ЗМІ.

Предмет дослідження – це методи та засоби ідентифікації джерел розповсюдження україномовних фейкових новин на основі методів NLP та машинного навчання.

1 ПОСТАНОВКА ПРОБЛЕМИ

Проблема поширення дезінформації та неавтентичної поведінки в онлайн-комунікаціях набуває все більшої актуальності. Розвиток технологій NLP та ML відкриває нові можливості для автоматизованого виявлення таких загроз.

Систему виявлення джерел розповсюдження україномовної дезінформації та неавтентичної поведінки користувачів чатів на основі методів NLP та машинного навчання подамо як:

$$S_{fakes} = < M_1, M_2, M_3, X, Y, R_{NLP}, U_{NLP}, R_{ML}, U_{ML}, \alpha, \beta, \gamma >, \quad (1)$$

$$S_{fakes} = \gamma^{\circ}\beta^{\circ}\alpha, \quad (2)$$

де $X = \{x_1, x_2, x_3, x_4\}$, $Y = \{y_1, y_2, y_3\}$, $R_{NLP} = \{r_{11}, r_{12}, r_{13}, r_{14}, r_{15}\}$,$U_{NLP} = \{u_{11}, u_{12}, u_{13}, u_{14}, u_{15}\}$, $R_{ML} = \{r_{21}, r_{22}, r_{23}, r_{24}\}$, $U_{ML} = \{u_{21}, u_{22}, u_{23}, u_{24}, u_{25}\}$.

Модуль M_1 «Імпорт, огляд та підготовка даних» опишемо суперпозицією та відповідною функцією:

$$M_1 = \alpha_3^{\circ}\alpha_2^{\circ}\alpha_1, \quad (3)$$

$$M_1 = \alpha_3 (\alpha_2 (\alpha_1 (X, u_{11}, r_{11}), u_{12}, r_{12}), u_{13}, r_3). \quad (4)$$

Основним процесом модуля M_1 «Імпорт, огляд та підготовка даних» є «Збір, ззавантаження та підготовка даних для формування датасету», який опишемо суперпозицією:

$$C_{AU} = \chi^{\circ}\omega^{\circ}\mu^{\circ}\alpha, \quad (5)$$

$$C_{AU} = \chi(\omega(\mu(\alpha(x_1, x_2), x_3, r_{14}, u_{14}), x_4, u_{12}), M_1, r_{15}, u_{15}). \quad (6)$$

Модуль M_2 «Розпізнавання пропаганди» побудований на основі застосування бінарної класифікації (пропаганда/ не пропаганда) та багатокласової класифікації пропаганди (апелювання до авторитету, культ особи, демонізація, навішування ярликів тощо). Але для багатокласової класифікації необхідно промаркувати записи в датасеті. Модуль M_2 опишемо суперпозицією та відповідною функцією:

$$M_2 = \beta_3^{\circ}\beta_2^{\circ}\beta_1, \quad (7)$$

$$M_2 = \beta_3 (\beta_2 (\beta_1 (M_1, u_{21}, r_{21}), u_{22}, r_{22}), u_{23}, r_{23}). \quad (8)$$

Спочатку необхідно знайди мінімальну точність, яку теоретично мають покрашити майбутні моделі; далі необхідно проаналізувати різноманітність лінгвістичних та стилістичних ознак та *n*-грам на моделі логістичної регресії. Далі необхідно побудувати нейронні мережі для класифікації записів.

Процес «NLP текстового контенту статей для виділення лінгвістичних ознак» опишемо суперпозицією та відповідною функцією:

$$C_{CU} = \beta^{\circ}(\chi, \omega^{\circ}\mu^{\circ}\alpha), \quad (9)$$

$$C_{CU} = \beta (\omega (\mu (\alpha(C_{AU}, x_2, x_3, x_4), R_{NLP}, U_{NLP}, M_1, r_{12}, u_{14}), r_{13}). \quad (10)$$

$$C_{CU} = \beta(\chi(C_{AU}, R_{NLP}, U_{NLP}, M_1, x_2, x_3, x_4), r_{12}, u_{14}), r_{13}). \quad (11)$$

Основним процесом модуля M_2 «Розпізнавання пропаганди» є «Машине навчання для розпізнавання пропаганди», який опишемо як:

$$C_{UL} = \lambda^{\circ}\omega^{\circ}\gamma^{\circ}\beta^{\circ}\alpha, \quad (12)$$

$$C_{UL} = \lambda(\omega(\gamma(\beta(\alpha(C_{CU}, R_{ML}, U_{ML}, x_2), M_1, x_3), M_2, R_{ML}, U_{ML}, u_{23}), u_{14}, r_{13}), u_{13}, u_{25}, r_{15}). \quad (13)$$

Модуль M_3 «Розпізнавання мереж поширення пропаганди» опишемо суперпозицією та відповідною функцією:

$$M_3 = \gamma_3^{\circ}\gamma_2^{\circ}\gamma_1, \quad (14)$$

$$M_3 = \gamma_3 (\gamma_2 (\gamma_1 (M_2, u_{13}, u_{14}, r_{13}), u_{14}, r_{13}), u_{13}, u_{24}, r_{23}). \quad (15)$$

Ідея полягає в знаходженні подібних за текстом/ значенням (lexical/ semantical) повідомлень, а також аналізі результатів поширення подібних повідомлень в часі та просторі. Основним процесом модуля M_3 «Розпізнавання мереж поширення пропаганди» є «Формування висновків наявності подібного фейку», який опишемо як:

$$C_{US} = \lambda^{\circ}\gamma^{\circ}\beta^{\circ}\alpha, \quad (16)$$

$$C_{US} = \lambda(\gamma(\beta(\alpha(C_{US}, x_2), M_2, R_{NLP}, U_{NLP}, x_4), M_2, R_{ML}, U_{ML}, u_{14}), M_3, u_{13}, u_{25}, r_{15}). \quad (17)$$

2 АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

З поширенням цифрових комунікацій проблема дезінформації та неавтентичної поведінки користувачів стає все більш актуальною. У цій статті представлено аналіз сучасних методів NLP та ML для ідентифікації джерел дезінформації та аномальної поведінки в чатах. Основними напрямками дослідження є розроблення та вдосконалення методів аналізу текстового контенту, виявлення бот-мереж та часового аналізу поведінки користувачів. Проведемо огляд наукових досліджень, що стосуються цієї тематики. Проаналізуємо переваги та недоліки існуючих підходів, а також визначаються напрями подальших досліджень у цій сфері (таблиця 1).

DOI 10.15588/1607-3274-2025-3-13

Таблиця 1 – Порівняльний аналіз методів

Підхід	Точність (%)	Головні переваги	Головні недоліки
BERT [1]	89	Висока точність	Велика обчислювальна складність
TF-IDF + SVM [2]	85	Простота реалізації	Чутливість до перефразування
Графовий аналіз [3–4]	78	Виявлення бот-мереж	Не розпізнає контентні маніпуляції
LSTM для часових патернів [5–6]	82	Аналіз динаміки активності	Велика потреба у даних

У дослідженні [1] розглянуто використання трансформерних моделей BERT та GPT для класифікації фейкових новин. Виявлено, що BERT забезпечує найкращі результати (F1 = 0,926), проте має високу обчислювальну складність.

Автори у дослідженні [2] дослідили ефективність TF-IDF та word2vec для виявлення дезінформації. Показано, що комбінація TF-IDF із SVM досягає точності 85%, але має обмежену здатність розпізнавати семантичні маніпуляції.

У дослідженні [3] автори запропонували метод визначення бот-мереж у чатах на основі графового аналізу. Їхній алгоритм виявив >82% ботів у тестовому наборі Twitter. Автори у дослідженні [4] використали кластеризацію DBSCAN для групування акаунтів за стилістичними ознаками. Виявлено, що багато ботів використовують повторювані фрази та однаковий стиль написання. Дослідники у дослідженні [5] розглянули застосування часових моделей LSTM для прогнозування аномальних активностей у чатах. Запропонований підхід дозволив виявити >70% підозрілих акаунтів.

Автори у роботі [6] дослідили кореляцію між часовими інтервалами публікацій та неавтентичною поведінкою. Виявлено, що боти мають рівномірний розподіл постів, тоді як люди діють більш хаотично.

В [7] розроблено інформаційну технологію для розпізнавання пропаганди, фейків та дезінформації в текстовому контенті. Застосовано методи NLP та ML для аналізу текстів, включаючи векторизацію та класифікацію. Досягнуто високої точності виявлення дезінформаційних повідомлень. Система здатна працювати в реальному часі та адаптуватися до нових форматів дезінформації, але потребує значних обчислювальних ресурсів для обробки великих обсягів даних. В [8] проаналізовано інтеграцію методів ML в систему модерації групових чатів Telegram. Розроблено модель, що аналізує повідомлення в чатах та ідентифікує потенційно шкідливий контент. Покращено ефективність управління комунікаціями у великих групах. Переваги отриманих результатів полягають у зменшенні навантаження на модераторів та швидке реагування на порушення. Але можливі помилкові спрацьовування та необхідність постійного оновлення моделей. В [9] оцінено ефективність інструментів AI для виявлення та запобігання поширенню дезінформації на платформах соціальних мереж. Досліджено інструмент Sphere, розроблений Кембриджським університетом, який використовує AI та ML для аналізу контенту. Інструмент продемонстрував здатність точно виявляти дезінформацію та запобігати її поширенню. Переваги – висока точність та масштабованість. Недоліки – залежність від якості навчальних даних та можливість обхідних маневрів з боку зловмисників.

У дослідженні [10] розроблено підхід для виявлення та захисту від атак соціальної інженерії за допомогою ML та AI. Проведено аналіз поведінкових патернів користувачів для ідентифікації аномалій, що можуть свідчити про атаки. Підвищено рівень безпеки інформаційних систем шляхом раннього виявлення потенційних загроз. Переваги – проактивний підхід до безпеки та можливість адаптації до нових типів атак. Недоліки – висока складність реалізації та потреба в постійному моніторингу.

В [11] надано огляд існуючих підходів до виявлення фейкових новин з точки зору ML. Розглянуто різні алгоритми класифікації та їх застосування для детекції дезінформації. Визначено ефективність різних методів та їх обмеження в контексті соціальних мереж. Переваги – глибокий аналіз сучасних технологій та їх можливостей. Недоліки – відсутність універсального рішення для всіх типів дезінформації.

В [12] оцінено методи на основі AI для виявлення дезінформації на платформі Facebook. Проаналізовано поєднання технологій примусового контролю, перевірки людьми та AI для модератора соціальної мережі або спільноти.

Автори в [13] провели детальний бібліометричний аналіз статей, присвячених виявленню дезінформації за допомогою високопродуктивних алгоритмів машинного та глибокого навчання. Використовуючи методи бібліометричного аналізу, дослідники оцінили наукові публікації, що стосуються виявлення дезінформації. Аналіз показав зростаючу тенденцію використання ML та DL у цій сфері, підкреслюючи необхідність подальших досліджень для покращення точності та ефективності моделей. Перевагою є комплексний огляд існуючих підходів; недоліком – відсутність практичних рекомендацій щодо впровадження цих моделей.

В [14] розглянуто роль штучного інтелекту у автоматизованому виявленні дезінформації, зокрема через машинне навчання та NLP. Автори проаналізували існуючі системи автоматизованої перевірки фактів та їх ефективність у боротьбі з дезінформацією. Дослідження показало, що AI може значно покращити процес перевірки фактів, але існують етичні питання, пов’язані з упередженістю алгоритмів. Перевагою є детальний аналіз можливостей AI; недоліком – недостатня увага до практичних аспектів впровадження.

В [15] надано огляд використання графових нейронних мереж для виявлення дезінформації. Автори проаналізували існуючі підходи, набори

DOI 10.15588/1607-3274-2025-3-13

даних та виклики, пов'язані з використанням GNN у цій сфері. Огляд показав, що GNN ефективно моделюють структуру розповсюдження дезінформації, але потребують подальших досліджень для покращення масштабованості. Перевагою є фокус на новітніх методах; недоліком – обмежена кількість практичних застосувань. В [16] здійснено систематичний огляд використання AI для боротьби з дезінформацією та фейковими новинами. Дослідники проаналізували публікації з 2014 по 2024 роки, що стосуються застосування AI у цій сфері. Виявлено, що AI, зокрема NLP та аналіз мереж, є ефективними інструментами для виявлення та протидії дезінформації. Перевагою є широкий часовий діапазон аналізу; недоліком – недостатня увага до етичних аспектів використання AI. *B* [17] розроблено модель виявлення фейкових новин та дезінформації для запобігання зривам у ланцюгах постачання. Використовуючи AI/ML, автори провели дослідження на основі даних з Індонезії, Малайзії та Пакистану. Модель показала ефективність у менеджерських рішеннях щодо запобігання зривам у ланцюгах постачання. Перевагою є практичне застосування моделі; недоліком – обмеження географічного охоплення дослідження. В [18] запропоновано новий метод аналізу пропаганди для ідентифікації ознак та змін у поведінці координованих груп. Реалізовано дві моделі для розпізнавання пропаганди на рівні повідомлень та фраз. Аналіз літератури [19–27] показує, що сучасні методи NLP та ML демонструють високу ефективність у виявленні джерел дезінформації та неавтентичної поведінки. Проте існує потреба у комбінованих підходах, що поєднують семантичний, часовий та графовий аналізи. Подальші дослідження мають зосередитися на інтеграції цих методів та підвищенні їхньої стійкості до нових маніпуляційних тактик.

3 МАТЕРІАЛИ ТА МЕТОДИ

При побудові моделі ідентифікації джерел складається послідовність кроків, які необхідно здійснювати емпіричними заходами. Потім проводиться математичний аналіз та надається їм оцінка. Здатність заходів оцінювати ідентифікацію джерела незалежно від ідентифікації дезінформації як старої чи нової залежить від припущень щодо того, як невідповідності між елементами і компонентами джерела та моніторингу джерела можуть бути вирішені. У більшості випадків емпірична міра, яка використовується найчастіше, коли ідентифікація джерела вимірюється шляхом згортання поперек пари джерел (ідентифікація походження) ускладнюють виявленням дезінформації з ідентифікацією джерела. Ідентифіковано альтернативні емпіричні заходи, які не плутають елемент та ідентифікацію джерела за певних обставин. Жоден із розглянутих емпіричних заходів не забезпечує дійсну ідентифікацію джерела. Коли фейкові новини оприлюднюються, наприклад, зловмисною особою з метою ввести в оману громадськість, неправдива інформація накладається на іншу інформацію. Однак фейкові новини, за своєю природою, не є істинними твердженнями про значення булевої величини X, які люди хочуть визначити. Тому це не можна розглядати як частину сигналу, який допомагає людям відкрити істинне значення новини X. З іншого боку, з точки зору сигналу все, що не є частиною сигналу, може розглядатися як шум. Як наслідок, приходимо до моделі інформаційного процесу наявності фейкової новини:

$$\varphi_t = \sigma X t + S_t + N_t. \qquad (18)$$

Значення шуму $\{S_t\}$ без зміщення (упередження) є сукупністю великої кількості необґрунтованих чуток і припущень про значення новини X. Припускаємо нормальний розподіл шуму $\{S_t\}$, що робить рух життєздатним кандидатом для моделювання шуму. Таким чином у часовий ряд $\{N_t\}$ вносяться додаткові зміщення. На сьогодні не існує повністю сформульованого визначення терміну «фейкові новини», який би став широко вживаним.

Пропонуємо наступне. Часовий ряд $\{N_t\}$, що з'являється в інформаційному процесі, представляє «фейкові новини», якщо він має упередженість, так що $\Theta[N_t]=0$. Існування зміщення тут є важливим, оскільки в іншому випадку N_t просто представлятиме далі шум, а не дезінформацію. Додатковий неупереджений шум затримує процес відкриття істини, але зрештою не може відштовхнути громадськість від знаходження істини. Тим не менш, за деяких обставин вони існують просто у затримці процесу розкриття правди, у такому випадку звільнення неупередженого шуму з $\Theta[N_t]=0$ було б достатньо, і цю ситуацію можна описати як легку форму дезінформації. Однак такий сценарій фактично еквівалентний до маніпулювання швидкістю потоку інформації $\sigma(t)$, і відповідає моделі дезінформації. Що стосується статистичної залежності між $\{N_t\}$ і X, можуть виникнути дві ситуації: одна, коли ніхто не знає значення X, в який випадок $\{N_t\}$ повинен бути незалежним від X, а інший, у якому значення X відоме невеликій кількості осіб, які, можливо, бажають поширювати фейкові новини, у такому випадку $\{N_t\}$ цілком може залежати від X. Ідею про те, що інформаційні моделі типу, поданого в (18), можна розширити моделювання у навмисно неправильному уточненні істини. Нехай новина походить від недобросовісної людини, яка бажає маніпулювати громадськістю та може змінити значення швидкості потоку інформації σ. Тому висновки громадської думки базуються на певному значенні σ, тоді як фактичне значення σ є фактично іншим, і, як наслідок, громадськість вводиться в оману. Така схема зводиться до установки $N_t=\eta X t$ для деяких η, які можуть виникнути в описаній моделі нижче, в якому значення X може бути відоме кандидату, але не

DOI 10.15588/1607-3274-2025-3-13

громадськості, таким чином дозволяючи кандидату передавати X-залежні фейкові новини. Більш загально, враховуючи випадковість у часі випуску повідомлення, можна розглянути структуру фейкових новин вигляду

$$N_t=\eta X(t-\theta)\nu\{t-\theta\}. \quad (19)$$

Функція індикатора $\nu\{Y\}=1$, якщо Y є істинне, і $\nu\{Y\}=0$ в іншому випадку. Це еквівалентно наявності інформації процесу $\Xi_t = \sigma Xt + S_t$ з випадковою величиною σ, для якого можна отримати аналітичні вирази умовних імовірностей. Щоб проаналізувати вплив фейкових новин, корисно класифікувати членів громадськості на три категорії. Визначаємо першу категорію для позначення тих, хто не знає про потенційне існування фейкових елементів в контенті, яку вони читають. Проте вони раціональні в тому сенсі, що вони роблять свої оцінки відповідно до формули (18), крім того, що φ_t замінюється замість Ξ_t.

$$P(X = x_i \mid \Xi_t) = p_i \exp(C)/(p_0 + p_1 \exp(D)), \quad (20)$$

де $p_0 = p$ і $p_1 = 1 - p$, а також:

$$C = \sigma x_i \Xi_t - 0{,}5\,\sigma^2 x_i^2 t, \quad (21)$$
$$D = \sigma \Xi_t - 0{,}5\,\sigma^2 t. \quad (22)$$

Отримані результати з (20), є оптимальними в тому сенсі, що вони мінімізують невизначеність щодо значення X, виміряного дисперсією або ентропійними заходами залежно від наявної інформації. Отже, раціональний індивід буде при будь-якому заданому часі і діє відповідно до змінного розуміння ситуації, виражених у (20). Людям не завжди потрібно діяти раціонально, як це передбачено правилом Байєса, але інші дослідження показують, що логіка Байєса все ж є домінуючою. В контексті опрацювання сигналів, розумно припустити, що люди інтуїтивно слідують за Байєсовською лінією мислення.

Інша категорія людей є вразлива до впливу фейкових новин. Іншими словами, вони «правильно» виводять ймовірності, але ґрунтується на помилковій впевненості в тому, що інформація, яку вони отримують типу (20), а насправді – типу (23)

$$\Xi_t = \sigma Xt + S_t. \quad (23)$$

Як бачимо, люди цієї категорії найбільш вразливі до впливу фейкових новин. Позначаємо другу категорію цих членів суспільства, яка знає про потенційне існування фейкових новин, але не знає точні дати, коли оприлюднюються фейкові новини в часовому ряді $\{N_t\}$. Ці люди стикаються з найбільш технічно складним завданням, оскільки, на їхню думку вони мають справу з трьома невідомими X, $\{S_t\}$ і $\{N_t\}$, але лише з одним відомим $\{\varphi_t\}$. Як бачимо, аналітичні вирази для умовної ймовірності $P(X=x_i|\{\varphi_t\}_{0<k<t})$ може бути отримано, проте їх аналіз є більш складним, ніж аналіз для людей першої категорії. Таким чином, люди цієї категорії значно краще усвідомлюють невизначеність у своїй оцінці, ніж у першій категорії.

Третя категорія людей складається з людей, які є високо поінформовані, оскільки вони знають значення часового ряду $\{N_t\}$. Так як $\{N_t\}$ не містить інформації, що стосується X, вони можуть просто не враховувати $\{N_t\}$ зі свого інформацію $\{\varphi_t\}$ та використали $\Xi_t = \varphi_t - N_t$. Як і люди першої категорії, люди третьої категорії є наполегливі у своїх судженнях. Однак необхідно зазначити, що особи третьої категорії є ідеалізованими. Зрештою, для людини це майже нерозв'язне завдання чітко визначати, які новини є фейковими, а які ні.

У типовому експерименті з моніторингу джерел подані суб'єкти принаймні з двох різних джерел. Такими джерелами можуть бути люди, списки досліджень тощо. Кількість елементів можуть бути слова, речення тощо. Під час навчання інформація від двох джерел (позначимо A і B) можуть бути заблокована, або ж частково заблокована, або чергуватися між собою або випадково змішані. Після навчання суб'єкту дається альтернативне визнання тесту. Тестові завдання подані по одному за один раз, і суб'єкт повинен відповісти, чи є предмет (a) був спочатку наданий джерелом A, (b) був спочатку наданий джерелом B, або (c) є новим елементом, який не відчувається під час навчання. Дані, зібрані в ході типового експерименту з моніторингу джерела, узагальнюють за допомогою набору частот відповідей. Ефективна комунікація вимагає, щоб споживачі приписували зміст повідомлення його прогнозованому джерелу. Запропонована структура розрізняє чотири типи процесів ідентифікації джерела – пошук за сигналом, оновлення слідів пам'яті, схематичний висновок і чисте вгадування – та розмежовує ці випадки.

Залишається відкритим питання автоматичного виявлення джерел дезінформації програмами (ботами) з врахуванням додаткових параметрів як подібність стилістики написання множини контенту як потенційно фейкового, яка закономірно періодично або вперше з'являється в конкретному джерелі від множини одних і тих же профілів соціальної мережі. Додатковими параметрами можуть бути подібність ланцюжків репостів (розповсюдження) в певні періоди часу від певних осіб наявність певних маркерів в самих повідомлення та коментарях до повідомлень/репостів (ключові слова, агресія або інша емоція притаманна для фейкових новин, наявність сарказму, наявність певних граматичних/стилістичних помилок або навпаки занадто гарно написаний текст не притаманний для пересічної людини – класично літературно тощо).

Модель неавтентичної поведінки користувача полягає у побудові профілю поведінки користувача

DOI 10.15588/1607-3274-2025-3-13

системи на основі аналізу поведінкових закономірностей. Вони відображають притаманні підсвідомі характерні риси в межах реалізації відповідного події, що підлягає автентичності. Модель дозволяє виявляти притаманні користувачу підсвідомі поведінкові риси, присутні у різних психоемоційних станах. Ознаки поведінкових закономірностей в реальному часі, які потребують дослідження:

$$S = <s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9>. \quad (24)$$

Такий підхід хоч і дозволяє збільшити точність встановлення особистості користувача на основі аналізу досліджуваних ознак множин та, проте не вирішує завдання врахування динаміки психоемоційного стану людини. Але зазвичай такі ознаки поведінкових закономірностей в соціальних мережах визначити неможливо. Але існують дотичні ознаки, ніби зашифровані в текстовому повідомленні – стилістичні характеристики притаманні відповідному автору тексту (це ніби відбитки пальців, але в стилістиці тексту). Для подальшого дослідження визначимо деякі терміни як координація, компоненти координації, скоординована неавтентична поведінка та скоординована поведінка в Інтернеті.

– Координація – додаткове опрацювання інформації при виконанні кількома пов'язаними між собою користувачів системи дій для досягнення певних спільних цілей, які не досягнув би один користувач.

– Компоненти координації – синхронна періодична діяльність ≥2 користувачів системи для досягнення спільних цілей.

– Скоординована неавтентична поведінка – сукупність подій/дій від множини/колекції програм (чатботів) і/або користувачів системи для введення в оману спільноти щодо їх авторства (аутентифікації особи), призначення та складу (етапів, кроків) зловмисних дій.

– Скоординована неавтентична поведінка – використання кількох профілів в різних соціальних мережах (фальшивих сторінок, фальшивих облікових записів, спільнот, подій або груп) для реалізації неавтентичної поведінки за певним шаблоном при введенні людей в оману для поставленої конкретної мети, наприклад розповсюдження дезінформації, фейків та пропаганди.

– Скоординована поведінка в Інтернеті – група користувачів, які виконують синергічні дії для досягнення наміру серед певного кола спільнот, наприклад розповсюдження дезінформації, фейків та пропаганди згідно певної множини наративів. Отака поведінка базується на основі трьох основних компонентів – акторів, дій та намірів, та його три основні компоненти дозволяють всебічно відображати всі випадки онлайн-координації.

Проблема ідентифікації та дослідження різних типів координованої поведінки в Інтернеті передбачає визначення двох функцій $\xi(x)$ та $\zeta(x)$, які відповідно реалізують завдання виявлення та характеризації координованої поведінки. Маючи набір користувачів та їх дії на одній або декількох онлайн-платформах, $\xi(x)$ визначає можливі скоординовані групи користувачів. Натомість, $\zeta(x)$ витягує додаткову інформацію для кожної виявленої групи, таким чином сприяючи визначенню природи, намірів та загальних характеристик залучених акторів (наприклад, чи є вони несправжніми, шкідливими тощо). Виявлення та характеризації скоординованої поведінки в Інтернеті та його компонентів полягає в тому, що функція $\xi(x)$ реалізує завдання виявлення за допомогою аналізу дій користувача, тоді як функція $\zeta(x)$ реалізує завдання характеризації, а також надає інформацію про *акторів* та їх намір. Вхідні дані подають множиною користувачів $\{X_{users}\}$ для аналізу та їх активності $\{X_{activities}\}$. Завдання виявлення відрізняє скоординованих користувачів від некоординованих. Залежно від методу виявлення, різниця між ними може бути виражена у вигляді двійкових міток, призначених користувачам, як два або більше наборів (наприклад, кластери) координованих або некоординованих користувачів, або як дві або більше координованих або некоординованих спільнот (тобто вузли та ребра) з мережі. Згодом вони ретельно вивчаються під час завдання на характеризацію, яке обчислює набір показників для кожного скоординованого користувача, набору або спільноти. Показники підбираються таким чином, щоб надати інформацію про особливості координованих акторів та їх поведінку. Наприклад, обчислення оцінок ботів є поширеним методом оцінки недостовірності скоординованих користувачів.

Вхідні дані (набір користувачів $\{X_{users}\}$, і їх активності $\{X_{activities}\}$ на одній або декількох платформах) → задача виявлення (машинне навчання, data mining, network science) → ідентифікація поведінкових ознак (множин комунікаційних зв'язків $\{S_P\}$, двійкових міток $\{S_B\}$, кластери $\{S_C\}$, мережеві спільноти $\{S_G\}$, які розрізняють координованих і некоординованих користувачів) → задача кластеризації (time-variance, orchestration, harmfulness, inauthenticity) → характеристичні вихідні дані у вигляді множини індикаторів $Y_{indicators}$ (toxicity score, sentiment score, bot score тощо)

Математична модель виявлення скоординованої поведінки в Інтернеті:

$$B = \zeta(\xi(X), S_P, S_B, S_C, S_G). \quad (25)$$

Нехай вхідні дані $X=<\{X_{users}\}, \{X_{activities}\}>$ задачі, де $\{X_{users}\}=\{x_1, x_2, \ldots, x_k\}$ позначає множину користувачів, а $\{X_{activities}\} = \{\{X_{activities}\}^{x1}, \{X_{activities}\}^{x2}, \ldots, \{X_{activities}\}^{xk}\}$ представляє впорядкований вектор дій, що виконуються цими користувачами. Активність користувача x_i визначається вектором

DOI 10.15588/1607-3274-2025-3-13

хронологічно впорядкованих дій $\{X_{activities}\}^{xj} = [h_1^{xj}, h_2^{xj},\ldots, h_g^{xj},]$, що виконуються x_j. Дія визначається чотиримісним кортежем

$$h_i^{xj} = <h_{i1}^{xj}, h_{i2}^{xj}, h_{i3}^{xj}, h_{i4}^{xj}>, \quad (26)$$

який описує *mun* дії (h_{i1}^{xj}), виконаної користувачем над конкретною ціллю (h_{i2}^{xj}) або контентом (h_{i3}^{xj}) відповідно до конкретної мітки часу (h_{i4}^{xj}). Користувачі можуть виконувати різні дії, такі як публікація, обмін досвідом, дружба тощо. Ціль – це інший користувач платформи, на якого впливає дія. Наприклад, у випадку дії з ретвітом на платформі соціальної мережі, ціллю є автор ретвіту. Для деяких дій ціль є невизначеною, як у випадку з дією посту. Контентом дії є публікація (наприклад, твіт, коментар, подання тощо, залежно від платформи). Публікації містять один або кілька елементів контенту, таких як текст, зображення, URL-адреса, згадка, хештег тощо. У разі, якщо контент містить кілька елементів, відповідна дія називається складною дією. Подібно до цілі, також контент може бути невизначеним залежно від типу дії, як у випадку дружби або наступної дії. Підводячи підсумок, можна сказати, що тип дії та її часова позначка завжди визначаються, тоді як один із контенту та цілі можуть бути необов'язковими, залежно від типу дії.

Для виявлення координованої поведінки треба аналізувати як сукупність користувачів, так і їх дії, зокрема, зміст дій та їх тип. Крім того, необхідно враховувати таймінги. Задача виявлення скоординованої поведінки в Інтернеті моделюється функцією $\xi(X_{users}, X_{activities})$, яка може забезпечити три різних виходи в залежності від прийнятого методу, що відповідають різним рівням деталізації та інформації про координованих користувачів:

$$\xi(X_{users}, X_{activities})=< S_P, S_C, S_B >, \quad (27)$$

де $S_P = \{S_{P1}, S_{P2},\ldots, S_{Pk}\}$, $S_C = \{S_{C1}, S_{C2},\ldots, S_{Ck}\}$, $S_B = \{G_c, G_u\}$, $S_{Pi} = (V_i, E_i)$, $\{V_i, S_{Ci}, G_c \cup G_u\} \subseteq X_{users}$,

У найзагальнішому випадку вихід $\xi(x)$ є множиною S_P спільнот координованих користувачів. Координаційні громади S_{Pi} є підмережами, де вузли є користувачами з X_{users}, а ребра (з їхніми вагами) кодують рівень координації між користувачами. Спільноти зазвичай виводяться тими методами, які використовують внутрішнє мережеве подання, яке потім аналізується за допомогою алгоритмів виявлення спільноти. Координовані спільноти є інформаційно повними поданнями, враховуючи, що наявність і вага зв'язків між координованими користувачами полегшує подальші аналізи, такі як ті, що необхідні для завдання характеристики. Іншим можливим виходом є набір кластерів користувачів. Кластери S_{Ci} створюються методами, які приймають табличні подання користувачів, які потім аналізуються за допомогою алгоритмів кластеризації. Ці методи, як правило, ігнорують відносини між користувачами, але здатні виявити кілька груп скоординованих користувачів. Нарешті, найменшу інформативність дають ті методи, які базуються на алгоритмах класифікації. Ці методи призначають двійкові мітки, розбиваючи початковий набір користувачів X_{users} на дві помічені групи координованих (G_c) і некоординованих (G_u) користувачів. Ці позначені групи не надають інформації ні про відносини між користувачами, ні про існування декількох координованих груп користувачів в X_{users}.

Для визначення характеристик координованої поведінки в Інтернеті задача характеризації моделюється функцією $B=\zeta(Y, X_{activities})$, вхідними даними якої є групи координованих користувачів, що є результатом завдання *виявлення* $Y=\xi(X_{users}, X_{activities})$, де $Y \in \{S_P, S_B, S_C\}$, визначених в (27), з їх активністю $\{X_{activities}\}$. Завдання характеризації спрямоване на обчислення набору кількісних показників B для вимірювання відмінних властивостей виявлених координованих моделей поведінки в термінах визначальних розмірів: автентичність (authenticity), шкідливість (harmfulness), оркестрація (orchestration – взаємодія сервісів, в тому числі бізнес-логіка та послідовність дій) та дисперсія в часі (time-variance). Показники, які використовують в характеризації, частково залежать від методів і вихідних даних задачі виявлення. Наприклад, асортативність вимірює ступінь, в якій вузли з високим ступенем в мережі з'єднані з іншими вузлами з високим ступенем, і навпаки. Цей показник використовувався для отримання уявлення про внутрішню структуру та організацію певних координованих спільнот. Однак асортативність обчислюється тільки в тому випадку, якщо метод виявлення координації виводить спільноти, а не кластери або двійкові мітки. Навпаки, інші показники обчислюють незалежно від методу виявлення, такі як вищезгадані оцінки ботів, які зазвичай використовуються як оцінка недостовірності скоординованих користувачів. Корисність завдання характеризації не обмежується розпізнаванням характеристик виявлених координованих форм поведінки або розрізненням різних випадків явища. Фактично, вихідні дані характеризації також використовують для перевірки результату виявлення, як у тих частих випадках, коли обґрунтування скоординованих користувачів недоступне.

Координаційні методи виявлення класифікують на дві основні категорії залежно від підходу, що лежить в їх основі: мережева наука або машинне навчання. Основні етапи мережевої науки: методи виявлення координованої поведінки в Інтернеті.

1. Вибрані користувачі стають вузлами мережі.

2. Подібність користувача обчислюється функцією подібності з призначенням ваг меж мережі.

3. Мережа фільтрується для збереження лише подібність із заданими властивостями.

4. Виявлення спільноти виконується для виявлення груп строго координованих користувачів.

DOI 10.15588/1607-3274-2025-3-13

4 ЕКСПЕРИМЕНТИ

Систему виявлення джерел розповсюдження україномовної дезінформації та неавтентичної поведінки користувачів чатів на основі методів NLP та машинного навчання подамо як:

$$S_{fakes} = < M_1, M_2, M_3, X, Y, R_{NLP}, U_{NLP}, R_{ML}, U_{ML}, \alpha, \beta, \gamma, \lambda >, \quad (28)$$

$$S_{fakes} = \lambda \circ \gamma \circ \beta \circ \alpha, \quad (29)$$

де $X = \{x_1, x_2, x_3, x_4\}$, $Y = \{y_1, y_2, y_3\}$, $R_{NLP} = \{r_{11}, r_{12}, r_{13}, r_{14}, r_{15}\}$, $U_{NLP} = \{u_{11}, u_{12}, u_{13}, u_{14}, u_{15}\}$, $R_{ML} = \{r_{21}, r_{22}, r_{23}, r_{24}\}$, $U_{ML} = \{u_{21}, u_{22}, u_{23}, u_{24}, u_{25}\}$.

Модуль M_1 «Імпорт, огляд та підготовка даних» опишемо суперпозицією та відповідною функцією:

$$M_1 = \lambda \circ \alpha_3 \circ \alpha_2 \circ \alpha_1, \quad (30)$$

$$M_1 = \lambda(\alpha_3(\alpha_2(\alpha_1(X, u_{11}, r_{11}), u_{12}, r_{12}), u_{13}, r_3)). \quad (31)$$

Основними процесами модуля M_1 «Імпорт, огляд та підготовка даних» є «Збір, завантаження та підготовка даних для формування датасету», «Дослідження унікальних символів», «Функція пошуку підрядків» та «Попередня обробка тексту», які опишемо суперпозицією:

$$C_{AU} = \chi \circ \omega \circ \mu \circ \alpha, \quad (32)$$

$$C_{AU} = \chi(\omega(\mu(\alpha(x_1, x_2), x_3, r_{14}, u_{14}), x_4, u_{12}), M_1, r_{15}, u_{15}). \quad (33)$$

Програма працює з датасетом, що містить твіти, позначені як пропагандистські або нейтральні. Вхідні дані представлені у форматі CSV-файлу, який містить такі основні поля:

– text – вміст твіту,

– label – мітка класу (0 – не пропаганда, 1 – пропаганда),

– додаткові технічні параметри (наприклад, ідентифікатор твіту, дата публікації тощо).

Для обробки даних використовується бібліотека pandas. Спочатку виконуються наступні дії:

1. Видалення непотрібних колонок (Unnamed: 0, id), які не мають значення для аналізу.

2. Перевірка балансу класів, щоб визначити, чи рівномірно представлені обидві категорії. Якщо виявляється значна диспропорція, можуть застосовуватися методи балансування, такі як oversampling або undersampling. Поточний стан розподілу класів в датасеті складається з 17,5% фейкових новин та 82,5% правдивої текстової інформації із онлайн ЗМІ. Проведено 9 експериментів, опис який подано в таблиці 2.

3. Перевірка наявності пропущених значень у колонці text. Якщо такі значення виявляються, вони видаляються або заповнюються, залежно від контексту.

4. Додавання колонки з довжиною твіту, що дає змогу оцінити можливий вплив коротких або довгих повідомлень на ефективність моделі.

Таблиця 2 – Опис експериментів

№	Cleanup	Вектори-зація	ML
1	– Remove HTML tags	TF-IDF	ComplementNB
2	– Remove Special Characters – Convert to Lowercase - Normalize Whitespace – Tokenize – Stem Words (UkrStemmer lib)	FastText	GaussianNB
3	– Convert to Lowercase	TF-IDF	ComplementNB
4	– Tokenize – Remove stopwords – Lemmatize(spaCy lib)	W2V	GaussianNB
5	– Remove punctuation	TF-IDF	ComplementNB
6	– Replace numbers with words – Convert to Lowercase	Glove	HistGradient Boosting Classifier
7	– Remove stopwords – Translates English words to Ukrainian		RandomForest
8	– Remove stopwords	Glove	MultinominalNB
9	– Lemmatize – Remove emojis		RandomForest

Оскільки Twitter дозволяє використовувати широкий набір символів, включаючи емодзі та спеціальні знаки, важливо розуміти їхню присутність у текстах. Для цього створюється множина унікальних символів, яка допомагає виявити потенційні проблеми під час обробки тексту.

Аналіз показує, що у твітерах часто зустрічаються:

– Емодзі, які можуть нести емоційне забарвлення повідомлення.

– Символи інших алфавітів, що може вказувати на багатомовність датасету.

– Спеціальні символи та знаки пунктуації, які можуть впливати на токенізацію.

Виходячи з цього аналізу, приймається рішення щодо подальшої обробки таких символів (видалення, заміна або врахування під час аналізу).

Для виявлення тематичних ключових слів, пов’язаних із пропагандою, реалізовано функцію substring_check(substring). Вона дозволяє знаходити певні слова або фрази у твітерах та аналізувати їхню частотність у різних класах. Це дає змогу:

– Визначити патерни вживання ключових слів у пропагандистських текстах.

– Аналізувати вплив певних термінів на класифікацію.

– Вдосконалювати модель шляхом розширення набору ознак.

Тексти твітерів проходять кілька етапів обробки для підготовки до подальшого аналізу:

– Видалення спеціальних символів, посилань, емодзі, пунктуації.

– Токенізація – поділ тексту на окремі слова.

– Заміна всіх слів на нижній регістр.

– Видалення стоп-слів (наприклад, «і», «це», «або»).

– Лематизація – приведення слів до їхньої основної форми.

DOI 10.15588/1607-3274-2025-3-13

Ці кроки допомагають зробити текст більш стандартизованим, що покращує точність моделі.

Модуль M_2 «Розпізнавання пропаганди» побудований на основі застосування бінарної класифікації (пропаганда/ не пропаганда) та багатокласової класифікації пропаганди (апелювання до авторитету, культ особи, демонізація, навішування ярликів тощо). Але для багатокласової класифікації необхідно промаркувати записи в датасеті. Модуль M_2 опишемо суперпозицією та відповідною функцією:

$$M_2 = \lambda \circ \beta_3 \circ \beta_2 \circ \beta_1, \quad (34)$$

$$M_2 = \lambda(\beta_3(\beta_2(\beta_1(M_1, u_{21}, r_{21}), u_{22}, r_{22}), u_{23}, r_{23})). \quad (35)$$

Спочатку необхідно знайди мінімальну точність, яку теоретично мають покрашити майбутні моделі; далі необхідно проаналізувати різноманітність лінгвістичних та стилістичних ознак та *n*-грам на моделі логістичної регресії. Далі необхідно побудувати нейронні мережі для класифікації записів.

Процес «NLP текстового контенту статей для виділення лінгвістичних ознак» опишемо суперпозицією та відповідною функцією:

$$C_{CU} = \beta \circ (\chi, \omega \circ \mu \circ \alpha), \quad (36)$$

$$C_{CU} = \beta(\omega(\mu(\alpha(C_{AU}, x_2, x_3, x_4), R_{NLP}, U_{NLP}, M_1, r_{12}, u_{14}), r_{13}). \quad (37)$$

$$C_{CU} = \beta(\chi(C_{AU}, R_{NLP}, U_{NLP}, M_1, x_2, x_3, x_4), r_{12}, u_{14}), r_{13}). \quad (38)$$

Основними процесами модуля M_2 «Розпізнавання пропаганди» є «Векторизація тексту», «Машине навчання моделі для розпізнавання пропаганди» та «Оцінка ефективності моделі», який опишемо як:

$$C_{UL} = \lambda \circ \omega \circ \gamma \circ \beta \circ \alpha, \quad (39)$$

$$C_{UL} = \lambda(\omega(\gamma(\beta(\alpha(C_{CU}, R_{ML}, U_{ML}, x_2), M_1, x_3), M_2, R_{ML}, U_{ML}, u_{23}), u_{14}, r_{13}), u_{13}, u_{25}, r_{15}). \quad (40)$$

Для перетворення текстів у числові вектори використовується метод TF-IDF (TfidfVectorizer). Основна ідея – оцінка важливості слів у контексті всього датасету. Формула TF-IDF виглядає так:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \text{IDF}(t), \quad (41)$$

– TF(t, d) – частота терміна t у документі d;

– IDF(t) – інверсна частота документа, що зменшує вагу загальновживаних слів.

Для інших експериментів застосовані моделі векторизації FastText, W2V та Glove (таблиця 2).

Для навчання моделі використовуються різні алгоритми машинного навчання (таблиця 3):

– Complement Naïve Bayes (ComplementNB) – це варіант Multinomial Naïve Bayes, спеціально розроблений для обробки незбалансованих класів у задачах текстової класифікації.

– Gaussian Naïve Bayes – це варіант Naïve Bayes, що припускає нормальний розподіл (Гаусса) ознак.

– HistGradientBoostingClassifier – це потужний алгоритм градієнтного бустингу, заснований на побудові ансамблю дерев рішень із використанням гістограмного біннінгу.

– Multinomial Naïve Bayes – це алгоритм Naïve Bayes, який підходить для текстової класифікації.

– Дерева рішень (RandomForest) – здатні знайти складні нелінійні залежності.

Таблиця 3 – Порівняльна таблиця алгоритмів

Алгоритм	Підходить для	Переваги	Недоліки
ComplementNB	Текстові дані	Стійкий до незбалансованих класів	Не для числових ознак
GaussianNB	Числові дані	Простий, швидкий	Погано працює з негаусовими розподілами
HistGradient Boosting	Великі набори даних	Швидкий, стійкий	Складна настройка
Random Forest	Різні типи ознак	Добре масштабується, гнучкий	Важкий для інтерпретації
Multinomialnb	Текстова класифікація	Швидкий, добре працює на частотах слів	Не підтримує числові ознаки

Навчальні та тестові вибірки формуються у співвідношенні 80:20. Для текстових даних найкраще підходять MultinomialNB та ComplementNB. Для числових ознак варто використовувати GaussianNB або ансамблеві методи (RandomForest). Для великих наборів даних найефективнішим буде HistGradientBoosting. RandomForest підходить для змішаних ознак (числових + категоріальних).

Ефективність моделі оцінюється за метриками: Accuracy – загальна точність передбачень; Recall – частка коректнот передбачених пропагандистських твітів; F1-міра – середнє між точністю та повнотою.

Запропонований модуль демонструє високу ефективність у виявленні пропаганди. Подальше вдосконалення можливе шляхом розширення датасету та адаптації моделі до багатомовного аналізу. Модуль M_3 «Розпізнавання мереж поширення пропаганди» опишемо суперпозицією та відповідною функцією:

$$M_3 = \lambda \circ \gamma_3 \circ \gamma_2 \circ \gamma_1, \quad (42)$$

$$M_3 = \lambda(\gamma_3(\gamma_2(\gamma_1(M_2, u_{13}, u_{14}, r_{13}), u_{14}, r_{13}), u_{13}, u_{24}, r_{23})). \quad (43)$$

Ідея полягає в знаходженні подібних за текстом/ значенням (lexical/ semantical) повідомлень, а також аналізі результатів поширення подібних повідомлень в часі та просторі. Основним процесом модуля M_3 «Розпізнавання мереж поширення пропаганди» є «Формування висновків наявності подібного фейку», який опишемо як:

$$C_{US} = \lambda \circ \gamma \circ \beta \circ \alpha, \quad (44)$$

$$C_{US} = \lambda(\gamma(\beta(\alpha(C_{US}, x_2), M_2, R_{NLP}, U_{NLP}, x_4), M_2, R_{ML}, U_{ML}, u_{14}), M_3, u_{13}, u_{25}, r_{15}). \quad (45)$$

DOI 10.15588/1607-3274-2025-3-13

5 РЕЗУЛЬТАТИ

Отримані результат подано на рис. 1–9. Ці зображення містять по три матриці (Confusion Matrix, Matrix by Recall та Matrix by Precision), що використовуються для оцінки продуктивності класифікаційної моделі. Матриця помилок відображає кількість правильно та неправильно класифікованих зразків. По діагоналі (верхній лівий і нижній правий квадранти) розташовані правильні передбачення для випадкового набору (вибірки) записів з датасету. Позадіагональні значення показують помилки.

Матриця за Повнотою відображає значення, які вказують на рівень повноти (recall) для кожного класу. Наприклад, для рис. 1 для класу 0 recall є 0,38 (тобто лише 38% зразків класу 0 було правильно класифіковано). Для класу 1 recall є 0,91 (91% зразків класу 1 правильно класифіковано). Висока повнота для класу 1 означає, що модель добре знаходить позитивні зразки, але для класу 0 вона неефективна.

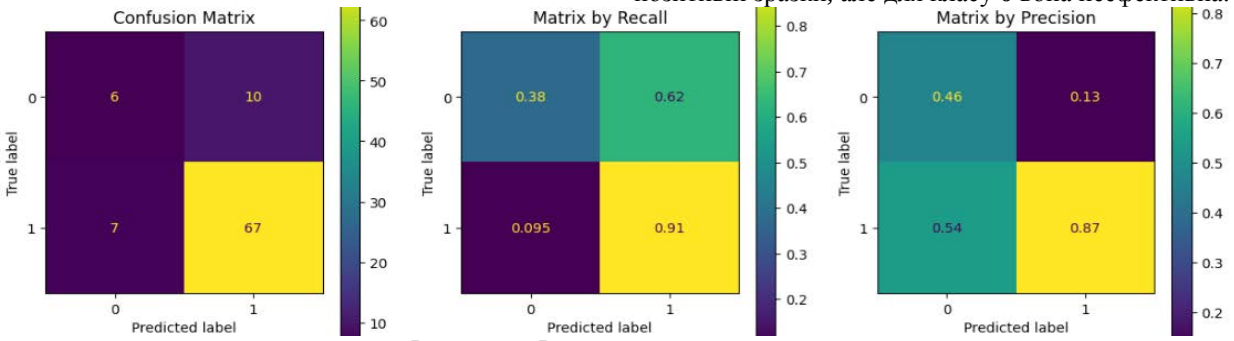


Рисунок 1 – Результати експерименту 1

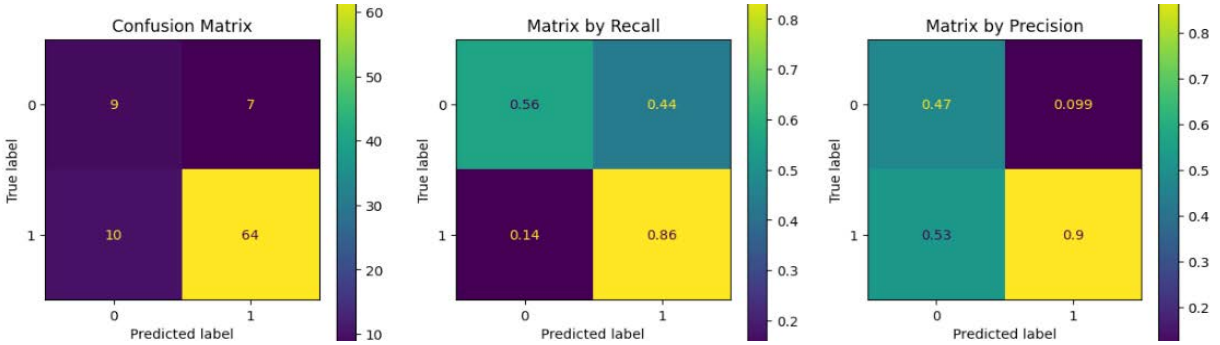


Рисунок 2 – Результати експерименту 2

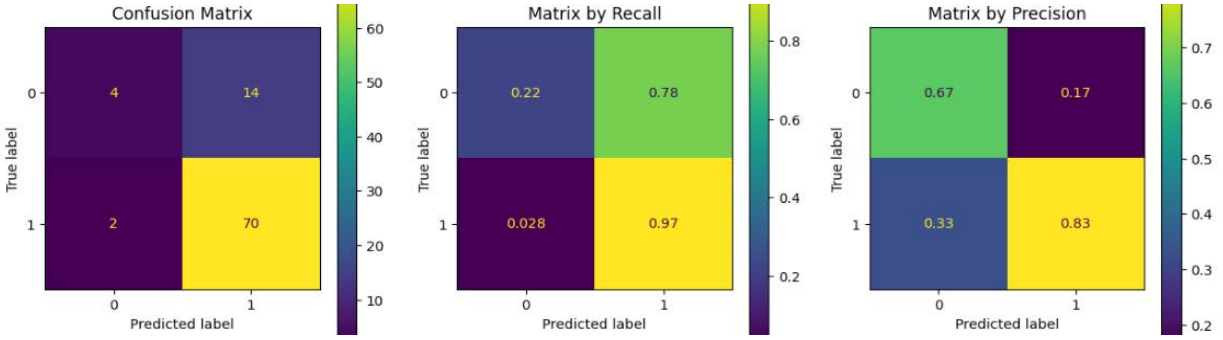


Рисунок 3 – Результати експерименту 3

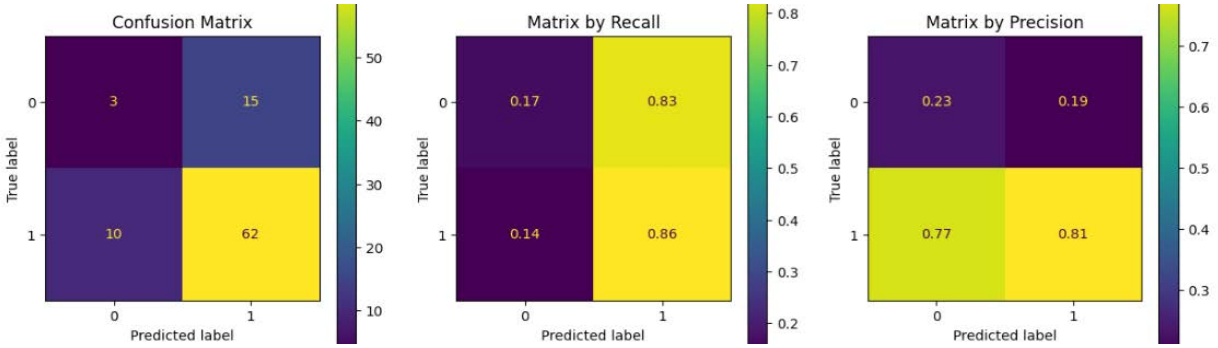


Рисунок 4 – Результати експерименту 4

DOI 10.15588/1607-3274-2025-3-13

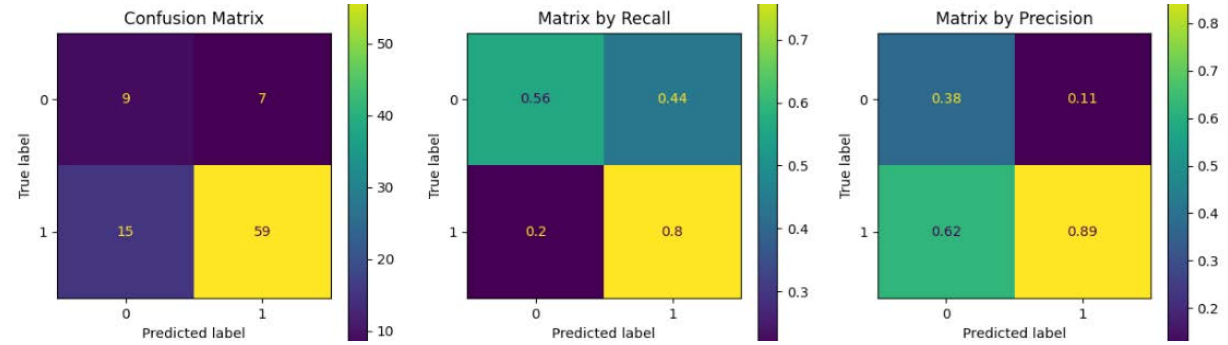


Рисунок 5 – Результати експерименту 5

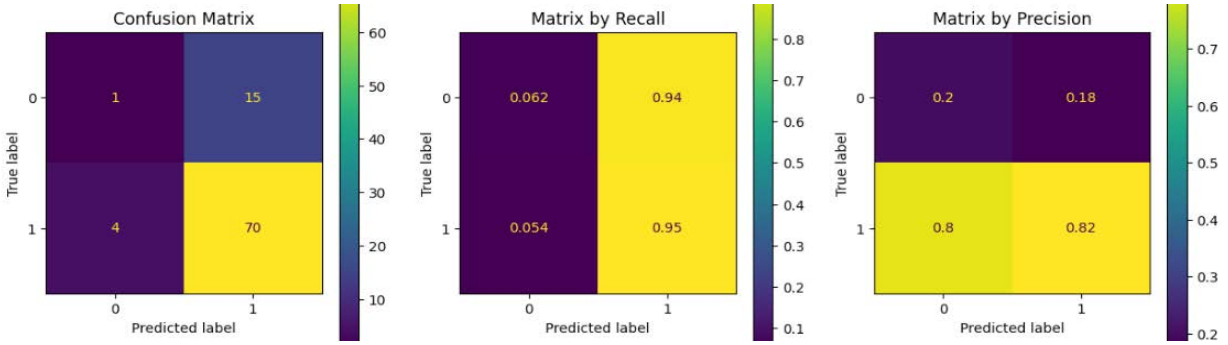


Рисунок 6 – Результати експерименту 6

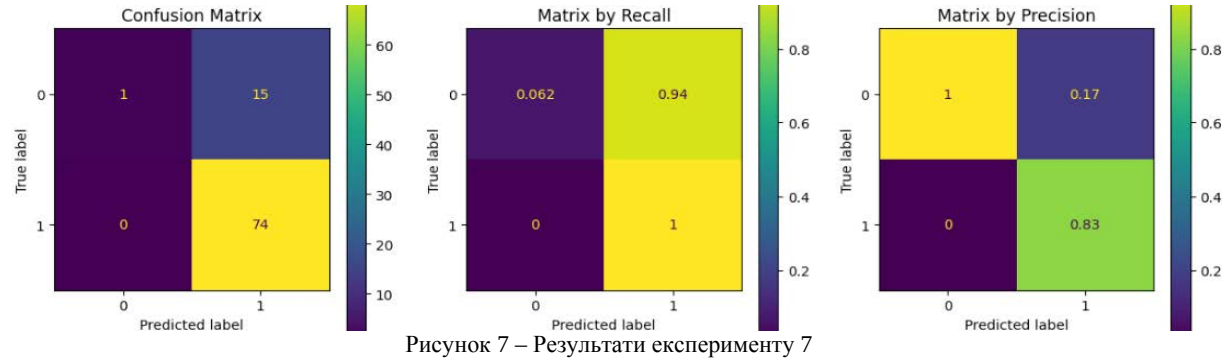


Рисунок 7 – Результати експерименту 7

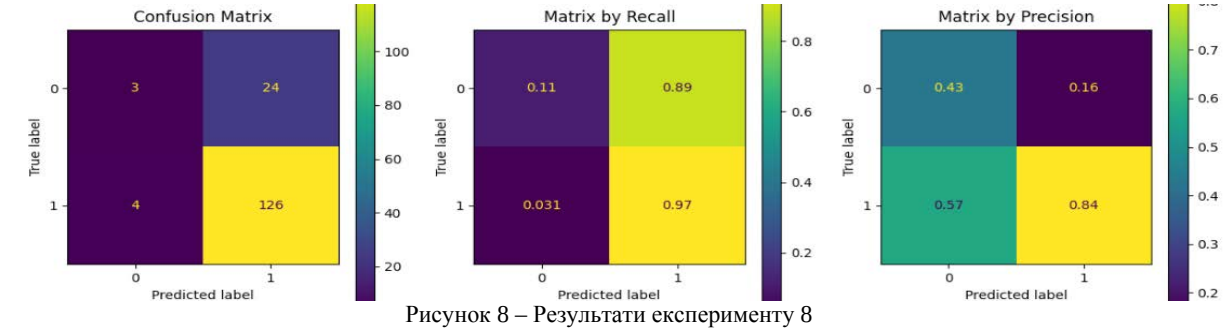


Рисунок 8 – Результати експерименту 8

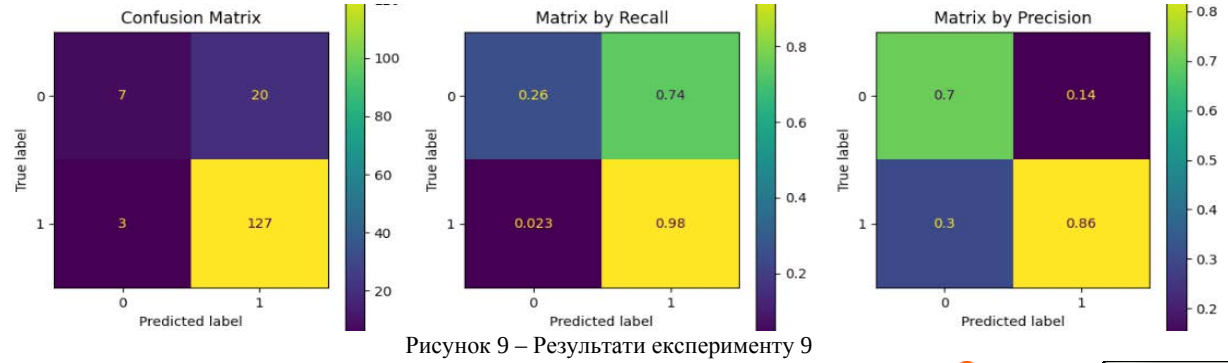


Рисунок 9 – Результати експерименту 9

DOI 10.15588/1607-3274-2025-3-13

Матриця за precision відображає значення, які показують рівень точності, наприклад, для рис. 1 для класу 0 precision є 0,46 (46% передбачень класу 0 були правильними). Для класу 1 precision є 0,87 (87% передбачень класу 1 були правильними). Висока точність для класу 1 означає, що більшість зразків, передбачених як клас 1, є правильними. Низька точність для класу 0 (0,46) вказує на велику кількість помилкових передбачень. Моделі добре передбачає клас 1 (високі recall та precision). Для класу 0 точність і повнота значно нижчі, що може свідчити про класовий дисбаланс або необхідність покращення моделі. У разі потреби можна використати стратегії балансування класів або оптимізувати поріг прийняття рішень для покращення продуктивності.

6 ОБГОВОРЕННЯ

Найкращі результати на даний момент показує експеримент 5 на основні TF-IDF+ ComplementNB (рис. 5). Для класу 0 recall є 0,56 (тобто лише 56% зразків класу 0 було правильно класифіковано). Для класу 1 recall є 0,8 (80% зразків класу 1 правильно класифіковано). Висок а повнота для класу 1 означає, що модель добре знаходить позитивні зразки, але для класу 0 вона менш ефективна (рис. 10). Для рис. 5 для класу 0 precision є 0.38 (38% передбачень класу 0 є правильними). Для класу 1 precision є 0.89 (89% передбачень класу 1 є правильними). Висока точність для класу 1 означає, що більшість зразків, передбачених як клас 1, є правильними. Низька точність для класу 0 (0.38) вказує на велику кількість помилкових передбачень. При цьому в серії експериментів спостерігаються певні аномалії (зокрема в експерименті 7 на основі Glove+ RandomForest – рис. 7), які потребують подальшого дослідження. Підсумовуючі результати по класу F (Фейк) подані на рис. 10–12. Проведення наступних експериментів (комбінацій методів які себе краще показали) а також конструювання нових фічерів (зокрема оцінки сентименту).

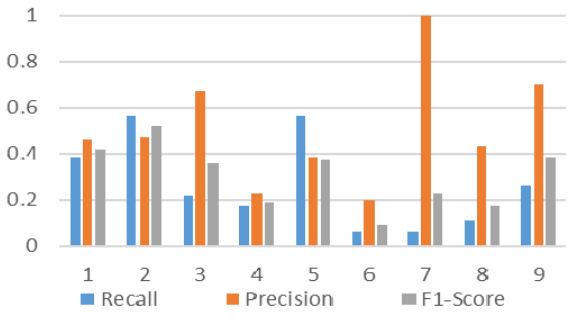


Рисунок 11 – Recall, Precision та F1-Score

ВИСНОВКИ

Стаття описує алгоритм роботи програми, що виконує автоматичне виявлення пропагандистських повідомлень у Twitter, джерел розповсюдження езінформації та неавтентичної поведінки чатів. Основна увага приділяється методам збору та підготовки даних, попередній обробці тексту, векторизації, навчанню моделі та оцінці її ефективності. Методи NLP та ML дозволяють виявляти такі загрози шляхом аналізу стилю авторів, часових закономірностей публікацій та графових зв'язків між користувачами. Описано процес збору, підготовки та очищення даних, а також розглянуто різні підходи до навчання моделі та оцінки її ефективності. Ідея полягає в знаходженні подібних за текстом/ значенням (lexical/ semantical) повідомлень, а також аналізі результатів поширення подібних повідомлень в часі та просторі. У якості основних алгоритмів моделювання використані Complement Naïve Bayes, HistGradientBoostingClassifier, Gaussian Naïve Bayes, Multinomial Naïve Bayes та RandomForest для виявлення джерел розповсюдження дезінформації та неавтентичної поведінки чатів. Основна увага приділяється методам попередньої обробки текстів, векторизації даних та машинному навчанню для автоматичної класифікації повідомлень. Проведено 9 експериментів для різних методів попереднього опрацювання даних, моделей векторизації та алгоритмів моделювання. Найкращі результати на даний момент показує експеримент 5 на основні TF-IDF+ComplementNB. При цьому в серії проведених експериментів спостерігаються певні аномалії (зокрема в експерименті 7 на основі Glove+ RandomForest), які потребують подальшого дослідження. Отримані результати можуть бути використані для подальшого вдосконалення методів виявлення джерел розповсюдження дезінформації, неавтентичної поведінки чатів та шкідливого контенту для збільшення обороздатності країни.

Наукова новизна полягає у розробленні методів:

– ідентифікація схожих за стилістикою фейкових новин для виявлення шляхів розповсюдження дезінформації в часі та просторі;

– стилістичного опрацювання фейкових новин для виявлення спільних лінгвістичних характеристик текстового контенту на основі NLP;

– виявлення неавтентичної поведінки ботів в чатах на основі аналізу скоординованої поведінки користувачів у соціальних мережах та онлайн ЗМІ.

Практична цінність полягає у розробленні системи підтримки прийняття рішення для пошуку та виявлення джерел розповсюдження україномовних фейкових новин, пропаганди, та дезінформації у соціальних мережах та онлайн ЗМІ, а також експериментальна апробація для розрахунку точності отриманих результатів на основі реалізації модуля:

– інтелектуального пошуку, збору, лінгвістичного аналізу, попереднього опрацювання, маркування та класифікації текстового контенту для формування датасету та підготовки даних для виявлення дезінформації та джерел розповсюдження;

– розпізнавання україномовної дезінформації, фейкових новин та пропаганди для виявлення стилістично та змістовно подібного текстового

DOI 10.15588/1607-3274-2025-3-13

контенту при ідентифікації джерел розповсюдження та неавтентичної поведінки ботів;

– розпізнавання мереж поширення пропаганди на основі знаходженні подібних за текстом/ значенням повідомлень, а також аналізі результатів поширення подібних повідомлень в часі та просторі.

Очікувані результати виконання проєкту:

– розроблено метод стилістичного аналізу та лінгвістичного опрацювання текстового контенту на основі NLP та ML для формування інформаційного портрету генератора фейкового повідомлення та подібних до нього за множиною наративів.

– запропоновано моделі та основні принципи інформаційної технології виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів, що дозволить своєчасно виявляти інформаційні загрози в кіберпросторі країни.

– запропоновано параметри та критерії поведінки користувачів чатів для моделі виявлення неавтентичної поведінки ботів, характерні для відповідної групи. Модель неавтентичної поведінки користувача полягає у побудові профілю поведінки користувача системи на основі аналізу поведінкових закономірностей. Вони відображають притаманні підсвідомі характерні риси в межах реалізації відповідного події, що підлягає автентичності. Модель дозволяє виявляти притаманні користувачу підсвідомі поведінкові риси, присутні у різних психоемоційних станах.

ПОДЯКИ

Дана стаття підготована завдяки грантової підтримки Національного Фонду Досліджень України, реєстраційний номер проєкту 187/0012 від 1/08/2024 (2023.04/0012) «Розроблення інформаційної системи автоматичного виявлення джерел дезінформації та неавтентичної поведінки користувачів чатів» за конкурсом «Наука для зміцнення обороноздатності України».

ЛІТЕРАТУРА

1. BERT Based Fake News Detection Model / [Y. Zhang, Y. Shao, X. Zhang et al.] // Training. – 2022. – Vol. 1530. – P. 383.
2. Cahyani D. E. Performance comparison of TF-IDF and word2vec models for emotion text classification / D. E. Cahyani, I. Patasik // Bulletin of Electrical Engineering and Informatics. – 2021. – Vol. 10(5). – P. 2780–2788. DOI: 10.11591/eei.v10i5.3157
3. Bhosale S. Identifying Bots on Twitter with Benford's Law / S. Bhosale. – Access mode: https://scholarworks.sjsu.edu/etd_projects/1041/
4. Ghaemi Z. A Varied Density-based Clustering Approach for Event Detection from Heterogeneous Twitter Data / Z. Ghaemi, M. Farnaghi // ISPRS International Journal of Geo-Information. – 2019. – 8(2). – P. 82. DOI: 10.3390/ijgi8020082
5. Lazebnik T. Temporal graphs anomaly emergence detection: benchmarking for social media interactions / T. Lazebnik, O. Iny // Applied Intelligence. – 2024. – Vol. 54. – P. 12347–12356. DOI: 10.1007/s10489-024-05821-3
6. Do Social Bots (Still) Act Different to Humans? – Comparing Metrics of Social Bots with Those of Humans / [S. Stieglitz, F. Brachten, D. Berthelé et al.] // Lecture Notes in Computer Science. – 2017. – Vol. 10282. – P. 379–395. DOI: 10.1007/978-3-319-58559-8_30
7. Vysotska V. Information technology for recognizing propaganda, fakes and disinformation in textual content based on nlp and machine learning methods / V. Vysotska // Radio Electronics, Computer Science, Control. – 2024. – Vol. 2. – P. 126. DOI: 10.15588/1607-3274-2024-2-13
8. Мокрицька О. В. Використання алгоритмів машинного навчання для автоматизації процесу модерації контенту в групових чатах месенджерів / О. В. Мокрицька, Ю. М. Мочернюк // Scientific Bulletin of UNFU. – 2024. – Том 34(7). – С. 52–59. DOI: 10.36930/40340707
9. Дмитроца Л. П. Аналіз інструментів штучного інтелекту для виявлення дезінформації в новинах Facebook / Л. П. Дмитроца, С. В. Дацик // Інформаційні моделі, системи та технології. – 2023. – С. 35–36. – Access mode: https://elartu.tntu.edu.ua/bitstream/lib/44384/2/IMSTT_2023_Dmytrotsa_L_P-Analysis_of_artificial_35-36.pdf
10. Семенюк А. В. Використання методів машинного навчання та штучного інтелекту для захисту від влпиву соціальної інженерії при кібератаках / А. В. Семенюк. – Access mode: http://ir.lib.vntu.edu.ua/handle/123456789/41797
11. Аналіз методів виявлення дезінформації в соціальних мережах за допомогою машинного навчання / [М. С. Марценюк, В. А. Козачок, О. Богданов, З. М. Бржевська] // Кібербезпека: освіта, наука, техніка. – 2023. – Том 2(22). – С. 148–155. – Access mode: https://elibrary.kubg.edu.ua/id/eprint/48271/
12. Дмитроца Л. П. Застосування методів штучного інтелекту для виявлення та протидії дезінформації у Facebook / Л. П. Дмитроца, С. В. Дацик // Інформаційні моделі, системи та технології. – 2023. – С. 37–38. – Access mode: https://elartu.tntu.edu.ua/bitstream/lib/44385/2/IMSTT_2023_Dmytrotsa_L_P-Application_of_artificial_37-38.pdf
13. Machine Learning and Deep Learning Applications in Disinformation Detection: A Bibliometric Assessment / A. Sandu, L.-A. Cotfas, C. Delcea et al.] // Electronics. – 2024. – Vol. 13(22). – P. 4352. DOI: 10.3390/electronics13224352
14. Santos F. C. C. Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis / F. C. C. Santos // Journalism and Media. – 2023. – Vol. 4(2). – P. 679–687. DOI: 10.3390/journalmedia4020043
15. Lakzaei B. Disinformation detection using graph neural networks: a survey / B. Lakzaei, Haghir M. Chehreghani, A. Bagheri // Artificial Intelligence Review. – 2024. – Vol. 57. – P. 52. DOI: 10.1007/s10462-024-10702-9
16. Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches / [H. R. Saeidnia, E. Hosseini, B. Lund et al.] // Knowledge and Information Systems. – 2025. DOI: 10.1007/s10115-024-02337-7
17. Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions / [P. Akhtar, A.M. Ghouri, H.U.R. Khan et al.] // Annals of Operations Research. – 2023. – Vol. 327. – P. 633–657. DOI: 10.1007/s10479-022-05015-5
18. Disinformation, Fakes and Propaganda Identifying Methods in Online Messages Based on NLP and Machine Learning

DOI 10.15588/1607-3274-2025-3-13

Methods / [V. Vysotska, K. Przystupa, L. Chyrun et al.] // International Journal of Computer Network and Information Security(IJCNIS). – 2024. – Vol. 16(5). – P. 57–85. DOI:10.5815/ijcnis.2024.05.06
19. Prokipchuk O. Ukrainian Language Tweets Analysis Technology for Public Opinion Dynamics Change Prediction Based on Machine Learning / O. Prokipchuk, V. Vysotska // Radio Electronics, Computer Science, Control. – 2023. – № 2(65). – P. 103–116. DOI: 10.15588/1607-3274-2023-2-11
20. Vysotska V. NLP Tool for Extracting Relevant Information from Criminal Reports or Fakes/Propaganda Content / V. Vysotska, S. Mazepa, L. Chyrun, O. Brodyak, I. Shakleina, V. Schuchmann, // Computer Sciences and Information Technologies : 17th International Conference, Lviv, 2022, November. – Lviv: IEEE, 2021. – P. 93–98. DOI: 10.1109/CSIT56902.2022.10000563
21. Information technology for identifying disinformation sources and inauthentic chat users' behaviours based on machine learning / [V. Vysotska, L. Chyrun, S. Chyrun, I. Holets] // CEUR Workshop Proceedings. – 2024. – Vol. 3723. – P. 466–483.
22. Іосіфов Є. Порівняльний аналіз методів, технологій, сервісів та платформ для розпізнавання голосової інформації в системах забезпечення інформаційної безпеки / Є. Іосіфов, В. Соколов // Кібербезпека: освіта, наука, техніка. – 2024. – Том 1(25). – С. 468–486. DOI: 10.28925/2663-4023.2024.25.468486
23. Аналіз методів виявлення дезінформації в соціальних мережах за допомогою машинного навчання / [М. Марценюк, В. Козачок, О. Богданов та ін.] // Кібербезпека: освіта, наука, техніка. – 2023. – Том 2(22). – С. 148–155. DOI: 10.28925/2663-4023.2023.22.148155
24. Інтелектуальний метод виявлення джерел мультилінгвальної дезінформації / [М. Комар, Х. Ліп'яніна-Гончаренко, І. Кіт та ін.] // Measuring and computing devices in technological processes. – 2023. – Том 2. – С. 221–230. DOI: 10.31891/2219-9365-2023-74-31
25. Prytula M. Detection of aggressive rhetoric in text using machine learning algorithms / M. Prytula, I. Olenych // Electronics and information technologies. – 2023. – Vol. 22. DOI: 10.30970/eli.22.4
26. Deep learning for misinformation detection on online social networks: a survey and new perspectives / [M. R. Islam, S. Liu, X. Wang, G. Xu] // Social Network Analysis and Mining. – 2020. – Vol. 10. – P. 82. DOI: 10.1007/s13278-020-00696-x
27. Detecting and responding to hostile disinformation activities on social media using machine learning and deep neural networks / [B. Cartwright, R. Frank, G. Weir, K. Padda] // Neural Computing and Applications. – 2022. – Vol. 34. – P. 15141–15163. DOI: 10.1007/s00521-022-07296-0

Стаття надійшла до редакції 07.03.2025.
Після доробки 09.06.2025.

UDC 004.9

INFORMATION TECHNOLOGY FOR DETECTION OF DISINFORMATION SOURCES AND INAUTHENTICAL BEHAVIOR OF CHAT USERS BASED ON NLP AND MACHINE LEARNING METHODS

Vysotska V. – PhD, Professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. In the modern digital environment, the spread of disinformation and inauthentic behaviour of users in chat rooms poses a serious threat to society. Natural language processing and machine learning methods offer effective approaches to detecting and countering such threats.

Objective of the study is to develop information technology for automatically detecting the spread of sources of Ukrainian-language fake news and inauthentic behaviour of chat users, which is built using natural language processing methods and implemented, based on machine learning technologies.

Method. To implement the project, such feature construction methods as the TF-IDF statistical indicator, the Bag of Words vectorization model, and part-of-speech mark-up were used. For other experiments, the FastText, W2V, and Glove word2vec vectorization models were used to obtain vector representations of words, as well as to recognize trigger words (reinforcing words, absolute pronouns, and "shiny" words). The idea is to find similar messages in terms of text/meaning (lexical/semantical), as well as analyse the results of the distribution of similar messages in time and space. Complement Naïve Bayes, Gaussian Naïve Bayes, HistGradientBoostingClassifier, MultinomialNB and Random Forest were used as the main modelling algorithms to identify sources of disinformation and inauthentic chat behavior.

Results. This article discusses the development of software for detecting propaganda messages in social networks based on the analysis of Twitter text data. The main attention is paid to the methods of text pre-processing, data vectorization and machine learning for message classification. The process of collecting, preparing and cleaning data is described, and various approaches to training the model and evaluating its effectiveness are considered. 9 experiments were conducted for the selected methods of post-processing data, vectorization models and modelling algorithms.

Conclusions. The created models show excellent results in recognizing sources of propaganda, fakes and disinformation in social networks and online media. The best results so far are shown by experiment 5 on the main TF-IDF + Complement Naïve Bayes. The high recall value for class 1 (0.8) means that the model finds positive samples well, but for class 0 it is less effective (0.56). The correspondingly high precision value for class 1 (0.89) means that most of the samples predicted as class 1 are correct. The low precision for class 0 (0.38) indicates a large number of false predictions. At the same time, certain anomalies are observed in the series of experiments (in particular, in experiment 7 based on Glove + Random Forest), which require further research. The results obtained can be used to further improve the algorithms for detecting sources of disinformation, inauthentic chat behaviour and malicious content to increase the country's transparency.

KEYWORDS: disinformation, source of disinformation, way of disinformation dissemination, disinformation dissemination network, fake, propaganda, natural language processing, stylistic analysis.

DOI 10.15588/1607-3274-2025-3-13

REFERENCES

1. Zhang Y., Shao Y., Zhang X., Wan W., Li J., Sun J. BERT Based Fake News Detection Model, *Training*, 2022, Vol. 1530, P. 383.
2. Cahyani D. E., Patasik I. Performance comparison of TF-IDF and word2vec models for emotion text classification, *Bulletin of Electrical Engineering and Informatics*, 2021, Vol. 10(5), pp. 2780–2788. DOI: 10.11591/eei.v10i5.3157
3. Bhosale S. Identifying Bots on Twitter with Benford's Law. Access mode: https://scholarworks.sjsu.edu/etd_projects/1041/
4. Ghaemi Z., Farnaghi M. A Varied Density-based Clustering Approach for Event Detection from Heterogeneous Twitter Data, *ISPRS International Journal of Geo-Information*, 2019, 8(2), P. 82. DOI: 10.3390/ijgi8020082
5. Lazebnik T., Iny O. Temporal graphs anomaly emergence detection: benchmarking for social media interactions, *Applied Intelligence*, 2024, Vol. 54, pp. 12347–12356. DOI: 10.1007/s10489-024-05821-3
6. [Stieglitz S., Brachten F., Berthelé D., Schlaus M., Venetopoulou C., Veutgen D. Do Social Bots (Still) Act Different to Humans? – Comparing Metrics of Social Bots with Those of Humans, *Lecture Notes in Computer Science*, 2017, Vol. 10282, pp. 379–395. DOI: 10.1007/978-3-319-58559-8_30
7. Vysotska V. Information technology for recognizing propaganda, fakes and disinformation in textual content based on nlp and machine learning methods, *Radio Electronics, Computer Science, Control*, 2024, Vol. 2, P. 126. DOI: 10.15588/1607-3274-2024-2-13
8. Mokrytska O. V., Mochernyuk YU. M. Using machine learning algorithms to automate the content moderation process in messenger group chats, *Scientific Bulletin of UNFU*, 2024, Vol. 34(7), pp. 52–59. DOI: 10.36930/40340707
9. Dmytrotsa L. P., Datsyk S. V. Analysis of artificial intelligence tools for detecting disinformation in Facebook news, *Information models, systems and technologies*, 2023. pp. 35–36. Access mode: https://elartu.tntu.edu.ua/bitstream/lib/44384/2/IMSTT_2023_Dmytrotsa_L_P-Analysis_of_artificial_35-36.pdf
10. Semenyuk A. V. Using machine learning and artificial intelligence methods to protect against social engineering in cyberattacks. Access mode: http://ir.lib.vntu.edu.ua/handle/123456789/41797
11. Martsenyuk M. S., Kozachok V. A., Bogdanov O., Brzhevska Z. M. Analysis of methods for detecting disinformation in social networks using machine learning, *Cybersecurity: education, science, technology*, 2023, Vol. 2(22), pp. 148–155. Access mode: https://elibrary.kubg.edu.ua/id/eprint/48271/
12. Dmytrotsa L. P., Datsyk S. V. Application of artificial intelligence methods to detect and counter disinformation on Facebook, *Information models, systems and technologies*, 2023, pp. 37–38. Access mode: https://elartu.tntu.edu.ua/bitstream/lib/44385/2/IMSTT_2023_Dmytrotsa_L_P-Application_of_artificial_37-38.pdf
13. Sandu A., Cotfas L.-A., Delcea C., Ioanăș C., Florescu M.-S., Orzan M. Machine Learning and Deep Learning Applications in Disinformation Detection: A Bibliometric Assessment, *Electronics*, 2024, Vol. 13(22), P. 4352. DOI: 10.3390/electronics13224352
14. Santos F. C. C. Artificial Intelligence in Automated Detection of Disinformation: A Thematic Analysis, *Journalism and Media*, 2023, Vol. 4(2), p. 679–687. DOI: 10.3390/journalmedia4020043
15. Lakzaei B., Chehreghani Haghir M., Bagheri A. Disinformation detection using graph neural networks: a survey, *Artificial Intelligence Review*, 2024, Vol. 57, P. 52. DOI: 10.1007/s10462-024-10702-9
16. [Saeidnia H.R., Hosseini E., Lund B., Tehrani M. A., Zaker S., Molaei S. Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches, *Knowledge and Information Systems*, 2025. DOI: 10.1007/s10115-024-02337-7
17. Akhtar P., Ghouri A. M., Khan H. U. R., Haq M. A., Awan U., Zahoor N., Khan Z., Ashraf A. Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions, *Annals of Operations Research*, 2023, Vol. 327, pp. 633–657. DOI: 10.1007/s10479-022-05015-5
18. Vysotska V., Przystupa K., Chyrun L., Vladov S., Ushenko Y., Uhryn D., Hu Z. Disinformation, Fakes and Propaganda Identifying Methods in Online Messages Based on NLP and Machine Learning Methods, *International Journal of Computer Network and Information Security(IJCNIS)*, 2024, Vol.16(5), pp. 57–85. DOI:10.5815/ijcnis.2024.05.06
19. Prokipchuk O., Vysotska V. Ukrainian Language Tweets Analysis Technology for Public Opinion Dynamics Change Prediction Based on Machine Learning, *Radio Electronics, Computer Science, Control*, 2023, № 2(65), pp. 103–116. DOI: 10.15588/1607-3274-2023-2-11
20. Vysotska V., Mazepa S., Chyrun L., Brodyak O., Shakleina I., Schuchmann V. NLP Tool for Extracting Relevant Information from Criminal Reports or Fakes/Propaganda Content, *Computer Sciences and Information Technologies : 17th International Conference, Lviv, 2022, November*. Lviv, IEEE, 2021, pp. 93–98. DOI: 10.1109/CSIT56902.2022.10000563
21. Vysotska V., Chyrun L., Chyrun S., Holets I. Information technology for identifying disinformation sources and inauthentic chat users' behaviours based on machine learning, *CEUR Workshop Proceedings*, 2024, Vol. 3723, pp. 466–483.
22. Iosifov E., Sokolov V. Comparative analysis of methods, technologies, services and platforms for voice information recognition in information security systems, *Cybersecurity: education, science, technology*, 2024, Vol. 1(25), pp. 468–486. DOI: 10.28925/2663-4023.2024.25.468486
23. Martsenyuk M., Kozachok V., Bogdanov O., Iosifov E., Brzhevska Z. Analysis of methods for detecting disinformation in social networks using machine learning, *Cybersecurity: education, science, technology*, 2023, Vol. 2(22), pp. 148–155. DOI: 10.28925/2663-4023.2023.22.148155
24. Komar M., Lipyanina-Honcharenko H., Kit I., Madarash R., Yurkiv H. An intellectual method for identifying sources of multilingual disinformation, *Measuring and computing devices in technological processes*, 2023, Vol. 2, pp. 221–230. DOI: 10.31891/2219-9365-2023-74-31
25. Prytula M., Olenych I.Detection of aggressive rhetoric in text using machine learning algorithms, *Electronics and information technologies*, 2023, Vol. 22. DOI: 10.30970/eli.22.4
26. Islam M.R., Liu S., Wang X., Xu G.Deep learning for misinformation detection on online social networks: a survey and new perspectives, *Social Network Analysis and Mining*, 2020, Vol. 10, P. 82. DOI: 10.1007/s13278-020-00696-x
27. Cartwright B., Frank R., Weir G., Padda K. Detecting and responding to hostile disinformation activities on social media using machine learning and deep neural networks, *Neural Computing and Applications*, 2022, Vol. 34, pp. 15141–15163. DOI: 10.1007/s00521-022-07296-0

DOI 10.15588/1607-3274-2025-3-13

UDC 004.93

CARDIAC SIGNAL PROCESSING WITH ALGORITHMS USING VARIABLE RESOLUTION

Kalmykov V. G. – PhD, Senior Researcher of the Institute of Mathematical Machines and Systems Problems, Kyiv Ukraine.

Sharypanov A. V. – PhD, Chief of Laboratory of Medical and Biological Informatics, Institute of Mathematical Machines and Systems Problems, Kyiv Ukraine.

Vishnevskey V. V. – PhD, Leading Researcher of the Institute of Mathematical Machines and Systems Problems, Kyiv Ukraine.

ABSTRACT

Context. The proposed paper relates to the field of cardiac signal processing, in particular, to the segmentation of the cardiac signal into cardiac cycles, as well as one of the most important features definition used in cardiac diagnosis, the *T*-wave end.

Objective. The purpose and object of study is to develop an algorithm for processing the cardiac signal in the presence of interference that allows the identification of features necessary for diagnosis and, at the same time, does not distort the original signal as is usually the case when it is processed by band-pass digital filters to exclude interference, which leads to the original signal distortion and, possibly, loss of diagnostic features.

The proposed **Method** involves representing the cardiac signal as part of some image contour. Cardiac signal processing consists first of all in segmentation into cardiac cycles. Usually, *R*-waves are used to segment the cardiac signal into cardiac cycles, i.e., the sequence of *R*-waves in the processed part of the cardiac signal is determined. When determining the *R*-wave, a model is used that assumes an increase in the signal followed by a decrease, and the increase (decrease) rate must be greater in absolute value than a certain predetermined value. For a selected segment of the cardiac signal, the sequence of *R*-waves is determined at different resolutions. The answer is the sequence that is repeated for the largest number of resolutions and that is used to segment the cardiac signal into cardiac cycles. The *T*-wave model can be represented as a sequence of curved arcs without breaks. In one of the common cases, the *T*-wave is determined by the largest maximum of the cardiac signal within the cardiac cycle, following the R-wave. The end of the *T*-wave is determined by the first minimum following the already determined maximum for the *T*-wave. As in the case of cardiac signal segmentation, the maximum of the *T*-wave and the *T*-wave end are determined at different resolutions, and the answer is considered to be those values that coincide at the largest number of used resolutions.

Results. Algorithms for cardiac signal processing using variable resolution have been developed and experimentally verified, namely, the algorithm for segmentation of the cardiac signal into cardiac cycles and the algorithm for *T*-wave end detection, which is of great importance in cardiac diagnostics. Means of cardiac signal processing, using the proposed algorithms, do not change the processed cardiac signal, unlike traditional means that use filtering of the cardiac signal, distorting the cardiac signal itself, which leads to distortion of the processing result.

Conclusions. Scientific novelty consists in the fact that algorithms of cardiac signal processing in the presence of interference using variable resolution typical of visual perception are proposed.

The practical significance consists in the fact that the means of cardiac signal processing, using the proposed algorithms, do not change the processed cardiac signal, unlike traditional means that use filtering of the cardiac signal, distorting the cardiac signal itself, which leads to distortion of the processing result. The use of the presented tools in practical medical practice will lead to an improvement in the quality of cardiac diagnostics and, as a result, the quality of treatment.

KEYWORDS: cardiac signal, segmentation, cardiac cycles, *T*-wave end, variable resolution.

ABBREVIATIONS

ECG is an electrocardiogram.

NOMENCLATURE

P-wave is an electrocardiogram element;
Q-wave is an electrocardiogram element;
R-wave is an electrocardiogram element;
S-wave is an electrocardiogram element;
T-wave is an electrocardiogram element;
ST-segment is an electrocardiogram element;
QRS is a complex of three waves – *Q*, *R*, *S*;
Tpeak and Tend are some *T*-wave points;
(x_i, y_i) is a sample pair number i from a sequence that are discrete realization of the function $y(x)$, representing cardiosignal;
$x_1=a$, $x_I=b$, (a,b) is a domain of signal definition;
I is a number of samples in the domain definition;
$\{R(e)\}$ is a final list of *R*-waves;
q is a number samples in the "coarse" sample, which determine the current resolution;
J is a number of "coarse" samples at a given resolution m;
z_j is a value of the "coarse" count j ;
w_j is a sample sequence of the "coarse" count j, at a given resolution m;
m is an one of the the current resolutions;
$Z^{(m)} = \{z_j^{(m)}\}$ is a discrete realization of the function $y(x)$ at a resolution m;
$i_Q^{(j)}$ is a sample number of the sequence I, corresponding to the "coarse" sample j at a resolution m, which represent *Q*-wave;
$i_R^{(j)}$ is a sample number of the sequence I, corresponding to the "coarse" sample j at a resolution m, which represent *R*-wave;

DOI 10.15588/1607-3274-2025-3-14

$i_S^{(j)}$ is a sample number of the sequence I, corresponding to the "coarse" sample j at a resolution m, which represent S-wave;

Q is a value of Q-wave at sample $i_Q^{(j)}$;

R is a value of R-wave at sample $i_R^{(j)}$;

Q is a value of Q-wave at sample $i_S^{(j)}$;

θ is a predetermined threshold to recognize R-wave;

$R^{(m)} = i_{R_1}^{(m)}, i_{R_2}^{(m)}, ..., i_{R_n}^{(m)}, ...$ is a preliminary answer for R-wave sequence samples at resolution m;

d is a predetermined threshold to recognize T-wave.

INTRODUCTION

The electrocardiogram (ECG) reflects the electrical activation of the heart and is an important biomedical signal for determining the functional state of the heart. The ECG consists of a repetitive sequence of P, QRS and T waves associated with each heartbeat (Fig. 1). The detection of QRS complexes in the cardiogram is used to break it down into a sequence of individual cardiac cycles, i.e., primary segmentation. After determining the location of QRS complexes in the cardiogram, this information can be used to construct ECG-derived signals, such as amplitude and rhythmograms, or to further find individual components of the cardiac cycle.

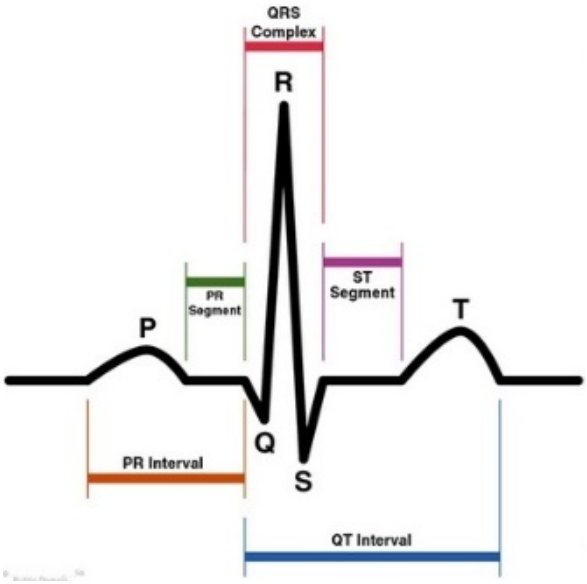


Figure 1 – Cardiac cycle graph image

Various features of the ECG signal are useful for diagnosing heart disease. Reliable detection of the P and T waves is more difficult than the QRS waves for several reasons, including their low amplitude, low signal-to-noise ratio, amplitude and morphological variability, and even possible overlap of the P-wave with the QRS complex. A flattened or negative T wave is interpreted as a symptom of coronary heart disease. P-wave prolongation can be used to detect atrial fibrillation.

Traditionally, the electrocardiac signal is filtered to exclude interference. Since the frequency parameters of interference cannot be determined, filtering often leads to distortion of the electrocardiogram and, consequently, to a decrease in the quality of diagnostics.

The human visual system processes the visual image of a cardiac signal almost instantaneously, even in the presence of interference.

The implementation of variable resolution means processing an image, in particular an electrocardiac signal, using a certain number of resolutions. This paper deals with the development of algorithms for segmentation and detection of cardiac signal elements using variable resolution.

The subject of study is the processing of electrocardiograms as a special case of an image. The proposed method uses a variable resolution, which should provide satisfactory results in the presence of interference, without the use of pre-filtering the electrocardiogram signal. Pre-filtering with predefined and specified filter parameters is typical for most electrocardiogram processing methods, which significantly limits the application in the presence of interference.

The purpose of the work is to develop an algorithm for processing an electrocardiogram as a special case of an image, in particular, to determine the end of the T wave, which is an important diagnostic factor. Variable resolution should be used in the development of the algorithm to ensure satisfactory performance in the presence of interference without using filtering of the electrocardiogram signal with predefined and specified filter parameters.

1 PROBLEM STATEMENT

Let the sequence of measurements of the cardiac signal be given as $(x_i, y_i), i = \overline{1,I}$; $x_1=a$, $x_I=b$, where (a,b) is the signal detection domain, I is the number of measurements in the detection domain.

The first task is to segment the cardiac signal into cardiac cycles by determining the sequence of R-waves $\{R(e)\}$, $(e=1,...,E)$, where E is the number of R-waves.

The main task is to determine for each cardiac cycle in the $(R(e), R(e+1))$, $(e=1,...,E-1)$ interval the number of the count corresponding to the end of the T-wave, that is, the minimum signal value immediately following the maximum signal value for the T-wave.

Commonly used approaches to cardiac signal segmentation, T-wave recognition, and T-wave endpoint detection involve the use of bandpass filters to eliminate interference. There is an empirical understanding that useful ECG signals belong to the frequency range of 0.5 Hz – 10 Hz. All frequencies outside of this range are considered interference. However, it is possible that the interference appears in the same frequency range as the useful ECG signal, as well as for certain parts of the ECG signal, frequencies greater than 10 Hz are specific. As a result, the processed signal may be distorted. In this case, either the ECG signal processing may be refused, or errors in

DOI 10.15588/1607-3274-2025-3-14

cardiac signal processing are possible, and, as a result, it is impossible to provide the patient with a high-quality ECG diagnosis.

It is relevant to develop methods and algorithms that are capable of processing an ECG signal in the presence of interference without the use of bandpass filters. In particular, methods and algorithms that use variable resolution do not affect the signal at all, and the cardiac signal to be processed.

This paper presents an algorithm for segmenting the cardiac signal, recognizing *T*-waves and *T*-wave endpoints with variable resolution.

2 REVIEW OF THE LITERATURE

Various methods for detecting *T*- and *P*-waves can be found in the literature: Discrete Fourier Transform, Discrete Cosine Transform, and adaptive filters. The distinction between *P*- and *T*-waves is considered in [1–3]. An algorithm based on fractional-order digital differentiation for detecting *P*- and *T*-waves is proposed [4]. A method for detecting monophasic *P*- and *T*-waves is described in [5]. A generalized and robust method for *P*-and *T*-waves detection is described in [6]. The identification of *P*-and *T*-waves based on fuzzy theory is discussed in [7]. A multi-stage methodology using wavelet transform is used to determine the *P*-waves, as proposed in [8]. The Discrete Wavelet Transform, which uses the Haar wavelet to detect the peak of the *T*-waves and the end of the *T*-wave, is considered in [9]. A mathematical model based on *T*-waves recognition is proposed in [10]. The classification and identification of *T* and *P* waves based on the support vector method is discussed in [11]. In recent years, there has been a significant use of systems based on field programmable gate arrays for ECG processing [12] and *QRS* detection [13].

In [14, 15], the algorithm for detecting *P* and *T* waves implemented in real time is considered. The *P* and *T* waves are identified based on their location in the non-*QRS* region and the corresponding *T* and *P* waves search zones are formed. The waves detection algorithm is divided into two parts, namely, the training period and the detection period. During the initial training period, the characteristic of the *R*-wave, the *R–R* interval, the polarity and the maximum slope of the *T* and *P* waves are determined (Fig. 2). If *i* (point A) is the current reading, then points B and C correspond to the *i*-30 and *i*-60 positions of the readings In Fig. 2 shows the probable positions A, B, and C located in the *T*-wave. The point B, which corresponds to the (*i*-30th) count, is checked for the presence of a valid *T* or *P* peak after the *i*-th count. The slopes of segments BA and BC relative to point B must correspond to certain predefined values.

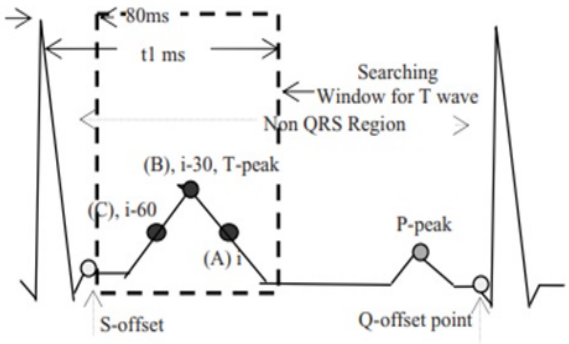


Figure 2 – *T*-peak detection

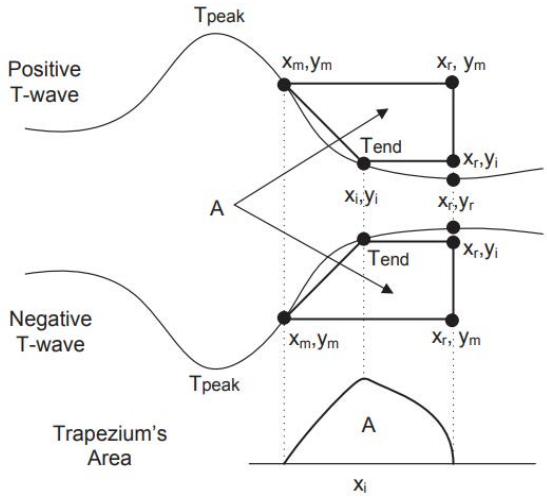


Figure 3 – Determination of the *T*- wave end (for a monophasic wave) by calculating the areas of several trapezoids formed by three fixed points and one moving point (x_i, y_i). The trend corresponds to the point where the area A is maximized

Paper [16] presents a robust and numerically efficient method based on two moving average filters developed to detect P and T waves in electrocardiograms (ECGs). The algorithm that implements the method detects *P*-and *T*-waves in the presence of interference.

It uses preliminary information about the duration of the *P*-and *T*-waves to make decisions. Bandpass filtering is applied to eliminate baseline drift and high frequencies.

Detection of *T*-wave endpoints on an ECG is a basic procedure for ECG processing and analysis[17,18].

In [19], an algorithm for detecting the end point of the *T*-wave on the ECG in the presence of broadband noise was investigated.

To detect the *Tend* point, various methods have been proposed based on: line intersection [20], thresholding by *T*-wave amplitude [21], thresholding by the first derivative of the ECG signal [22], calculating distances [23], angles [24], and areas [25], correlation with a pattern [26], mathematical models of the ECG [27], wavelet transform [28], and other methods. All of them have certain advantages and disadvantages due to complexity, computational costs, morphological variations of

DOI 10.15588/1607-3274-2025-3-14

waveforms, sensitivity to noise, and dependence of *Tend* on the threshold.

The trapezoidal area method assumes that *T*-peaks are found by searching for maxima and local minima in a window that starts with the previous *R*-wave peak.

The trapezoidal area method method is based on the calculation of successive areas of a rectangular trapezoid with three fixed vertices and one moving vertex: (x_i, y_i), which is shifted under the influence of the signal from point (x_m, y_m) to point (x_r, y_i), and the total area is calculated. The *T*-wave is defined as the point where the area A of the trapezoid is maximized (Fig. 3).

3 MATERIALS AND METHODS

In terms of image processing within the structural model, the part of the cardiac signal to be processed is part of a contour that delimits some imaginary object. Like any contour, the cardiac signal can be described as a cellular complex with a sequence of 1-cells and 0-cells [29]. Cardiac signal processing consists primarily in segmentation into cardiac cycles. Usually, *R*-waves are used to segment the cardiac signal into cardiac cycles, i.e., the sequence of *R*-waves in the processed part of the cardiac signal is determined. An *R*-wave can be represented by a model (Fig. 4), which is determined by two events: a rise along the *QR* line and a fall along the *RS* line. The derivative of the rise and fall must be greater than a certain value of θ in absolute value.

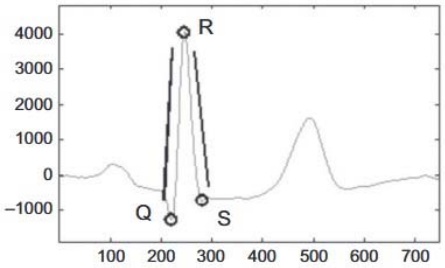


Figure 4 – The *R*- wave as a sequence of two events: 1) growth along the QR line; 2) decline along the RS line.
The derivative of growth and decline in absolute value must be greater than a certain value of θ

Further processing of the cardiac signal is performed within cardiac cycles. A cardiac cycle is a part of the cardiac signal between two adjacent *R*-waves. In particular, the *T*-wave is a part of the cardiac cycle located next to the *QRS* complex or next to the *ST*-segment, which is not always present in the cardiac signal. The most accurate determination of the *T*-wave end is considered particularly important for diagnosis. The *T*-wave can be monophasic – one wave is positive or negative, or biphasic (positive-negative, or vice versa). In this study, we consider the monophasic positive case. It is believed that other configurations can be considered as variants of this basic case. The complexity of determining the *T*-wave end increases due to the presence of interference.

It should be noted that the *T*-wave as a detection object has certain differences from other cardiac signal objects, in particular, from the *R*-wave. Usually, contour objects that are sequences of line segments and curve arcs have breakpoints in their composition, for example, like an *R*-wave. A *T*-wave, on the other hand, usually does not contain any breakpoints. Therefore, a *T*-wave is an object without any specific points, in particular, without clearly defined start and end points. Since we are considering the monophasic positive case, the *T*-wave is defined by the following two events:

1. the presence of the first signal maximum after the *R*-wave and

2. a minimum or a certain interval following it, the signal level of which is close to the baseline level.

Determination of *R*-wave and *T*-wave features is complicated by the presence of interference. In the presence of interference, the events that define the elements of the cardiac signal are determined using variable resolution [30].

Thus, the detection of both *R*-waves and *T*-waves is conditioned by the determination of two events characteristic for each of them.

The proposed algorithm defines a separate event for the discrete realization of the ECG signal at variable resolution. That is, two types of events should be defined – rise and fall – to determine the *R*-wave, and two types of events – maximum and minimum – to determine the *T*-wave. All types of events must be determined by the proposed algorithm.

A discrete realization is a sequence of measurement pairs $(x_i, y_i), i=\overline{1,I}$; $x_1=a$, $x_I=b$, where (a,b) is the signal detection domain, I is the number of measurements in the detection domain.

The definition area is divided into equal intervals, the values of which are set and changed during the operation of the algorithm. Each interval contains the same number of samples q, which determines the resolution The number J of parts, containing q samples, corresponding to the function definition domain determines the number of "coarse" samples at a given resolution. For each of the parts $q_j, j=\overline{1,J}$, the value of the "coarse" counts $z_j = g(w_j)$ is calculated from the sequence of counts $w_j = \{y_{q*(j-1)+1}, y_{q*(j-1)+2}, \ldots, y_{q*(j-1)+q}$, that are part of this interval. All values of the coarse samples form a discrete realization of the function $Z^{(m)} = \{z_j^{(m)}\}, j=\overline{1.J}$ at a given resolution m, where $m=\overline{1,M}$, is the total number of resolutions used to solve the problem using this algorithm.

An *R*-wave is characterized by two events. Event (R1) is an increase in the signal from the value of *Q* corresponding to the sample $i_Q^{(j)}$ to the value of R corresponding to the sample $i_R^{(j)}$. Event (R2) is a decrease in the signal from the value of R corresponding to the sample

DOI 10.15588/1607-3274-2025-3-14

$i_R^{(j)}$ to the value of S corresponding to the sample $i_S^{(j)}$. The value of growth and decline (first derivative) by absolute value must exceed a predetermined threshold $(R-Q)/(i_R^{(j)}-i_Q^{(j)})>\theta$, and $(R-S)/(i_R^{(j)}-i_S^{(j)})>\theta$.

The list of sample numbers for *R*-waves that form a preliminary answer for resolution *m*: $R^{(m)}=i_{R_1}^{(m)}, i_{R_2}^{(m)},..., i_{R_n}^{(m)},...$ Two preliminary answers for resolutions *m* and *m* + 1 are the same if:

1. The number of *R*-waves for $R^{(m)}$ is equal to the number of *R*-waves for $R^{(m+1)}$, i.e. the lengths of the both lists are equal.

2. The condition $(i\,R_n^{(m)}-q^{(m)})\le i\,R_n^{(m+1)}\le(i\,R_n^{(m)}+q^{(m)})$ is fulfilled. That is, the count number of the R_n-wave from the list $R^{(m+1)}$ corresponds to the count number of the R_n-wave from the list $R^{(m)}$, if the count number $i\,R^{(m+1)}$ of the *R*-wave does not differ from the count number $i\,R^{(m)}$ of the *R*- wave by more than $q^{(m)}$.

The allocation of the *T*-wave end immediately following the *QRS* complex is considered on the case of a positive monophasic *T*-wave. It should be noted that the maximum value of the *T*-wave is the largest after the *R*-wave, and the end of the *T*-wave is determined by the beginning of the minimum value following the maximum of the *T*-wave.

The *T*-wave is the part of the cardiac cycle following the *ST*-segment and is characterized, in particular, by two events. The event (T_1) is defined as the determination of the maximum signal value in the definition area. Event (T_1) is characterized by the corresponding "exact" count $i^{(j)}=j*q$ and the bounds of the cardiac cycle at which the event was recorded. The event (T2) is defined as the determination of the minimum signal value within the current cardiac cycle or one that does not exceed a predetermined threshold *d*, with $i_{(T2)}>> i_{(T1)}$, i.e., the event (T_2) occurs significantly after the event (T_1).

Next, we present an algorithm for segmenting the ECG into cardiac cycles and an algorithm for determining the *T*-wave endpoints for discrete realization of the ECG signal at variable resolution.

The main steps of the algorithm are:

1. Determine the number of "coarse" samples *J*, the number of resolutions *M*, and the values of the variables at the initial resolution *m*=1.

2. For each value of the resolution *m* (*m*=1,...,*M*), the lists of events {$R_1(m)$}, {$R_2(m)$} are formed, which determine the list of *R*-waves – {*R*(*m*)}, more precisely the list of their sample numbers $\{i_{R^{(m)}}\}$.

3. Determine the lists of *R*-waves – {*R*(*m*)}, which are appropriate, that is, the same for the largest number of resolutions, which form the answer to the final list of *R*-wave samples – {*R*(*e*)}, (*e*=1,...,*E*), more precisely the list of their sample numbers $\{i_{R^{(e)}}\}$, where *e* is the number of *R*-waves in the final list.

4. For each of the cardiac cycles having an interval $(i\,R^{(e)}, i\,R^{(e+1)})$, (*e*=1,...,*E*–1), the maximum signal value located in the sequence of samples after the *QRS* complex at all resolution values is determined. These values are preliminary answers to the maximum value of the *T* wave.

5. For each of the cardiac cycles having an interval $(i\,R^{(e)}, i\,R^{(e+1)})$, (*e*=1,...,*E*–1), determine the minimum signal value (or one that is close to the baseline) located in the sequence of samples after the maximum value determined in the previous step at all resolution values. These values are preliminary answers to the end of the *T* wave.

6. For each of the cardiac cycles, among the preliminary responses to the maximum value and the end of the *T*-wave, determine those that are appropriate, that is, the same for the greatest number of resolutions. The obtained values of the *T*-wave ends are considered final and a list of *T*-wave end values is formed.

4 EXPERIMENTS

In the process of experimental verification of the proposed algorithm using variable resolution, more than 100 fragments of cardiograms were processed, each containing about 30 cardiac cycles. Although the selected fragments of cardiograms were distorted by noise, no preprocessing of the cardiac signal, in particular, filtering, was used. Cardiograms with monophasic *T*-waves were selected for the experiment, which does not affect the generalizability of the results to other types of *T*-waves. An example of cardiac signal fragment segmentation in the presence of interference is shown in Fig. 5. Fig. 6 shows examples the *T*-waves endpoint determining.

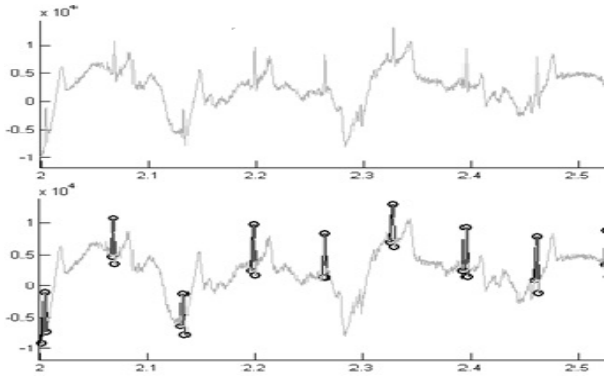


Figure 5 – An example of cardiac signal fragment segmentation in the presence of interference

DOI 10.15588/1607-3274-2025-3-14

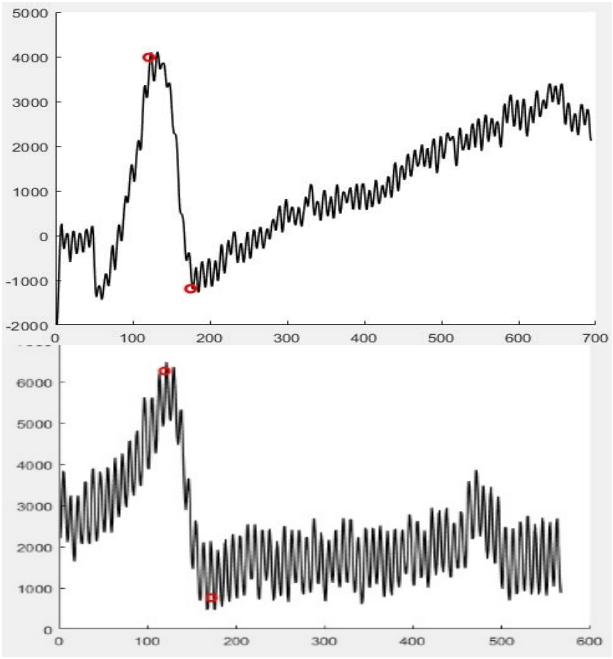


Figure 6 – Examples the *T*-waves endpoint determining

5 RESULTS

Algorithms for cardiac signal processing using variable resolution have been developed and experimentally verified, namely, the algorithm for segmentation of the cardiac signal into cardiac cycles and the algorithm for *T*-wave end detection, which is of great importance in cardiac diagnostics.

Means of cardiac signal processing, using the proposed algorithms, do not change the processed cardiac signal, unlike traditional means that use filtering of the cardiac signal, distorting the cardiac signal itself, which leads to distortion of the processing result.

6 DISCUSSION

ECG is most often distorted by noise in the measurement process and analog-to-digital conversion. The predominant noise in ECG is baseline wander, power line noise and electromyogram. Baseline wander is caused by the patient's movements due to breathing; the frequency range of baseline wander is usually below 0.5 Hz, which is in the same frequency range of ST segments. Power line noise is the 50 Hz/60 Hz component caused by parasitic electromagnetic fields from power lines and interferes with the analysis of low-amplitude components. It is necessary to place power lines as far away as possible or shield them, since improper electrical insulation will cause such noise. Electromyogram is a signal caused by muscle activity in the body, its frequency band is in the range of (5 Hz – 2 kHz) or (1 Hz – 5 kHz), its influence is difficult to exclude because its frequency band overlaps with the frequency band of ECG. In addition, there are other sources of interference due to motion artifacts, electrode contacts and electronic devices.

There are 4 typical filter processes in an ECG device: (a) anti-aliasing and upper-frequency cutoff, (b) baseline wander suppression and lower-frequency cutoff, (c) line-frequency rejection, and (d) muscle artifact reduction.

All types of frequency filtering affect the original cardiac signal to one degree or another, which leads to its distortion and loss of diagnostic features. At the same time, when processing the cardiac signal by algorithms using variable resolution, the original cardiac signal does not change in principle, all diagnostic features are preserved, can be identified and used.

Algorithms for cardiac signal processing using variable resolution have been developed and experimentally verified, namely, the algorithm for segmentation of the cardiac signal into cardiac cycles and the algorithm for *T*-wave detection, which is of great importance in cardiac diagnostics. The probability of a correct answer for traditional algorithms using pre-filtering is up to 98%. No errors were found in the experimental validation of the proposed algorithms. This makes it possible to eliminate distortion of the cardiac signal during ECG acquisition and improve the quality of cardiac diagnostics.

CONCLUSIONS

The article deals with algorithms of cardiac signal processing.

Scientific novelty consists in the fact that algorithms of cardiac signal processing in the presence of interference using variable resolution typical of visual perception are proposed.

The practical significance consists in the fact that the means of cardiac signal processing, using the proposed algorithms, do not change the processed cardiac signal, unlike traditional means that use filtering of the cardiac signal, distorting the cardiac signal itself, which leads to distortion of the processing result. The use of the presented tools in practical medical practice will lead to an improvement in the quality of cardiac diagnostics and, as a result, the quality of treatment.

Prospects for further research are as follows. It is supposed to develop tools for selecting all objects of the cardiac cycle.

ACKNOWLEDGEMENTS

The work is supported by the state budget scientific research project of the Institute of Mathematical Machines and Systems Problems “Structural methods of processing cyclic biomedical signals and cloud services based on them” (state registration number 0121U110584).

REFERENCES

1. Murthy I. S. N., Niranjan U. C. Component wave delineation of ECG by filtering in the fourier domain, *Medical & Biological Engineering & Computing*, 1992, Vol. 30, pp. 169–176. DOI: 10.1007/bf0244-6127
2. Murthy I. S. N., Prasad G. S. S. D. Analysis of ECG from pole-zero models, *IEEE Transactions on Biomedical Engineering*, 1992, Vol. 39, №7, pp. 741–751. DOI: 10.1109/10.142649
3. Thakor N. V., Zhu Y. S. Application of adaptive filtering to ECG analysis: Noise cancellation and arrhythmia detection, *IEEE Transactions on Biomedical Engineering*, 1991, Vol. 38, № 8, pp. 785–793. DOI: 10.1109/10.83591

DOI 10.15588/1607-3274-2025-3-14

4. Goutas F. Y., Herbeuval J. P., Boudraa M. et al. Digital fractional order differentiation-based algorithm for P and T-waves detection and delineation, *ITBM-RBM*, 2005, Vol. 26, pp. 127–132. DOI: 10.1016/j.rbmret.2004.11.022
5. Li C., Zheng C., Tai C. Detection of ECG characteristic points using wavelet transforms, *IEEE Transactions on Biomedical Engineering*, 1995, Vol. 42, №1, pp. 21–28. DOI: 10.1109/10.362922
6. Martínez J.P., Almeida R., Olmosat S. et al. A Wavelet-Based ECG Delineator: Evaluation on Standard Databases, *IEEE Transactions on Biomedical Engineering*, 2004, Vol. 51, № 4, pp. 570–581. DOI: 10.1109/TBME.2003.821031
7. Mehta S. S., Saxena S. C., Verma H. K. Recognition of P and T waves in electrocardiograms using fuzzy theory, *Biomedical Engineering Society of India: 14th Conference, New Delhi, 01–08 February 1995: proceedings.* Los Alamitos: IEEE 1995, pp. 15–18. DOI: 10.1109/RCEMBS.1995.511733
8. Sovilj S., Jeras M., Magjarevic R. Real Time P-wave Detector Based on Wavelet Analysis, *IEEE: 12th IEEE Mediterranean Electrotechnical Conference, Dubrovnik*, Croatia, May 12–15 2004: proceedings, 2004, pp. 403–406. DOI:10.1109/MELCON.2004.–1346895
9. Wong S., Francisco N., Mora F. et al. QT Interval Time Frequency Analysis using Haar Wavelet, *Computers in Cardiology*, 1998, Vol. 25, pp. 405–408. DOI: 10.1109/CIC.1998.731888
10. Vila J.A., Gang Y., Presedo J. M. R. et al. A new approach for TU complex characterization, *IEEE Transactions on Biomedical Engineering*, 2000, Vol. 47, №6, pp. 764–772. DOI: 10.1109/10.844227
11. Mehta S. S., Lingayat N. S. Detection of P and T-waves in Electrocardiogram, *Engineering and Computer Science: Proceedings of theWorld Congress*. San Francisco, USA, October 22–24 2008, pp. 22–24. DOI: 10.1109/ICCIMA.2007.25
12. Ieong C. I., Vai M. I., Mak P. E. et al. QRS recognition with programmable hardware, *Bioinformatics and Biomedical Engineering: 2nd. Annual conference: proceedings*. – Shanghai, 16–18 May 2008, pp. 2028–2031. DOI: 10.1109/ICBBE.2008.836
13. Shukla S. and Macchiarulo L. A fast and accurate FPGA based QRS detection system, *Engineering in Medicine and Biology: 30th. annual IEEE international conference: proceedings*. Vancouver, Canada, 20–25 August, 2008, pp. 4828–4831. DOI: 10.1109/IEMBS.2008.4650294
14. Chatterjee H. K., Gupta R., Bera J. N. et al. An FPGA implementation of real-time QRS detection algorithm, *Computer and Communication Technology: IEEE 2nd International conference*. Allahabad, India, Sept 15–17 2011, pp. 274–279. DOI: 10.1109/ICCCT.2011.6075114
15. Chatterjee H. K., Gupta R., Mitra M. et al. Real time P and T wave detection from ECG using FPGA, *Procedia Technology*, 2012, № 4, pp. 840–844. DOI: 10.1016/j.protcy.2012.05.138
16. Elgendi M., Meo M., Abbott D. et al. A Proof-of-Concept Study: Simple and Effective Detection of P and T Waves in Arrhythmic ECG Signals, *Bioengineering*, 2016, Vol.26, №3, pp. 1–14. DOI: 10.3390/bioengineering3040026
17. Clifford G. D., Azuaje F., McSharry P. et al. Advanced Methods And Tools for ECG Data Analysis. Norwood, USA: Artech House Publishers, 2006, 400 p. DOI: 10.1186/1475-925X-6-18
18. Vázquez-Seisdedos C. R., Neto J. E., Reyes E.J.M. at al. New approach for T-wave end detection on electrocardiogram: Performance in noisy conditions, *BioMedical Engineering OnLine*, 2011. http://www.biomedical-engineering-online.com/content /10/1/77 DOI: 10.1186/1475-925X-10-77
19. Friesen G. M., Jannette T. C., Jadallah M. A. et al. A comparison of the noise sensitivity of nine QRS detection algorithms, *IEEE Transactions on Biomedical Engineering*, 1990, Vol. 37, pp. 85–98. DOI: 10.1109/10.43620
20. Ferreti G. F. Re L., Zayat M. et al. A New Method for the Simultaneous Measurement of the RR and QT Intervals in Ambulatory ECG Recordings, *Computers in Cardiology, IEEE Computer Society*, 1992, pp. 171–174. DOI: 10.1109/CIC.1992.269419
21. McLaughlin N. B., Campbell R. W., Murray A. Comparison of automatic QT measurement techniques in the normal 12 lead electrocardiogram, *Br Heart J.*, 1995, Vol. 74, pp. 84–89. DOI: 10.1136/hrt.74.1.84
22. Laguna P., Thakor N. V., Caminal P. New algorithm for QT interval analysis in 24- hour Holter ECG: performance and applications, *Medical&Bioljgical Engineering&Computing*, 1990, Vol. 28, pp. 67–73. DOI: 10.1007/BFO2441680
23. Helfenbein E. D., Zhou S. H., Lindauer J. M. et al. An algorithm for continuous real-time QT interval monitoring, *Journal of Electrocardiology*, 2006, Vol. 39, pp. 123–127. DOI: 10.1016/j.jelectrocard.-2006.05.18
24. Daskalov I. K., Christov I. I. Automatic detection of the electrocardiogram T-wave end, *Medical&Bioljgical Engineering&Computing*, 1999, Vol. 37, pp. 348–353. DOI: 10.1007/BFO2513311
25. Zhang Q., Manriquez A. Illanes, Médigue C. et al. An Algorithm for Robust and Efficient Location of TWave Ends in Electrocardiograms, *IEEE Transactions Biomedical Engineering*, 2006, Vol. 53, pp. 2544–2552. DOI: 10.1109/TBME.2006.884644
26. Last T., Nugent C. D., Owens F. J. Multi-component based cross correlation beat detection in electrocardiogram analysis, *Biomedical Engineering Online*, 2004, Vol. 3, P. 26 [http://www.biomedical-engineering-online.com/content/3/1/26]. Doi: 10.1186/1475-925X-3-26
27. Vila J., Gang Y., Presedo J. et al. A new approach for TU complex characterization, *IEEE Transactions Biomedical Engineering*, 2000, Vol. 47, pp. 764–772. DOI:10.1109/10.844227
28. MartAtnez J. P., Almeida R., Olmos S. et al. A Wavelet-Based ECG Delineator: Evaluation on Standard Databases, *IEEE Transactions Biomedical Engineering*, 2004, Vol. 51, pp. 570–581. DOI: 10.1109/TBME.2003.821031
29. Kalmykov V. G., Sharypanov A. V., Vishnevskey V. V. The curve arc as a structure element of an object contour in the image to be recognized, *Radio Electronics, Computer Science, Control*, 2023, №1, pp. 89–98. DOI:10.15588/1607-3274-2023-1-9. WOS:001066631000009
30. Kalmykov V., Sharypanov A. Segmentation of Experimental Curves Distorted by Noise, *Journal of Computer Science Systems Biology*, 2017, Vol. 10, № 3, pp. 50–59. doi:10.4172/jcsb.1000248

Accepted 10.01.2025.
Received 12.06.2025.

DOI 10.15588/1607-3274-2025-3-14

УДК 004.93

ОБРОБЛЕННЯ КАРДІОСИГНАЛУ АЛГОРИТМАМИ, ЩО ВИКОРИСТОВУЮТЬ ЗМІННУ РОЗДІЛЬНУ ЗДАТНІСТЬ

Калмиков В. Г. – канд. техн. наук, старший науковий співробітник Інституту проблем математичних машин і систем, Київ, Україна.

Шарипанов А. В. – канд. техн. наук, завідувач лабораторії медичної і біологічної інформатики Інституту проблем математичних машин і систем, Київ, Україна.

Вишневський В. В. – канд. техн. наук, провідний науковий співробітник Інституту проблем математичних машин і систем, Київ, Україна.

АНОТАЦІЯ

Актуальність. Запропонована робота відноситься до області обробки кардіосигналів, зокрема, до сегментації кардіосигналу на серцеві цикли, а також до визначення однієї з найважливіших ознак, що використовується в кардіодіагностиці, – кінця зубця *T*.

Мета. Метою і завданням дослідження є розробка алгоритму обробки кардіосигналу в присутності завад, який дозволяє виділити необхідні для діагностики ознаки і, в той же час, не спотворює вихідний сигнал, як це зазвичай відбувається при його обробці смуговими цифровими фільтрами для виключення завад, що призводить до спотворення і можливої втрати діагностичних ознак.

Запропонований **метод** полягає у представленні серцевого сигналу як частини контуру певного уявного зображення. Обробка серцевого сигналу полягає перш за все у сегментації на серцеві цикли. Зазвичай *R*-зубці використовують для сегментації серцевого сигналу на серцеві цикли, тобто визначають послідовність *R*-зубців в частині серцевого сигналу, що підлягає обробленню. При визначенні *R*-зубця використовується модель, яка передбачає збільшення сигналу з наступним його зменшенням, причому швидкість збільшення (зменшення) має бути за абсолютною величиною більшою, ніж певне задане значення. Для обраного сегмента серцевого сигналу послідовність *R*-зубців визначається з різними роздільними здатностями. Відповіддю є послідовність, яка повторюється для найбільшої кількості роздільних здатностей з тих, що були задіяні для сегментації серцевого сигналу на серцеві цикли. Модель *T*-зубця можна представити як послідовність дуг кривих без розривів. В одному з поширених випадків *T*-зубець визначається найбільшим максимумом серцевого сигналу в межах серцевого циклу, наступним за *R*-зубцем. Кінець *T*-зубця визначається першим мінімумом, наступним за вже визначеним максимумом для *T*-зубця. Як і у випадку сегментації серцевого сигналу, максимум *T*-зубця і кінець *T*-зубця визначаються при різних роздільних здатностях, а відповіддю вважаються ті значення, які збігаються при найбільшій кількості роздільних здатностей з тих, що були використані.

Результати. Розроблено та експериментально перевірено алгоритми оброблення кардіосигналу, що використовують змінну роздільну здатність, а саме алгоритм сегментації серцевого сигналу на кардіоцикли та алгоритм виявлення кінця зубця T, що має велике значення в кардіологічній діагностиці. Засоби оброблення серцевого сигналу, що використовують запропоновані алгоритми, не змінюють оброблений кардіосигнал, на відміну від традиційних засобів, які використовують фільтрацію серцевого сигналу, спотворюючи сам серцевий сигнал, що призводить до спотворення результату оброблення.

Висновки. Наукова новизна полягає в тому, що запропоновано алгоритми обробки серцевого сигналу за наявності перешкод із використанням змінної роздільної здатності, характерної для зорового сприйняття. Практичне значення полягає в тому, що засоби оброблення кардіосигналу, які використовують запропоновані алгоритми, не спотворюють кардіосигнал, що оброблюється, на відміну від традиційних засобів, які використовують фільтрацію серцевого сигналу, що призводить до спотворення результату оброблення. Використання представленого інструментарію в практичній медичній практиці призведе до підвищення якості кардіологічної діагностики і, як наслідок, якості лікування.

КЛЮЧОВІ СЛОВА: кардіосигнал, сегментація, кардіоцикли, кінець зубця *T*, змінна роздільна здатність.

ЛІТЕРАТУРА

1. Murthy I. S. N. Component wave delineation of ECG by filtering in the fourier domain / I. S. N. Murthy, U. C. Niranjan // Medical & Biological Engineering & Computing. – 1992. – Vol. 30. – P. 169–176. DOI: 10.1007/bf0244-6127
2. Murthy I. S. N. Analysis of ECG from pole-zero models / I. S. N. Murthy, G. S. S. D. Prasad // IEEE Transactions on Biomedical Engineering. – 1992. – Vol. 39, №7. – P. 741–751. DOI: 10.1109/10.142649
3. Thakor N. V. Application of adaptive filtering to ECG analysis: Noise cancellation and arrhythmia detection / N. V. Thakor, Y. S. Zhu // IEEE Transactions on Biomedical Engineering. – 1991. – Vol. 38, №8. – P. 785–793. DOI: 10.1109/10.83591
4. Digital fractional order differentiation-based algorithm for P and T-waves detection and delineation / [F. Y. Goutas, J. P. Herbeuval, M. Boudraa at al.] //ITBM-RBM. – 2005. – Vol. 26. – P. 127–132. Doi: 10.1016/j.rbmret.2004.11.022
5. Li C. Detection of ECG characteristic points using wavelet transforms / C. Li, C. Zheng, C.Tai // IEEE Transactions on Biomedical Engineering. – 1995. – Vol. 42, №1. – P. 21–28. DOI: 10.1109/10.362922
6. A Wavelet-Based ECG Delineator: Evaluation on Standard Databases / [J. P. Martínez, R. Almeida, S. Olmosat at al.]// IEEE Transactions on Biomedical Engineering. – 2004. – Vol. 51, № 4. – P. 570–581. DOI: 10.1109/TBME.2003.821031
7. Mehta S. S. Recognition of P and T waves in electrocardiograms using fuzzy theory / S. S. Mehta, S. C. Saxena, H. K. Verma // Biomedical Engineering Society of India: 14th Conference, New Delhi, 01–08 February 1995: proceedings. – Los Alamitos : IEEE 1995. – P. 15–18. DOI: 10.1109/RCEMBS.1995.511733
8. Sovilj S. Real Time P-wave Detector Based on Wavelet Analysis/ S. Sovilj, M. Jeras, R. Magjarevic // IEEE: 12th IEEE Mediterranean Electrotechnical Conference, Dubrov-

DOI 10.15588/1607-3274-2025-3-14

nik, Croatia, May 12–15 2004: proceedings. – 2004. – P. 403–406. DOI:10.1109/MELCON.2004.-1346895
9. Wong S. QT Interval Time Frequency Analysis using Haar Wavelet/ [S.Wong, N. Francisco, F. Mora et al.] // Computers in Cardiology. – 1998. – Vol. 25. – P. 405–408. DOI: 10.1109/CIC.1998.731888
10. Vila J.A. A new approach for TU complex characterization/ [J. A. Vila, Y. Gang, J. M. R. Presedo et al.] // IEEE Transactions on Biomedical Engineering. – 2000. – Vol. 47, №6. – P. 764–772. DOI: 10.1109/10.844227
11. Mehta S. S. Detection of P and T-waves in Electrocardiogram/ S. S. Mehta, N. S. Lingayat// Engineering and Computer Science: Proceedings of theWorld Congress, San Francisco, USA, October 22–24. – 2008. – P. 22–24. DOI: 10.1109/ICCIMA.2007.25
12. Ieong C. I. QRS recognition with programmable hardware/ [C. I. Ieong, M. I. Vai, P. E. Mak et al.]// Bioinformatics and Biomedical Engineering: 2nd. Annual conference: proceedings. – Shanghai, 16–18 May 2008. – P. 2028–2031. DOI: 10.1109/ICBBE.2008.836
13. Shukla S. A fast and accurate FPGA based QRS detection system / S. Shukla, and L. Macchiarulo // Engineering in Medicine and Biology: 30th. annual IEEE international conference: proceedings. – Vancouver, Canada, 20–25 August, 2008. – P.4828–4831. DOI: 10.1109/IEMBS.2008.4650294
14. Chatterjee H. K. An FPGA implementation of real-time QRS detection algorithm / [H. K. Chatterjee, R. Gupta, J. N. Bera et al.] // Computer and Communication Technology: IEEE 2nd International conference, Allahabad, India, Sept 15–17 2011. – P. 274–279. DOI: 10.1109/ICCCT.2011.6075114
15. Real time P and T wave detection from ECG using FPGA/ [H. K. Chatterjee, R. Gupta, M. Mitra et al.]// Procedia Technology. – 2012. – № 4. – P. 840–844. DOI: 10.1016/j.protcy.2012.05.138
16. A Proof-of-Concept Study: Simple and Effective Detection of P and T Waves in Arrhythmic ECG Signals / [M. Elgendi, M. Meo, D. Abbott et al.] // Bioengineering. – 2016. – Vol. 26, №3. – P. 1–14. DOI: 10.3390/bioengineering3040026
17. Advanced Methods And Tools for ECG Data Analysis / [G. D. Clifford, F. Azuaje, P. McSharry et al.]. – Norwood, USA: Artech House Publishers, 2006. – 400 p. DOI: 10.1186/1475-925X-6-18
18. Vázquez-Seisdedos C. R. New approach for T-wave end detection on electrocardiogram: Performance in noisy conditions/ [C. R. Vázquez-Seisdedos, J. E. Neto, E. J. M Reyes et al.] // BioMedical Engineering OnLine. – 2011. – http://www.biomedical-engineering-online.com/content /10/1/77 Doi: 10.1186/1475-925X-10-77
19. Friesen G. M. A comparison of the noise sensitivity of nine QRS detection algorithms/ [G. M. Friesen, T. C. Jannette, M. A. Jadallah et al.] // IEEE Transactions on Biomedical Engineering. – 1990. – Vol. 37. – P. 85–98. DOI: 10.1109/10.43620
20. Ferreti G. F. A New Method for the Simultaneous Measurement of the RR and QT Intervals in Ambulatory ECG Recordings/ [G. F. Ferreti, L. Re, M. Zayat, et al.] // Computers in Cardiology, IEEE Computer Society. – 1992. – P. 171–174. DOI: 10.1109/CIC.1992.269419
21. McLaughlin N. B. Comparison of automatic QT measurement techniques in the normal 12 lead electrocardiogram / N. B. McLaughlin, R. W. Campbell, A. Murray// Br Heart J. – 1995. – Vol. 74. – P. 84–89. DOI: 10.1136/hrt.74.1.84
22. Laguna P. New algorithm for QT interval analysis in 24-hour Holter ECG: performance and applications / P. Laguna, N. V. Thakor, P. Caminal // Medical&Bioljgical Engineering&Computing. – 1990. – Vol. 28. – P. 67–73. DOI: 10.1007/BFO2441680
23. Helfenbein E. D. An algorithm for continuous real-time QT interval monitoring/ [E. D. Helfenbein, S. H. Zhou, J. M. Lindauer at al.] Journal of Electrocardiology. – 2006. – Vol. 39. – P. 123–127. DOI: 10.1016/j.jelectrocard.-2006.05.18
24. Daskalov I. K. Automatic detection of the electrocardiogram T-wave end / I. K. Daskalov, I. I. Christov // Medical&Bioljgical Engineering&Computing. – 1999. – Vol. 37. – P. 348–353. Doi: 10.1007/BFO2513311
25. Zhang Q. An Algorithm for Robust and Efficient Location of TWave Ends in Electrocardiograms/ [Q. Zhang, A. Illanes Manriquez, C. Médigue et al.] // IEEE Transactions Biomedical Engineering. – 2006. – Vol. 53. – P. 2544–2552. Doi: 10.1109/TBME.2006.884644
26. Last T. Multi-component based cross correlation beat detection in electrocardiogram analysis / T. Last, C. D. Nugent, F. J. Owens // Biomedical Engineering Online. – 2004. – Vol. 3. – P. 26. [http://www.biomedical-engineering-online.com/content/3/1/26]. Doi: 10.1186/1475-925X-3-26
27. A new approach for TU complex characterization/ [J. Vila, Y. Gang, J. Presedo et al.] // IEEE Transactions Biomedical Engineering. – 2000. – Vol. 47. – P.764–772. DOI:10.1109/10.844227
28. A Wavelet-Based ECG Delineator: Evaluation on Standard Databases/ [J. P. MartAtnez, R. Almeida, S. Olmos et al.] // IEEE Transactions Biomedical Engineering. – 2004. – Vol. 51. – P. 570–581. DOI: 10.1109/TBME.2003.821031
29. Kalmykov V. G. The curve arc as a structure element of an object contour in the image to be recognized / V. G. Kalmykov, A. V. Sharypanov, V. V. Vishnevskey// Radio Electronics, Computer Science, Control. – 2023 . – № 1. – P. 89–98. DOI:10.15588/1607-3274-2023-1-9. WOS:001066631000009
30. Kalmykov V. Segmentation of Experimental Curves Distorted by Noise / V. Kalmykov, A. Sharypanov // Journal of Computer Science Systems Biology. – 2017. – Vol. 10. – №3. – P. 50–59. DOI:10.4172/jcsb.1000248

DOI 10.15588/1607-3274-2025-3-14

UDC 004.05

METHODS FOR EVALUATING SOFTWARE ACCESSIBILITY

Kuz Mykola – Dr. Sc., Professor of the Department of Information Technology Vasyl Stefanyk Precarpathian National University, Ivano-Frankivsk, Ukraine.

Yaremiy Ivan – Dr. Sc., Professor of the Department of Applied Physics and Materials Science, Vasyl Stefanyk Precarpathian National University, Ivano-Frankivsk, Ukraine.

Yaremii Hanna – MSc., Vasyl Stefanyk Precarpathian National University, Ivano-Frankivsk, Ukraine.

Pikuliak Mykola – PhD, Associate Professor of the Department of Information Technology Vasyl Stefanyk Precarpathian National University, Ivano-Frankivsk, Ukraine.

Lazarovych Ihor – PhD, Associate Professor of the Department of Information Technology Vasyl Stefanyk Precarpathian National University, Ivano-Frankivsk, Ukraine.

Kozlenko Mykola – PhD, Associate Professor of the Department of Information Technology Vasyl Stefanyk Precarpathian National University, Ivano-Frankivsk, Ukraine.

Vekeryk Denys – MSc., Vasyl Stefanyk Precarpathian National University, Ivano-Frankivsk, Ukraine.

ABSTRACT

Context. The development and enhancement of methods for evaluating software accessibility is a relevant challenge in modern software engineering, as ensuring equal access to digital services is a key factor in improving their efficiency and inclusivity. The increasing digitalization of society necessitates the creation of software that complies with international accessibility standards such as ISO/IEC 25023 and WCAG. Adhering to these standards helps eliminate barriers to software use for individuals with diverse physical, sensory, and cognitive needs. Despite advancements in regulatory frameworks, existing accessibility evaluation methodologies are often generalized and fail to account for the specific needs of different user categories or the unique ways they interact with digital systems. This highlights the need for the development of new, more detailed methods for defining metrics that influence the quality of user interaction with software products.

Objective. Building a classification and mathematical model and developing accessibility assessment methods for software based on it.

Methods. A method for assessing the quality subcharacteristic "Accessibility", which is part of the "Usability" quality characteristic, has been developed. This enabled the analysis of a website's inclusivity for individuals with visual impairments, and the formulation of specific recommendations for further improvements, which is a crucial step toward creating an inclusive digital environment.

Results. Comparing to standardized approaches, a more detailed and practically oriented accessibility assessment methodology has been proposed. Using this methodology, an analysis of the accessibility of the main pages of Vasyl Stefanyk Precarpathian National University's website was conducted, and improvements were suggested to enhance its inclusivity.

Conclusions. This study presents the development of a classification and mathematical model, along with an accessibility assessment methodology for websites based on the ISO 25023 standard, and an analysis of the main pages of the university's web portal. The identified quantitative accessibility indicators enable an evaluation of the web resource's compliance with modern inclusivity requirements and provide recommendations for its improvement.

The scientific novelty of this research lies in the development of assessment methods for the "Accessibility" quality subcharacteristic by introducing new subproperties and attributes of software quality, based on clearly defined metrics specifically adapted for evaluating the accessibility level of digital products for individuals with visual impairments. This approach ensures a more precise and objective determination of web resources' compliance with inclusivity requirements, contributing to their effectiveness and usability for this user group.

The practical significance of the obtained results lies in their applicability for objectively evaluating the accessibility of software products and web resources.

KEYWORDS: accessibility, inclusivity, quality subproperty, quality attribute, perceptiveness, operability, understandability, localization.

ABBREVIATIONS

ISO is an International Organization for Standardization;
IEC is an International Electrotechnical Commission;
DSTU is a National Standard of Ukraine;
WCAG is a Web Content Accessibility Guidelines.

NOMENCLATURE

$X_{UAC-1-G}$ is an "Accessibility for users with disabilities" quality property;

$A_{UAC-1-G}$ is a number of functions successfully used by the users with a specific disability;

$B_{UAC-1-G}$ is a number of functions implemented;

$X_{UAC-2-S}$ is a "Supported languages adequacy" quality property;

$A_{UAC-2-S}$ is a number of languages actually supported;

$B_{UAC-2-S}$ is a number of languages needed to be supported;

$X_{UAC-1.1.1-G}$ is an "Alternative text" quality attribute;

DOI 10.15588/1607-3274-2025-3-15

$A_{UAC-1.1.1-G}$ is a number of multimedia elements with meaningful text alternatives;

$B_{UAC-1.1.1-G}$ is a total number of multimedia elements in the system;

$X_{UAC-1.1.2-G}$ is a "Color contrast" quality attribute;

$A_{UAC-1.1.2-G}$ is a contrast level;

$B_{UAC-1.1.2-G}$ is a total number of elements;

$B_{UAC-1.1.2-G^+}$ is a number of elements that meet the following conditions: contrast level ≥ 4.5:1 for main text and ≥ 3:1 for auxiliary text;

$X_{UAC-1.1.3-G}$ is a "Subtitles and audio descriptions" quality attribute;

$A_{UAC-1.1.3-G}$ is a number of videos with subtitles or audio descriptions;

$B_{UAC-1.1.3-G}$ is a total number of videos in the system;

$X_{UAC-1.2.1-G}$ is a "Keyboard navigation" quality attribute;

$A_{UAC-1.2.1-G}$ is a number of interactive elements accessible via keyboard navigation;

$B_{UAC-1.2.1-G}$ is a total number of interactive elements in the system;

$X_{UAC-1.2.2-G}$ is a "Structured navigation" quality attribute;

$A_{UAC-1.2.2-G}$ represents presence or absence of breadcrumbs (1 if present, 0 if absent);

$B_{UAC-1.2.2-G}$ is a number of skipped heading levels;

$C_{UAC-1.2.2-G}$ is a total number of titles;

$X_{UAC-1.3.1-G}$ is a "Clear instructions" quality attribute;

$A_{UAC-1.3.1-G}$ is a number of instructions rated as clear;

$B_{UAC-1.3.1-G}$ is a total number of instructions;

$X_{UAC-1.3.2-G}$ is a "Input assistance" quality attribute;

$A_{UAC-1.3.2-G}$ is a number of fields with autocomplete or hint functions;

$B_{UAC-1.3.2-G}$ is a total number of fields that could have autocomplete or hint functions;

$X_{UAC-1.3.3-G}$ is a "Correct input support" quality attribute;

$A_{UAC-1.3.3-G}$ is a number of forms with error messages;

$B_{UAC-1.3.3-G}$ is a total number of forms in the system;

$X_{UAC-1.1-G}$ is a "Perceptiveness" quality subproperty;

$w_{UAC-1.1.1-G}$ is a weight of "Alternative text" quality attribute;

$w_{UAC-1.1.2-G}$ is a weight of "Color contrast" quality attribute;

$w_{UAC-1.1.3-G}$ is a weight of "Subtitles or audio descriptions" quality attribute;

$X_{UAC-1.2-G}$ is an "Operability" quality subproperty;

$w_{UAC-1.2.1-G}$ is a weight of "Keyboard navigation" quality attribute;

$w_{UAC-1.2.2-G}$ is a weight of "Structured navigation" quality attribute;

$X_{UAC-1.3-G}$ is a "Understandability" quality subproperty;

$w_{UAC-1.3.1-G}$ is a weight of "Clear instructions" quality attribute;

$w_{UAC-1.3.2-G}$ is a weight of "Input assistance" quality attribute;

$w_{UAC-1.3.3-G}$ is a weight of "Correct input support" quality attribute;

$X_{UAC-2.1-S}$ is a "Localization" quality subproperty;

w_1 is weight of state language;

$A_{UAC-2.1-S}$ represents presence of the state language;

w_2 is a weight of English language;

$B_{UAC-2.1-S}$ represents presence of the English language;

w_3 is a weight of popular European languages;

$C_{UAC-2.1-S}$ represents presence of one of the popular European languages

w_4 is weight of other languages;

$D_{UAC-2.1-S}$ represents presence of other languages;

$w_{UAC-1.1-G}$ is a weight of "Perceptiveness" quality subproperty;

$w_{UAC-1.2-G}$ is a weight of "Operability" quality subproperty;

$w_{UAC-1.3-G}$ is a weight of "Understandability" quality subproperty;

$w_{UAC-2.1-S}$ is a weight of "Localization" quality subproperty;

$w_{UAC-1-G}$ is a weight of "Accessibility for users with disabilities" quality property;

$w_{UAC-2-S}$ is a weight of "Supported languages adequacy" quality property.

INTRODUCTION

The development and assessment of software products in today's environment require consideration of a wide range of characteristics that define their quality. Among these, the subcharacteristic "Accessibility" is one of the most critical, as it determines how easily a software system can be used by the broadest possible range of users.

Accessibility is particularly relevant in the context of increasing focus on inclusivity, as ensuring that software

DOI 10.15588/1607-3274-2025-3-15

products can be used by individuals with diverse physical, sensory, and cognitive needs not only promotes social equity but also expands the potential user base [1].

According to the ISO 25023 standard [2], the assessment of software product quality is based on metrics that quantitatively reflect their efficiency, usability, and accessibility.

The accessibility criterion for users with disabilities determines how successfully individuals with specific physical, cognitive, or sensory limitations can complete tasks within a system. The level of accessibility is measured as the ratio of successfully utilized functions to the total number of implemented functions. This approach provides an objective assessment of a system's usability for users with various disabilities. At the same time, applying this criterion highlights the need to consider the diverse requirements of individuals with different types of impairments, as their needs can vary significantly. For instance, users with sensory impairments may require enhanced visual elements, while those with physical disabilities may prioritize adaptive system functionalities to improve interaction.

The system's compliance with users' language support needs is assessed by evaluating the level of language support, which is determined as the proportion of actually supported languages relative to those required for effective use. The language support metric is crucial for evaluating the internationalization of a software product, particularly in multilingual regions or among users with diverse linguistic preferences. It helps determine how well the system accommodates linguistic diversity and whether it can provide a seamless user experience in a multilingual environment.

Although these metrics provide fundamental guidelines for assessing system quality, their application often encounters limitations due to their generalized nature. They do not always account for the specific use cases or the unique operating conditions of the software. This highlights the need for their expansion and refinement to better address users' specific needs and usage contexts. Such an approach would enable a more objective and effective accessibility evaluation, ultimately improving the overall quality of the software product.

The object of this study is the process of evaluating software accessibility.

The subject of the study is the methods for determining the qualitative assessment of accessibility.

The aim of this work is to develop a classification and mathematical model and, based on it, design methods for evaluating software accessibility.

1 PROBLEM STATEMENT

The problem of modeling software accessibility indicators for individuals with visual impairments is considered.

Let the subcharacteristic "Accessibility", in accordance with the ISO 25023 standard [2], include the quality attribute "Accessibility for users with disabilities" (UAC-1-G), which defines the extent to which potential users, including those with impairments or disabilities, can successfully use a software system or product. This quality attribute ensures that users with visual, auditory, motor, or cognitive impairments can complete their tasks within the system, potentially with the aid of assistive technologies. As specified in [2], this quality attribute is determined using the following formula:

$$X_{UAC-1-G} = \frac{A_{UAC-1-G}}{B_{UAC-1-G}}. \quad (1)$$

However, the methodology for determining parameter $A_{UAC-1-G}$ is not provided in standard [2].

The second component of the "Accessibility" subcharacteristic is the quality attribute "Supported languages adequacy" (UAC-2-S). This attribute evaluates the issue that users often encounter operational errors when attempting to use software in a language different from their native one. Misinterpretation of descriptions and messages leads to a decrease in accessibility. According to [2], this quality attribute is determined using the following formula:

$$X_{UAC-2-S} = \frac{A_{UAC-2-S}}{B_{UAC-2-S}}. \quad (2)$$

The methodology for determining parameter $B_{UAC-2-S}$ is not provided in standard [2] as well.

The following tasks are to be addressed: 1) the development of a classification and mathematical model of software accessibility indicators by introducing a new group of indicators – quality attributes – based on which the parameters $A_{UAC-1-G}$ and $B_{UAC-2-S}$ can be determined; 2) the development of methods for qualitative assessment of accessibility; 3) the validation of the proposed methodology using the example of a university web portal.

2 REVIEW OF THE LITERATURE

This study focuses on users with visual impairments, as they represent the most active group among individuals with disabilities using the internet [3]. The further analysis centers on adapting the ISO 25023 standard [2] for accessibility evaluation and expanding its metrics to better address the needs of this user category.

Currently, software quality assessment according to ISO 25010 [4] is conducted based on the quality properties defined in ISO 25023 [2]. This process involves evaluating quality characteristics and subcharacteristics, which collectively contribute to the overall assessment of a software product's quality. Specifically, ISO 25023 [2] provides a set of measures that enable the quantitative evaluation of software quality. The application of metric analysis methods in this process allows for the calculation of numerical metric values that describe the degree to

DOI 10.15588/1607-3274-2025-3-15

which a software product meets defined requirements. According to ISO 24765 [5], a metric is defined as a numerical measure of the extent to which a product possesses a specific property, making it a crucial tool in the quality assurance process.

The existing standardized criteria for evaluating quality properties, as outlined in [2], lack a sufficient methodological framework for assessment. Specifically, standard [2] provides formulas for determining various quality properties and descriptions of formula parameters but does not include clear methodologies for their determination. This also applies to the "Accessibility" subcharacteristic. Additionally, the document does not define any criteria for establishing weighting coefficients for quality metrics.

Study [6] initiated research aimed at improving standardized methodologies for software quality assessment. The applied aspects of software quality assessment methods are presented in studies [7–9]. These works provide examples of the practical validation of software quality assessment methodologies and qualitative evaluation of a web forum, thereby demonstrating the effectiveness of the developed approaches.

An essential component of the methodology for evaluating software quality measures is the determination of weighting coefficients. One such method is presented in study [10]. It is based on expert evaluation of quality measures and includes tools for verifying the accuracy and reliability of expert decisions. The method described in [10] served as the foundation for the weighting coefficient determination methodology (significance levels) for software quality metrics, including characteristics, subcharacteristics, and quality attributes.

Based on the analysis of existing software quality evaluation methods, it was determined that, despite the availability of a developed methodology for assessing software quality, including methods for determining weighting coefficients, the primary drawback is the absence of methods for defining specific quality attributes, which represent the lowest level of software product quality metrics.

As a result, this work presents the development of a method for evaluating one of the quality subcharacteristics, "Accessibility", along with its constituent quality attributes: "Accessibility for users with disabilities" and "Supported languages adequacy".

3 MATERIALS AND METHODS

An effective accessibility assessment will ensure the inclusivity of the system, making it intuitive and functional for the widest possible range of users.

By breaking down the quality attributes of the "Accessibility" subcharacteristic into subproperties, four key subproperties have been identified:

– perceptiveness (UAC-1.1-G);
– operability (UAC-1.2-G);
– understandability (UAC-1.3-G);
– localization (UAC-2.1-S).

This approach enables a more detailed examination of various aspects of accessibility that directly impact the ability of users with disabilities to successfully interact with the system.

"Perceptiveness" is critical for ensuring accessibility, as users with visual or hearing impairments rely on text alternatives, color contrast, subtitles, and audio descriptions to effectively perceive information.

"Operability" defines how well the interface can be used with different input methods, including keyboard, mouse, and assistive devices, which is especially important for users with motor impairments.

"Understandability" is fundamental to creating an inclusive experience, as it ensures clarity of instructions, accessibility of error correction features, and support for proper data input, reducing cognitive load.

"Localization" ensures the adaptation of textual content and interface elements to the linguistic and cultural characteristics of users, enabling diverse audiences to interact effectively with the web resource. This is especially relevant in multilingual environments, where accurate translation and proper content structure play a key role in information perception and comprehension.

These quality subproperties were identified as priorities due to their direct impact on users' ability to successfully complete tasks within the system. They reflect the fundamental principles of accessibility outlined in standards such as WCAG [11] and ISO 25023 [2], allowing for a more detailed examination of aspects that pose the greatest challenges for users with specific limitations.

Let's take a closer look at each subproperty, starting with "Perceptiveness", one of the most critical aspects influencing how users perceive information.

The first critically important quality attribute is "Alternative text" (UAC-1.1.1-G), which defines the proportion of images that have correct textual descriptions. The formula for calculating this indicator is as follows:

$$X_{UAC-1.1.1-G} = \frac{A_{UAC-1.1.1-G}}{B_{UAC-1.1.1-G}}. \quad (3)$$

The second quality attribute is "Color contrast" (UAC-1.1.2-G), which evaluates the compliance of text-to-background contrast levels with established requirements. The formula for calculating this indicator is as follows:

$$X_{UAC-1.1.2-G} = \frac{\sum_{i=1}^{n}\left(A_{UAC-1.1.2-Gi} \cdot B_{UAC-1.1.2-Gi^+}\right)}{\sum_{i=1}^{n}\left(A_{UAC-1.1.2-Gi} \cdot B_{UAC-1.1.2-Gi}\right)}. \quad (4)$$

The final parameter of perceptiveness is "Subtitles and audio descriptions" (UAC-1.1.3-G), which assesses the proportion of video content that includes subtitles or audio descriptions, focusing on the accessibility of multime-

DOI 10.15588/1607-3274-2025-3-15

dia content for users with visual impairments. This parameter is determined using the following formula:

$$X_{UAC\text{–}1.1.3\text{–}G} = \frac{A_{UAC\text{–}1.1.3\text{–}G}}{B_{UAC\text{–}1.1.3\text{–}G}}. \quad (5)$$

The next accessibility subproperty is “Operability” (UAC-1.2-G), which evaluates how easily users can interact with the interface using different input methods, such as a keyboard or mouse. This subproperty is critical for ensuring accessibility for users with visual impairments, who often rely on the keyboard as their primary navigation tool.

To evaluate the proportion of interactive elements that are accessible via keyboard navigation, the quality attribute “Keyboard navigation” (UAC-1.2.1-G) is introduced. This attribute is calculated using the following formula:

$$X_{UAC\text{–}1.2.1\text{–}G} = \frac{A_{UAC\text{–}1.2.1\text{–}G}}{B_{UAC\text{–}1.2.1\text{–}G}}. \quad (6)$$

The second quality attribute is “Structured navigation” (UAC-1.2.2-G). This parameter evaluates the proportion of pages with well-structured navigation, such as breadcrumbs, and a clear heading hierarchy, and is described by one of the following formulas:

$$X_{UAC\text{–}1.2.2\text{–}G} = 0.5 \cdot A_{UAC\text{–}1.2.2\text{–}G} + \\ + 0.5 \cdot \left(1 - \frac{B_{UAC\text{–}1.2.2\text{–}G}}{C_{UAC\text{–}1.2.2\text{–}G}}\right), \quad (7)$$

or

$$X_{UAC\text{–}1.2.2\text{–}G} = 1 - \frac{B_{UAC\text{–}1.2.2\text{–}G}}{C_{UAC\text{–}1.2.2\text{–}G}}. \quad (8)$$

The variable A takes a value of 1 or 0, depending on the presence or absence of breadcrumbs, respectively. Formula (7) is used if the page is located at a deep hierarchical level; otherwise, Formula (8) is applied.

The third identified subproperty of “Accessibility” is “Understandability”, which evaluates how easily users can comprehend information and interface elements. This subproperty is critical for ensuring intuitive interaction with a website, particularly for users with visual impairments or cognitive disabilities. A clear and well-structured interface reduces errors, improves system usability, and enhances the overall user experience for a diverse audience.

To evaluate the clarity of form-filling instructions, we define the quality attribute “Clear Instructions” (UAC-1.3.1-G). This attribute analyzes whether users are provided with clear, specific, and understandable guidelines for completing forms. It is determined using the following formula:

$$X_{UAC\text{–}1.3.1\text{–}G} = \frac{A_{UAC\text{–}1.3.1\text{–}G}}{B_{UAC\text{–}1.3.1\text{–}G}}. \quad (9)$$

The next “Understandability” attribute is “Input Assistance” (UAC-1.3.2-G), which evaluates the proportion of forms that include features to prevent or correct input errors. These features include autofill, interactive hints, real-time data validation, and format input notifications. This quality attribute is designed to reduce errors and enhance user interaction with forms. It is calculated using the following formula:

$$X_{UAC\text{–}1.3.2\text{–}G} = \frac{A_{UAC\text{–}1.3.2\text{–}G}}{B_{UAC\text{–}1.3.2\text{–}G}}. \quad (10)$$

Another quality attribute of the “Understandability” subproperty is “Correct input support” (UAC-1.3.3-G), which evaluates the presence of functional error messages that help users identify and correct mistakes when entering data into forms. This attribute is determined using the following formula:

$$X_{UAC\text{–}1.3.3\text{–}G} = \frac{A_{UAC\text{–}1.3.3\text{–}G}}{B_{UAC\text{–}1.3.3\text{–}G}}. \quad (11)$$

The next step is to calculate the values of the quality subproperties.

The “Perceptiveness” subproperty is determined based on the quality attributes “Alternative Text”, “Color contrast” and “Subtitles and audio descriptions” weighted by their respective weighting coefficients:

$$X_{UAC\text{–}1.1\text{–}G} = w_{UAC\text{–}1.1.1\text{–}G} \cdot X_{UAC\text{–}1.1.1\text{–}G} + \\ + w_{UAC\text{–}1.1.2\text{–}G} \cdot X_{UAC\text{–}1.1.2\text{–}G} + \\ + w_{UAC\text{–}1.1.3\text{–}G} \cdot X_{UAC\text{–}1.1.3\text{–}G}. \quad (12)$$

The “Operability” subproperty is determined based on the quality attributes “Keyboard navigation” and “Structured navigation”, weighted by their respective weighting coefficients:

$$X_{UAC\text{–}1.2\text{–}G} = w_{UAC\text{–}1.2.1\text{–}G} \cdot X_{UAC\text{–}1.2.1\text{–}G} + \\ + w_{UAC\text{–}1.2.2\text{–}G} \cdot X_{UAC\text{–}1.2.2\text{–}G}. \quad (13)$$

The “Understandability” subproperty is determined based on the quality attributes “Clear instructions”, “Input assistance” and “Correct input support” weighted by their respective weighting coefficients:

$$X_{UAC\text{–}1.3\text{–}G} = w_{UAC\text{–}1.3.1\text{–}G} \cdot X_{UAC\text{–}1.3.1\text{–}G} + \\ + w_{UAC\text{–}1.3.2\text{–}G} \cdot X_{UAC\text{–}1.3.2\text{–}G} + \\ + w_{UAC\text{–}1.3.3\text{–}G} \cdot X_{UAC\text{–}1.3.3\text{–}G}. \quad (14)$$

DOI 10.15588/1607-3274-2025-3-15

The "Localization" subproperty (UAC-2.1-S) is included as part of the "Supported languages adequacy" quality attribute. It measures the number of available language versions of the interface and is determined using the following formula:

$$X_{UAC-2.1-S} = w_1 \cdot A_{UAC-2.1-S} + w_2 \cdot B_{UAC-2.1-S} + \\ + w_3 \cdot C_{UAC-2.1-S} + w_4 \cdot D_{UAC-2.1-S}. \quad (15)$$

To evaluate "Localization" the weighting coefficient method is applied, which assigns significance to each language based on the linguistic environment of the country where the website was developed, as well as the resource's target audience in international markets.

In this study, the evaluation of the "Localization" subproperty is based on weighting coefficients specifically determined for a Ukrainian website. These coefficients take into account Ukraine's linguistic context, where Ukrainian is the primary language, English serves as an international communication medium, and German and French are among the most widely spoken languages in Europe. This approach ensures that the model is adapted to real-world conditions and meets the needs of the target audience. Thus, Ukrainian has the highest weighting coefficient (0.6) as it is the state language for the target audience. English is assigned a coefficient of 0.2 due to its importance as an international language of communication. German and French each have a coefficient of 0.08, reflecting their relevance for European users where these languages are widely spoken. Other languages receive the lowest coefficient (0.04), as they cover less significant audience segments.

The indicators $A_{UAC-2.1-S}$, $B_{UAC-2.1-S}$, $C_{UAC-2.1-S}$, $D_{UAC-2.1-S}$ take a value of 1 if the corresponding language is available on the website and 0 if it is absent.

The quality attribute values are derived from their respective subproperties, weighted by their corresponding weighting coefficients.

The "Accessibility for users with disabilities" attribute is determined using the following formula:

$$X_{UAC-1-G} = w_{UAC-1.1-G} \cdot X_{UAC-1.1-G} + \\ + w_{UAC-1.2-G} \cdot X_{UAC-1.2-G} + \\ + w_{UAC-1.3-G} \cdot X_{UAC-1.3-G}. \quad (16)$$

The "Supported languages adequacy" attribute is determined using the following formula:

$$X_{UAC-2-S} = w_{UAC-2.1-S} \cdot X_{UAC-2.1-S}. \quad (17)$$

The integral measure of the "Accessibility" quality subcharacteristic is determined as the sum of the products of each metric's value and its corresponding weighting coefficient:

$$X_{UAC} = w_{UAC-1-G} \cdot X_{UAC-1-G} + \\ + w_{UAC-2-S} \cdot X_{UAC-2-S}. \quad (18)$$

The integral indicator provides an assessment of the overall compliance of the university web portal with modern inclusivity requirements.

For a clearer representation of the structure of accessibility indicators, Figure 1 presents a classification model illustrating the distribution of key quality attributes, subproperties, and properties. This model visually highlights the essential aspects of web accessibility described earlier.

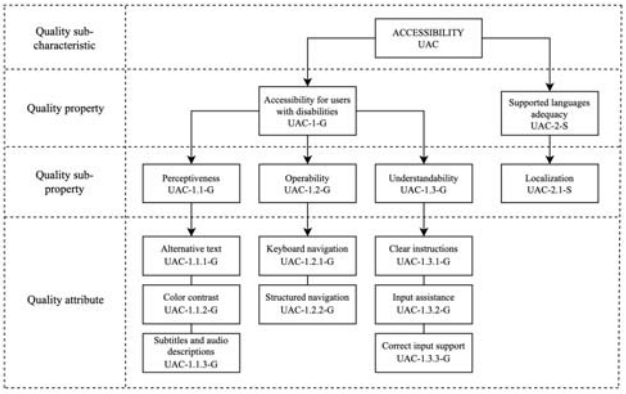


Figure 1 – Classification model of accessibility measures

A comprehensive approach to accessibility evaluation, covering perceptiveness, operability, understandability, and localization, enables a detailed analysis of all key aspects of user interaction with a web resource. The use of quantitative indicators and weighting coefficients ensures an objective assessment, allowing not only to determine the current level of compliance with accessibility standards but also to develop specific recommendations for improvement.

Thus, the proposed evaluation system not only reflects the overall level of accessibility but also helps identify the most problematic areas that require improvement. This contributes to enhancing the user experience for individuals with diverse needs, making the web portal more inclusive and user-friendly.

4 EXPERIMENTS

The methods developed in this study were applied to assess the accessibility of the website of Vasyl Stefanyk Precarpathian National University. University portals play a crucial role in providing access to information and services for a wide audience. Therefore, the website must be clear and user-friendly not only for students and faculty but also for prospective applicants, parents, international partners, and individuals with disabilities.

The accessibility of a web resource ensures equal access to educational materials, registration services, and general university information. Additionally, it enhances the institution's reputation, demonstrating its commitment to inclusivity and openness. In today's digital environ-

DOI 10.15588/1607-3274-2025-3-15

ment, compliance with accessibility standards is not only a technical necessity but also a social imperative.

Next, the user interaction with key pages of the university website was evaluated.

Some accessibility metrics can be assessed more efficiently using automated tools and methodologies, which help streamline the evaluation process and provide objective results quickly [12]. For example, the Image Alt Checker service was used to analyze the alternative text of images (Figure 2), significantly reducing the workload involved in manual verification [13]. The color contrast assessment was conducted using the WCAG Contrast Checker extension (Figure 3), which automatically detects problematic elements on a page and ensures compliance with accessibility standards [14].

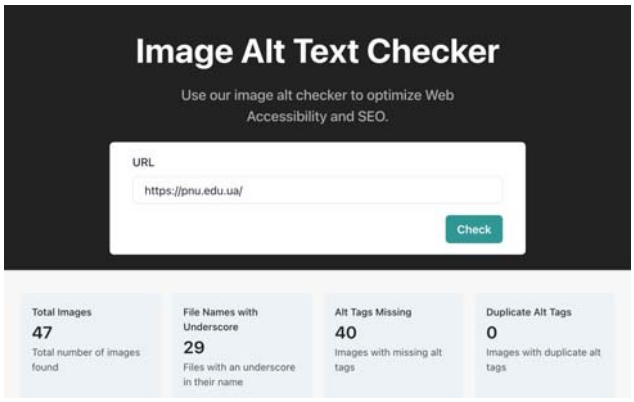


Figure 2 – Analysis results of the university website's homepage using the Image Alt Checker Service

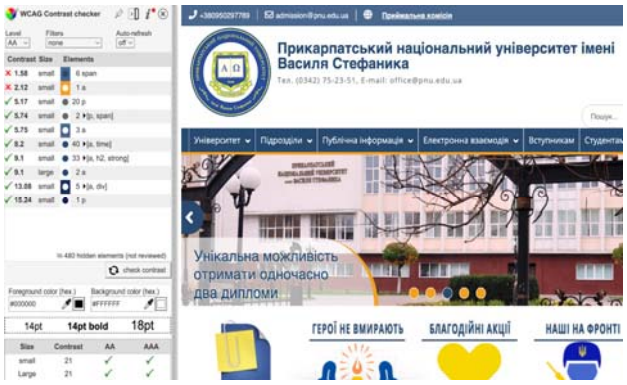


Figure 3 – Usage of the WCAG Contrast Checker plugin for color contrast evaluation

5 RESULTS

After calculating the values of each individual accessibility metric, their contribution to the overall assessment was determined through expert analysis. The expert evaluation method was used to establish weighting coefficients based on the importance of each quality indicator in shaping the overall accessibility level of the web resource.

To ensure clarity and ease of further calculations, the obtained values and weighting coefficients are presented in Table 1, allowing for a clear visualization of each metric's contribution to the final result.

Table 1 – Quality measures and their weights

ID j	Value X_j	Weight w_j
Quality attributes		
UAC-1.1.1-G	0.15	0.3
UAC-1.1.2-G	0.99	0.3
UAC-1.1.3-G	0	0.4
UAC-1.2.1-G	1	0.6
UAC-1.2.2-G	0.47	0.4
UAC-1.3.1-G	1	0.4
UAC-1.3.2-G	0	0.3
UAC-1.3.3-G	0.83	0.3
Quality subproperties		
UAC-1.1-G	0.342	0.3
UAC-1.2-G	0.788	0.3
UAC-1.3-G	0.649	0.4
UAC-2.1-S	0.8	1.0
Quality properties		
UAC-1-G	0.5986	0.6
UAC-2-S	0.8	0.4
Quality subcharacteristic		
UAC	0.67916	

Based on the data presented in Table 1, several recommendations can be made to improve the accessibility indicators of the Vasyl Stefanyk Precarpathian National University website. Key areas for improvement include the "Subtitles and audio descriptions" and "Input assistance" attributes, which have zero values, indicating a lack of implementation and the need for significant enhancement. The "Alternative text" attribute has a low value of 0.15, suggesting that many images lack proper descriptions. The "Structured navigation" attribute has a value of 0.47, meaning improvements in page hierarchy and breadcrumb navigation would be beneficial. Several quality attributes already meet high accessibility standards, with values either equal to 1 or close to 1 (0.99 and 0.83), indicating no immediate need for refinement. Increasing the numerical values of the weaker quality attributes will, in turn, improve the values of subproperties, properties, and the overall "Accessibility" subcharacteristic of the website.

For a more precise analysis, it is necessary to interpret the results based on a defined scale that classifies the level of accessibility and establishes minimum acceptable compliance thresholds. The scale, shown in Figure 4, is used for this purpose. This scale is developed based on Fibonacci numbers (the "golden ratio") and enables the determination of the quality level of the evaluated software product. This scale ensures objectivity in assessment by allowing results to be classified as "very poor", "poor", "satisfactory", "good", or "excellent". It also defines the minimum acceptable values, which are crucial for determining whether the web resource meets modern accessibility standards.

DOI 10.15588/1607-3274-2025-3-15

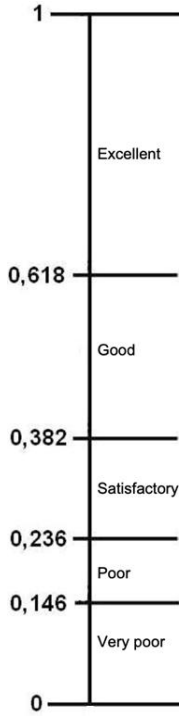


Figure 4 – Quality level scale for software products

According to the presented scale, the obtained overall accessibility assessment value of ≈0.68 falls within the "excellent" range, indicating a very high level of accessibility.

6 DISCUSSION

The developed mathematical model and accessibility evaluation methods demonstrate significant potential for analyzing and improving web resources in accordance with modern inclusivity standards.

Introducing new parameters into the subproperties of the "Accessibility" subcharacteristic allows for a more detailed assessment of this metric's quality.

The use of the proposed metrics enables a quantitative evaluation of accessibility at different stages of a software product's lifecycle, contributing to its improvement and enhancing the user experience.

The proposed methodology takes a systematic approach to analyzing web resources, focusing on key aspects of user interaction for individuals with disabilities.

Implementing the developed evaluation scale not only helps determine a website's compliance with modern accessibility requirements but also identifies specific areas that require improvement.

CONCLUSIONS

As a result of analyzing existing standardized accessibility evaluation metrics, it was found that while these metrics serve as a foundation for determining accessibility levels, they do not fully reflect real-world usage conditions and user needs due to their generalized approach. This highlights the necessity for further refinement and adaptation to enable more precise evaluation of specific digital products.

A classification and mathematical model was developed, and based on them, methods for evaluating software accessibility were designed.

The scientific novelty of this study lies in the development of evaluation methods for the "Accessibility" quality subcharacteristic by introducing new subproperties and quality attributes for software products. These are based on clearly defined metrics specifically adapted to assess the accessibility level of digital products for users with visual impairments. This approach ensures a more precise and objective assessment of web resources' compliance with inclusivity requirements, enhancing their effectiveness and usability for this user group.

An accessibility assessment was conducted for the main pages of the Vasyl Stefanyk Precarpathian National University website, allowing for an evaluation of its compliance with the expanded requirements of ISO 25023 [2]. The analysis covered key aspects such as content perceptiveness, interface operability, information clarity, and localization. The application of the proposed methodology not only enables an assessment of the current state of website accessibility but also provides practical recommendations for its improvement.

The practical significance of the obtained results lies in their application for objective accessibility evaluation of software products and web resources. This contributes to improving quality, ensuring compliance with international standards such as ISO 25023 [2] and WCAG [11], and promoting inclusivity, thereby expanding the user audience, including people with disabilities.

ACKNOWLEDGMENTS

The research was conducted as part of the scientific project "Development of a Methodology for Assessing the Quality of Software Products for Measuring Instruments" (state registration number 0116U002344).

REFERENCES

1. Gartland S., Flynn P., Carneiro M. A., Holloway G., Fialho J.d.S., Cullen J., Hamilton E., Harris A., Cullen C. The State of Web Accessibility for People with Cognitive Disabilities: A Rapid Evidence Assessment. Behav. Sci., 2022, 25 p. DOI: 10.3390/bs12020026.
2. System and software engineering. System and software Quality Requirements and Evaluation (SQuaRE). Measurement of system and software product quality: ISO/IEC 25023:2016. [Effective from 2016-06-15]. Geneve, ISO, 2016, 54 p.
3. Barbosa Natã M., Hayes Jordan, Kaushik Smirity, and Wang Yang. "Every Website Is a Puzzle!": Facilitating Access to Common Website Features for People with Visual Impairments. ACM Trans. Access. Comput., 2022, 35 p. DOI: 10.1145/3519032.
4. System and software engineering. System and software Quality Requirements and Evaluation (SQuaRE). Product quality model: ISO/IEC 25010:2023. [Effective from 2023-11-15]. Geneve, ISO, 2023, 35 p.

DOI 10.15588/1607-3274-2025-3-15

5. Systems and software engineering. Vocabulary: ISO/IEC/IEEE 24765:2017. [Effective from 2017-09-15]. Geneve, ISO, 2017, 15 p.
6. Nisrina Nurhuda, Eko Darwiyanto, Sri Widowati / Implementation of Analytical Hierarchy Process (AHP) for Determining Priority of Software Assessment in West Java Provincial Government Based on ISO/IEC 25010 (Case Study: Sapawarga Application), *Ind. Journal on Computing*, 2021, Vol. 6, Issue. 1, pp. 23–40. DOI: 0.34818/indojc.2021.6.1.525.
7. Yarshini Thamilarasan, Raja Rina Raja Ikram, Mashanum Osman, Lizawati Salahuddin, Wan Yaakob Wan Bujeri, Kasturi Kanchymalay. Enhanced System Usability Scale using the Software Quality Standard Approach, *Engineering, Technology & Applied Science Research*, 2023, Vol. 13, Number 5, pp. 11779–11784. DOI: 10.48084/etasr.5971.
8. Pikuliak M. V., Kuz M. V., Lazarovych I. M., Kuzyk Y. M., Skliarov V. V. Method of determining the parameter of qualitative evaluation of a web forum, *Radio Electronics, Computer Science, Control*, 2024, No. 3, pp. 151–159. DOI: 10.15588/1607-3274-2024-3-13.
9. Aditia Arga Pratama, Achmad Benny Mutiara. Software Quality Analysis for Halodoc Application using ISO 25010:2011, *International Journal of Advanced Computer Science and Applications*, 2021, Vol. 12, Number 8, pp. 383-392. DOI: 10.14569/IJACSA.2021.0120844.
10. Kuz M., Kozlenko M., Lazarovych I., Rysniuk O., Novak V., Novak M. Method of weights determination based on ratings of software quality metrics, *Proceedings of the IV International Scientific and Practical Conference on Applied scientific and technical research, Ivano-Frankivsk, 1–3 April 2020: proceedings*. Ivano-Frankivsk, Vol. 2, pp. 37–39. DOI: 10.6084/m9.figshare.28381274.
11. Web Content Accessibility Guidelines (WCAG) 2.0 [Electronic resource]. Access mode: https://www.w3.org/TR/WCAG20/.
12. Abascal J., Arrue M., Valencia X. Tools for Web Accessibility Evaluation. Springer, 2019, pp. 479–503. DOI: 10.1007/978-1-4471-7440-0_26.
13. ContentForest – Image Alt Checker [Electronic resource]. Access mode: https://contentforest.com/tools/image-alt-checker.
14. WCAG Color contrast checker [Electronic resource]. Access mode: https://chromewebstore.google.com/detail/wcag-color-contrast-check/plnahcmalebffmaghcpcmpaciebdhgdf.

Received 02.04.2025.
Accepted 30.06.2025.

УДК 004.05

МЕТОДИ ОЦІНКИ ДОСТУПНОСТІ ПРОГРАМНИХ ПРОДУКТІВ

Кузь М. В. – д-р техн. наук, професор, професор кафедри інформаційних технологій, Прикарпатський національний університет імені Василя Стефаника, Івано-Франківськ, Україна.

Яремій І. П. – д-р фіз.-мат. наук, професор, професор кафедри матеріалознавства і новітніх технологій, Прикарпатський національний університет імені Василя Стефаника, Івано-Франківськ, Україна.

Яремій Г. І. – магістр, Прикарпатський національний університет імені Василя Стефаника, Івано-Франківськ, Україна.

Пікуляк М. В. – канд. техн. наук, доцент, доцент кафедри інформаційних технологій, Прикарпатський національний університет імені Василя Стефаника, Івано-Франківськ, Україна.

Лазарович І. М. – канд. техн. наук, доцент, доцент кафедри інформаційних технологій, Прикарпатський національний університет імені Василя Стефаника, Івано-Франківськ, Україна.

Козленко М. І. – канд. техн. наук, доцент, доцент кафедри інформаційних технологій, Прикарпатський національний університет імені Василя Стефаника, Івано-Франківськ, Україна.

Векерик Д. В. – магістр, Прикарпатський національний університет імені Василя Стефаника, Івано-Франківськ, Україна.

АНОТАЦІЯ

Актуальність. Розробка та вдосконалення методів оцінювання доступності програмних продуктів є актуальною задачею сучасної програмної інженерії, оскільки забезпечення рівного доступу до цифрових сервісів є ключовим фактором підвищення їхньої ефективності та інклюзивності. Зростаюча цифровізація суспільства вимагає створення програмного забезпечення, яке відповідає міжнародним стандартам доступності, таким як ISO/IEC 25023 та WCAG. Це дозволяє усувати бар'єри у використанні програмних продуктів людьми з різними фізичними, сенсорними та когнітивними потребами. Незважаючи на розвиток нормативних документів, існуючі методики оцінювання доступності часто мають узагальнений характер і не враховують специфічні потреби різних категорій користувачів, або особливості їх взаємодії з цифровими системами. Це створює необхідність розробки нових, більш деталізованих методів визначення показників, які впливають на якість взаємодії користувача із програмним продуктом.

Мета. Побудова класифікаційної та математичної моделі і розробка на її основі методів оцінювання доступності програмного забезпечення.

Методи. Розроблено метод оцінки підхарактеристики якості «Доступність», яка входить до складу характеристики якості «Зручність використання», що дало можливість виконати аналіз вебсайту на предмет інклюзивності для осіб із вадами зору та на його основі сформулювати конкретні рекомендації для подальшого вдосконалення, що є важливим кроком у напрямку створення інклюзивного цифрового середовища.

Результати. Запропоновано більш деталізовану та практично орієнтовану методику оцінювання доступності, у порівнянні із стандартизованими методиками. Використовуючи розроблену методику здійснено аналіз доступності основних сторінок вебсайту Прикарпатського національного університету імені Василя Стефаника та запропоновано вдосконалення вебсайту для підвищення його інклюзивності.

DOI 10.15588/1607-3274-2025-3-15

Висновки. У даному дослідженні виконано побудову класифікаційної та математичної моделі і розроблено методику оцінювання доступності вебсайтів на основі стандарту ISO 25023 та проведено аналіз основних сторінок вебпорталу університету. Визначені кількісні показники доступності дозволяють оцінити відповідність вебресурсу сучасним вимогам інклюзивності та сформувати рекомендації щодо його вдосконалення.

Наукова новизна полягає в розробці методів оцінки підхарактеристики якості «Доступність» шляхом введення нових підвластивостей та атрибутів якості програмних продуктів, що ґрунтуються на чітко визначених метриках, спеціально адаптованих для оцінювання рівня доступності цифрових продуктів для осіб із порушеннями зору. Такий підхід забезпечує більш точне та об'єктивне визначення відповідності вебресурсів вимогам інклюзивності, що сприяє підвищенню їхньої ефективності та зручності використання для зазначеної категорії користувачів.

Практичне значення отриманих результатів полягає в можливості їх застосування для об'єктивного оцінювання доступності програмних продуктів та веб-ресурсів.

КЛЮЧОВІ СЛОВА: доступність, інклюзивність, підвластивість якості, атрибут якості, перцептивність, керованість, зрозумілість, локалізація.

ЛІТЕРАТУРА

1. Gartland S. The State of Web Accessibility for People with Cognitive Disabilities: A Rapid Evidence Assessment / [S. Gartland, P. Flynn, M. A. Carneiro, et al.]. – Behav. Sci., 2022. – 25 p. DOI: 10.3390/bs12020026.
2. System and software engineering. System and software Quality Requirements and Evaluation (SQuaRE). Measurement of system and software product quality: ISO/IEC 25023:2016. – [Effective from 2016-06-15]. – Geneve: ISO, 2016. – 54 p.
3. "Every Website Is a Puzzle!": Facilitating Access to Common Website Features for People with Visual Impairments / [Natã M. Barbosa, Jordan Hayes, Smirity Kaushik, and Yang Wang]. – ACM Trans. Access. Comput., 2022. – 35 p. DOI: 10.1145/3519032.
4. System and software engineering. System and software Quality Requirements and Evaluation (SQuaRE). Product quality model: ISO/IEC 25010:2023. – [Effective from 2023-11-15]. – Geneve: ISO, 2023. – 35 p.
5. Systems and software engineering . Vocabulary: ISO/IEC/IEEE 24765:2017. – [Effective from 2017-09-15]. – Geneve: ISO, 2017. – 15 p.
6. Nisrina Nurhuda. Implementation of Analytical Hierarchy Process (AHP) for Determining Priority of Software Assessment in West Java Provincial Government Based on ISO/IEC 25010 (Case Study: Sapawarga Application) / Nisrina Nurhuda, Eko Darwiyanto, Sri Widowati // Ind. Journal on Computing. – 2021. – Vol. 6, Issue. 1. – P. 23–40. DOI: 0.34818/indojc.2021.6.1.525.
7. Enhanced System Usability Scale using the Software Quality Standard Approach / [Yarshini Thamilarasan, Raja Rina Raja Ikram, Mashanum Osman et al.] // Engineering, Technology & Applied Science Research. – 2023. – Vol. 13, № 5. – P. 11779–11784. DOI: 10.48084/etasr.5971.
8. Pikuliak M. V. Method of determining the parameter of qualitative evaluation of a web forum / [M. V. Pikuliak, M. V. Kuz, I. M. Lazarovych et al.] // Radio Electronics, Computer Science, Control. – 2024. – № 3. – P. 151–159. DOI: 10.15588/1607-3274-2024-3-13.
9. Aditia Arga Pratama. Software Quality Analysis for Halodoc Application using ISO 25010:2011 / Aditia Arga Pratama, Achmad Benny Mutiara // International Journal of Advanced Computer Science and Applications. – 2021. – Vol. 12, No. 8. – P. 383–392. DOI: 10.14569/IJACSA.2021.0120844.
10. Method of weights determination based on ratings of software quality metrics [M. Kuz, M. Kozlenko, I. Lazarovych et al.] // Proceedings of the IV International Scientific and Practical Conference Applied scientific and technical research, Ivano-Frankivsk 1–3 April 2020: proceedings. – Ivano-Frankivsk. – 2020. – Vol. 2. – P. 37–39. DOI: 10.6084/m9.figshare.28381274.
11. Web Content Accessibility Guidelines (WCAG) 2.0 [Electronic resource]. – Access mode: https://www.w3.org/TR/WCAG20/.
12. Abascal J. Tools for Web Accessibility Evaluation / J. Abascal, M. Arrue, X. Valencia. – Springer, 2019. – P. 479–503. DOI: 10.1007/978-1-4471-7440-0_26.
13. ContentForest – Image Alt Checker [Electronic resource]. – Access mode: https://contentforest.com/tools/image-alt-checker.
14. WCAG Color contrast checker [Electronic resource]. – Access mode: https://chromewebstore.google.com/detail/wcag-color-contrast-check/plnahcmalebffmaghcpcmpaciebdhgdf.

DOI 10.15588/1607-3274-2025-3-15

p-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2025. № 3
e-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2025. № 3

UDC 004.94

REDUNDANT ROBOTIC ARM PATH PLANNING USING RECURSIVE RANDOM INTERMEDIATE STATE ALGORITHM

Medvid A. Y. – Post-graduate student of the Department of Artificial Intelligence Systems, Lviv Polytechnic National University, Lviv, Ukraine.

Yakovyna V. S. – Dr. Sc., Professor of the Department of Artificial Intelligence Systems, Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. Collision-free path planning in joint space for redundant robotic manipulators remains a challenging task due to the high-dimensional configuration space and dynamically changing environments. Existing methods often struggle to balance search time and path quality, which is crucial for real-time applications.

Objective. The aim of this study is to develop a new method to plan efficient, collision-free trajectories in real time for redundant robotic manipulators.

Method. A novel sampling-based algorithm for collision-free joint space path planning for redundant robotic manipulators presented in this study. The algorithm is called the Recursive Random Intermediate State (RRIS). The RRIS algorithm primarily works by generating a set of random intermediate states and iteratively selecting the optimal one based on the number of collisions along the discretized path. Furthermore, the paper proposes an axis-aligned bounding box generation strategy and an early exit strategy to improve algorithm speed. Finally, repeated calls of the algorithm are proposed to improve its reliability. The performance of the RRIS algorithm is evaluated through a set of comprehensive tests and compared with the popular RRT Connect algorithm implemented in Open Motion Planning Library.

Results. Experimental evaluations show that the RRIS algorithm under the test conditions produces collision-free paths with significantly shorter average lengths and reduces search time by approximately three times compared to the RRT Connect algorithm.

Conclusions. The proposed RRIS algorithm demonstrates a promising approach to real-time path planning for redundant robotic manipulators. By combining strategic intermediate state sampling with efficient collision evaluation and early termination mechanisms, the algorithm offers a robust alternative to known methods.

KEYWORDS: path planning, redundant robotic manipulator, collision avoidance.

ABBREVIATIONS

RRIS is a Recursive Random Intermediate State;

RRT is a Rapidly-Exploring Random Trees;

OMPL is an Open Motion Planning Library [17];

DOF is degrees of freedom.

NOMENCLATURE

J^i is the angle of the i-th joint of robotic arm;

S is a state of the robotic arm;

S_i is the i-th state of discretized path between arm states;

J_i^j is the angle of j-th joint of an i-th state of discretized path between arm states;

$|Path|$ is a norm of the path between two arm states;

S_S is a start arm state;

S_F is a final arm state;

S_I is an intermediate arm state;

S_M is a middle state between start and final;

d is a step of displacement of the middle state in the joint space for generating intermediate states;

Dir_i is the step of displacement direction being made for the i-th joint (can take values –1, 0, 1);

BB is a bounding box in joints space;

M_{BB} is a bounding box margin;

N_r is a number of random intermediate states generated;

$step$ is a discretization step in radians used to trace between two arm states;

$Collisions_{S->F}$ is a number of collisions on discretized path between start and final states;

$Collisions_{S->I}$ is a number of collisions on discretized path between start and intermediate states;

$Collisions_{I->F}$ is a number of collisions on discretized path between intermediate and final states;

$Path_{S->I}$ is a path between start and intermediate states;

$Path_{I->F}$ is a path between intermediate and final states;

$States_{Collides}$ is a list of states of discretized path between the start state and the final state which has collisions;

$Collisions_{Best}$ is a current minimal number of collisions on the discretised path from start state to intermediate and from intermediate state to final;

$Collisions_{EE}$ is a thershold number of collisions used to check early exit condition during selecting intermediate state;

$|delta|_{\text{Max}}$ is a maximum absolute difference between joint angles in two arm states;

$|delta|_{Manhattan}$ is a sum of absolute differences between joint angles in two arm states.

INTRODUCTION

Articulated robots are currently used for a variety of automation tasks. Industrial robotic arms with revolute (articulated) joints are widely employed for tasks such as palletizing, material handling, welding, quality inspection, picking and placing objects [1], and many others.

Despite the fact that robotic arms perform different types of tasks, in general, we can summarize the task for a robotic arm: to move to a certain place at a specific time without causing damage to surrounding objects or itself. From this arises the problem of planning a collision-free path for the robotic arm.

DOI 10.15588/1607-3274-2025-3-16

The control of the arm takes place by specifying rotation angles for each joint of the arm, that is, the control program specifies coordinates in the state space of the joints (or simply "joint space"). Meanwhile, the position of the manipulator's end point (the last link of the robotic arm, if counting from the base, also known as the "end effector") is determined in Cartesian space.

An articulated robotic arm requires at least 6 joints to achieve full 6-degree-of-freedom movement in Cartesian space [2]. But despite the fact that a "6-joint arm with 6 independent joints can specify any position and orientation of the manipulator" [3], robotic arms with a larger number of joints (such robotic arms are also called redundant) and degrees of freedom, respectively, are widely used in the industry. The reasons for this are the ability to avoid the problem of singularities in the robot's workspace and the solution to the problem of joint restrictions [3].

The large dimensionality of the state of the robotic arm and the complexity of the inverse kinematics problem (search for arm coordinates in joint space given coordinates in the Cartesian space), as well as the presence of many possible solutions to inverse kinematics due to the redundant joint, make path planning task difficult. There are plenty of parameters to optimize in path planning algorithms while the most important among them are path planning time and success rate. At present, there is no universal path planning algorithm for redundant robotic arms that would guarantee finding the optimal path, or guarantee finding any path if it exists.

The object of study is the process of path planning for redundant robotic manipulators.

The subject of study is the reliability and execution speed of path planning algorithms.

The purpose of the work is to develop a new sampling-based algorithm for path planning for redundant robotic arms that selects an optimal intermediate point based on the number of collisions along the path passing through it.

1 PROBLEM STATEMENT

Before describing the developed algorithm, let's clarify the task it has to solve. Suppose we have a robotic arm operating in an environment that changes its state during the robot's operation. We need to plan the path of the robotic arm from one state in joint space to another state in joint space. The arm should avoid collisions with the surrounding environment along the found path.

The algorithm can use third-party tools for collision checking at a certain state of the arm. The path segment collision checking will be based on discretization of the path with a given step and checking all discrete states for collisions.

The state of the robotic arm is described by a vector of rotation angles for each joint, starting from the arm's fixation point. We can write it down in the next way:

$$S = \left\{ J^0, J^1, ..., J^{DOF-1} \right\}. \quad (1)$$

The norm of the difference between two states will be defined as the maximum absolute difference in rotation angles for each joint of the robotic arm:

$$|S_1 - S_2| = MAX\left(J_1^0 - J_2^0, J_1^1 - J_2^1, ..., J_1^{DOF-1} - J_2^{DOF-1}\right). \quad (2)$$

The length of the path is the sum of the norms of the differences between all neighboring states on the path:

$$|Path| = \sum_{i=1}^{N-1} |S_{i+1} - S_i|. \quad (3)$$

We'd like to get the result of the planning as soon as possible, because we are building trajectories real-time. So, the main criteria for evaluating a planning algorithm are the success rate and the average time to build the path. An additional parameter is the length of the found path, so the shorter the path, the better is the solution.

2 REVIEW OF THE LITERATURE

The well-known path planning algorithms for a redundant robotic arm among others include the following algorithms:

– Probabilistic Roadmaps: it's a sampling-based method for path planning where random samples from the configuration space are used to create nodes, which are then connected to create a roadmap [4];

– Rapidly-Exploring Random Trees (RRT): this algorithm is particularly useful for high dimensional spaces and real-time applications [5];

– Artificial Potential Fields: this method treats the robot as a particle moving under the influence of artificial forces. The goal and obstacles generate attractive and repulsive forces, respectively [6];

– Deep Reinforcement Learning based approaches: recent works have proposed learning-based methods for path planning, which can effectively handle redundant manipulators [7].

Among others, worth noting one of the recent works where Khan et al. proposed a model-free kinematic tracking controller for redundant robotic manipulators using Zeroing Neural Networks (ZNN) and Beetle Antennae Search (BAS). The ZNNBAS algorithm avoids traditional Jacobian-based approaches by leveraging a meta-heuristic optimization method in continuous time, eliminating the need for precise kinematic modeling. Tested on a 7-DOF manipulator, it achieved real-time redundancy resolution with minimal tracking errors, demonstrating the potential of hybrid optimization techniques for real-time path planning [8].

The RRT algorithm is the most popular solution today. There are many variations of it. In particular, the following should be mentioned:

RRT-Connect: this variation of RRT makes aggressive attempts to connect the tree directly to the goal, leading to faster solutions [9];

Bidirectional RRT (Bi-RRT): in this method, two RRTs grow towards each other, one from the initial state and the other from the goal. This can be more efficient in some problem spaces [10, 18];

DOI 10.15588/1607-3274-2025-3-16

RRT* (RRT Star): This variant of RRT introduces the idea of an “optimal” path, gradually improving the path quality by selectively rewiring nodes in the tree to minimize total path cost [11].

The disadvantages of the RRT-based algorithms described above can include their lack of evaluation for the currently generated states in the tree. As a result, even if the algorithm has almost found a collision-free path (i.e., one of the tree vertices can reach the target state with minimal collision), it will not attempt to complete the path, but will continue to generate states randomly without additional changes [12].

Ganesan et al. propose Hybrid-RRT, a novel path-planning algorithm that combines uniform and non-uniform sampling to improve the performance of RRT*-based motion planning. The hybrid approach balances exploration and exploitation by dynamically selecting between uniform and goal-directed non-uniform sampling. Experimental results demonstrate that Hybrid-RRT* achieves faster convergence, higher success rates, and reduced node exploration compared to baseline algorithms, including RRT*, Informed RRT*, and RRT*-N. The method is particularly effective in complex environments, addressing limitations of both traditional uniform and non-uniform sampling strategies [13].

One of the algorithms that changes its behavior based on the current state assessment is Informed RRT* [14]. This is a further improvement to RRT*, it takes into account the best current path to guide the sampling process, leading to faster convergence towards an optimal solution. Despite the advantages of this approach, the idea of assessing a specific state is not widely used today.

3 MATERIALS AND METHODS

The main idea of the newly developed Recursive Random Intermediate State (RRIS) algorithm is that a set of random intermediate states is generated. Then we iterate through all of the states, and if a state has collisions, then we need to skip it. If a state has no collisions, then we need to calculate a penalty for this state. This penalty is based on the number of states that have collisions on the discretized path from initial state to intermediate state and from intermediate state to final state.

Among all intermediate states we choose the one with the lowest penalty. Then the task of finding a path from the initial state to the final state is reduced to the task of finding a path from the initial state to the intermediate state and from the intermediate state to the final state. So the algorithm calls itself recursively.

Algorithm will stop current recursion step execution and in one of two cases:

– we found intermediate state, which creates path without collisions;

– we checked all intermediate states and none of them creates a path that is better than a straight one. Which means that the number of states with collisions on discretized path doesn’t get smaller on any checked intermediate state.

These conditions may vary depending on the chosen strategy and we will return to this later.

In the end, if both parts of the path are successfully found, then we can build the whole path by merging the path from initial state to intermediate state and path from intermediate state to final state. And if we fail to find a safe path in at least one of the parts, then we fail to find a safe path.

The generation of intermediate states can depend on the specific implementation. In particular, such options can be used:

– select the step of displacement in the joint space based on the length of the direct path, and for each joint consider 3 displacement options: clockwise, counter-clockwise, or zero displacement. Then the intermediate state can be calculated using formula (4):

$$S_I = \left\{ J_M^0 + d \cdot Dir^0, J_M^1 + d \cdot Dir^1, ..., J_M^{DOF-1} + d \cdot Dir^{DOF-1} \right\} \quad (4)$$

Thus, we will have $N = 3^{DOF}$ intermediate states;

– build an axis-aligned bounding box around all colliding states on the straight path from initial to final state (as described in Algorithm 4), generate a fixed number of intermediate states randomly and uniformly within the bounding box (as described in Algorithm 5).

Intermediate states generation using a bounding box showed better results as can be seen in Table 3.

To compare paths that goes through different intermediate states, we can minimize the number of collisions in at least two ways:

1. Minimize the total number of collisions on two path parts. In this case, we assume that a smaller total number of collisions means that we will need to expend fewer efforts to avoid collisions in the subsequent steps.

2. Minimize the maximum number of collisions on two path parts. In this case, we believe that even if the number of collisions through the intermediate path increases, but they are both less than the maximum, then we will have to circumvent fewer in each of the two parts of the path, making it easier to bypass them.

Both approaches show good results and it’s shown in Table 3.

Depending on the input data and the sequence of generated intermediate states, the algorithm may not get an intermediate state that would lead us to the goal without collisions. However, a sufficiently “good” state, a path through which contains a small number of collisions, may appear among the first. During the recursive descent we often can build a collision-free path through a “good” state quite quickly.

Therefore, instead of always iterating through the entire set of intermediate states, a check for quick exit from the iteration can be introduced. We propose the following condition for early exit: if both parts of the path through the intermediate state have fewer than half collisions of the direct path, then we choose this intermediate state and exit the iteration.

DOI 10.15588/1607-3274-2025-3-16

The use of the early exit strategy significantly improved the algorithm's speed, which is evident in the results section in Table 3.

The algorithm described above, despite its high speed, still has one major drawback. Due to the fact that the algorithm has no backtracking tool, in some cases it will get stuck in local minima. As shown in Table 3 the failure rate of an algorithm on a test set with a single run is about 90 percent. And the most common reason for an algorithm to fail is stucking in local minima.

In Table 3 can be seen that the failure rate of different variations of this algorithm is much higher than that of the RRTConnect algorithm from OMPL [17]. Therefore, we decided to add a simple way to escape from the local minima. Specifically – rerunning the algorithm a certain number of times until a path is found.

In the results section in Table 4 the testing results of the algorithm that initiates the search path up to 5 times in case of failures in previous steps are presented.

The pseudo code description of an optimal version (based on test results) of the RRIS algorithm is described below together with additional algorithms used by main algorithm. Intermediate states generated uniformly random in the axis-aligned bounding box. Intermediate states comparison is based on minimizing the maximum number of collisions on two path parts. And an early exit strategy applied.

Data: S_S, S_F, Algorithm options (M_{BB}, N_r, $step$)
Result: Path from S_S to S_F or failure
$Collisions_{S\to F} \leftarrow$ `CountCollisions`$(S_S, S_F, step, \infty)$
if $Collisions_{S\to F} = 0$ **then**
  return direct path (S_S, S_F) ; `/* Path is collision-free */`
end
$States_{Collides} \leftarrow$ `CollidingStatesList`$(S_S, S_F, step)$;
$BB \leftarrow$ `ComputeBoundingBox`$(States_{Collides}, M_{BB})$;
$States_I \leftarrow$ `GenerateRandomStates`(BB, N_r);
Sort $States_I$ by total path distance (ascending);
$S_{best} \leftarrow$ `NULL`;
$Collisions_{best} \leftarrow Collisions_{S\to F}$;
$Collisions_{EE} \leftarrow Collisions_{S\to F}/2$;
foreach $S_I \in States_I$ **do**
  if S_I *has collisions* **then**
    continue;
  end
  $Collisions_{S\to I} \leftarrow$ `CountCollisions`$(S_S, S_I, step, Collisions_{best})$;
  $Collisions_{I\to F} \leftarrow$ `CountCollisions`$(S_I, S_F, step, Collisions_{best})$;
  if $Collisions_{S\to I} < Collisions_{EE}$ *and* $Collisions_{I\to F} < Collisions_{EE}$ **then**
    $S_{best} \leftarrow S_I$;
    break;
  end
  if $Collisions_{S\to I} + Collisions_{I\to F} < Collisions_{best}$ **then**
    $Collisions_{best} \leftarrow Collisions_{S\to I} + Collisions_{I\to F}$;
    $S_{best} \leftarrow S_I$;
  end
end
if $S_{best} = NULL$ **then**
  return failure ; `/* No valid path found */`
end
$Path_{S\to I} \leftarrow$ `CollisionFreePathPlanning`$(S_S, S_{best}, M_{BB}, N_r, step)$;
$Path_{I\to F} \leftarrow$ `CollisionFreePathPlanning`$(S_{best}, S_F, M_{BB}, N_r, step)$;
if $Path_{S\to I}$ *and* $Path_{I\to F}$ *found* **then**
  return concatenated path $(Path_{S\to I}, Path_{I\to F})$;
end
return failure;

Algorithm 1 – Collision-Free Path Planning

To speed up the path planning additional parameter $Collisions_{\max}$ is passed to *CountCollisions* and *CollidingStatesList* methods. This parameter used to interrupt algorithm if collisions count exceeds the collisions limit.

Data: $S_a, S_b, step, Collisions_{max}$
Result: Collisions count
return $|CollidingStatesList(S_a, S_b, step, Collisions_{max})|$;

Algorithm 2 – Count Collisions on Discretized Path

Data: $S_a, S_b, step, Collisions_{max}$
Result: Set of colliding states $States$
$States \leftarrow \emptyset$;
$Path \leftarrow LinearDiscretization(S_a, S_b, step)$;
foreach $S_i \in Path$ **do**
  if S_i *collides in PyBullet* **then**
    Add S_i to $States$;
  end
  if $|States| > Collisions_{max}$ **then**
    return $States$;
  end
end
return $States$;

Algorithm 3 – Colliding States List

Data: $States_{Collides}$, M_{BB}
Result: BB
foreach *joint* j *in the arm* **do**
  foreach *state* S_i *in* $States_{Collides}$ **do**
    $BB_{min}[j] \leftarrow \min(S_i[j] - M_{BB}, BB_{min}[j])$;
    $BB_{max}[j] \leftarrow \max(S_i[j] + M_{BB}, BB_{max}[j])$;
  end
  Clip $BB_{min}[j], BB_{max}[j]$ to joint limits;
end
return BB;

Algorithm 4 – Compute a Bounding Box in a Joint Space

Data: BB, N_r
Result: Set of random states $States$
$States \leftarrow \emptyset$;
for $i = 1$ *to* N_r **do**
  Generate S_i uniformly in BB;
  Add S_i to $States$;
end
return $States$;

Algorithm 5 – Generate Random States

Data: S_a, S_b, $step$
Result: List of discretized states $Path$
$Path \leftarrow [S_a]$;
$n \leftarrow \lceil |S_b - S_a|/step \rceil$;
for $i = 1$ *to* $n - 1$ **do**
  $S_i \leftarrow S_a + i \cdot \frac{S_b - S_a}{n}$;
  Append S_i to $Path$;
end
Append S_b to $Path$;
return $Path$;

Algorithm 6 – Linear Dicretization of Path Between Two States

Please note, that separate runs of an algorithm are absolutely independent, and the rerunning process is not included in algorithm description.

DOI 10.15588/1607-3274-2025-3-16

4 EXPERIMENTS

Before starting to describe the results of the algorithm's work let's clarify which auxiliary software products were used, for what hardware the test trajectories were constructed and what is the working space of the robotic arm.

The auxiliary software used for algorithm development includes:

– Bullet Collision Detection & Physics Library – a library for collision detection and physics simulation, used in the algorithm for collision search [15];

– software code by Somatic Holdings LTD, which allows for quick simulation and visualization of the motion planning results of the robotic arm in a working environment.

The test trajectories were constructed for the following robotic arm:

UFACTORY xArm 7 Robotic Arm – a 7-degree-of-freedom robotic arm with revolute joints [16].

Visual representation of a robot with robotic arm installed and test working environment shown in Fig. 1 and Fig. 2 and the working range for each joint can be seen in Table 1.

The parameters of OMPL's RRT-Connect algorithm used for comparison in testing process are:

– state space: a 7-dimensional RealVectorStateSpace, corresponding to the 7 degrees of freedom (DOF) of the robotic arm;

– joint limits: the search space is bounded using a margin of 120 degrees and limited with arm joint limits;

– collision checking resolution: set the portion of 0.05 of the state space's maximum extent (0.05 / space->getMaximumExtent());

– planner time limit: the algorithm attempts to find a solution within maximum 20.0 seconds, but interrupts as soon as any solution is found.

5 RESULTS

In Table 2 we represent four versions of the RRIS algorithm that are tested and compared. Base version of the algorithm is the one with generating states in the bounding box, intermediate states paths comparison minimizing the maximum number of collisions on two path parts and early exit strategy applied. In three other versions we checked how changing states generating strategy, states comparison method or early exit usage affects algorithm performance. So, the first algorithm version is a base algorithm, in the second version states generating strategy changed to middle state displacement, in the third version states comparison method changed to minimize the sum of collisions on path parts, in the fourth version early exit strategy disabled.

Also, RRIS based algorithm versions compared with RRTConnect algorithm (one of the most efficient nowadays), presented in OMPL.

Table 1 – Joint limits for xArm7 robotic arm [16]

Joint number	0	1	2	3	4	5	6
Minimum angle	–360°	–118°	–360°	–11°	–360°	–97°	–360°
Maximum angle	360°	120°	360°	225°	360°	180°	360°

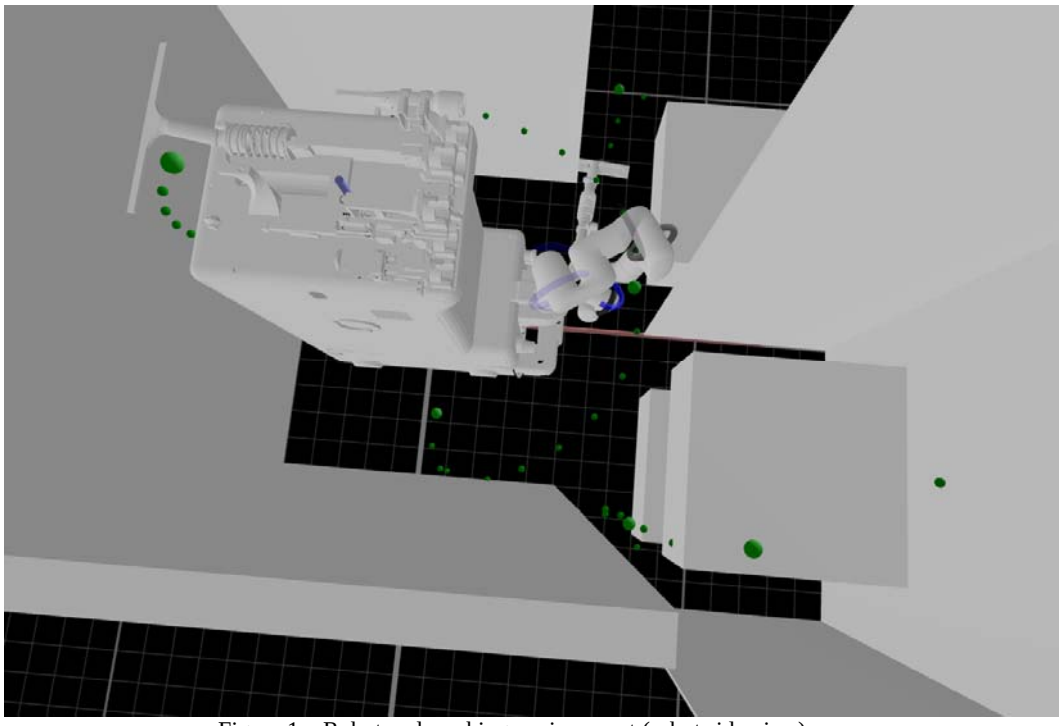


Figure 1 – Robot and working environment (robot side view)

DOI 10.15588/1607-3274-2025-3-16

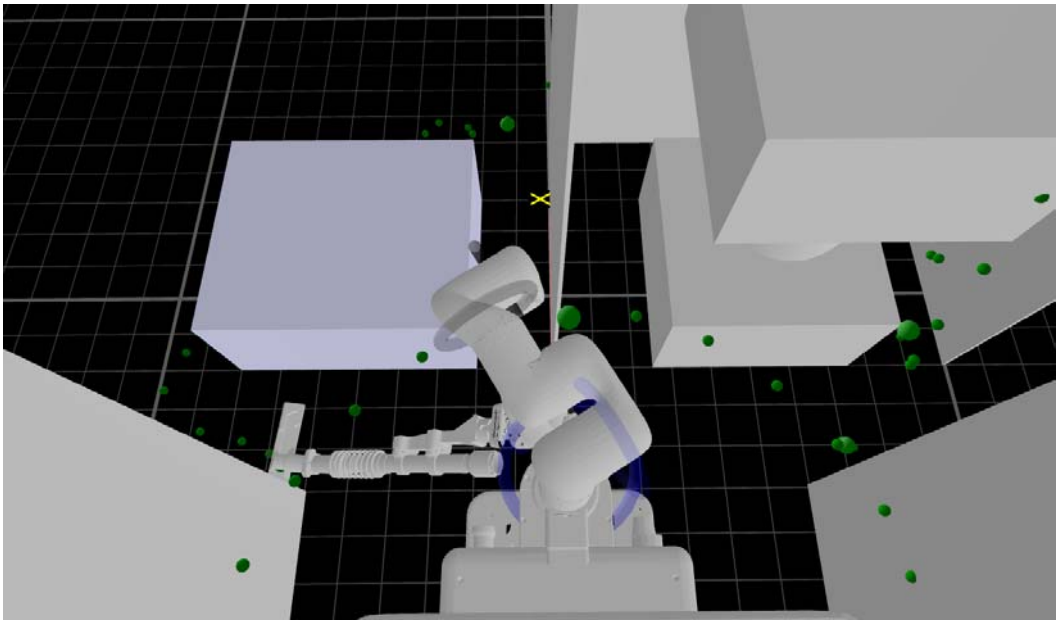


Figure 2 – Robot and working environment (robot back view)

Test set containing 104 pairs of states. It is guaranteed under experimental conditions that there is always a path without collisions between these pairs of states. Only 4 test cases have no collision on a straight path, requiring pathfinding in 96% of cases. There are three different types of tools installed to the arm wrist: sprayer, tip and vacuum. Vacuum (on Fig. 1 and Fig. 2) is much larger than two others and the wide majority of failures occurred with this tool.

Test cases contain various difficult situations like: arm should go from one side of the wall to another side, arm should move from one side of the robot to another side, arm should move between two boxes, and a lot of other complicated variations.

All the algorithm versions take 0.05 radians as a collision check step. The bounding box margin is set to 60 degrees for algorithm version 1, 3, and 4. Number of states generated on each recursive function call is 1000 for 1, 3, 4 versions and $3^7 = 2187$ for version number 2. Step for middle state displacement for version 2 calculates by formula (5).

$$d = 0.1 \cdot |S_F - S_S|_{\text{Max}} + 0.1 \cdot |S_F - S_S|_{Manhat\,\text{tan}}. \quad (5)$$

Here the constant 0.1 selected experimentally and can be configured for other robotic arms and work environments.

Table 2 – Tested algorithms versions

Version Num	States Generating	Comparison method	Early Exit
1	Bounding Box	Minimize Max	Yes
2	Middle State Displacement	Minimize Max	Yes
3	Bounding Box	Minimize Max	Yes
4	Bounding Box	Minimize Max	No

Table 3 – Testing algorithm versions results

Algorithm Version	Version 1	Version 2	Version 3	Version 4	RRTConnect from OMPL
Found Paths	92/104	92/104	91/104	94/104	104/104
Average straight path length* (radians)	3.474	3.605	3.508	3.603	3.731
Average path without collisions length* (radians)	5.750	5.857	5.652	5.756	44.167
Average search time (s)	0.555	1.341	0.582	1.825	1.109
Collision check count	4235.07	10164.01	4204.25	14421.64	10408.2

* – average straight path length and average path without collisions length calculated only for test cases where path was found

Table 4 – Comparing RRIS with repetitive calls and RRTConnect

Algorithm	RRIS algorithm with 5 attempts	RRTConnect from OMPL
Found Paths	104/104	104/104
Average straight path length (radians)	3.731	3.731
Average path without collisions length (radians)	7.427	44.167
Average search time (s)	0.36	1.109
Collision check count	2958.808	10408.2

DOI 10.15588/1607-3274-2025-3-16

Table 3 presents the summarized results of the performance of different versions of the algorithm.

We use the RRT Connect algorithm (OMPL implementation) to compare with the described algorithm. We tested RRT Connect with margins 60, 90, 120, 150 degrees and selected 120 degrees as it shows the best results with this margin value.

As can be seen in Table 3 RRIS algorithm with single run has a success rate of 87.5% – 90.4% depending on algorithm version.

In Table 4 we present the results of the RRIS algorithm that initiates the search path up to 5 times in case of failures in previous steps and compares it to the RRTConnect from OMPL. We are using version 1 (see Table 2) algorithm but reducing the number of generated states to 500. And we call it repeatedly until a path is found (but no more than 5 times).

As shown in Table 4 multi-run RRIS algorithm has 100% success rate as well as RRTConnect from OMPL, but it has 3.08 times smaller average path search time and 3.52 times smaller collisions check count. Also, multiple algorithm runs allowed to decrease the number of generated states from 1000 to 500, which decreased average search time from 0.555s to 0.36s.

6 DISCUSSION

As shown in Table 3 generating random states in the bounding box gives us better results than displacing the middle state. Probably, the reason for this may be better flexibility of this type of solution. It could generate states close or far from initial and final states and find the best option in most cases faster.

Also, results presented in Table 3 shows that early exit strategy has a great impact on algorithm productivity. It means that ideas described in section 3 are correct.

On the other hand, we can't see much difference between minimizing maximum collisions count and minimizing the sum of collisions count strategies. One strategy works better in one part of test cases and the other strategy works better for the other part.

Calling the algorithm multiple times significantly improved its reliability as shown in Table 4. The issue of local minima is significantly reduced now. Also, this allowed for a reduction in the number of generated states in a single iteration without degrading the algorithm's performance.

Compared to the RRTConnect algorithm implemented in the OMPL library, the algorithm proposed in this paper not only has better performance but also constructs a shorter path on average (as shown in Table 4). The OMPL library has an integrated path improvement system that works very well, but still the initial result path of the algorithm proposed in this paper is on average 5.947 times shorter.

However, the comparison between the performance of the RRIS algorithm and RRT-Connect depends significantly on the specific parameter settings of RRT-Connect. A deeper investigation is required to make a better comparison between these algorithms.

DOI 10.15588/1607-3274-2025-3-16

The core idea of the algorithm – to select an intermediate state based on collisions count criteria shows its effectiveness. Despite the fact that the test dataset contained many trajectory scenarios that were challenging to search for, still algorithms managed to find a path in these situations.

CONCLUSIONS

A new motion planning sampling-based algorithm was developed for solving the problem of collision-free path planning for redundant robotic manipulators in joint space in real-time mode. The algorithm is based on the principle of selecting an optimal intermediate point based on the number of collisions along the discretized path that passes from the initial to the final point through the intermediate point.

A strategy for generating intermediate points within an axis-aligned bounding box was proposed for this algorithm. Additionally, an early exit strategy was proposed to improve the algorithm's speed.

The algorithm demonstrated high efficiency. An implementation of this algorithm with iterated calls managed to find a path in test cases 3.08 times faster than the RRTConnect algorithm implemented in OMPL under the testing conditions. Also, the length of original paths found by algorithm is on average 5.947 times shorter than paths found by RRTConnect in the presented tests set.

The scientific novelty of obtained results is a newly developed sampling-based algorithm called the Recursive Random Intermediate State (RRIS) algorithm. This algorithm is able to plan the path in a dynamic environment in real time. Besides, we propose an axis-aligned bounding box generation strategy and an early exit strategy to improve algorithm speed.

The practical significance of this study lies in the development of the Recursive Random Intermediate State algorithm, which enables real-time path planning for redundant robotic arms.

Prospects for further research include enhancing the RRIS algorithm by incorporating machine learning techniques for adaptive intermediate state selection.

ACKNOWLEDGEMENTS

The authors would like to express their deep gratitude to Somatic Holdings LTD, whose codebase greatly facilitated the development of the algorithm presented in this paper.

REFERENCES

1. Data Center Solutions, IOT, and PC Innovation [Electronic resource]. Mode of access: https://www.intel.com/content/www/us/en/robotics/robotic-arm.html (date of access: 15 June 2023).
2. Pennestri E., Cavacece M., Vita L. On the Computation of Degrees-of-Freedom: A Didactic Perspective [Electronic resource], *ASME 2005 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference: proceedings of the conference, [Long Beach, California, USA]*, September 24–28, 2005, Mode of access: https://doi.org/10.1115/DETC2005-84109, pp. 1733–1741.

3. Ashitava G. Resolution of Redundancy in Robots and in a Human Arm, *Mechanism and Machine Theory*, 2018, Vol. 125, pp. 126–136.
4. Kavraki L. E., Svestka P., Latombe J.-C., Overmars M. H. Probabilistic Roadmaps for Path Planning in High-Dimensional Configuration Spaces, *IEEE Transactions on Robotics and Automation*, 1996, Vol. 12, No. 4, pp. 566–580. doi: 10.1109/70.508439.
5. LaValle S. M. Rapidly-Exploring Random Trees: A New Tool for Path Planning, *The Annual Research Report*, 1998.
6. Khatib O. Real-Time Obstacle Avoidance for Manipulators and Mobile Robots, *The International Journal of Robotics Research*, 1986, Vol. 5(1), pp. 90–98. doi:10.1177/027836498600500106.
7. Lillicrap T. P., Hunt J. J., Pritzel A., Heess N., Erez T., Tassa Y., Wierstra D. Continuous Control with Deep Reinforcement Learning, *arXiv preprint arXiv:1509.02971*, 2015.
8. Khan A. T., Cao X., Li Z., Li S. Evolutionary Computation Based Real-Time Robot Arm Path-Planning Using Beetle Antennae Search, *EAI Endorsed Transactions on AI and Robotics*, 2022, Vol. 1, P. e3.
9. Kuffner J. J., LaValle S. M. RRT-Connect: An Efficient Approach to Single-Query Path Planning, *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*. San Francisco, CA, USA, 2000, Vol. 2, pp. 995–1001, doi: 10.1109/ROBOT.2000.844730.
10. LaValle S. M., Kuffner J. J. Randomized Kinodynamic Planning, *The International Journal of Robotics Research*, 2001, Vol. 20(5), pp. 378–400. doi:10.1177/02783640122067453.
11. Karaman S., Frazzoli E. Sampling-Based Algorithms for Optimal Motion Planning, *International Symposium on Robotics Research*, 2011, May, Vol. 71, pp. 65–70.
12. Kang J.-G., Lim D.-W., Choi Y.-S., Jang W.-J., Jung J.-W. Improved RRT-Connect Algorithm Based on Triangular Inequality for Robot Path Planning, *Sensors*, 2021, Vol. 21, P. 333. https://doi.org/10.3390/s21020333.
13. Ganesan S., Ramalingam B., Mohan R. E. A Hybrid Sampling-Based RRT Path Planning Algorithm for Autonomous Mobile Robot Navigation, *Expert Systems with Applications*, 2024, Vol. 233, P. 125206. https://doi.org/10.1016/j.eswa.2024.125206.
14. Gammell J. D., Srinivasa S. S., Barfoot T. D. Informed RRT*: Optimal Sampling-Based Path Planning Focused via Direct Sampling of an Admissible Ellipsoidal Heuristic, *arXiv preprint arXiv:1404.2334*, 2014.
15. Bullet Collision Detection & Physics Library [Electronic resource]. Mode of access: https://pybullet.org/Bullet/BulletFull/index.html (date of access: 15 June 2023).
16. The Difference Between UFACTORY xArm5, UFACTORY xArm6 and UFACTORY xArm7 [Electronic resource], *UFACTORY*. Mode of access: http://help.ufactory.cc/en/articles/4491842-the-difference-between-ufactory-xarm5-ufactory-xarm6-and-ufactory-xarm7 (date of access: 15 June 2023).
17. Open Motion Planning Library, Version 1.4.2 [Electronic resource]. OMPL Development Team. Mode of access: http://ompl.kavrakilab.org (date of access: 8 August 2023).
18. Xin P., Wang X., Liu X., Y. Wang, Z. Zhai, X. Ma Improved Bidirectional RRT* Algorithm for Robot Path Planning, *Sensors*, 2023, Vol. 23, No. 2, P. 1041. https://doi.org/10.3390/s23021041.

Received 07.04.2025.
Accepted 30.06.2025.

УДК 004.94

ПЛАНУВАННЯ ШЛЯХУ ДЛЯ НАДЛИШКОВИХ РОБОРУК З ВИКОРИСТАННЯМ АЛГОРИТМУ РЕКУРСИВНОГО ВИПАДКОВОГО ПРОМІЖНОГО СТАНУ

Медвідь А. Я. – аспірант кафедри Систем Штучного Інтелекту, Національний університет «Львівська політехніка», Львів, Україна.

Яковина В. С. – д-р техн. наук, професор кафедри Систем Штучного Інтелекту, Національний університет «Львівська політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. Планування шляху без зіткнень в просторі суглобів для надлишкових роборук (роботизованих маніпуляторів) залишається складною задачею через високу вимірність конфігураційного простору і динамічну зміну середовища. Існуючі методи планування часто стикаються з труднощами у балансуванні між часом пошуку та якістю траєкторії, що є критично важливим для застосувань у режимі реального часу.

Мета роботи – розробка нового методу планування траєкторій без зіткнень в режимі реального часу для роборук з надлишковими суглобами.

Метод. У цьому дослідженні представлений новий алгоритм планування шляху без зіткнень у просторі суглобів для надлишкових роборук, що працює на основі генерації випадкових станів. Алгоритм отримав назву Рекурсивного Випадкового Проміжного Стану (РВПС). Принцип роботи алгоритму полягає у генерації набору випадкових проміжних станів із подальшим ітеративним вибором оптимального на основі кількості зіткнень уздовж дискретизованої траєкторії. Крім того, у статті пропонується стратегія побудови обмежувального прямокутного паралелепіпеда (bounding box) та стратегія раннього виходу для підвищення швидкості роботи алгоритму. Нарешті, для підвищення надійності пропонується повторне викликання алгоритму. Ефективність алгоритму РВПС оцінюється шляхом проведення комплексних тестів та порівнюється з популярним алгоритмом RRT Connect, реалізованим у бібліотеці Open Motion Planning Library.

Результати. Експериментальні дослідження показують, що алгоритм РВПС за умов тестування забезпечує траєкторії без зіткнень зі значно коротшою середньою довжиною та скорочує час пошуку приблизно у три рази порівняно з алгоритмом RRT Connect.

Висновки. Запропонований алгоритм РВПС демонструє перспективний підхід до планування траєкторій у режимі реального часу для надлишкових роботизованих маніпуляторів. Поєднуючи стратегічну вибірку проміжних станів із ефективною оцінкою зіткнень та механізмами раннього завершення, алгоритм пропонує надійну альтернативу відомим методам.

КЛЮЧОВІ СЛОВА: планування шляху, надлишковий роботизований маніпулятор, уникнення зіткнень.

DOI 10.15588/1607-3274-2025-3-16

ЛІТЕРАТУРА

1. Data Center Solutions, IOT, and PC Innovation [Електронний ресурс]. – Режим доступу: https://www.intel.com/content/www/us/en/robotics/robotic-arm.html (дата звернення: 15 червня 2023).
2. Pennestri E. On the Computation of Degrees-of-Freedom: A Didactic Perspective [Електронний ресурс] / E. Pennestri, M. Cavacece, L. Vita // ASME 2005 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference: proceedings of the conference, [Long Beach, California, USA], September 24–28, 2005. – Режим доступу: https://doi.org/10.1115/DETC2005-84109. – С. 1733–1741.
3. Ashitava G. Resolution of Redundancy in Robots and in a Human Arm / G. Ashitaya // Mechanism and Machine Theory. – 2018. – Том 125. – С. 126–136.
4. Probabilistic Roadmaps for Path Planning in High-Dimensional Configuration Spaces / [L. E. Kavraki, P. Svestka, J.-C. Latombe, M. H. Overmars] // IEEE Transactions on Robotics and Automation. – 1996. – Том 12, № 4. – С. 566–580. doi: 10.1109/70.508439.
5. LaValle S. M. Rapidly-Exploring Random Trees: A New Tool for Path Planning / S. M. LaValle // The Annual Research Report. – 1998.
6. Khatib O. Real-Time Obstacle Avoidance for Manipulators and Mobile Robots / O. Khatib // The International Journal of Robotics Research. – 1986. – Том 5(1). – С. 90–98. doi:10.1177/027836498600500106.
7. Lillicrap T. P. Continuous Control with Deep Reinforcement Learning / [T. P. Lillicrap, J. J. Hunt, A. Pritzel et al.] // arXiv preprint arXiv:1509.02971. – 2015.
8. Evolutionary Computation Based Real-Time Robot Arm Path-Planning Using Beetle Antennae Search / [A. T. Khan, X. Cao, Z. Li, S. Li] // EAI Endorsed Transactions on AI and Robotics. – 2022. – Том 1. – С. e3.
9. Kuffner J. J. RRT-Connect: An Efficient Approach to Single-Query Path Planning / J. J. Kuffner, S. M. LaValle // Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065), San Francisco, CA, USA, 2000. – Том 2. – С. 995–1001. DOI: 10.1109/ROBOT.2000.844730.
10. LaValle S. M. Randomized Kinodynamic Planning / S. M. LaValle, J. J. Kuffner // The International Journal of Robotics Research. – 2001. – Том 20(5). – С. 378–400. doi:10.1177/02783640122067453.
11. Karaman S. Sampling-Based Algorithms for Optimal Motion Planning / S. Karaman, E. Frazzoli // International Symposium on Robotics Research. – 2011, May. – Том 71. – С. 65–70.
12. Kang J.-G. Improved RRT-Connect Algorithm Based on Triangular Inequality for Robot Path Planning / [J.-G. Kang, D.-W. Lim, Y.-S. Choi et al.] // Sensors. – 2021. – Том 21. – С. 333. https://doi.org/10.3390/s21020333.
13. Ganesan S. A Hybrid Sampling-Based RRT Path Planning Algorithm for Autonomous Mobile Robot Navigation / S. Ganesan, B. Ramalingam, R. E. Mohan // Expert Systems with Applications. – 2024. – Том 233. – С. 125206. https://doi.org/10.1016/j.eswa.2024.125206.
14. Gammell J. D. Informed RRT*: Optimal Sampling-Based Path Planning Focused via Direct Sampling of an Admissible Ellipsoidal Heuristic / J. D. Gammell, S. S. Srinivasa, T. D. Barfoot // arXiv preprint arXiv:1404.2334. – 2014.
15. Bullet Collision Detection & Physics Library [Електронний ресурс]. – Режим доступу: https://pybullet.org/Bullet/BulletFull/index.html (дата звернення: 15 червня 2023).
16. The Difference Between UFACTORY xArm5, UFACTORY xArm6 and UFACTORY xArm7 [Електронний ресурс] / UFACTORY. – Режим доступу: http://help.ufactory.cc/en/articles/4491842-the-difference-between-ufactory-xarm5-ufactory-xarm6-and-ufactory-xarm7 (дата звернення: 15 червня 2023).
17. Open Motion Planning Library, Version 1.4.2 [Електронний ресурс] / OMPL Development Team. – Режим доступу: http://ompl.kavrakilab.org (дата звернення: 8 серпня 2023).
18. Improved Bidirectional RRT* Algorithm for Robot Path Planning / [P. Xin, X. Wang, X. Liu et al.] // Sensors. – 2023. – Том 23, № 2. – С. 1041. https://doi.org/10.3390/s23021041.

DOI 10.15588/1607-3274-2025-3-16

УДК 681.326

ІНЖЕНЕРНИЙ СОЦІАЛЬНИЙ КОМП'ЮТИНГ

Хаханов В. І. − д-р техн. наук, професор кафедри автоматизації проектування обчислювальної техніки, Харківський національний університет радіоелектроніки, Україна.

Чумаченко С. В. – д-р техн. наук, професор, завідувач кафедри автоматизації проектування обчислювальної техніки, Харківський національний університет радіоелектроніки, Україна.

Литвинова Є. І. – д-р техн. наук, професор кафедри автоматизації проектування обчислювальної техніки, Харківський національний університет радіоелектроніки, Україна.

Хаханова Г. В. – д-р техн. наук, професор кафедри автоматизації проектування обчислювальної техніки, Харківський національний університет радіоелектроніки, Україна.

Хаханов І. В. – канд. техн. наук, асистент кафедри автоматизації проектування обчислювальної техніки, Харківський національний університет радіоелектроніки, Україна.

Обрізан В. І. – канд. техн. наук, докторант кафедри автоматизації проектування обчислювальної техніки, Харківський національний університет радіоелектроніки, Україна.

Хаханова І. В. – д-р техн. наук, професор кафедри автоматизації проектування обчислювальної техніки, Харківський національний університет радіоелектроніки, Україна.

Максимова Н. Г. – аспірантка кафедри автоматизації проектування обчислювальної техніки, Харківський національний університет радіоелектроніки, Канада.

АНОТАЦІЯ

Актуальність. Актуальність дослідження зумовлена необхідністю усунення протиріч між менеджментом та виконавцями шляхом запровадження інженерного соціального комп'ютингу, що забезпечує моральне управління соціальними процесами на основі їх метричного моніторингу.

Мета. Мета дослідження – розробка інженерних архітектур моніторинга та управління соціальними процесами на основі векторної логіки.

Метод. Дослідження орієнтоване на розробку інженерних векторно-логічних схем та архітектур управління соціальними процесами на основі їх вичерпного метричного моніторингу з метою створення комфортних умов для творчої праці. Даються визначення основних понять AI-розвитку. Вводиться рівняння комп'ютингу, як транзитивне замикання у тріаді відносин – як помилки, що створює нові структури, процеси чи явища. Розробляються механізми інтелектуального комп'ютингу, які поєднують алгоритми та структури даних детермінованого та ймовірнісного AI-комп'ютингу. Пропонуються механізми побудови моделей на основі універсуму примітивів, які мають Similarity по відношенню до їх використання для моделювання процесів (in-hardware synthesis, in-software programing, in neural network training, in-qubit quantization, in-memory modeling, in-truth table logic generation). Запроваджується метрика інтелектуального комп'ютингу, яка використовується для вибору архітектури та моделей обчислювальних процесів з метою отримати ефективні рішення практичних завдань.

Результати. Запропоновано: 1) рівняння комп'ютингу як транзитивне замикання у тріаді відносин – як помилка, що створює нові структури, процеси чи явища; 2) механізми інтелектуального комп'ютингу, орієнтовані на істотне зниження часових та енергетичних витрат при вирішенні практичних завдань за рахунок обнулення алгоритмів обробки великих даних завдяки експоненційній надмірності розумних та надмірних AI-моделей; 3) механізми побудови моделей на основі універсуму примітивів, які мають Similarity по відношенню до їх використання для моделювання процесів.

Висновки. Наукова новизна полягає у розробці наступних інноваційних рішень: 1) запропоновано тріаду відносин на основі хог-операції для вимірювання процесів та явищ у кіберсоціальному світі; 2) запропоновано архітектуру інтелектуального комп'ютингу для управління соціальними процесами на основі їх вичерпного моніторингу; 3) реалізації схем в архітектурі in-memory комп'ютингу, що дає можливість не використовувати інструкції процесора, тільки read-write транзакції на логічних векторах, що економить час та енергію для виконання алгоритмів аналізу великих даних; 4) запропоновано механізми синтезу векторно-логічних моделей соціальних процесів або явищ на основі унітарного кодування патернів на універсумі примітивів, орієнтованих на верифікацію, моделювання та тестування прийнятих рішень. Практична значимість дослідження полягає в тому, що запропонована метрика інтелектуального комп'ютингу використовується як метод для вибору архітектури та моделей обчислювальних процесів для одержання ефективних рішень практичних завдань. Інженерний соціальний комп'ютинг покликаний сприяти побудові миролюбних, справедливих і відкритих суспільств задля досягнення Цілей сталого розвитку ООН (ЦСР 16).

КЛЮЧОВІ СЛОВА: мозок людства, інтернет-інфраструктура, інтелектуальні комп'ютинг, artificial intelligence, розумні структури даних, моделі комп'ютера, історія комп'ютера, метрика комп'ютера, AI-Industry, Modeling for simulation, цілі сталого розвитку, peaceful society.

DOI 10.15588/1607-3274-2025-3-17

АБРЕВІАТУРИ

AI – Artificial Intelligence;
AMD – Advanced Micro Devices;
API – Application Programming Interface;
ASIC – Application-Specific Integrated Circuit;
CAD – Computer-Aided Design;
CNN – Convolutional Neural Networks;
RNNs – Recurrent Neural Networks;
EDA – Electronic Design Automation;
EWDTS – East-West Design and Test Symposium;
FPGA – Field-Programmable Gate Array;
GenAI – Generative AI;
GPU – Graphics Processing Unit;
IEEE – Institute of Electrical and Electronics Engineers;
IP-core – Semiconductor Intellectual Property Core;
ML – Machine Learning;
NN – Nueral Network;
NoSQL – Not Only SQL databases;
NVMe (NVM Express) – Non-Volatile Memory Host Controller Interface Specification;
SSD – Solid State Drives;
LLM – Large Lingustic Models;
RAG – Retrieval Augmented Generation;
RISC-V (risk-five) – Reduced Instruction Set Computer;
SECT – Standard для Embedded Core Test;
SoC – System-on-Chip;
SQL – Structured Query Language;
TSMC – Taiwan Semiconductor Manufacturing Company;
VLC – Vector-Logical Computing;
VLSI – Very-Large-Scale Integration;
VVV – Volume, Velocity, Variety (обсяг, швидкість генерування та різноманітність);
VHDL – Very high speed integrated circuits Hardware Description Languag);
АЛП – арифметико-логічний пристрій;
ШІ – штучний інтелект;
ПК – персональний комп'ютер;
ТІ – таблиця істинності;
ЦСР – Цілі сталого розвитку.

НОМЕНКЛАТУРА

X – вхідні набори;
X_i – вхідний ефект;
Y – вихідні сигнали;
n – кількість логічних змінних;
2^n – кількість бітів;
T – тест (реалізація, актуальна модель);
A – алгоритм;
M – модель (пам'ять);
M_i – біт логічного вектора;
F – помилки (виміряні помилки);
L – ідея (векторна логіка, специфікація);
xor – сума за модулем два або хог-операція;
d_i – відстані між компонентами.

ВСТУП

Будь-яке цілеспрямоване відношення є комп'ютинг. Інтелектуальний комп'ютинг – це гармонійні відносини між ресурсами та метою, що використовують надмірність розумних структур даних. Модель комп'ютингу вперше запропонував John von Neumann у 1945 році. Вона містила пристрої управління, виконання (АЛП) та пам'ять. Усього три компоненти, які працюють у просторі планети, і не лише, вже 80 років. Історично модель комп'ютингу можна розглядати як певну гармонію взаємодоповнюючих протилежностей (управління та виконання), що призводять до гармонії мети та витрат для її досягнення. Тут одна з цілей – це гармонія відносин між людиною та довкіллям. «Скажи мені – і я забуду, покажи мені – і я запам'ятаю, дай мені зробити – і я зрозумію», – Конфуцій, V ст. до нашої ери. Цією давньою фразою представлений сучасний комп'ютинг навчання та бізнесу у форматі «слово – образ – дія» або «алгоритм – модель – моніторинг-управління». Для людини образ – структурна простота наочної надмірності слова чи модель, що активує канал зору. Слово призначене для слуху, образ – для зору. Надмірність – благо для будь-якого комп'ютингу, зокрема, навчання. Людина має п'ять інформаційних каналів. Більше 90% інформації надходить до нас через зір та слух з них на зір припадає 90% інформації. Будь-яке цілеспрямоване відношення (взаємно-доповнюючих протилежностей) є комп'ютингом. Суть комп'ютингу – виявлення гармонії між витратами та метою.

Об'єкт дослідження – кібер-фізичний та кібер-соціальний комп'ютинг на основі векторно-логічних моделей фізичних та соціальних процесів та явищ.

Предмет дослідження – інженерне метричне управління соціальними процесами.

Проблема дослідження – це усунення протиріч між менеджментом та виконавцями шляхом запровадження інженерного соціального комп'ютингу, що забезпечує моральне управління соціальними процесами на основі їх метричного моніторингу.

Мета дослідження – розробка інженерних архітектур моніторинга та управління соціальними процесами на основі векторної логіки.

1 ПОСТАНОВКА ЗАДАЧІ

Існує три види комп'ютингу (рис. 1) в просторі і часі на основі кінцевого результату обчислень (inference): 1) печерний комп'ютинг – відображення f буття X (інформація) на будь-якому носії Y для прийняття рішень людиною $Y=f(X)$; 2) пасивний (імітаційний, simulation) комп'ютинг – моделювання вхідних ефектів X на модель L процесу або об'єкта з метою отримання висновку (inference) $Y=L(X)$ (прогнозування, діагностика, розпізнавання, кластеризація, класифікація). Ідеальним варіантом комп'ютингу є ситуація, коли при моделюванні вхідний ефект X_i зводиться до транзакцій читання-запису (read-write) за векторною логікою L в M-пам'яті $M_i=L[M(X_i)]$, без участі процесора. Ця технологія відповідає сучасному IT-тренду «Low Code/No Code» [1]; 3) активний (істинний) комп'ютинг – це генерація сигналів виконавчого механізму A для автоматичного управління процесом або об'єктом на основі двійкового кодування сигналів моніторингу X. Керуюча формула $A_i=f(X_i, A_{i-1})$ є моделлю автомата першого роду за В. М. Глушковим.

DOI 10.15588/1607-3274-2025-3-17

а б в

Рисунок 1 – Три типи комп'ютингу: а – печерний, б – пасивний, в – активний

Розглядається задача синтезу моделі (розумних структур даних) для моделювання будь-яких вхідних наборів – modeling for simulation.

Розв'язок задачі полягає у процесі синтезу (тренінгу) моделі за заданими вхідно-вихідними наборами сигналів. Simulation – процес визначення станів виходів моделі на заданих наборах вхідних сигналів

Нехай задані навчальні *XY*-набори, моделювання несправностей $f = XxorY$.

Практичне завдання полягає у синтезі ML-моделі (modeling) за допомогою *XY*-наборів $f=XxorY$ з метою подальшого тестування smart моделі та моделювання (simulation) на ній будь-яких вхідних *X*-наборів для визначення вихідних *Y*-сигналів, $Y=f(X)$.

Завдання дослідження: 1) аналіз інженерного комп'ютингу управління фізичними та соціальними об'єктами та процесами; 2) розробка концептуальних метрик, моделей та архітектур управління соціальними процесами на основі їх вичерпного моніторингу; 3) векторно-логічний метричний комп'ютинг соціальних процесів.

2 ОГЛЯД ЛІТЕРАТУРИ

Механізм комп'ютингу – це активація на основі моніторингу. Таблиця істинності – логічні відношення між вхідними та вихідними сигналами чи найуніверсальніша модель комп'ютингу на вирішення всіх завдань у всі часи. Таблиця істинності – це розумні структури даних, які мінімізують витрати виконання алгоритму їхньої обробки рахунок надмірності своєї простоти. Правильне вирішення проблеми – це завжди математично красиве і водночас просте, як цегла.

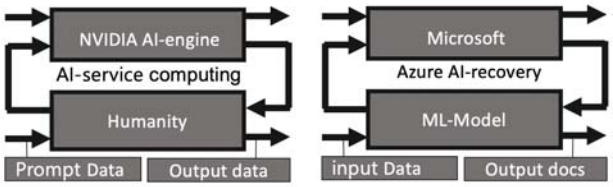
Бізнес-комп'ютинг – цілеспрямовані відношення між механізмами управління та виконання з функціями моніторингу, моделювання та актюації бізнес-процесу. Модель бізнес-процесу – це гра детермінованих та імовірнісних структур даних. Цифровізація – це цифрове (двійкове) кодування процесів та явищ на універсумі примітивів для точного управління фізичними та соціальними об'єктами на основі вичерпного моніторингу, що дозволяє скоротити час на обробку даних та підвищити якість сервісного обслуговування. Правильне рішення – це коли «просто, зрозуміло і красиво».

Промпт-комп'ютинг – цілеспрямовані відношення між механізмами управління та виконання з функціями інтелектуального формування вичерпної відповіді на правильно сформульований запит клієнта. Промпт-комп'ютинг формує відповіді на запитання.

Бот-комп'ютинг – це цілеспрямовані відношення між бот-додатком та клієнтом з функціями верифікації та корекції відповідей клієнта на правильно сформульовані запити бота. Бот-комп'ютинг формує питання на відповіді.

1. Актуальність створення RAG-бота-автомата для обслуговування клієнта в банку. GenAI та великі мовні моделі (LLM) представили нові можливості та ефективності. Розмір світового ринку LLM оцінювався в 4,35 млрд. доларів США в 2023 році і, за прогнозами, зростатиме з річним темпом зростання 36% з 2024 по 2030 рік [2].

2. Тисячі компаній і більше сотні урядів мають керівництва, структури або принципи використання штучного інтелекту (ШІ) для управління на основі моніторингу. Аналіз практики, що розвивається, використання AI-engine і використовуваних в них мов дає важливі відомості про те, як поширюються і змінюються AI-норми (рис. 2), управління суспільством і компаніями, та напрямок руху до інтелектуального prompt-computing [3].



а б

Рисунок 2 – Приклади Industry AI: а – AI-service computing (ШІ-сервіс комп'ютинг), де AI-engine – ШІ-двигун; Humanity – людство; б – Azure AI-recovery (Azure ШІ-оновлення), де Prompt Data – дані запитів; Output Data – вихідні дані; Input Data – вхідні (початкові) дані; Output docs – вихідні документи

Rufus LLM використовує генерацію доповненого пошуку (RAG) для отримання інформації з джерел, які вважаються надійними, таких як каталог продукції, відгуки клієнтів та повідомлення спільноти питань та відповідей; він також може викликати відповідні API магазинів Amazon. Система RAG надзвичайно складна як через різноманітність джерел даних, що використовуються, так і через різну релевантність кожного з них залежно від питання. Кожен LLM та кожне використання генеративного ШІ – це робота в процесі. Щоб Rufus ставав краще з часом, йому потрібно дізнатися, які корисні відповіді, а які можна поліпшити. Клієнти – найкраще джерело цієї інформації. Amazon заохочує клієнтів надавати Rufus зворотний зв'язок, даючи моделі знати, чи сподобалася або ні їм відповідь, і ці відповіді використовуються в процесі навчання з підкріпленням. Згодом Rufus навчається на зворотному зв'язку клієнтів та покращує свої відповіді, щоб надавати найбільш релевантну та корисну відповідь на будь-яке задане запитання. Іноді це означає розгорнуту текстову відповідь, але іноді це корот-

DOI 10.15588/1607-3274-2025-3-17

кий текст або «клікабельне» посилання для навігації магазином. Потрібно переконатися, що подана інформація слідує логічному потоку. Якщо не згрупувати і не відформатувати все правильно, можна отримати заплутану відповідь, яка не дуже корисна для клієнта. Саме тому Rufus використовує передову потокову архітектуру доставки відповідей. Клієнтам не потрібно чекати, поки довга відповідь буде повністю згенерована – натомість вони отримують першу частину відповіді, поки генерується решта її частини. Rufus заповнює потокову відповідь правильними даними (процес, званий гідратацією), надсилаючи запити у внутрішні системи. Крім генерації контенту для відповіді, він також генерує інструкції форматування, які визначають, як повинні відображатися різні елементи відповіді [4].

3) З іншого боку, є AI-продукти на ринку IT, які присвячені відновленню та генерації документів. Зразок такої системи – продукт компанії Microsoft, який вирішує завдання визначення даних та заголовків для швидкого навчання AI-механізму. Однак на сьогоднішній день ця Azure-система є пасивною і не вирішує завдання керування-бізнес процесом на основі аналізу документів та текстів [5].

У роботах [6, 7] розглядається інноваційне вирішення проблеми економіки при обробці великих даних та тестування цифрових пристроїв. Пропонується векторно-логічний енергозберігаючий комп'ютинг великих даних у пам'яті, вільний від команд процесора. Відмінність запропонованого комп'ютингу полягає у зменшенні обчислювальної складності алгоритмів до нуля шляхом збільшення надмірності пам'яті для суперпозиції розумних та явних структур даних. Логічний двійковий вектор розглядається як універсальна форма моделі уявлення процесів та явищ, структур та функціональностей. Логічний вектор на адресному просторі 2^n його бітів утворює таблицю істинності для n-змінних, яка використовується як основа структур даних для вирішення комбінаторних завдань комп'ютингу. Адреса розглядається як двійковий n-розрядний вектор, що ідентифікує належність елемента універсуму до кінцевого числа тестових n або інших патернів. Тому адресний вектор використовується для кодування великих даних (несправностей) з метою обробки на таблиці істинності без програмування. На основі адресного кодування даних на таблиці істинності в паралельному режимі обробки n змінних вирішуються завдання: моделювання тестових наборів та несправностей, логічний аналіз графових структур, кластеризація великих даних, ідентифікація та розпізнавання патернів. На логічному векторі синтезується матриця відстаней між тестом та таблицею істинності, що формує матрицю активності логічної функціональності. Ці матриці зводять алгоритми моделювання несправностей та побудови тесту до разової матричної процедури синтезу карти тестування логічної функціональності. У цьому зменшення складності алгоритмів синтезу відбувається рахунок експоненціальних витрат пам'яті, необхідні явної суперпозиції матриць і таблиць істинності. Розроблені моделі та методи тестування цифрових виробів орієнтовані на стандарт граничного сканування IEEE 1500 SECT, який зменшує складність обчислювального виробу завдяки розбиттю складної SoC на сукупність простих IP-core. Структура дослідження містить дві частини: 1) in-memory моделювання тестів та несправностей, як адрес, цифрових систем та елементів; 2) in-memory аналіз великих даних, як адрес, для розпізнавання патернів. Усі запропоновані моделі та алгоритми аналізу даних є інноваційними та не мають аналогів у світі. Валідність отриманих результатів підтверджується обґрунтованим доказом основних теоретичних положень та численними експериментами на програмних додатках, виконаних мовою Python. Запропоновані обчислювальні механізми можуть бути корисними для фахівців, які займаються розробкою, тестуванням та верифікацією цифрових виробів, а також студентів, які вивчають курси комп'ютерної інженерії.

Пам'ять та процесори. NVMe SSDs up to 4TB. Не варто впадати в оману: AMD та Intel – безумовно, найнадійніші конкуренти Nvidia. Вони поділяють історію розробки успішних чіпів та створення програмних платформ для них. Але серед менших, менш перевірених гравців виділяється один: Cerebras. Компанія, яка спеціалізується на ШІ для суперкомп'ютерів, викликала фурор у 2019 році з Wafer Scale Engine, гігантським шматком кремнію розміром із пластину, упакованим 1,2 трильйона транзисторів. Остання ітерація Wafer Scale Engine 3 підвищує ставки до 4 трильйонів транзисторів.

Домінування Nvidia у сфері ШІ. Заснована в 1993 році, Nvidia залишила свій слід у тоді ще новій області графічних процесорів (GPU) для персональних комп'ютерів. Але чіпи AI компанії, а не графічне обладнання для ПК, вивели Nvidia до лав найдорожчих компаній світу. Виявляється, графічні процесори Nvidia також добре підходять для ШІ. В результаті її акції стали більш ніж у 15 разів кориснішими, ніж на початку 2020 року; виручка зросла приблизно з 12 мільярдів доларів США у 2019 фінансовому році до 60 мільярдів доларів у 2024 році; а передові чіпи цієї електростанції (генератора) штучного інтелекту так само рідкісні і бажані, як вода в пустелі. Groq – це ідеальна форма для функції та з унікальним підходом до обладнання ШІ. Переваги: відмінна продуктивність виведення ШІ. Недоліки: в даний час програма обмежена виведенням. Підхід Groq зосереджений на тісному поєднанні пам'яті та обчислювальних ресурсів для прискорення швидкості, з якою велика мовна модель може відповідати на запити. «Їхня архітектура заснована на пам'яті. Пам'ять тісно пов'язана із процесором. Було б краще мати більше вузлів, але ціна за токен та продуктивність – це божевілля», – каже Мурхед. «Токен» – це базова одиниця даних, яку обробляє модель; у LLM це зазвичай слово або частина

DOI 10.15588/1607-3274-2025-3-17

слова. Продуктивність Groq ще більше вражає, враховуючи, що її чіп, званий Language Processing Unit Inference Engine, виготовлений з використанням 14-нанометрової технології GlobalFoundries, яка на кілька поколінь відстає від технології TSMC, що використовується в Nvidia H100. У липні Groq опублікував демонстрацію швидкості виведення свого чіпа, яка може перевищувати 1250 токенів за секунду, працюючи на LLM Meta's Llama 3 з 8 мільярдами параметрів. Це перевершує навіть демо Samba Nova, яке може перевищувати 1000 токенів за секунду [8].

Інтелектуальний комп'ютинг [9] – нова парадигма обчислень, яка з'єднує людину з комп'ютером, традиційні обчислення з перцептивним, когнітивним та автономним інтелектом та сприяє цифровій революції в епоху великих даних, ШІ та Інтернету речей за допомогою нових обчислювальних теорій, архітектур, методів, систем. Інтелектуальний комп'ютинг – це історія обчислень за час існування людства, що має на меті – управління процесами та явищами на основі їх моніторингу. Інтернет – це мозок людства з функціями комунікацій, моніторингу та управління. Глобальна мета інтелектуального комп'ютингу – створення мозку людства в рамках інтернет-інфраструктури.

Восени 2024 року в Єревані зібралися три конференції (East-West Design and Test Symposium, Micro Electronic Forum, INDUSTRY.AI), присвячені AI вирішенню актуальних технологічних проблем в області design and test комп'ютингу. Симпозіум IEEE East-West Design & Test 2024 року досліджує нові тенденції у тестуванні, діагностиці та ремонті мікроелектронних систем, а також кібербезпеці, автомобілебудуванні, Інтернеті речей та штучному інтелекті [9, 10]. Він об'єднав вчених планети з 27 країн світу. Було запрошено 17 ключових доповідачів (keynotes), що є рекордом за всю 20-річну історію EWDT-симпозіуму [10]. Тематика симпозіуму була представлена багатьма напрямками CAD і EDA, які дедалі більше пронизані інструментами AI. Кожен день роботи симпозіуму було відзначено круглим столом наукових дискусій, за яким експерти в галузі AI-комп'ютингу обговорювали питання технологічної досконалості у сфері управління якістю продуктів, виробів процесів та сервісів. Безумовним технологічним AI-лідером виступає спонсор EWDT-симпозіуму – компанія Synopsys-Вірменія під керівництвом її лідера та головного архітектора Yervant Zorian, який протягом 20 років несе технологічну досконаліть до університетів, компаній та держав Сходу та Заходу. «Through East-West-cooperation to the harmony of excellence» – девіз симпозіуму. Це знайшло відображення в засіданні вчених, бізнесменів і дослідників, присвячених інтелектуалізації освіти та індустрії у Вірменії, коли у величезній залі одночасно зібралося понад 1000 осіб, які хочуть впроваджувати AI-механізми у всі сфери людської діяльності маленької, але технологічно досконалої держави. Запрошені keynotes представили доповіді, що відображають сучасні тенденції в AI-комп'ютингу (якість мікроелектронних 3D-виробів, AI-computing in design and test; prompt AI-engeniring) [10–12]. Регулярні доповіді (70 papers) на симпозіумі були представлені відомими науковими школами: професора Vazgen Melikyan (80 захищених дисертацій), професора Samvel Shukurian, професорів Paolo Prinetto, H. Fatih Ugurdag, Zainalabedin Navabi).

3 МАТЕРІАЛИ І МЕТОДИ

Впровадимо наступні визначення, метрику та механізми комп'ютингу.

Intelligent Computing (Vladimir Hahanov) – це галузь знань, яка займається теорією, практикою та економікою гармонійних відношень між детермінованим та ймовірнісним (AI) комп'ютингом за метрикою <час – ресурси – якість> для управління процесами та явищами на основі моніторингу оцифрованих та розумно-пов'язаних між собою соціального, физичного та інформаційного просторів.

Big Data – великі дані – кількість та форма інформації, що важко сприймається свідомістю людини. Метрика великих даних VVV: volume, velocity, variety (обсяг, швидкість генерування та різноманітність).

Data Science – наука, що вивчає життєвий цикл та форми даних з метою отримання актуальної інформації для прийняття рішення. Використовує підходи: дискретна математика та статистика, штучний інтелект та хмарні обчислення для аналізу великих обсягів даних під час вирішення завдань класифікації, регресії, кластеризації. Ключові інструменти: R, Python, Apache Hadoop, MapReduce, Apache Spark, NoSQL Databases, Cloud computing, GitHub.

Machine learning – алгоритми пошуку закономірностей у вхідних даних без програмування на основі розумних механізмів їх структуризації з метою розпізнавання патернів та прийняття рішення. «Ніхто машину не навчає». Наприклад, висловлювання «я навчаю комп'ютер» або «ми навчили нейромережу» звучать щонайменше наївно, на думку авторитетів у даній галузі (Daniel Faggella) [13].

Розумні (пов'язані) структури даних дозволяють без програмування вирішувати завдання структуризації корисної інформації шляхом суперпозиції таблиць та матриць, векторів. Ця ідея є основною в механізмах технічної діагностики, машинного навчання, включаючи нейронні мережі як тип розумних структур даних.

Artificial Intelligence (штучний інтелект) – це комп'ютинг, що моделює людське мислення з метою прийняття раціональних рішень на основі обробки природної мови, розпізнавання мови та зображень.

Neural Networks (NN) – штучні нейронні мережі є механізми-алгоритми машинного навчання рекурентних чи конволюційних структур даних. Вони складаються з шарів нейронів, що мають ваги вхідних дуг, та поріг-активатор на виході. Якщо вихідні дані вузла перевищують поріг, він активується для передачі даних на наступний рівень мережі. Convolutional neural

DOI 10.15588/1607-3274-2025-3-17

networks (CNN) – це аналог комбінаційної схеми. Recurrent neural networks (RNNs) – аналог схеми з глобальними зворотними зв'язками. З допомогою NN вирішуються завдання машинного зору, розпізнавання мови, обробки природної мови.

Інженерний соціальний комп'ютинг – практичне впровадження ідей, наукових теорій, формул вченого у практику життя соціальної групи. Приклади інженерного комп'ютингу – промпт-інженерія (чат-боти) компаній NVIDIA, Microsoft, Google, DeepSeek AI.

Інтелектуальний комп'ютинг використовує метрику xor двійкових оцифрованих процесів, що конволюційно (згортково) замикають відстань між ними до нуля: $\sum_{xor} d_i=0$. Така метрика за трьома компонентами або процесами стає метрикою для будь-якого комп'ютингу. Всі моніторингові моделі управління соціальними і фізичними процесами на основі моніторингу використовують єдине рівняння інтелектуального комп'ютингу $TxorLxorF=0$ (T – актуальна модель, L – специфікація, F – помилки) для вирішення всього трьох завдань: 1) $F=LxorT$ – верифікація або тестування, $L=TxorF$ – синтез специфікації, $T=LxorF$ – синтез (тренінг) актуальної моделі. Метрика інтелектуального комп'ютингу враховує наступні чинники: 1) економіка обчислень; 2) надмірність моделей чи розумних структур даних; 3) мінімальна затримка під час обробки вхідних запитів; 4) мінімальне енергоспоживання під час обробки великих даних; 5) гармонія між метою та засобами для її досягнення; 6) обнуління алгоритму обробки структур даних (моделі) рахунок їх експоненційної надмірності; 7) використання транзитивного замикання бінарного відношення $TxorL=F$ для вдосконалення та генерації нових властивостей обчислювачів; 8) збереження та нарощування історії обчислень при обробці інших типів даних; 9) цифровізація будь-яких процесів та явищ на введеному універсумі примітивів для керування на основі моніторингу; 10) безперервна масштабованість комп'ютерних архітектур для обробки процесів та явищ на основі інфраструктури edge-fog-cloud; 11) використання однієї з двох процедур modeling або synthesis при створенні моделі для виконання simulation вхідного впливу з метою отримання вихідного результату; 12) використання всієї сучасної палітри обчислювальних архітектур (AI, quantum, classical, in-memory, processor, програмування алгоритмів, FPGA, RISC-V, ASIC, VLSI, matrix processor) для організації економних та швидких обчислень.

Механізми комп'ютингу. Механізм комп'ютингу являє собою гармонійне відношення між моделлю і алгоритмом, що дозволяє мінімізувати час обчислень за рахунок збільшення надмірності (redundancy) пам'яті і навпаки. Модель і алгоритм взаємодіють спрощено як дві сполучені сосудини $M+A=1$. Це співвідношення має два крайніх стани: $M=1$, $A=0$; $M=0$, $A=1$. Обчислювальний механізм – це гармонійне співвідношення між моделлю і алгоритмом, метою і витратами. Як модель сучасного комп'ютингу, вона має тенденцію до інтелектуалізації в бік надмірності (redundancy) моделі та обнулення алгоритму. Формулу «Інь-Ян механізму» комп'ютингу $MxA=$const важливо враховувати для синтезу нових архітектур і методів.

Модель функціональності – це система відношень між вхідними та вихідними сигналами. Встановлення таких відношень по кожному патерну – це шлях до автоматичної побудови моделі бізнес-процесу та його компонентів. Комп'ютинг – це цілеспрямовані відношення між структурами даних та алгоритмом. Мета – побудова моделі для комп'ютингу бізнес-процесу за форматом чат-бота або із залученням оператора. Автоматизація – це синтез моделі (алгоритму) зі стійкою тенденцією до повторення результату інших даних. Кожна міні-функція має заголовки, що повторюються на бізнес потоці, і унікальні неповторювані на бізнес-потоці дані, унітарно-закодовані на універсумі примітивів. AI-computing – це логічне відношення між алгоритмом та моделлю, що призводить у процесі навчання до створення смарт-моделі, що обнулює алгоритм за рахунок надмірності розумних структур даних. Smart model або ML-модель – це розумні структури даних, які створюють стійкі логічні зв'язки під час навчання вхід-вихідними даними для використання отриманої логіки для моделювання будь-якого іншого вхідного набору даних без алгоритму. Алгоритм – це логіка обробки вхідних даних на смарт-моделі для автоматичного визначення відносин між компонентами вхідної інформації. Тут computing алгоритм замінюється на data set для smart model. Навчання чи синтез моделі – це процес встановлення логічних зв'язків між компонентами розумних структур даних на навчальному data set з метою подальшого адекватного моделювання будь-якого вхідного набору даних. У процесі навчання логіка роботи modeling-алгоритму перетворюється на надмірність розумних структур даних під актюаторним впливом вхідних даних.

Існує пряма аналогія між процесом навчання ML-моделі та синтезом комбінаційної схеми. Обидві моделі не мають пам'яті та навчаються завдяки синтезу логічних зв'язків між вихідними та вихідними сигналами (рис. 3).

DOI 10.15588/1607-3274-2025-3-17

а б в

Рисунок 3 – Моделі AI-computing – три фази життєвого циклу: а – AI-моделювання; AI-тестування; AI-симуляція

Суть AI-computing – синтез ML-моделі (modeling) за допомогою навчальних (відомих) *XY*-наборів *f=XxorY* з метою подальшого тестування smart моделі та моделювання (simulation) на ній будь-яких вхідних X-наборів для визначення вихідних *Y*-сигналів *Y=f(X)*. Інакше: синтез моделі (розумних структур даних) для моделювання будь-яких вхідних наборів або modeling for simulation. Modeling – процес синтезу (тренінгу) моделі за заданими вхідно-вихідними наборами сигналів. Simulation – процес визначення станів виходів моделі на вхідних наборах сигналів, як зазначено у [6].

Тут йдеться про карту тестування, як продукт modeling and simulation (рис. 4). Карта синтезується за логічним вектором функціональності і є смарт-модель для design and test промпт-інжинірингу.

Рисунок 4 – Схема моделювання та симуляції для карти тестування

Таким чином, у загальному випадку intelligent computing – це синтез моделі на вхідних-вихідних наборах для подальшого моделювання будь-яких вхідних наборів. Intelligent computing is modeling for simulation. Modeling – це синтез моделі (логічного вектора) на вхідних наборах. Simulation – це визначення стану виходу моделі за вхідним набором та синтезованою моделлю (логічним вектором). Перевагою запропонованого дослідження є доведення моделей та алгоритмів до рівня простих механізмів інженерного використання.

4 ЕКСПЕРИМЕНТИ

Існує повна аналогія процесів modeling та simulation (рис. 5) на основі векторно-логічних та інтелектуальних моделей. Smart model – розумні структури даних для моделювання будь-якого вхідного набору сигналів для визначення стану виходів без програмування алгоритму. AI-modeling – це синтез моделі за відомими станами входів-виходів.

AI (forward) simulation – це процес визначення стану виходів за відомими станами входів. Це також робочий режим формування inference. AI backsimulation (backpropagation) – це процес визначення стану входів за відомими станами виходів, що застосовується при пошуку помилок та оптимізації смарт-моделі. Testing map – це карта тестування функціональності, що вирішує всі питання: моделювання та діагностування несправностей, синтезу мінімального тесту та оцінки його якості.

Всі зазначені процеси оперують трьома компонентами рівняння (*T xor L xor F = 0*), серед яких два повинні бути відомі для пошуку третього. Але є виняток. Якщо відома геном-модель помилок, можна по одному компоненту згенерувати два інших. Це промпт-комп'ютинг.

У [6, 7] запропоновано нову технологію верифікації та моделювання несправностей, засновану на in-memory промпт-комп'ютингу. Сенс запропонованих інновацій полягає у побудові карти тестування функціональності за її логічним вектором. При цьому алгоритм побудови карти практично обнулюється за рахунок експоненційної надмірності розумних структур даних (рис. 6). Modeling for simulation – це розумні прості та доступні для розуміння механізми щодо проблем тестування та верифікації.

Далі розглядаються 10 експериментів над різними схемами, що моделюють фізичні та соціальні процеси. Враховувався час: підготовки схем, тестування схем та пошуку трьох дефектів. Результати наведені в таблиці 1, де VHDL – це базовий варіант, заснований на описі схеми за допомогою HDL-мови; VLC – запропонований варіант, заснований на використанні векторної логіки для тестування цифрових схем і систем. При виконанні експерименту інформація про схему вводиться малюнком на екрані за допомогою GUI. Експерименти були проведені аспірантами та студентами факультету комп'ютерні інженерії ХНУРЕ.

DOI 10.15588/1607-3274-2025-3-17

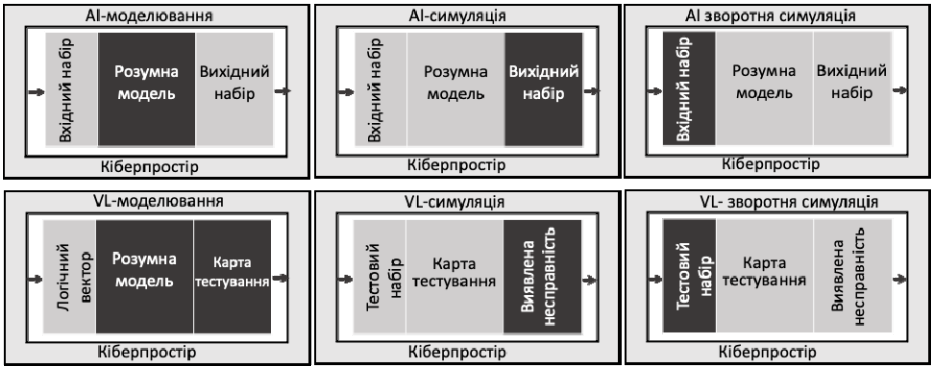


Рисунок 5 – Схема «Modeling for simulation»

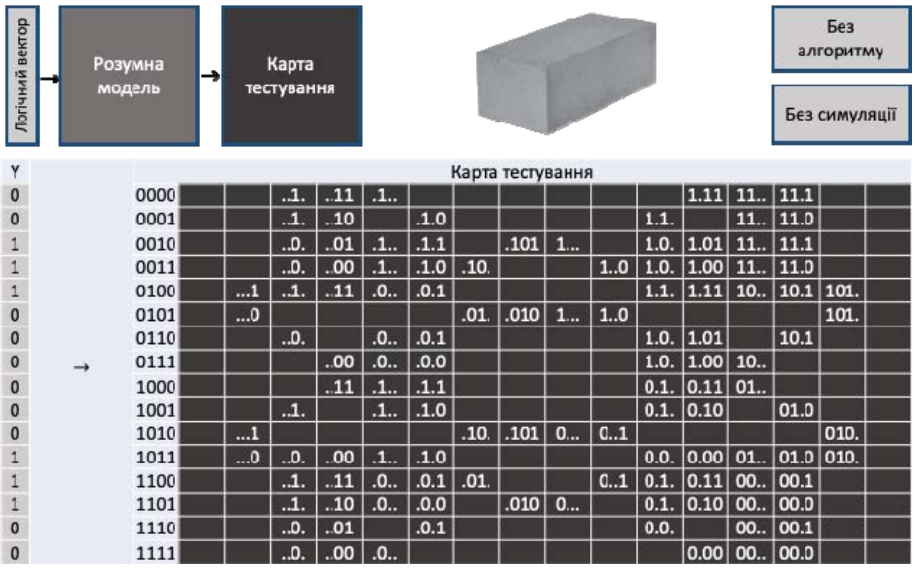


Рисунок 6 – Карта тестування несправностей для логіки 0011100000011100

Таблиця 1 – Результати експериментів

	Час, хвилини					
Експерименти	підготовки схем		тестування схем		пошуку дефектів	
Механізми	VHDL	VLC	VHDL	VLC	VHDL	VLC
Schneider	23	12	17	10	14	5
c17	20	6	15	14	13	4
c23	15	9	10	13	18	8
Trigger	19	10	15	16	11	7
M4-1	25	15	20	10	10	9
Coder 4-1	22	12	18	8	13	8
Adder	29	9	22	18	17	11
Recover	27	17	21	17	16	7
Dedec 4	25	20	20	10	20	6
Sell 4	30	19	27	15	24	9
S 432	36	25	32	17	13	9

5 РЕЗУЛЬТАТИ

Результати експериментів з оцінки часу підготовки та тестування схем, а також пошуку дефектів у них показані на графіках (рис. 7–9). Розрахунки проводяться для 11 пристроїв (вісь абсцис). По осі ординат відраховується час у хвилинах.

1. Підготовка схем (див. рис. 7) за допомогою мови опису апаратури VHDL (верхній графік) та на основі векторної логіки VLC (нижній графік).

2. Тестування схем (див. рис. 8) передбачає роботу додатка та роботу фахівця, який вручну визначає мінімальний тест перевірки одиночних несправностей за результатами моделювання.

3. Пошук дефектів (див. рис. 9) з використанням отриманої інформації про моделювання несправності на тестах: необхідно знайти три несправності у схемі.

DOI 10.15588/1607-3274-2025-3-17

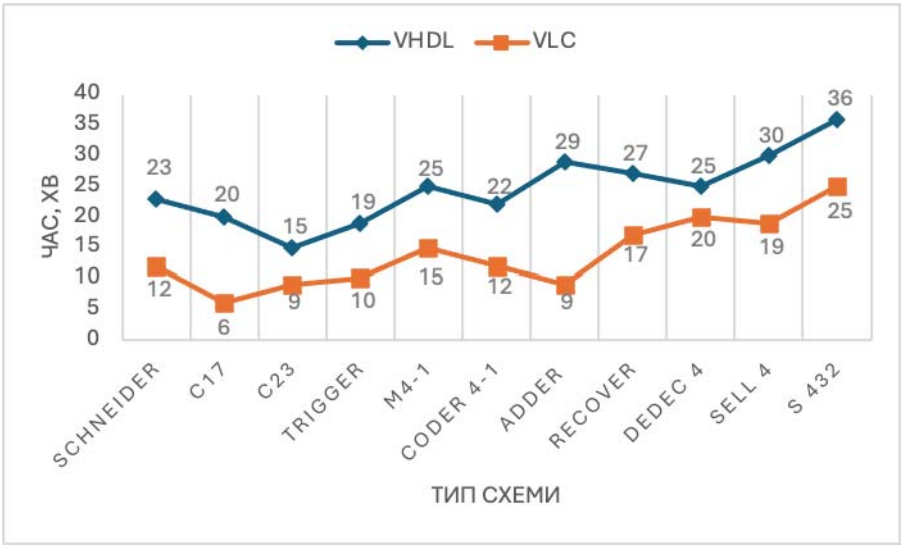


Рисунок 7 – Графіки часу підготовки схеми

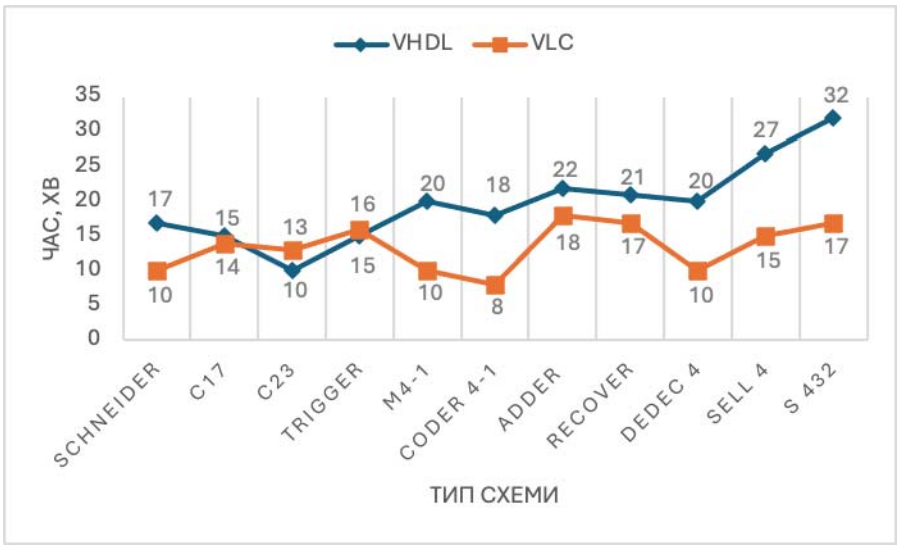


Рисунок 8 – Діаграма часу тестування схеми

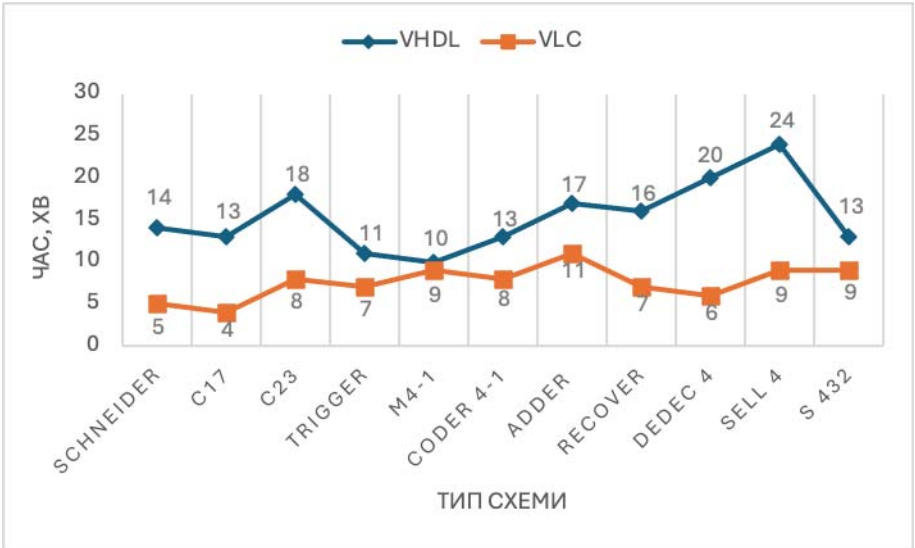


Рисунок 9 – Часові діаграми пошуку дефекту

DOI 10.15588/1607-3274-2025-3-17

Таким чином, запропоновано:

1) рівняння комп'ютингу як транзитивне замикання у тріаді відношень – як помилки, що створює нові структури, процеси чи явища;

2) механізми інтелектуального комп'ютингу орієнтовані на суттєве зниження часових та енергетичних витрат при вирішенні практичних завдань за рахунок обнулення алгоритмів обробки великих даних завдяки експоненційній надмірності розумних та надмірних AI-моделей;

3) механізми побудови моделей на основі універсуму примітивів, які мають Similarity по відношенню до їх використання для моделювання процесів (in-hardware synthesis, in-software programing, in neural network training, in-qubit quantization, in-memory modeling, in-truth table logic generation);

Наведено метричне порівняння механізмів моделювання несправностей, які є у промислових системах. Оцінено також векторне моделювання несправностей у цій метриці координат.

6 ОБГОВОРЕННЯ

Порівняння отриманих в ході експериментів результатів з аналогами демонструє наступне.

1. Під час підготовки схем переможцем є підхід, заснований на VLC (див. рис. 7). За рахунок наявності GUI він суттєво випереджає базовий варіант з малювання елементів та за верифікацією моделі.

2. Робота програми при тестуванні схем в обох випадках, як VHDL, так і VLC, займає менше секунди, інше – ручна обробка (див. рис. 8).

3. При пошуку дефектів на основі отриманої інформації про моделювання несправностей знову виграє новий варіант VLC, оскільки синтезована карта тестування – це явна форма завдання тестових наборів і комбінація несправностей, яка перевіряє кожен тестовий набір (див. рис. 9).

Таким чином, для відпрацювання навичок тестової верифікації необхідно використовувати прості технології введення інформації про схему та інженерні механізми побудови карт тестування цифрових логічних функціональностей. Мови опису апаратури (VHDL, Verilog) потрібно використовувати тільки у великих проектах, які потребують інтеграції великої кількості інженерів, які займаються проблемами тестування та верифікації. Більш технологічного та простого інженерного апарату, ніж механізми векторної логіки побудови карт тестування та моделювання, у світі на сьогоднішній день відсутні. Йдеться про дві альтернативи. Перша – це складні алгоритми ймовірнісної обробки текстів та даних для подальшого управління соціальним процесом. Друга – це детермінований vector-logic in-memory комп'ютинг, який використовують розумні моделі, лінійні алгоритми моделювання соціальних процесів. У результаті виходить інженерний комп'ютинг для моніторингу та управління соціальними процесами, який використовують добре напрацьовану теорію моделювання цифрових систем. Платою за таку стійку технологію modeling for simulation є препроцесор отримання двійкової моделі та постпроцесор дешифрації одержаного результату до форми, що є зручною для очей людини. Слід зазначити, що двійкова векторна модель соціального процесу легко верифікуються, завдяки напрацьованим простим інженерним механізмам побудови карт тестування з урахуванням промпт-інженерії [14, 15]. Тому новий перспективний науковий напрямок можна ідентифікувати як «VLC social modeling for simulation», де VLC – Vector Logic in-memory Computing, економіка якого строго спрямована на мінімізацію часу та енергії.

Тріада – це відношення між трьома компонентами. Виникає питання – навіщо потрібний третій компонент, яку функцію він виконує серед двох інших? Як не дивно, але третій компонент виконує дуже важливу функцію метричного виміру двох об'єктів, що утворюють якісну бінарність. Так виникає найпростіша модель сучасного комп'ютингу $LxorT=F$, L – ідея, T – реалізація, F – виміряні помилки між ідеальним та актуальним рішенням. Така проста тріада відношень дозволяє: 1) метрично оцінити одержаний актуальний розв'язок $F=LxorT$; 2) скоригувати ций розв'язок з метою усунення зафіксованих помилок $T=LxorF$; 3) скоригувати ідею для одержання нової якості комп'ютингу $L=TxorF$.

Тріада відношень формує процес синтезу (modeling) моделі як у детермінованому, так і ймовірнісному комп'ютингу. Помилки між ідеальним та актуальним розв'язком – це шлях до досконалості проекту шляхом їх виправлення, а також створення нового виробу, де помилка – це корисна інновація. Транзитивне замикання у тріаді відношень – це помилка, що створює нові структури, процеси чи явища у Всесвіті, природі, соціумі, комп'ютингу. Помилки призводять до нових результатів.

ВИСНОВКИ

Наукова новизна полягає у розробці наступних інноваційних рішень:

1) запропоновано тріаду відношень на основі хог-операції для вимірювання процесів та явищ у кіберсоціальному просторі;

2) запропоновані архітектури інтелектуального комп'ютингу для управління соціальними процесами на основі їх вичерпного моніторингу. Механізми інтелектуального комп'ютингу орієнтовані на істотне зниження часових і енергетичних витрат під час вирішення практичних завдань рахунок обнулення алгоритмів обробки великих даних, завдяки експоненційної надмірності розумних і надлишкових AI-моделей;

3) механізми побудови моделей на основі універсуму примітивів, які мають Similarity по відношенню до їх використання для моделювання процесів (in-hardware synthesis, in-software programing, in neural network training, in-qubit quantization, in-memory modeling, in-truth table logic generation). Експонентна

DOI 10.15588/1607-3274-2025-3-17

надмірність таблиці істинності 2^n від n змінних породжує супер-експоненційну надмірність $2^{\wedge}(2^n)$ логічних функцій, що обнуляє алгоритми обробки структур даних або робить їх лінійними.

Практична значимість дослідження. Реалізація запропонованих рішень в архітектурі in-memory комп'ютингу дає можливість не використовувати інструкції процесора, тільки read-write транзакції на логічних векторах, що економить час та енергію для виконання алгоритмів аналізу великих даних. Запропонована метрика інтелектуального комп'ютингу використовується для вибору архітектури та моделей обчислювальних процесів з метою отримання ефективних розв'язків практичних завдань.

Перспективи дослідження. Запропоновані механізми синтезу векторно-логічних моделей соціальних процесів або явищ на основі унітарного кодування патернів на універсумі примітивів, які орієнтовані на верифікацію, моделювання та тестування прийнятих рішень.

Інженерний соціальний комп'ютинг покликаний сприяти побудові миролюбних, справедливих та відкритих суспільств для досягнення цілей сталого розвитку. Для цього необхідні ефективні та засновані на активній участі широких верств населення державні інститути, здатні забезпечити якісну освіту та охорону здоров'я, справедливу економічну політику та всеосяжний захист довкілля [16].

ЛІТЕРАТУРА

1. Low Code/No Code Meets the Metaverse // [G. F. Hurlburt, G. K. Thiruvathukal, N. Kshetri and N. Ahmad] / Computer. – 2025. – Vol. 58, No. 03. – P. 22–28. DOI: 10.1109/MC.2024.3520883.
2. Shan R. Certifying Generative AI: Retrieval-Augmented Generation Chatbots in High-Stakes Environments / R. Shan // Computer. – 2024. – Vol. 57, No. 09. – P. 35–44. DOI: 10.1109/MC.2024.3401085.
3. The Evolution of AI Governance / [S. Chesterman, Y. Gao, J. Hahn, and V. Sticher] // Computer. – 2024. – Vol. 57, No. 09. – P. 80–92. DOI: 10.1109/MC.2024.3381215.
4. Chilimbi T. How We Built Rufus, Amazon's AI-Powered Shopping Assistant. A custom language model uses new techniques to answer shoppers' questions quickly [Electronic resource] / T. Chilimbi // IEEE Spectrum. – 04 Oct 2024. – Access mode: https://spectrum.ieee.org/shipt
5. Azure AI Document Intelligence [Electronic resource] / Access mode: https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence https://github.com/labelmeai/labelme
6. Vector Synthesis of Fault Testing Map For Logic / [V. Hahanov, W. Gharibi, S. Chumachenko, E. Litvinova] // IAES International Journal of Robotics and Automation (IJRA). – 2024. – Vol. 13, No. 3. – P. 293–306. DOI:10.11591/ijra.v13i3.pp293-306
7. In-Memory Intelligent Computing / [V. I. Hahanov, V. H. Abdullayev, S. V. Chumachenko et al.] // Radio Electronics, Computer Science, Control. – 2024. – №1. – P. 161–174. https://doi.org/10.15588/1607-3274-2024-1-15
8. Matthew S. Smith. Challengers are coming for Nvidia's Crown [Electronic resource] / S. Matthew // IEEE Spectrum. – 03 Sep 2024. https://spectrum.ieee.org/europa-clipper-2669391232
9. Intelligent Computing: The Latest Advances, Challenges, and Future [Electronic resource] / [Shiqiang Zhu, Ting Yu, Tao Xu et al.] // Intell Comput. – 2023. – 2:0006. DOI:10.34133/icomputing.0006
10. East-West Design & Test Symposium [Electronic resource] / Access mode: https://conf.ewdtest.com
11. Testing for Electromigration in Sub-5-nm FinFET Memories / [M. Mayahinia, M. Tahoori, G. Tshagharyan et al.] // IEEE Design & Test. – Dec. 2024. – Vol. 41, No. 6. – P. 54–61. – doi: 10.1109/MDAT.2024.3411527.
12. Tahoori M. Special Issue on Silicon Lifecycle Management / M. Tahoori, Y. Zorian // IEEE Design & Test. – Aug. 2024. – Vol. 41, No. 4. – P. 5–6. DOI: 10.1109/MDAT.2024.3392620.
13. Faggella D. What is Machine Learning? Comprehensive Overview [Electronic resource] / [D. Faggella] // February 26, 2020. – Access mode: https://emerj.com/ai-glossary-terms/what-is-machine-learning/.
14. Vector-Logical In-Memory Simulation of Faults as Truth Table Addresses / [V. Hahanov, E. Litvinova, H. Hahanova et al.] // 2024 IEEE East-West Design & Test Symposium (EWDTS), Yerevan, Armenia, 2024, P. 1–6. DOI: 10.1109/EWDTS63723.2024.10873615.
15. Prompt-Testing of Logic / [V. Hahanov, D. Devadze, I. Hahanov et al.] // 2024 IEEE East-West Design & Test Symposium (EWDTS). – Yerevan, Armenia, 2024. – P. 1–5. DOI: 10.1109/EWDTS63723.2024.10873774.
16. Sustainable Development Goals [Electronic resource] – Access mode: https://www.un.org/sustainabledevelopment/peace-justice/

Стаття надійшла до редакції 22.04.2025.
Після доробки 06.07.2025.

DOI 10.15588/1607-3274-2025-3-17

UDC 681.326

ENGINEERING SOCIAL COMPUTING

Hahanov V. I. – Dr. Sc., Professor of the Design Automation Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Chumachenko S. V. – Dr. Sc., Professor, Head of the Design Automation Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Lytvynova E. I. – Dr. Sc., Professor of the Design Automation Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Khakhanova H. V. – Dr. Sc., Professor of the Design Automation Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Hahanov I. V. – PhD, Assistant of the Design Automation Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Obrizan V. I. – PhD, Post-Doctoral Student of the Design Automation Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Hahanova I. V. – Dr. Sc., Professor of the Design Automation Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

Maksymova N. G. – Postgraduate student of the Design Automation Department, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine.

ABSTRACT

Context. The relevance of the study is due to the need to eliminate contradictions between management and performers by introducing engineering social computing, which ensures moral management of social processes based on their metric monitoring.

Objective. The goal of the investigation is to develop engineering architectures for monitoring and managing social processes based on vector logic.

Method. The research is focused on the development of engineering vector-logical schemes and architectures for management of social processes based on their comprehensive metric monitoring in order to create comfortable conditions for creative work. Definitions of the main concepts of AI development are given. Interesting fragments of the history of computing are given. The computing equation is introduced as a transitive closure in a triad of relations – in the form of an error that creates new structures, processes or phenomena. Mechanisms of intelligent computing are developed that combine algorithms and data structures of deterministic and probabilistic AI computing. Mechanisms for constructing models based on the universe of primitives that have Similarity in relation to their use for process modeling (in-hardware synthesis, in-software programming, in neural network training, in-qubit quantization, in-memory modeling, in-truth table logic generation) are proposed. An intelligent computing metric is introduced, which is used to select the architecture and models of computing processes in order to obtain effective solutions to practical problems.

Results. The following is proposed: 1) the computing equation as a transitive closure in a triad of relations – in the form of an error that creates new structures, processes or phenomena; 2) mechanisms of intelligent computing aimed at a significant reduction in time and energy costs in solving practical problems by zeroing out algorithms for processing big data, due to the exponential redundancy of smart and redundant AI models; 3) mechanisms for constructing models based on the universe of primitives that have Similarity in relation to their use for modeling processes.

Conclusions. Scientific novelty concludes the following innovative solutions: 1) a triad of relations based on the xor-operation for measuring processes and phenomena in the cyber-social world is proposed; 2) intelligent computing architectures are proposed for managing social processes based on their comprehensive monitoring; 3) the implementation of these schemes in the in-memory computing architecture makes it possible not to use processor instructions, only read-write transactions on logical vectors, which saves time and energy for the execution of big data analysis algorithms; 4) mechanisms for synthesizing vector-logical models of social processes or phenomena based on unitary coding of patterns on the universe of primitives are proposed, which are focused on verification, modeling and testing of decisions made. The practical significance of the study lies in the fact that the metric of intelligent computing is proposed, which is used as a method for selecting the architecture and models of computing processes to obtain effective solutions to practical problems. Engineering social computing is designed to contribute to the construction of peaceful, fair and open societies to achieve the Sustainable Development Goals (SDG 16).

KEYWORDS: human brain, internet infrastructure, intelligent computing, artificial intelligence, smart data structures, computing models, computing history, computing metrics, AI-Industry, Modeling for simulation, Sustainable Development Goals, peaceful society.

REFERENCES

1. Hurlburt G. F., Thiruvathukal G. K., Kshetri N. and Ahmad N. Low Code/No Code Meets the Metaverse, *Computer*, 2025, Vol. 58, No. 03, pp. 22–28. DOI: 10.1109/MC.2024.3520883.
2. Shan R. Certifying Generative AI: Retrieval-Augmented Generation Chatbots in High-Stakes Environments,

DOI 10.15588/1607-3274-2025-3-17

Computer, 2024, Vol. 57, No. 09, pp. 35–44. DOI: 10.1109/MC.2024.3401085.
3. Chesterman S., Gao Y., Hahn J., and V. Sticher. The Evolution of AI Governance, *Computer*, 2024, Vol. 57, No. 09, pp. 80–92. DOI: 10.1109/MC.2024.3381215.
4. Chilimbi T. How We Built Rufus, Amazon's AI-Powered Shopping Assistant. A custom language model uses new techniques to answer shoppers' questions quickly [Electronic resource], *IEEE Spectrum*, 04 Oct 2024. Access mode: https://spectrum.ieee.org/shipt
5. Azure AI Document Intelligence [Electronic resource] / Access mode: https://azure.microsoft.com/en-us/products/ai-services/ai-document-intelligence https://github.com/labelmeai/labelme
6. Hahanov V. Gharibi W., Chumachenko S., Litvinova E. Vector Synthesis of Fault Testing Map For Logic, *IAES International Journal of Robotics and Automation (IJRA)*, 2024, Vol. 13, No. 3, pp. 293–306. DOI:10.11591/ijra.v13i3.pp293-306
7. Hahanov V. I. Abdullayev V. H., Chumachenko S. V., E. I. Lytvynova, I. V. Hahanova In-Memory Intelligent Computing, *Radio Electronics, Computer Science, Control*, 2024, №1, pp. 161–174. – https://doi.org/10.15588/1607-3274-2024-1-15
8. Matthew S. Smith. Challengers are coming for Nvidia's Crown [Electronic resource], *IEEE Spectrum*, 03 Sep 2024. https://spectrum.ieee.org/europa-clipper-2669391232
9. Shiqiang Zhu, Ting Yu, Tao Xu, Hongyang Chen, Schahram Dustdar, Sylvain Gigan, Deniz Gunduz, Ekram Hossain, Yaochu Jin, Feng Lin, Bo Liu, Zhiguo Wan, Ji Zhang, Zhifeng Zhao, Wentao Zhu, Zuoning Chen, Tariq S. Durrani, Huaimin Wang, Jiangxing Wu, Tongyi Zhang, Yunhe Pan Intelligent Computing: The Latest Advances, Challenges, and Future [Electronic resource], *Intell Comput*, 2023, 2:0006. DOI:10.34133/icomputing.0006
10. East-West Design & Test Symposium [Electronic resource]. Access mode: https://conf.ewdtest.com
11. Mayahinia M., Tahoori M., Tshagharyan G., Amirkhanyan K., Ghukasyan A., Harutyunyan G., Zorian Y. Testing for Electromigration in Sub-5-nm FinFET Memories, *IEEE Design & Test*, Dec. 2024, Vol. 41, No. 6, pp. 54–61. DOI: 10.1109/MDAT.2024.3411527.
12. Tahoori M., Zorian Y. Special Issue on Silicon Lifecycle Management, *IEEE Design & Test*, Aug. 2024, Vol. 41, No. 4, pp. 5–6. doi: 10.1109/MDAT.2024.3392620.
13. Faggella D. What is Machine Learning? Comprehensive Overview [Electronic resource], February 26, 2020. Access mode: https://emerj.com/ai-glossary-terms/what-is-machine-learning/.
14. Hahanov V., Litvinova E., Hahanova H., Chumachenko S., Davitadze Z., Hahanova I., Kulak H., Ponomarova V., Abdullayev V. H. Vector-Logical In-Memory Simulation of Faults as Truth Table Addresses, *2024 IEEE East-West Design & Test Symposium (EWDTS)*. Yerevan, Armenia, 2024, pp. 1–6. DOI: 10.1109/EWDTS63723.2024.10873615.
15. Hahanov V., Devadze D., Hahanov I., Chumachenko S., Litvinova E., Obrizan V., Dmytro P., Mishchenko A., Maksymova N. Prompt-Testing of Logic, *2024 IEEE East-West Design & Test Symposium (EWDTS)*. Yerevan, Armenia, 2024, pp. 1–5. DOI: 10.1109/EWDTS63723.2024.10873774.
16. Sustainable Development Goals [Electronic resource] – Access mode: https://www.un.org/sustainabledevelopment/peace-justice/

DOI 10.15588/1607-3274-2025-3-17

p-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2025. № 3
e-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2025. № 3

УПРАВЛІННЯ У ТЕХНІЧНИХ СИСТЕМАХ

CONTROL IN TECHNICAL SYSTEMS

UDC 519.852.6

AN INNOVATIVE APPROXIMATE SOLUTION METHOD FOR AN INTEGER PROGRAMMING PROBLEM

Mamedov K. Sh. – Dr. Sc., Professor, Professor of Baku State University, Head of the Department, Institute of Control Systems, Azerbaijan.

Niyazova R. R. – Doctorant and Scientist, Institute of Control Systems, Azerbaijan.

ABSTRACT

Context. There are certain methods for finding the optimal solution to integer programming problems. However, these methods cannot solve large-scale problems in real time. Therefore, approximate solutions to these problems that work quickly have been given. It should be noted that the solutions given by these methods often differ significantly from the optimal solution. Therefore, the problem of taking any known approximate solution as the initial solution and improving it further arises.

Objective. Initially, a certain approximate solution is found. Then, based on proven theorems, the coordinates of this solution that do not coincide with the optimal solution are determined. After that, new solutions are found by sequentially changing these coordinates. The one that gives the largest value to the functional among these solutions is accepted as the final solution.

Method. The method we propose in this work is implemented as follows:

First, a certain approximate solution to the problem is established, then the numbers of the coordinates of this solution that do not coincide with the optimal solution are determined. After that, new solutions are established by sequentially assigning values to these coordinates one by one in their intervals. The best of the solutions found in this process is accepted as the final innovative solution.

Results. A problem was solved in order to visually illustrate the quality and effectiveness of the proposed method.

Conclusions. The method we propose in this article cannot give worse results than any approximate solution method, is simple from an algorithmic point of view, is novel, can be easily programmed, and is important for solving real practical problems.

KEYWORDS: integer programming problem, initial approximate solution, the interval in which the coordinates of the approximate solution may differ from the optimal solution, innovative approximate solution, computational experiments.

NOMENCLATURE

N is a number of issues resolved;

n is a number of unknowns;

a_{ij} is a given positive integer;

c_j is a given positive integer;

d_j is a positive integer;

b_i is a given positive integers;

x_j is an j-th unknown;

j_* is a value of the unknown;

X^0 is an initial approximate solution;

r is an initial solution is the number of the first coordinate that received a value of “0” when constructing the solution;

f^0 is a value of the approximate solution to the objective function;

$\tilde{X}$ is a solution found when the unknowns in the problem are not required to be integers;

$\frac{\alpha}{\beta}$ is a coordinate of the proper fraction $\tilde{X}$;

k is a number of the fractional coordinate in the numerical $\tilde{X}$ solution;

p is a positive integer;

q is a positive integer;

δ is a positive integer;

δ_k^1 is a certain integers expressed as a percentage;

δ_k^0 is a certain integers expressed as a percentage;

$n(0)$ is a minimal number of zeros coordinates in the optimal solution;

$n(d)$ is a minimal number of non-zero coordinates in the optimal solution;

$\underline{X^*}$ is an optimal solution of certain problems;

$\underline{X^{lp}}$ is an optimal solution of certain problems;

ω_1 is a certain set of number;

ω_2 is a certain set of number;

DOI 10.15588/1607-3274-2025-3-18

$n(\omega_1)$ is a number of element of certain sets ω_1;

$n(\omega_2)$ is a number of element of certain sets ω_2;

$\underline{n}(d)$ is an integer not greater than $n(d)$;

n_1 is a lower bound on the minimal number of non-zero coordinates in the approximate solution;

$\overline{n}$ is a maximum number of non-zero coordinates in the optimal solution;

$\underline{n}(0)$ is a lower bound on the minimum number of zeros in the optimal solution;

$\tilde{f}$ is a value of the functional with respect to the solution $\tilde{X}$;

X^{it} is an innovative improved final solution;

f^{it} is a value of the objective function according to the solution X^{it};

X^t is a certain intermediate approximate solution;

f^t is a value of the function according to the solution X^t.

INTRODUCTION

Let's look at the well-known integer programming problem given below:

$$\sum_{j=1}^{n} c_j x_j \to \max, \tag{1}$$

$$\sum_{j=1}^{n} a_{ij} x_j \le b_i \quad \left(i = \overline{1,m}\right) \tag{2}$$

$$0 \le x_j \le d_j \quad j = \overline{(1,n)}. \tag{3}$$

Here $c_j > 0$, $a_{ij} > 0$, $b_i > 0$ and $d_j > 0$ $(i = \overline{1,m},\ j = \overline{1,n})$ are integers.

Note that problem (1)–(3) is called an integer linear programming problem in the literature.

In this problem, any vector $X = (x_1, x_2, ..., x_n)$ that satisfies conditions (2)–(3) is called a possible solution to the problem. The optimal solution to the problem is understood to be the solution that gives the largest (maximum) value to the function (1) among the possible solutions.

Note that among the possible solutions to problem (1)–(3), the solution that gives a large value to function (1) based on certain criteria is called an approximate (suboptimal) solution.

It is known that the problem (1)–(3) belongs to the class of "hard-to-solve problems", that is, to the class "*NP*-complete" [1]. In other words, the maximum number of operations required by any of the known methods for finding the optimal solution to this problem (e.g., branch and bounds, combinatorial, etc.) is not limited by any polynomial that depends on the size of the problem. Therefore, approximate solutions of problems (1)–(3) of various nature and speed have been developed [2–7,13–16,21–25 etc.].

First, let us briefly explain one of the methods proposed in [13]. For this purpose, let us give a certain economic interpretation to the problem (1)–(3).

Suppose a certain enterprise must produce *n* different products, expressed in number. For this, *m* number of resources $b_i, \left(i = \overline{1,m}\right)$ are allocated accordingly. Let us assume that the production of one unit of the *j-th* $j = \overline{(1,n)}$ product requires the expenditure of the *i-th* $\left(i = \overline{1,m}\right)$ resource a_{ij}, $(i = \overline{1,m},\ j = \overline{1,n})$. In this case, products should be produced such that the total resources spent on their production do not exceed the given corresponding limit resources $b_i, \left(i = \overline{1,m}\right)$ and at the same time the total income from their sale is maximized.

If we denote the price of one unit of the *j-th* $j = \overline{(1,n)}$ product by c_j, $(j = \overline{1,n})$ and the quantity of the product to be produced by x_j, $(j = \overline{1,n})$, then we obtain model (1)–(3).

Suppose that a certain specified product *j-th*, $j = \overline{(1,n)}$ to be produced. Then, a_{ij}, $(i = \overline{1,m},\ j = \overline{1,n})$ amount of the *i*-th $\left(i = \overline{1,m}\right)$ resource must be spent on the production of one unit of this product. (If these resources are measured in monetary units, then an amount of a_{ij}, $(i = \overline{1,m},\ j = \overline{1,n})$ must be spent).

In this case, the worst-case cost per unit of product *j*-th $j = \overline{(1,n)}$ is equal to $\max\limits_i a_{ij}$.

Then the price increase for each unit of product *j*-th $j = \overline{(1,n)}$ mentioned above is equal to $\dfrac{c_j}{\max\limits_i a_{ij}}$ $j = \overline{(1,n)}$.

Naturally, we need to produce product number j_* such that the expression $\dfrac{c_j}{\max\limits_i a_{ij}}$ $j = \overline{(1,n)}$ is the largest.

Thus, we get the following selection criteria:

$$\max_j \frac{c_j}{\max\limits_i a_{ij}} = \frac{c_{j_*}}{\max\limits_i a_{ij_*}}$$

or

$$j_* = \arg\max_i \frac{c_j}{\max\limits_i a_{ij}}. \tag{4}$$

DOI 10.15588/1607-3274-2025-3-18

It should be noted that in studies [8, 10, 14], a criterion of type (4) was used for interval integer and mixed-integer programming problems of type (1)–(3).

When constructing an approximate solution to the problem (1)–(3), most methods initially assume $X=(0,0,...,0)$ as the starting value, and the unknown variable x_{j_*} with the index j_*, determined by a certain criterion, is assigned a value. After this process is carried out for all indices j, $j=\overline{(1,n)}$, a certain approximate solution $X^0=\left(x_1^0,x_2^0,...,x_n^0\right)$ is obtained.

It should be noted that the approximate solution method described above is one of the known methods. However, numerous computational experiments have shown that the solution given by known approximate solution methods can differ significantly from the optimal solution. Therefore, there is a need to develop an approximate solution method that does not give a worse solution than the solutions given by known methods, works quickly, is easy to implement, and does not cause difficulties from a programming point of view. It should be noted that such a solution method is called an innovative approximate solution method, and the solution found is called an innovative approximate solution [9, 10, 26]. In this work, we have developed such an approximate solution method.

1 PROBLEM STATEMENT

Without loss of generality, let us assume that in problem (1)–(3) the coefficients are numbered as follows based on condition (4). $\frac{c_1}{\max\limits_i a_{i1}}\ge\frac{c_2}{\max\limits_i a_{i2}}\ge...\ge\frac{c_k}{\max\limits_i a_{ik}}\ge...\ge\frac{c_n}{\max\limits_i a_{in}}$ In this case, the initial approximate solution $X^0=\left(x_1^0,x_2^0,...,x_n^0\right)$ of problem (1)–(3) is found analytically by the following formula: for each number j, $j=\overline{(1,n)}$

$$x_j^0=\begin{cases} d_j,\ if\ \forall i, i=\overline{(1,m)},\ a_{ij}d_j\le b_i-\sum\limits_{l=1}^{j-1}a_{il}x_l^0 \\ \min\limits_i\left[\left(b_i-\sum\limits_{l=1}^{j-1}a_{il}x_l^0\right)/a_{ij}\right],\ \text{otherwise}. \end{cases}$$

We can briefly write this formula as follows:

$$x_j^0=\min\left\{d_j;\ \min_i\left[\left(b_i-\sum_{l=1}^{j-1}a_{il}x_l^0\right)/a_{ij}\right]\right\}. \qquad (5)$$

Here, the symbol [z] denotes the integer part of the number z. It is clear that the solution found by formula (5) will have the following structure

$$X^0=\left\{x_1^0,x_2^0,...,x_{r-1}^0,0,x_{r+1}^0,...,x_n^0\right\}. \qquad (6)$$

In the problem under consideration, the value of the function (1) corresponding to the approximate solution of (6) is

$$f^0=\sum_{j=1}^{n}c_jx_j^0\ .$$

Note that in formula (6) we must have $x_j^0>0$, $x_r^0=0$ for the $j=\overline{1,r-1}$ numbers. The remaining coordinates $x_{r+1}^0,...,x_n^0$ can be 0 or positive integers.

2 REVIEW OF THE LITERATURE

Integer programming problems, as well as their individual classes, have been known since the middle of the last century. Since these problems are of great practical importance, various exact methods for their solution have been developed. However, it soon became clear that these problems belong to the "NP-complete class", that is, to the class of "hard-to-solve problems" [1]. In other words, there are no polynomial-time methods for finding optimal solutions to these problems. Therefore, methods have been developed to find various types of approximate (suboptimal) solutions to these problems. [2–5, 13, 15–17, 21, 22, 24, 25]. However, although these methods work quickly, the solution they provide may differ significantly from the optimal solution. On the other hand, more general classes of integer programming problems, namely problems with initial data in the form of intervals, have begun to be studied. [6–12, 14, 18 etc]. In addition, various models of the Boolean programming problem, as well as some integer programming problems, have been investigated in [19, 20, 23]. However, there is a need to develop new and more efficient approximate solution methods. Because better approximate solutions to real practical problems must be developed. Such innovative methods have been proposed in [9, 10, 26] for the knapsack problem and the integer knapsack problem. Here, generalization means that the given coefficients are located in certain intervals. However, a new innovative approximate solution method for the more general integer programming problem is implemented in this work. Note that the method proposed in [26] is a special case of the method presented in this paper.

3 MATERIALS AND METHODS

First, let us determine a certain number k that falls within the interval where the coordinates of the optimal solution and the approximate solution differ. For this purpose, let us construct a certain $\tilde{X}=(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_n,)$ solution of the problem as follows, by taking $0\le x_j\le d_j$, $\left(j=\overline{1,n}\right)$ instead of the condition (3). For each number j, $j=\overline{(1,n)}$

DOI 10.15588/1607-3274-2025-3-18

$$\tilde{x}_j = \begin{cases} a_{ij} d_j \le b_i - \sum_{l=1}^{j-1} a_{il}\tilde{x}_l, \forall i, i = \overline{1,m}, \\ \min_i \left(b_i - \sum_{l=1}^{j-1} a_{il}\tilde{x}_l \Big/ a_{ij} \right), \exists i \;\; a_{ij} d_j > b_i - \sum_{l=1}^{j-1} a_{il}\tilde{x}_l \;\; (k := j), \\ 0, \;\; j = k+1,...,n \end{cases}$$

Obviously, this solution will have the following structure.

$$\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_k, ..., \tilde{x}_n) = \left(d_1, d_2, ..., \; d_{k-1}, \frac{\alpha}{\beta}, 0, ..., 0 \right). \qquad (7)$$

Here, should be $0 \le \frac{\alpha}{\beta} < d_k$.

From experiments conducted on numerous random problems of various sizes, it becomes clear that the approximate solution $X^0 = \left(x_1^0, x_2^0, ..., x_n^0\right)$ found by formula (6) and the coordinates of the optimal solution to the problem can differ around a certain $(k-p;\, k+p)$, the number k in the notation (7). (The choice of numbers p and q will be discussed below.) Because, for numbers j in that neighborhood, the $\frac{c_j}{\max\limits_i a_{ij}}, \left(j \in (k-p; k+q)\right)$ ratios are close to each other. Therefore, when constructing a solution using formula (5), the advantage of which coordinate in that interval is selected and evaluated is of great importance. However, since these advantages do not differ significantly, new solutions can be constructed using formula (5) by assigning separate values to the unknowns in that interval. We can accept the one that gives the largest value to the function (1) among the obtained solutions as the final solution. Note that, based on the principle we have shown, a certain approximate solution method for the problem (1)–(3) was given in [7, 10]. These works differ from each other in the choice of the numbers p and q. In those works, $p=q$ was assumed and the procedure for finding them was shown. It is clear that if the number k determined by expression (7) is close to the last number n, or to the first number, then the numbers k-p and k+p may go beyond the interval [1, n]. Therefore, the methods given in works [7–10] may not work. In work [10], the numbers p and q are chosen as a certain percentage of the number k, with $p=q$. Naturally, this method may not work if the number k is close to the ends of the interval [1, n]. Taking all this into account, we have given a new and more universal method in this work for choosing the numbers p and q. Thus, we have estimated the corresponding minimal number of zeros and non-zero coordinates in the optimal solution of the problem (1)–(3) and denoted them by $n(0)$ and $n(d)$. Therefore, there is no need to change the values of the first $n(d)$ and last $n(0)$ coordinates in the solution of (6). Because these coordinates are the same as the coordinates of the optimal solution. Thus, it is important to find the numbers $n(0)$ and $n(d)$, and in this work we have proved certain theorems that allow us to find these numbers.

Thus, for each number j, $\left(j \in [k-p, k+q]\right)$ we can record the possible values of the coordinate x_j and construct the remaining coordinates using formula (5). It is clear that in this case, some coordinate will coincide with the coordinate of the optimal solution.

It should be noted that the numbers p and q can be found, respectively, through the minimal number $n(d)$ of non-zero coordinates and the minimal number $n(0)$ of zeros in the optimal solution of the problem (1)–(3). In other words, must be $p = k - n(d)$, $q = n - n(0) - k$. We will show below the difficulty of finding the numbers $n(d)$ and $n(0)$ used here. Therefore, we will also give the process of evaluating these numbers. Note that such a problem was considered in papers [8, 10, 13, 14]. For example, in paper [13], the neighborhood of $[k-p, k+p]$ is considered and the number p here is found from the relation

$$p = \arg\left\{ \max_i \left(\frac{c_{k-i}}{a_{k-i}} - \frac{c_{k+i}}{a_{k+i}} \right) \le \delta \right\}.$$

The number δ used here is a positive integer and must be given in advance. If the number k found by formula (7) is close to the last number n or the first number, then specifying the interval $[k-p, k+p]$ becomes uncertain. Instead of the $[k+p, k-p]$ neighborhood of the k-th coordinate, the interval $\left[\delta_k^1, \delta_k^0\right]$ was determined in works [8, 14].

Here, the numbers δ_k^1 and δ_k^0 are found from the relation $\delta_k^1 = \left[k \cdot \frac{q}{100}\right]$, $\delta_k^0 = \left[(n-k) \cdot \frac{q}{100}\right]$, and the number q is the minimum number of units or zeros in the optimal solution. However, in this case, uncertainties may also arise. Because, when the minimum number of units or zeros is chosen for the number q, the interval $\left[\delta_k^1, \delta_k^0\right]$ may be different.

Note that in Work [10], the interval integer bag problem was reduced to the corresponding known integer bag problem, and a certain neighborhood of the k-th coordinate taking a fractional value in the corresponding continuous problem was selected.

However, in this work, the interval $[k-p, k+p] = [n(d), n-n(0)]$ is adopted as the circumference of the k-th coordinate. It is clear that the condition $n(d) \le k \le n - n(0)$ will be satisfied for the k-th number. As can be seen, in order to obtain a better solution than the initial approximate solution by performing a small number of calculation operations, we

DOI 10.15588/1607-3274-2025-3-18

need to find the numbers $n(d)$ and $n(0)$ for the problem under consideration. In this case, in the process of constructing an innovative approximate solution, we need to keep the first $n(d)$ and last $n(0)$ coordinates in the solution (7) as they are.

Thus, it is necessary to solve the following problems.

$$\sum_{j=1}^{n} x_j \to \min, \tag{8}$$

$$\sum_{j=1}^{n} a_{ij} x_j \le b_i, \ (i=\overline{1,m}), \tag{9}$$

$$\sum_{j=1}^{n} c_j x_j \ge f^0, \tag{10}$$

$$0 \le x_j \le d_j, \quad j=\overline{1,n}, \tag{11}$$

$$x_j - \text{int}\,eger, \ \left(j=\overline{1,n}\right) \tag{12}$$

and

$$\sum_{j=1}^{n} x_j \to \max, \tag{13}$$

$$\sum_{j=1}^{n} a_{ij} x_j \le b_i, \ (i=\overline{1,m}), \tag{14}$$

$$\sum_{j=1}^{n} c_j x_j \ge f^0, \tag{15}$$

$$0 \le x_j \le d_j, \quad j=\overline{1,n}, \tag{16}$$

$$x_j - \text{int}\,eger \quad \left(j=\overline{1,n}\right). \tag{17}$$

We can solve problem (18)–(12) and find its optimal solution $\underline{X}^* = \left(\underline{x}_1^*, \underline{x}_2^*, ..., \underline{x}_n^*\right)$. Then the set $\omega_1 = \left\{j \mid \underline{x}_j^* > 0\right\}$ is easily determined. Then the minimal number $n(d)$ of non-zero coordinates in the optimal solution of problem (1)–(3) is $n(d) = n(\omega_1)$. Here, $n(\omega_1)$ is the number of elements of the set ω_1.

Since the problem (8)–(12) shown above belongs to the class of hard-to-solve problems, it may not be possible to solve it in real time. To alleviate this problem, it is necessary to solve the problem (8)–(11) which has a larger domain. By solving the obtained linear programming problem, we can find its optimal solution $\underline{X}^{lp} = \left(\underline{x}_1^{lp}, \underline{x}_2^{lp}, ..., \underline{x}_n^{lp}\right)$. For the $\underline{n}(d)$ number of non-zero coordinates in this solution, the $\underline{n}(d) = n(\omega_2) \le n(d)$ relation is satisfied. Here $\omega_2 = \left\{j \mid \underline{x}_j^{lp} > 0\right\}$. Because, the number $\underline{n}(d)$ is taken from the problem with a larger domain.

Note that to find the number $\underline{n}(d)$, the optimal solution of the linear programming problem is used. Naturally, when solving such large-scale problems, time and memory problems may still arise. To overcome this problem, we can solve the problem

$$\sum_{j=1}^{n} x_j \to \min,$$

$$\sum_{j=1}^{n} c_j x_j \ge f^0,$$

$$0 \le x_j \le d_j, \quad j=\left(\overline{1,n}\right).$$

which has a larger possible solution region, instead of the problem (8)–(12).

In this problem, assuming that the conditions $c_1 \ge c_2 \ge ... \ge c_{n-1} \le c_n$ are satisfied, then we can find the minimal number n_1 of nonzero coordinates in its optimal solution from the relation

$$\sum_{j=1}^{n_1} c_j d_j \le f^0 \le \sum_{j=1}^{n_1+1} c_j d_j.$$

Naturally, it will be $n_1 \le \underline{n}(d) \le n(d)$.

Therefore, the obtained number n_1 can be a lower bound on the minimal number of non-zero coordinates in the optimal solution of problem (1)–(3).

Note that to find the minimal number of zeros in the optimal solution of problem (1)–(3), we must first solve problem (13)–(17) or (13)–(16), respectively, and finally solve problem

$$\sum_{j=1}^{n} x_j \to \max, \tag{18}$$

$$\sum_{j=1}^{n} a_{ij} x_j \le b_i, \ (i=\overline{1,m}), \tag{19}$$

$$0 \le x_j \le d_j, \quad j=\left(\overline{1,n}\right), \tag{20}$$

$$x_j - l \quad is\ \text{int}\,eger. \tag{21}$$

In this case, we find the maximum number of non-zero coordinates in the optimal solution of problem (1)–(3). However, since we need to find the minimum number of zeros in the optimal solution, we need to subtract the maximum number of non-zero coordinates from the number of unknowns *n*. For this purpose, let us assume that the conditions $a_{i1} \le a_{i2} \le ... \le a_{in}$ are satisfied separately for each *i*-th inequality in the system (19).

Then, in problem (1)–(3), we can find the maximum number of nonzero coordinates $\bar{n}$ of the optimal solution based on the conditions

DOI 10.15588/1607-3274-2025-3-18

$$\sum_{j=1}^{n_i} a_{ij} d_j \le b_i \le \sum_{j=1}^{n_i+1} a_{ij} d_j, \ (i=\overline{1,m})$$

as $\bar{n} = \min_i n_i$.

Then, we can accept the number $\underline{n}(0) = n - \bar{n} \le n(0)$ as the minimal number $\underline{n}(0)$ of zeros in the optimal solution of that problem.

So the following theorem has been proven.

Theorem: If the numbers $n(d)$ and $n(0)$ denote the minimum numbers of ones and zeros, respectively, in the optimal solution of problem (1)–(3), $\underline{n}(d)$ and $\underline{n}(0)$ denotes their corresponding lower bounds, then the relations $\underline{n}(d) \le n(d)$ and $\underline{n}(0) \le n(0)$ are true.

As a result, we can conclude that the coordinates of the optimal solution to problem (1)–(3) and the coordinates of the solution to problem (7) can only take different values in the interval $[\underline{n}(d), \underline{n}(0)]$.

Note that in the solution (7) there are zeros to the right of the k-th coordinate, and d_j numbers to the left. Therefore, in order to construct new solutions using (5), we need to write $x_j = 1,2,...,d_j$ for each j=k+1, k+2,... $\underline{n}(0)$ and $x_j = d_{j-1}, d_{j-2},...,0$ for each $j = \underline{n}(d), \underline{n}(d)+1,...,k$ number. We select the best one from the obtained solutions and call it the innovative approximate solution.

Thus, we can write the algorithm for the process of constructing the innovative approximate solution that we propose.

A L G O R I T H M

Step 1. Must give the numbers $n, c_j, a_{ij}, b_i, d_j \ (i=\overline{1,m}, \ j=\overline{1,n})$ and assimilate $bb_i := b_i; \ (i=\overline{1,m})$.

Step 2. In problem (1)–(3), without considering the condition that the unknowns are complete, let us find its solution

$$\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_k, ..., \tilde{x}_n) = \left(d_1, d_2, ..., d_{k-1}, \frac{\alpha}{\beta}, 0, ..., 0\right)$$

using the following well-known formula. For each number $j, (j=1,2,...,n)$

$$\tilde{x}_j = \begin{cases} for\ d_j, if\ \forall i, i=\overline{1,m}\ , \ a_{ij} d_j \le b_i - \sum_{l=1}^{j-1} a_{il} \tilde{x}_l \\ \min_i \left(b_i - \sum_{i=1}^{j-1} a_{il} \tilde{x}_l \Big/ a_{ij} \right), if\ \exists i\ , a_{ij} d_j > b_i - \sum_{l=1}^{j-1} a_{il} \tilde{x}_l \ (k := j), \\ 0, \ j = k+1, ..., n. \end{cases}$$

and accept $kk := k$, $r := 0$;

Step 3. If the obtained $\tilde{x}_k$ coordinate is an integer, then the solution $\tilde{X}$ coincides with the optimal $X^* = (x_1^*, x_2^*, ..., x_n^*)$ solution of problem (1)–(3). Then, $f^* := \sum_{j=1}^{n} c_j \tilde{x}_j$, must accept $X^* = (x_1^*, x_2^*, ..., x_n^*) = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_k, ..., \tilde{x}_n)$, print and go to step 21.

Step 4. Let's find the approximate solution of the problem $X^0 = (x_1^0, x_2^0, ..., x_n^0)$ using the following formula

$$x_j^0 = \begin{cases} d_j, \ \forall i, i = (\overline{1,m}), \ a_{ij} d_j \le b_i - \sum_{l=1}^{j-1} a_{il} x_l^0 \\ \min_i \left[\left(b_i - \sum_{l=1}^{j-1} a_{il} x_l^0 \right) / a_{ij} \right], \text{otherwise} \end{cases}$$

for each number $(j = 1,2,...,n)$.

Step 5. Calculate the numbers $\tilde{f}$ and f^0 as follows.

$$\tilde{f} = \sum_{j=1}^{n} c_j \tilde{x}_j, \quad f^0 = \sum_{j=1}^{n} c_j x_j^0.$$

Accept $f^{it} := f^0$ $X^{it} = (x_1^0, x_2^0, ..., x_n^0)$ and remembered the solution $X^{it} = (x_1^0, x_2^0, ..., x_n^0)$ with f^{it}.

Step 6. To find the minimum number n_1 of non-zero coordinates, let's convert the $c_j, \left(j=\overline{1,n}\right)$ numbers to $c_1 \ge c_2 \ge ... \ge c_n$ form and use the following relation:

$$\sum_{j=1}^{n_1} c_j d_j \le f^0 \le \sum_{j=1}^{n_1+1} c_j d_j.$$

Step 7. To find the $n_i, (i=\overline{1,m})$ numbers, we need to arrange the coefficients of $a_{ij}, (i=\overline{1,m}, \ j=\overline{1,n})$ for each number $i, (i=\overline{1.m})$, separately in increasing order, like $a_{i1} \le a_{i2} \le ... \le a_{in}$. Then, we need to use the following relations:

$$\sum_{j=1}^{n_i} a_{ij} d_j \le b_i \le \sum_{j=1}^{n_i+1} a_{ij} d_j, \ (i=\overline{1,m}).$$

Finally, should be write and remember $\bar{n} = \min_i n_i$ and $\underline{n}(0) = n - \bar{n}$.

Step 8. Should be accept the following prices

$$x_k^t := [\tilde{x}_k]; \ b_i := bb_i - a_{ik} x_k^t, \ (i=\overline{1.m}).$$

DOI 10.15588/1607-3274-2025-3-18

Step 9. For each numbers j, $j=1,2,...,n$ and $j\neq k$, let's find the approximate solution of the interval $X^t=\left(x_1^t,x_2^t,...,x_n^t\right)$ using the following formula.

$$x_j^t=\begin{cases} for\ d_j,\ if\ \forall i,\ i=\overline{(1,m)}\ \ a_{ij}d_j\le b_i-\sum\limits_{l=1}^{j-1}a_{il}x_l^t \\ \min\limits_i\left[\left(b_i-\sum\limits_{l=1}^{j-1}a_{il}x_l^t\right)/a_{ij}\right],\ \ else. \end{cases}$$

Step 10. Let's calculate the number f^t

$$f^t:=\sum_{j=1}^{n}c_jx_j^t.$$

If $f^t>f^{it}$, then should be accept $f^{it}:=f^t$, $X^{it}=(x_1^t,x_2^t,...,x_n^t)$.

Step 11. If r=0, then should be accept $x_k^t:=\left\lfloor x_k^t\right\rfloor+1;\ b_i:=bb_i-a_{ik}x_k^y,\ \ (i=\overline{1,m});\ r:=1$ and go to step 9.

Step 12. Should be accept $r:=0$

Step 13. To find the intermediate solution $X^t=\left(x_1^t,x_2^t,...,x_n^t\right)$, let's use the following rule:

For each $j,\ j=(1,2,...,n;\quad j\neq k)$

$$x_j^t=\begin{cases} for\ d_j,\ if\ \ \forall i,\ i=\overline{(1,m)},\ \ a_{ij}d_j\le b_i-\sum\limits_{l=1}^{j-1}a_{il}x_l^t \\ \min\limits_i\left[\left(b_i-\sum\limits_{l=1}^{j-1}a_{il}x_l^t\right)/a_{ij}\right],\ \ else. \end{cases}$$

Then, the corresponding value of the functional is

$$f^t:=\sum_{j=1}^{n}c_jx_j^t.$$

Step 14. If $f^t>f^{it}$, $f^{it}:=f^t$, $X^{it}=(x_1^t,x_2^t,...,x_n^t)$ should be written and memorized. If $r:=1$ skip to Step 18.

Step 15. If $x_k^t<d_k$, then skip to Step 17.

Step 16. Should accept $k:=k+1$; If $k>n(0)$, then accept $k:=kk$ and skip to the Step 18.

Step 17. $x_k^t:=1,2,...,d_k$ and accordingly by taking $b_i:=bb_i-a_{ik}x_k^t\ \ (i=\overline{1.m})$ skip to the Step 13.

Step 18. $r:=1$; If $x_k^t=d_k$, then $k:=k-1$; If $k<n(1)$ skip to the Step 20.

Step 19. $x_k^t:=1,2,...,d_k$ k values corresponding to $b_i:=bb_i-a_{ik}x_k^t\ \ (i=\overline{1.m})$ and go to Step 13.

Step 20. Should be print f^{it}, $X^{it}=(x_1^{it},x_2^{it},...,x_n^{it})$ $\delta=\left(\left[\tilde{f}\right]-f^t\right)/\left[\tilde{f}\right]$ and $\delta^i=\left(\left[\tilde{f}\right]-f^{it}\right)/\left[\tilde{f}\right]$.

Step 21. STOP.

4 EXPERIMENTS

Let's solve a numerical example using the algorithm we wrote above. Here, we assume that the following conditions are met.

$$\frac{c_1}{\max\limits_i a_{i1}}\ge\frac{c_2}{\max\limits_i a_{i2}}\ge...\ge\frac{c_k}{\max\limits_i a_{ik}}\ge...\ge\frac{c_n}{\max\limits_i a_{in}},$$

$$9x_1+10x_2+8x_3+6x_4+7x_5\to\max,$$
$$3x_1+3x_2+1x_3+4x_4+2x_5\le18,$$
$$1x_1+5x_2+4x_3+2x_4+3x_5\le17,$$
$$4x_1+2x_2+3x_3+2x_4+5x_5\le20,$$
$$0\le x_j\le d_j,\quad \text{are integers},\ \ (j=\overline{1,5}).$$

Here $d_1=2,\ d_2=2,\ d_3=3,\ d_4=4,\ d_5=3$. İn order words we accepted $0\le x_1\le2,\ 0\le x_2\le2,\ 0\le x_3\le3,\ 0\le x_4\le4,\ 0\le x_5\le3$. In this problem, let's first construct the initial $\tilde{X}=\left(\tilde{x}_1,\tilde{x}_2,...,\tilde{x}_k,...,\tilde{x}_n\right)=\left(d_1,d_2,...,d_{k-1},\frac{\alpha}{\beta},0,...,0\right)$ solution using the above algorithm, without considering the completeness condition on the unknowns. Then we can get $\tilde{X}=(2,2,5/4,0,0)$ and $\tilde{f}=48$. From here we get $\tilde{x}_k=\tilde{x}_3=5/4$ and $k=3$. Then we need to take $kk:=k$, $r:=0$ according to the algorithm.

Now let's construct an initial approximate solution $X^0=\left(x_1^0,x_2^0,...,x_n^0\right)$ to this problem. Then we can get $X^0=(2,2,1,0,0)$ and $f^0=46$. Then, let's write the results in the following tables.

5 RESULTS

Since $k=3$, let's write the approximate solutions and f^t values found sequentially, noting $x_3^t=3,\ x_3^t=2,\ x_3^t=1,\ x_3^t=0$, in Table 1.

Table 1 – Evaluating the coordinate x_3^t

X^t	x_1^t	x_2^t	x_3^t	x_4^t	x_5^t	f^t
X_1^t	2	0	**3**	1	0	f_2^t =48
X_2^t	2	1	**2**	0	0	f_3^t =44
X_3^t	2	2	**1**	0	0	f_4^t =46
X_4^t	2	2	**0**	1	1	f_5^t =51

Now, let's write the successive approximate solutions and f^t values in Table 2 below, assuming $k=k-1=2$ and writing $x_2^t=2,\ x_2^t=1,\ x_2^t=0$ accordingly.

DOI 10.15588/1607-3274-2025-3-18

Table 2 – Evaluating the coordinate x_2^t

X^t	x_1^t	x_2^t	x_3^t	x_4^t	x_5^t	f^t
X_5^t	2	**2**	1	0	0	f_6^t =44
X_6^t	2	**1**	2	1	0	f_7^t =50
X_7^t	2	**0**	2	2	0	f_8^t =46

Let us write the approximate solutions and f^t values obtained by giving successive values of $x_1^t = 2,\ x_1^t = 1,\ x_1^t = 0$ by assuming $k = 1$ in Table 3.

Table 3 – Evaluating the coordinate x_1^t

X^t	x_1^t	x_2^t	x_3^t	x_4^t	x_5^t	f^t
X_8^t	**2**	2	1	0	0	f_9^t =46
X_9^t	**1**	2	1	1	0	f_{10}^t =43
X_{10}^t	**0**	2	1	1	0	f_{11}^t =34

Then, let's continue the solution process for the unknowns x_4 and x_5 located to the right of the k-th coordinate. First, let's find the appropriate approximate solutions and f^t values by noting $x_4^t = 4,\ x_4^t = 3,\ x_4^t = 2,\ x_4^t = 1,\ x_4^t = 0$ and write them in Table 4.

Table 4 – Evaluating the coordinate x_4^t

X^t	x_1^t	x_2^t	x_3^t	x_4^t	x_5^t	f^t
X_{11}^t	0	0	2	**4**	0	f_{11}^t =40
X_{12}^t	2	0	0	**3**	0	f_{12}^t =36
X_{13}^t	2	1	1	**2**	0	f_{13}^t =48
X_{14}^t	2	2	0	**1**	1	f_{14}^t =51
X_{15}^t	2	2	1	**0**	0	f_{15}^t =46

Finally, let's find the appropriate approximate solutions and f^t values by noting $x_5^t = 3,\ x_5^t = 2,\ x_5^t = 1$ and $x_5^t = 0$ and write them in Table 5.

Table 5 – Evaluating the coordinate x_5^t

X^t	x_1^t	x_2^t	x_3^t	x_4^t	x_5^t	f^t
X_{16}^t	1	0	0	0	**3**	f_{16}^t =30
X_{17}^t	2	1	0	0	**2**	f_{17}^t =32
X_{18}^t	2	2	0	1	**1**	f_{18}^t =51
X_{19}^t	2	2	1	0	**0**	f_{19}^t =46

Note that initially, the approximate solution found by the known method was obtained as $X^0 = (2,2,1,0,0)$ and $f^0 = 46$. However, as can be seen from the tables above, the value of $f^0 = 46$ increased to 48, 50 and 51.

Therefore, by choosing the one corresponding to the largest of these values, the innovative approximate solution will be $X^{it} = \left(x_1^{it}, x_2^{it}, x_3^{it}, x_4^{it}, x_5^{it}\right) = (2,2,0,1,1)$ and the innovative approximate value $f^{it} = 51$.

Note that this solution is also the optimal solution to the problem under consideration.

6 DISCUSSION

The main essence and novelty of the method proposed in this article is that a solution provided by any of the known methods is initially accepted as a starting point. Then, this solution is gradually improved**.** For this purpose, the numbers of coordinates at which the approximate solution accepted as the initial solution and the optimal solution can differ are determined. Finding these numbers is based on proven criteria. It is clear that if we know which of the coordinates of the approximate solution are not the same as the coordinates of the optimal solution and we give these coordinates one by one values in their variation interval, at a certain step those coordinates will coincide with the coordinates of the optimal solution. So a better solution can be obtained than the initial approximate solution. Note that we can find the optimal solution by giving values to all of these coordinates at the same time. However, in this case we will have to look at an exponential amount of coordinates. This would require unrealistic computer time. Therefore, we will get a new, better solution by changing these coordinates one by one. Therefore, the solution obtained through the application of this method will be better than the one provided by known approximate methods. The principle of improving the initial solution is rigorously mathematically justified through proven theorems. A specific problem was solved to clearly demonstrate the sequence of implementation of the proposed method, as well as to determine the quality of that method.

CONCLUSIONS

In the article, a new approximate solution method for the integer programming problem has been developed. In all known approximate solution methods, the number of the unknown is found based on certain criteria and the unknown is evaluated. After that, the unknown is removed from the list. As a result, a certain, approximate solution is obtained. In most cases, this solution differs significantly from the optimal solution. Therefore, it is necessary to accept any approximate solution as an initial solution and improve it further.

The scientific novelty of this article is that the numbers of coordinates at which the initially taken approximate solution may differ from the optimal solution are determined using proven theorems. Then, new solutions are constructed by assigning values to these coordinates in their variation interval.

DOI 10.15588/1607-3274-2025-3-18

The best of these solutions is accepted as an innovative approximate solution. From this it is immediately clear that the proposed method should be more effective. The mathematical model considered in the article arises in the problems of optimizing the production of products manufactured by number. Therefore, the method proposed in the article can be effectively applied to solving real practical problems of this type. This shows the practical value of the article. It is clear that if the initial approximate solution differs little from the optimal solution, then the method proposed in this article will provide few improvements. Therefore, in the future, it is planned to develop a new approximate solution method so that the absolute or relative error of the solution it provides is not greater than the optimal solution. It should be noted that in order to clarify the essence of the proposed method in the article, a specific issue was solved.In that problem, the initial value of the functional found by the known method was 46. In the subsequent steps, this value was 48, 50 and 51. More precisely, the initial solution was improved 3 times.

ACKNOWLEDGEMENTS

This article, i.e. “An Innovative approximate solution method for the integer programming problem” was carried out at the Institute of Control Systems of the Ministry of Science and Education of the Republic of Azerbaijan at the expense of the state budget (State Registration No. 0101 Az 00736).

REFERENCES

1. Garey M. R., Jhonson D. S. Computers and Intractability : a Guide to the Theory of NP-Completeness. San Francisco, Freeman, 1979, P. 314.
2. Martello S., Toth P. Knapsack problems: Algorithm and Computers İmplementations. New York, John Wiley & Sons, 1990, P. 296.
3. Erlebach T., Kellerer H., Pferschy U. Approximating multi-objective knapsack problems, *Management Science*, 2002, № 48, pp. 1603–1612. DOI: 10.1287/mnsc.48.12.1603.445
4. Kellerer H., Pferschy U., Pisinger D. Knapsack problems. Berlin, Heidelberg, New-york, Springer-Verlag, 2004, P.546 DOI:10.1007/978-3-540-24777-7
5. Vazirani V. V. Approximation algorithms. Berlin, Springer, 2001, P. 378. DOI:10.1057/palgrave.jors.2601377
6. Libura M. Integer programming problems with inexact objective function, *Control Cyber*, 1980, Vol. 9, No. 4. pp. 189–202.
7. Bukhtoyarov S. E., Emelichev V. A.Stability aspects of Multicriteria integer linear programming problem, Journal of Applied and Industrial mathematics, 2019, Vol. (13), №1, pp. 1–10. DOI:10.33048/daio.2019.26.624
8. Mamedov K. Sh., Mammadli N. O. Two methods for construction of suboptimistic and subpessimistic solutions of the interval problem of mixed-Boolean programming, *Radio Electronics, Computer Science, Control*, 2018, № 3 (46), pp. 57–67.
9. Niyazova R. R., Huseynov S. Y. An İnnovative İmproved Approximate Method for the knapsack Problem with coefficients Given in the Interval Form, *8-th International Conference on Control and Optimization with Industrial Applications*, Baku, 24–26 August, 2022, Vol. II, pp. 210–212.
10. Mammadov K. Sh., Niyazova R. R., Huseynov S. Y. Innovative approximate method for solving Knapsack problems with interval coefficients, *International Independent scientific journal*, 2022, №44, pp. 8–12. doi.org/10.5281/zenodo.7311206. DOI: 10.15588/1607-3274-2018-3-7.
11. Emelichev V. A., Podkopaev D. Quantitative stability analysis for vector problems of 0–1 programming, *Discrete Optimization*, 2010, Vol. 7, pp. 48–63. DOI: 10.1016/j.disopt.2010.02.001.
12. Li W., Liu X., Li H. Generalized solutions to interval linear programmers and related necessary and sufficient optimality conditions, *Optimization Methods Software*, 2015, Vol. 30, №3, pp. 516–530. DOI: 10.1080/10556788.2014.940948
13. Mamedov K. Sh., Huseinov S. Y. Method of Constructing Suboptimal Solutions of Integer Programming Problems and Successive Improvement of these Solutions, *Automatic Control and Computer Science*, 2007, Vol. 41, № 6, pp. 20–31. DOI: 10.3103/S014641160706003X
14. Mamedov K. Sh., Mamedova A. H. Ponyatie suboptimisticheskoqo i subpestimisticheskoqo resheniy i postroeniya ix v intervalnoy zadache Bulevoqo programmirovania, *Radio Electronics Computer Science Control*, 2016, No. 3, pp. 99–108. DOİ: 10.15588/1607-3274-2016-3-13
15. Hifi M., Sadfi S., Sbihi A.. An efficient algorithm for the knapsack sharing problem, *Computational Optimization and Applications*, 2002, № 23, pp. 27–45. DOI:10.1023/A:1019920507008
16. Hifi M., Sadfi S. The knapsack sharing problem: An exact algorithm, *Journal of Combinatorial Optimization*, 2002, № 6, pp. 35–54. DOI:10.1023/A:1013385216761
17. Hladik M. On strong optimality of interval linear programming, *Optimization Letters*, 2017, No. 11 (7), pp. 1459–1468. DOI:10.1007/s11590-016-1088-3
18. Devyaterikova M. V., Kolokolov A. A., Kolosov A. P. L-class enumeration algorithms for one discrete production planning problem with interval input data, *Computers and Operations Research*, 2009, Vol. 36, №2, pp. 316–324. DOI:10.1016/j.cor.2007.10.005
19. Babayev D. A., Mardanov S. S. Reducing the number of variables in integer and linear programming problems, *Computational Optimization and Applications*, 1994, № 3, pp. 99–109. DOI: https://doi.org/10.1007/BF01300969.
20. Basso A., Viscolani B. Linear programming selection of internal financial laws and a knapsack problem, *Calcolo*, 2000, Vol. 37, № 1, pp. 47–57. DOI:10.1007/s100920050003
21. Bertisimas D., Demir R. An approximate dynamic programming approach to multidimensional knapsack problems, *Management Science*, 2002, № 48, pp. 550–565. DOI:10.1287/mnsc.48.4.550.208

DOI 10.15588/1607-3274-2025-3-18

22. Billionnet A. Approximation algorithms for fractional knapsack problems, *Operation Research Letters*, 2002, № 30, pp. 336–342. DOI:10.1016/S0167-6377(02)00157-8
23. Broughan K., Zhu Nan An integer programming problem with a linear programming solution, *Journal American Mathematical Monthly*, 2000, Vol. 107, № 5, pp. 444–446. DOI:10.1080/00029890.2000.12005218
24. Calvin J. M., Leung Y. T. Average-case analysis of a greedy algorithm for the 0–1 knapsack problem, *Operation Research Letters*, 2003, № 31, pp. 202–210. DOI:10.1016/S0167-6377(02)00222-5
25. Mamedov K. Sh., Musaeva T. M. Metodi postroeniya priblijennix resheniy mnoqomernoy zadachi o rance i naxojdenie verxney ocenki optimuma, *Avtomatika i Vichislitelnaya Texnika*, 2004, № 5, pp. 72–82.
26. Mamedov K. Sh., Niyazova R. R. Innovative Improved Approximate Solution Method For the Integer Knapsack Problem, Error Compression and Computational Experiments, *Radio Electronics Computer Science Control*, 2024, № 4, pp. 64–74. DOI: 10.15588/1607-3274-2024-4-6

Received 20.05.2025.
Accepted 11.07.2025.

УДК 519.852.6

ІННОВАЦІЙНИЙ МЕТОД НАБЛИЖЕНОГО РОЗВ'ЯЗАННЯ ЗАДАЧІ ЦІЛОЧИСЛОВОГО ПРОГРАМУВАННЯ

Мамедов К. Ш. – д-р фіз.-мат. наук, професор Бакинського державного університету та завідувач відділу Інституту систем управління Міністерства науки і освіти.
Ніязова Р. Р. – докторант, науковий співробітник Інституту систем управління Міністерства освіти і науки.

АНОТАЦІЯ

Актуальність. Існують певні методи знаходження оптимального розв'язку задач цілочисельного програмування. Однак ці методи не можуть вирішувати масштабні задачі в режимі реального часу. Тому було запропоновано наближені розв'язки цих задач, які працюють швидко. Слід зазначити, що розв'язки, отримані цими методами, часто суттєво відрізняються від оптимального розв'язку. Тому виникає проблема прийняття будь-якого відомого наближеного розв'язку як початкового розв'язку та його подальшого вдосконалення.

Мета роботи Спочатку знаходиться певний наближений розв'язок. Потім, на основі доведених теорем, визначаються координати цього розв'язку, які не збігаються з оптимальним. Після цього, послідовно змінюючи ці координати, знаходять нові розв'язки. За остаточний розв'язок приймається той, який дає найбільше значення функціоналу серед цих розв'язків.

Метод. Метод, який ми пропонуємо в цій роботі, реалізується наступним чином:

Спочатку встановлюється певний наближений розв'язок задачі, потім визначаються номери координат цього розв'язку, які не збігаються з оптимальним розв'язком. Після цього встановлюються нові розв'язки шляхом послідовного присвоєння значень цим координатам по одному в їхніх інтервалах. Найкраще з розв'язків, знайдених у цьому процесі, приймається як остаточне інноваційне рішення.

Результати. Було вирішено задачу з метою візуальної ілюстрації якості та ефективності запропонованого методу.

Висновки. Метод, який ми пропонуємо в цій статті, не може дати гірших результатів, ніж будь-який метод наближеного рішення, простий з алгоритмічної точки зору, є новим, його можна легко програмувати та важливий для вирішення реальних практичних завдань.

КЛЮЧОВІ СЛОВА: задача цілочисельного програмування, вихідний наближений розв'язок, інтервал, в якому координати наближеного розв'язку можуть відрізнятися від оптимального розв'язку, інноваційний наближений-розв'язок, обчислювальні експерименти.

ЛІТЕРАТУРА

1. Garey M. R. Computers and Intractability : a Guide to the Theory of NP-Completeness / M. R. Garey, D. S. Jhonson. – San Francisco, Freeman, 1979. – P. 314.
2. Martello S. Knapsack problems: Algorithm and Computers İmplementations. / S. Martello, P. Toth. – New York : John Wiley & Sons, 1990. – P. 296.
3. Erlebach T. Approximating multi-objective knapsack problems / T. Erlebach, H. Kellerer, U. Pferschy // Management Science. – 2002. – № 48. – P. 1603–1612. DOI: 10.1287/mnsc.48.12.1603.445
4. Kellerer H. Knapsack problems / H. Kellerer, U. Pferschy, D. Pisinger. – Berlin, Heidelberg, New-york : Springer-Verlag, 2004. – P. 546 DOI:10.1007/978-3-540-24777-7
5. Vazirani V. V. Approximation algorithms / V. V. Vazirani. – Berlin : Springer, 2001. – P. 378. DOI:10.1057/palgrave.jors.2601377
6. Libura M. Integer programming problems with inexact objective function / M. Libura // Control Cyber. – 1980. –Vol. 9, No. 4. – P. 189–202.
7. Bukhtoyarov S. E. Stability aspects of Multicriteria integer linear programming problem / S. E. Bukhtoyarov, V. A. Emelichev // Journal of Applied and Industrial mathematics. – 2019. – Vol. (13), № 1. – P. 1–10. DOI:10.33048/daio.2019.26.624

DOI 10.15588/1607-3274-2025-3-18

8. Mamedov K. Sh. Two methods for construction of suboptimistic and subpessimistic solutions of the interval problem of mixed-Boolean programming / K. Sh. Mamedov, N. O. Mammadli // Radio Electronics, Computer Science, Control. – 2018. – № 3 (46). – P. 57–67.
9. Niyazova R. R. An İnnovative İmproved Approximate Method for the knapsack Problem with coefficients Given in the Interval Form / R. R. Niyazova, S. Y. Huseynov // 8-th International Conference on Control and Optimization with Industrial Applications, Baku : 24 – 26 August, 2022. Vol II. – P. 210–212.
10. Mammadov K. Sh. Innovative approximate method for solving Knapsack problems with interval coefficients. / K. Sh. Mammadov, R. R. Niyazova, S. Y. Huseynov // International Independent scientific journal. – 2022. – № 44. – P. 8–12. doi.org/10.5281/zenodo.7311206. DOI: 10.15588/1607-3274-2018-3-7.
11. Emelichev V. A. Quantitative stability analysis for vector problems of 0–1 programming // V. A. Emelichev, D. Podkopaev // Discrete Optimization. –2010. – Vol. 7. – P. 48–63. DOİ: 10.1016/j.disopt.2010.02.001.
12. Li W. Generalized solutions to interval linear programmers and related necessary and sufficient optimality conditions / W. Li, X. Liu, H. Li // Optimization Methods Software. –2015. – Vol. 30, № 3. – P. 516–530. DOI: 10.1080/10556788.2014.940948
13. Mamedov K. Sh. Method of Constructing Suboptimal Solutions of Integer Programming Problems and Successive Improvement of these Solutions / K. Sh. Mamedov, S. Y. Huseinov // Automatic Control and Computer Science. – 2007. – Vol. 41, № 6. – P. 20–31. DOI: 10.3103/S014641160706003X
14. Mamedov K. Sh. Ponyatie suboptimisticheskoqo i subpestimisticheskoqo resheniy i postroeniya ix v intervalnoy zadache Bulevoqo programmirovania / K. Sh. Mamedov, A. H. Mamedova // Radio Electronics, Computer Science, Control. – 2016. – No. 3. – P. 99–108. DOİ: 10.15588/1607-3274-2016-3-13
15. Hifi M. An efficient algorithm for the knapsack sharing problem / M. Hifi, S. Sadfi, A. Sbihi // Computational Optimization and Applications. – 2002. – № 23. – P. 27–45. DOI:10.1023/A:1019920507008
16. Hifi M. The knapsack sharing problem: An exact algorithm / M. Hifi, S. Sadfi // Journal of Combinatorial Optimization. – 2002. – № 6. – P. 35–54. DOI:10.1023/A:1013385216761
17. Hladik M. On strong optimality of interval linear programming / M. Hladik // Optimization Letters. – 2017. – No. 11(7). – P. 1459–1468. DOI:10.1007/s11590-016-1088-3
18. Devyaterikova M. V. L-class enumeration algorithms for one discrete production planning problem with interval input data / M. V. Devyaterikova, A. A. Kolokolov, A. P. Kolosov // Computers and Operations Research. – 2009. – Vol. 36, № 2. – P. 316–324. DOI:10.1016/j.cor.2007.10.005
19. Babayev D. A. Reducing the number of variables in integer and linear programming problems / D. A. Babayev, S. S. Mardanov // Computational Optimization and Applications. – 1994. – № 3. – P. 99–109. DOI: https://doi.org/10.1007/BF01300969.
20. Basso A. Linear programming selection of internal financial laws and a knapsack problem / A. Basso, B. Viscolani // Calcolo. – 2000. – Vol. 37, № 1. – P. 47–57. DOI:10.1007/s100920050003
21. Bertisimas D. An approximate dynamic programming approach to multidimensional knapsack problems / D. Bertisimas, R. Demir // Management Science. – 2002. – № 48. – P. 550–565. DOI:10.1287/mnsc.48.4.550.208
22. Billionnet A. Approximation algorithms for fractional knapsack problems / A. Billionnet // Operation Research Letters. – 2002. – № 30. – P. 336–342. DOI:10.1016/S0167-6377(02)00157-8
23. Broughan K. An integer programming problem with a linear programming solution./ K. Broughan, Nan Zhu // Journal American Mathematical Monthly. – 2000. – Vol. 107, № 5. – P. 444–446. DOI:10.1080/00029890.2000.12005218
24. Calvin J. M. Average-case analysis of a greedy algorithm for the 0–1 knapsack problem / J. M. Calvin, Y. T. Leung // Operation Research Letters. – 2003. – № 31. – P. 202–210. DOI:10.1016/S0167-6377(02)00222-5
25. Mamedov K. Sh. Metodi postroeniya priblijennix resheniy mnoqomernoy zadachi o rance i naxojdenie verxney ocenki optimuma. / K. Sh. Mamedov, T. M. Musaeva // Avtomatika i Vichislitelnaya Texnika. –2004. – № 5. – P. 72–82.
26. Mamedov K. Sh. Innovative Improved Approximate Solution Method For the Integer Knapsack Problem, Error Compression and Computational Experiments / K. Sh. Mamedov, R. R. Niyazova // Radio Electronics Computer Science Control. – 2024. – № 4. – P. 64–74. DOI: 10.15588/1607-3274-2024-4-6.

DOI 10.15588/1607-3274-2025-3-18

Наукове видання

Радіоелектроніка,

інформатика,

управління

№ 3/2025

Науковий журнал

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 27.08.2025. Формат 60×84/8.
Папір офс. Різогр. друк. Ум. друк. арк. 23,94.
Тираж 300 прим. Зам. № 758.

69063, м. Запоріжжя, НУ «Запорізька політехніка», друкарня, вул. Жуковського, 64

Свідоцтво суб’єкта видавничої справи
ДК № 6952 від 22.10.2019.